Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων
Εργαστηριο Τεχνητης Νοημοσυνης και Μηχανικης Μαθησης

Δημιουργία αντιπαραδειγμάτων και εμπλουτισμός πολυτροπικών συστημάτων μέσω εξωτερικής γνώσης

Διδακτορικη Διατριβη

Μαρία Λυμπεραίου

Συμβουλευτική Επιτροπή: Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούνιος 2025

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Τεχνητής Νοημοσύνης και Μηχανικής Μάθησης

Δημιουργία αντιπαραδειγμάτων και εμπλουτισμός πολυτροπικών συστημάτων μέσω εξωτερικής γνώσης

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Μαρία Λυμπεραίου

Συμβουλευτική Επιτροπή Γεώργιος Στάμου
Αθανάσιος Βουλόδημος
Μιχαήλ Βαζιργιάννης

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή τη 11η Ιουνίου, 2025.

....
Γεώργιος Στάμου
Καθηγητής Ε.Μ.Π.

....
Αθανάσιος Βουλόδημος
Επ. Καθηγητής Ε.Μ.Π.

....
Μιχαήλ Βαζιργιάννης
Καθηγητής Ecole Polytechnique

....
Δημήτριος Φωτάκης
Καθηγητής Ε.Μ.Π.

....
Αριστείδης Παγουρτζής
Καθηγητής Ε.Μ.Π.

....
Αθανάσιος Ροντογιάννης
Καθηγητής Ε.Μ.Π.

....
Χρυσούλα Ζέρβα
Ass. Professor
Instituto Superior Tecnico

Αθήνα, Ιούνιος 2025

……………………………………………

Μαρια Λυμπεραιου
Διδάκτωρ Ηλεκτρολόγος Μηχανικός
και Μηχανικός Υπολογιστών Ε.Μ.Π.

Copyright © – All rights reserved Maria Lymperaiou, 2025.
Με επιφύλαξη παντός δικαιώματος.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Περίληψη

Τα τελευταία χρόνια, τα μοντέλα φυσικής γλώσσας και κατ' επέκταση τα πολυτροπικά συστήματα έχουν επιδείξει αξιοσημείωτη επιτυχία σε ένα ευρύ φάσμα εργασιών, αξιοποιώντας κοινές αναπαραστάσεις οπτικών και κειμενικών εισόδων. Παρά τα συχνά εντυπωσιακά αποτελέσματα, τα μοντέλα αυτά είναι αδιαφανή, καθιστώντας την ερμηνεία της λήψης αποφάσεων αρκετά ασαφή, ενώ παράλληλα οι δυνατότητές τους σε συγκεκριμένες εργασίες είναι περιορισμένες, οδηγώντας σε περιορισμένη γενίκευση. Οι δύο αυτοί περιορισμοί μπορούν να επιλυθούν με την αξιοποίηση εξωτερικών πηγών γνώσης, οι οποίες παραδοσιακά συμπεριλαμβάνουν γράφους γνώσης, ενώ πιο πρόσφατα βασίζονται σε μεγάλα γλωσσικά μοντέλα. Η παρούσα διατριβή εξετάζει τη συνεισφορά της εξωτερικής γνώσης στην αντιμετώπιση των προαναφερθέντων περιορισμών αξιοποιώντας αντιπαραδείγματα: η αντικατάσταση εννοιών σε κειμενικές περιγραφές με καλά ορισμένο τρόπο δύναται να αποκαλύψει τον τρόπο λήψης αποφάσεων των μοντέλων χωρίς να είναι απαραίτητη η εισχώρηση στο εσωτερικό τους. Συγκεκριμένα, προτείνουμε την εφαρμογή αντιπαραδειγμάτων οδηγούμενων από γράφους γνώσης σε πληθώρα μοντέλων που αφορούν τόσο την ανάκτηση εικόνων από κείμενο, την παραγωγή εικόνων από κείμενο και την παραγωγή κειμενικών περιγραφών από εικόνες. Επιπλέον, η χρήση των μεγάλων γλωσσικών μοντέλων ως βάσεις γνώσης παρέχει εμπλουτισμό της υπάρχουσας γνώσης τους, επιτρέποντάς τους να αποδίδουν σε νέες εργασίες με καλύτερα αποτελέσματα. Στην παρούσα διατριβή επιτελείται ο εμπλουτισμός γνώσης σε πολυτροπικά μοντέλα ανάκτησης εικόνων από κείμενο, αποδεικνύοντας τη λειτουργικότητα της μεθόδου. Τέλος, γίνεται μια ανάλυση των περιορισμών χρήσης των μεγάλων γλωσσικών μοντέλων ως βάσεις γνώσης όσον αφορά τις ικανότητες συλλογιστικής και την εσφαλμένη παραγωγή απαντήσεων.

Λέξεις-κλειδιά Εξηγήσιμη τεχνητή νοημοσύνη, Εξηγήσεις μέσω αντιπαραδειγμάτων, Πολυτροπικά συστήματα, Βάσεις γνώσης, Μεγάλα γλωσσικά μοντέλα

Abstract

In recent years, natural language models and by extension multimodal systems have demonstrated remarkable success in a wide range of tasks, exploiting common representations of visual and textual inputs. Despite often impressive results, these models are opaque, making the interpretation of decision making quite fuzzy, while their potential in specific tasks is limited, leading to limited generalizability. These two limitations can be resolved by exploiting external knowledge sources, which traditionally include knowledge graphs, and more recently are based on large language models. This thesis examines the contribution of external knowledge in addressing the aforementioned constraints by exploiting counterexamples: replacing concepts in textual descriptions with well-defined ones can reveal the decision making of models without the need to penetrate inside them. Specifically, we propose the application of graph-driven knowledge counterexamples to a variety of models involving both image retrieval from text, image generation from text and the generation of textual descriptions from images. In addition, the use of large linguistic models as knowledge bases provides enrichment of their existing knowledge, allowing them to perform new tasks with better results. In this thesis, knowledge enrichment is performed on multimodal models for retrieving images from text, demonstrating the functionality of the method. Finally, an analysis of the limitations of using large language models as knowledge bases in terms of reasoning abilities and incorrect generation of answers is provided.

Keywords Explainable artificial intelligence, Counterfactual explanations, Multimodal systems, Knowledge bases, Large Language Models

Acknowledgements

Reaching the completion of this dissertation has been both a demanding and rewarding journey. It would not have been possible without the guidance, support, and kindness of many people, to whom I owe my deepest gratitude. First and foremost, I wish to express my profound gratitude to my supervisor, Giorgos Stamou, whose guidance, insight, and patience shaped every stage of this work. His ability to challenge my thinking while offering continuous support has been invaluable, and I am deeply thankful for the trust he placed in me throughout this process. I would also like to acknowledge him for establishing an outstanding group of research collaborators, whose contributions significantly enhanced the quality and depth of this dissertation.

I am also honored regarding the members constituting my committee, validating the scientific value and contribution of my work. First, I would like to thank Athanasios Voulodimos, who stood by me throughout this journey, not only guiding me in scientific matters, but also encouraging me and engaging in countless hours of discussion. I would also like to express my gratitude to Chryssoula Zerva, whose research expertise I had the privilege to learn from during our collaboration. Her guidance was enlightening and greatly contributed to the advancement of my research skills. I would also like to extend my heartfelt thanks to the rest of my dissertation committee: Michalis Vazirgiannis, Dimitrios Fotakis, Aristeidis Pagourtzis and Athanasios Rontogiannis.

This thesis was funded by H.F.R.I. 3rd call for PhD students, which allowed me to fully engage to my research.

This thesis is the outcome of extensive collaboration, and I sincerely thank all my co-authors for their meaningful contributions and support throughout this work. During our PhD journeys we became good friends, sharing endless hours together inside and outside the lab, while also traveling around the world. Especially, my heartfelt thanks go to Dr. Giorgos Filandrianos, with whom we have co-authored most of our papers. His collaborative spirit, insightful comments and dedication greatly enriched our work, leading to many successful outcomes that will hopefully continue in the future. I also extend my appreciation to Dr. Eddie Dervakos who greatly guided me during my first research steps, significantly shaping my understanding on the field and the way I approached my consequent research steps. Many thanks also to Angeliki Dimitriou and Konstantinos Thomas, two of my closest collaborators, who greatly inspire me and open novel research avenues for our collaborative research.

My sincere gratitude goes to all members of AILS laboratory, with whom we have shared happy moments and several interesting conversations: Nikos Chaidos, Nikos Spanos, Orfeas Menis-Mastromichalakis, Vassilis Lyberatos, Paraskevi Theofilou, Spyros Kantarelis, Ilias Mitsouras, Natalia Grigoriadou, Vassilis Karampinis, Jason Liartis, Christos Papadimitriou, Angelos Vlachos, Dido Stoikou, Anastasia Kritharoula, Panagiotis Giadikiaroglou, Giannis Panagiotopoulos, Voula Pavlaki, Lefteris Tsonis. I am also grateful to my diploma students with whom I had the privilege of co-authoring publications: Georgia Argyrou, Dimitris Lymperopoulos, Alexandros Koulakos, Nikitas Theodoropoulos, Elena Stringli, Andreas Evangelatos, Dimitra Karkani, Iraklis Premptis, Petros Raptopoulos— thank you for your dedication, enthusiasm, and teamwork.

Last but not least, I also would like to express my gratitude to my family and friends for the encouragement in the course of this research and beyond. Their supportive presence and unwavering belief in me provided the motivation and strength to persevere through difficult moments, and their love and understanding created the foundation upon which I could pursue this work with confidence and dedication. I am truly fortunate to have such a steadfast support network by my side.

Contents

Contents x

List of Figures xiv

1 Εκτεταμένη Περίληψη στα Ελληνικά 1
- 1.1 Συνεισφορά της διατριβής 1
- 1.2 Τύποι γνώσης 2
- 1.3 Αντιπαραδειγματικές εξηγήσεις 4
- 1.4 Ρητές βάσεις γνώσης για εξηγησιμότητα 6
  - 1.4.1 Επιτάχυνση εξαγωγής σημασιολογικών αντιπαραδειγμάτων 6
  - 1.4.2 Μη ελάχιστες παρεμβάσεις 7
- 1.5 Γνώση σε εξηγήσεις εικόνας-γλώσσας 9
- 1.6 Ενίσχυση συστημάτων κειμένου-εικόνας μέσω γνώσης 12
- 1.7 Ενίσχυση γνώσης μέσω Μεγάλων Γλωσσικών Μοντέλων 15
- 1.8 Συμπεράσματα 17

2 Introduction 19
- 2.1 Thesis contributions 20
  - 2.1.1 Survey and taxonomy of knowledge types 20
  - 2.1.2 Background in Counterfactual Explanations 20
  - 2.1.3 Explicit knowledge bases for explainability 21
  - 2.1.4 Knowledge in vision-language explanations 21
  - 2.1.5 Knowledge enhancement in vision-language tasks 22
  - 2.1.6 LLMs as knowledge bases 22

3 Types of knowledge 23
- 3.1 Implicit knowledge 25
  - 3.1.1 Large Language Models as Knowledge Bases 25
- 3.2 Explicit knowledge 25
- 3.3 Web-crawled knowledge 26
- 3.4 Internal knowledge 26

4 Counterfactual explanations 27
- 4.1 From Causality to Counterfactuals 27
- 4.2 Desiderata for Counterfactual Explanations 27
- 4.3 Counterfactual Explanations vs. Other XAI Methods 28
  - 4.3.1 Feature attribution methods 28
  - 4.3.2 Surrogate model explanations 29
  - 4.3.3 Prototype and example-based explanations 29
  - 4.3.4 Saliency and gradient-based methods 29
  - 4.3.5 Rule-based and symbolic explanations 30
  - 4.3.6 Complementarity and integration of methods 30
- 4.4 Knowledge and semantics in counterfactual generation 30

4.5 Semantic distances and conceptual interventions 31

5 Explicit knowledge bases for explainability 33
5.1 Matching algorithms for counterfactual explanations 33
5.1.1 Concept Matching using bipartite graph structures 33
5.1.2 Bipartite matching as an assignment problem 34
5.1.3 Algorithms for Bipartite Matching 35
5.1.4 Graph matching 36
5.1.5 Exact versus Inexact Graph Matching 36
5.1.6 Formulations and Optimization Objectives 36
5.1.7 Major Approaches to Graph Matching 37
5.1.8 Graph Matching as an Assignment Problem 38
5.1.9 Graph Matching and Semantics 39
5.2 Accelerating semantic counterfactual computations 40
5.2.1 Graph Neural Networks 40
5.2.2 GNN Architectures 41
5.2.3 Connection with knowledge 42
5.2.4 Concept-based counterfactuals using GNNs 43
5.2.5 Graph-based counterfactuals using GNNs 44
5.2.6 Non-minimal interventions 52

6 Knowledge in vision-language explanations 67
6.1 Explainability in Visual Question Answering 68
6.1.1 Introduction to Visual Question Answering (VQA) 68
6.2 Explainability in Text to Image generation 70
6.2.1 Background in Conditional Image Generation 70
6.2.2 Conceptual edits for generative evaluation 73
6.2.3 Method 73
6.2.4 Generative evaluation 74
6.2.5 Experiments on Story Visualization 76
6.2.6 Experiments on Scene Generation 78
6.2.7 Key takeaways on explainable image generation 81
6.3 Explainability in image captioning hallucinations 82
6.3.1 Background in Image Captioning 82
6.3.2 Explainable hallucinations evaluation 85
6.3.3 Hallucination detection framework 88
6.3.4 Experiments 90
6.3.5 Linguistic metrics may be misleading 94

7 Knowledge enhancement in vision-language tasks 99
7.1 Knowledge in Visual Question Answering (K-VQA) 100
7.1.1 Datasets 100
7.1.2 Methods 101
7.1.3 Evaluation 106
7.2 Knowledge in Visual Reasoning (K-VR) 106
7.2.1 Datasets 106
7.2.2 Methods 107
7.2.3 Evaluation 108
7.3 Knowledge in Visual Commonsense Reasoning (K-VCR) 108
7.3.1 Datasets 108
7.3.2 Methods 109
7.3.3 Evaluation 109
7.4 Knowledge in Image Captioning (K-IC) 110
7.4.1 Datasets 110
7.4.2 Methods 110
7.4.3 Evaluation 112

7.5 Knowledge in Visual Dialog (K-VD) . . . 113
  7.5.1 Datasets . . . 113
  7.5.2 Methods . . . 113
  7.5.3 Evaluation . . . 113
7.6 Knowledge in Visual Storytelling (K-VIST) . . . 113
  7.6.1 Datasets . . . 114
  7.6.2 Methods . . . 114
  7.6.3 Evaluation . . . 115
7.7 Knowledge in Image Generation . . . 115
  7.7.1 Datasets . . . 115
  7.7.2 Methods . . . 115
  7.7.3 Metrics . . . 116
7.8 Multi-task transformers with knowledge . . . 117
  7.8.1 Methods . . . 117
  7.8.2 Evaluation . . . 118
7.9 The future of knowledge enhancement in multimodality . . . 118

8 Knowledge enhancement using Large Language Models **121**
8.1 Background in LLM-enhanced vision-language tasks . . . 122
  8.1.1 LLM-as-KB for K-VQA . . . 122
  8.1.2 LLM-as-KB for K-VCR . . . 123
  8.1.3 LLM-as-KB for K-VCR . . . 123
8.2 Large Language Models for Visual Word Sense Disambiguation . . . 124
  8.2.1 Method . . . 125
  8.2.2 Experimental results . . . 128
8.3 Reasoning and Hallucinations in Large Language Models . . . 131
  8.3.1 Abstract reasoning in LLMs . . . 133
  8.3.2 Hallucinations in LLMs . . . 138
8.4 When trust in LLMs falters . . . 141
  8.4.1 Inverse tasks . . . 142
  8.4.2 Cognitive biases . . . 144
  8.4.3 Are LLMs as knowledge bases a good idea? . . . 148

9 Conclusion **151**

10 Bibliography **157**

List of Figures

3.0.1 Overview of knowledge sources. [206] 24

5.2.1 The pipeline of our method. In the first stage, we construct a bipartite graph using words as nodes, and in the second stage we utilize a GNN to get feasible substitutions that approximately solve the RLAP. In the final stage, we use beam search to change appropriate words of the original dataset, thus getting a new counterfactual dataset. [208] 43

5.2.2 This image illustrates a real-world scenario where scene graphs meet counterfactual explanations: what delineates the class "safe" from "not safe" can be attributed to certain graph edits [68]. 45

5.2.3 Results for Rusty → Brewer Blackbird. **Bold** denotes best results (lowest number of edits and GED scores). [69] 47

5.2.4 P@k of GCN variant for different training pairs p on CUB and VG. [69] 50

5.2.5 An overview of the *representation-ranking-evaluation* procedure to explain semantic similarity models. Both the default stream (blue) and the intervened one (pink) are provided. 55

5.2.6 A closer look at ranking procedure. Green lines denote ground truth matchings, while red lines indicate matchings selected from maximum cosine similarity scores between query and corpus embeddings [204]. 56

5.2.7 Conventional causal graph (left) and counterfactual intervention causal graph (right) when $q \rightarrow q^*$. 56

5.2.8 Example of semantically non-minimal interventions on attributes. With blue color we denote the initial attributes, while the pink nodes correspond to the attributes that substitute the initial ones, so that the intervention is performed. The bold edges between nodes denote the matched concepts-which actually consist the best matches available on this graph. 57

5.2.9 A herd of zebras grazing in a lush green field 61

5.2.10 ***Successful alternative.***Many people are relaxing under their umbrellas on the beach. . . . 62

5.2.11 ***Object color.***A vase sitting on a table with white flowers in it. 62

5.2.12 ***Object enumeration.*** A dirt bike rider performing a stunt while in the air. 62

5.2.13 ***Detailed semantics.***A cat playing with a shoe in a grassy field. 63

6.1.1 Overview of our proposed knowledge-based counterfactual VQA framework. [320] 69

6.2.1 Outline of the proposed semantic edits framework for image generation evaluation. [203] . . . 74

6.2.2 Ground truth vs generated story frames using [340] for L=4. [203] 78

6.2.3 An image sample generated by Stable Diffusion 2 to extract local explanations. [203] 81

6.3.1 Example of hallucination on image captioning. The generated caption c misses an accurate relationship between the man and the dog. The concept "laptop" should replace the concept "dog" in the generated caption, while the relationship "next to" should be added to connect the concepts "dog" and "man". [202] 87

6.3.2 An example of detected hallucination of objects in image captioning from our framework is presented, depicting each phenomenon along with the proposed metrics. Objects in yellow represent an overspecialized phenomenon, in purple a replacement, and in red a removal. Those in green are correct objects, and those in blue are the underspecialized objects (which do not constitute hallucinations, as the caption contains a more generic concept to the ground truth one). As shown, the hallucination rate is calculated as the sum of the rate of each hallucination phenomenon independently. [202] 88
6.3.3 An demonstration of the edge integration into HalCECE. The edges are highlighted in **bold**, and the different colors correspond to those of Figure 6.3.2. 89
6.3.4 Statistics of our proposed explainable metrics on ***object hallucinations*** by GiT-base on the $VG \cap COCO$ validation set. 93
6.3.5 Statistics of our proposed explainable metrics on ***role hallucinations*** by GiT-base on the $VG \cap COCO$ validation set. 94

7.0.1 Taxonomy of VL tasks on which knowledge enhancement via KGs has been performed. [206] 99

8.0.1 External knowledge is required to answer these visual questions [221, 301, 126, 205]. 121
8.2.1 An example of the VWSD task: given a full phrase, comprised of a target word and its context, the most appropriate image candidate must be retrieved. 124
8.2.2 Candidate images for the phrase "rowing dory". 130
8.2.3 Candidate images for the phrase t "football goal". 131
8.3.1 Categorization of puzzles [88]. 134
8.3.2 Examples of vertical and lateral thinking riddles. Vertical thinking may require commonsense knowledge ('How do you flood a room?') or straightforward, while abstract thinking steps ('I have five fingers but I am not alive. What am I?'). On the other hand, lateral thinking demonstrates tricky questions, where straightforward thinking fails, including sentence puzzles ('A man shaves everyday, yet he keeps his beard long') and word puzzles ('What type of cheese is made backwards?'). [135] 136
8.4.1 Redefined reasoning pathways. [321] 142
8.4.2 Number of anchored responses for varying LLM sizes in the Llama family (MC response format). 144
8.4.3 Examples of all implemented cognitive biases, used as adversarial attacks. 145
8.4.4 The MRR values for each product in the coffee machines dataset, for a positive and a negative influential attack for: (a) Claude 3.7, (b) LLaMA-405b. 147
8.4.5 Number of products that became the most frequently recommended due to the attack (not most recommended before). Only the biases with non-zero values are shown. *exp* stands for *expert attacks*, contrasting the *generated* ones. 147
8.4.6 MRR values pre- and post-attack in the coffee machines dataset, for various sizes of the LLaMA model. 147

Chapter 1

Εκτεταμένη Περίληψη στα Ελληνικά

Η Τεχνητή Νοημοσύνη (ΤΝ) έχει εξελιχθεί ραγδαία από θεωρητική έννοια σε καταλυτικό παράγοντα αλλαγής σε τομείς όπως η υγεία, τα χρηματοοικονομικά, η εκπαίδευση και η ψυχαγωγία. Η ικανότητά της να μαθαίνει από δεδομένα και να εκτελεί πολύπλοκες διεργασίες έχει εντείνει το ενδιαφέρον για τις κοινωνικές της επιπτώσεις. Καθώς τα συστήματα τεχνητής νοημοσύνης επηρεάζουν ολοένα και πιο κρίσιμες αποφάσεις, ζητήματα εμπιστοσύνης, λογοδοσίας και διαφάνειας αποκτούν κεντρική σημασία. Έτσι ενισχύεται η ανάγκη για Εξηγήσιμη Τεχνητή Νοημοσύνη (Explainable AI - XAI), η οποία επιδιώκει να κάνει τις αποφάσεις των μοντέλων κατανοητές στους ανθρώπους. Η εξηγησιμότητα είναι κρίσιμη για την οικοδόμηση εμπιστοσύνης, την ανίχνευση προκαταλήψεων, τη συμμόρφωση με κανονισμούς και τη διασφάλιση ηθικής χρήσης.

Οι αντιπαραδειγματικές εξηγήσεις (counterfactual explanations - CEs) αποτελούν μια ιδιαίτερα κατανοητή προσέγγιση, περιγράφοντας τις ελάχιστες αλλαγές στην είσοδο που οδηγούν σε διαφορετική πρόβλεψη. Ωστόσο, μεγάλο μέρος της υπάρχουσας βιβλιογραφίας βασίζεται σε μη ερμηνεύσιμα χαμηλού επιπέδου χαρακτηριστικά, περιορίζοντας τη χρησιμότητα των εξηγήσεων. Η χρήση υψηλού επιπέδου εννοιών (αντικειμένων, ιδιοτήτων, σχέσεων) παρέχει εξηγήσεις πιο κοντά στην ανθρώπινη κατανόηση και επαναληψιμότητα.

Πρόσφατες εργασίες αναδεικνύουν την αξία εξωτερικών πηγών γνώσης για την επιβολή εννοιολογικότητας στο XAI, ενώ ταυτόχρονα μπορούν να ενισχύσουν και την απόδοση των μοντέλων. Τα συστήματα ενισχυμένα με γνώση αποκτούν καλύτερες ικανότητες συλλογισμού και κατανόησης σύνθετων εννοιών. Με την έλευση των Μεγάλων Γλωσσικών Μοντέλων (Large Language Models - LLMs), η ενσωμάτωση εκτεταμένης γνώσης γίνεται ευκολότερη. Ωστόσο, η αδιαφάνειά τους, οι ψευδαισθήσεις και τα σφάλματα συλλογισμού εγείρουν ανησυχίες.

Η παρούσα διατριβή μελετά πώς εξωτερικές πηγές γνώσης μπορούν να παράγουν αξιόπιστες εννοιολογικές αντιμεταθετικές εξηγήσεις και να ενισχύσουν την απόδοση των συστημάτων τεχνητής νοημοσύνης.

1.1 Συνεισφορά της διατριβής

Στην παρούσα διατριβή, μελετάμε τον ρόλο των βάσεων γνώσης σε εργασίες Επεξεργασίας Φυσικής Γλώσσας (Natural Language Processing - NLP) και Πολυτροπικής Μάθησης (Multimodal Learning), τόσο ως προς την εξηγησιμότητα όσο και ως προς τη βελτίωση απόδοσης.

Επισκόπηση και ταξινόμηση τύπων γνώσης

Αρχικά κατηγοριοποιούμε τους τύπους γνώσης που μπορούν να ενσωματωθούν σε συστήματα τεχνητής νοημοσύνης (Κεφάλαιο 3). Οι δομημένες πηγές, όπως οι γράφοι γνώσης (knowledge graphs - KGs), προσφέρουν καθαρές σημασιολογικές σχέσεις, ενώ τα LLMs παρέχουν ευρεία αλλά λιγότερο εντοπίσιμη γνώση.

Υπόβαθρο στις αντιπαραδειγματικές εξηγήσεις (CEs)

Συνδέουμε τις πηγές γνώσης με τη δημιουργία CEs (Κεφάλαιο 4). Οι σημασιολογικές αποστάσεις επιτρέπουν εννοιολογικές αντικαταστάσεις ως ελάχιστες, κατανοητές παρεμβάσεις.

Δομημένες βάσεις γνώσης για εξηγησιμότητα

Στο Κεφάλαιο 5 προτείνουμε μεθόδους μετα-ανάλυσης (post-hoc) βασισμένες σε αντιστοίχιση γράφων για να εξάγουμε βέλτιστες εννοιολογικές επεξεργασίες. Οι συνεισφορές περιλαμβάνουν:

- Προσεγγιστικά ελάχιστες παρεμβάσεις για NLP και οπτικούς ταξινομητές (Κεφάλαια 5.2.4, 5.2.5).
- Μη ελάχιστες παρεμβάσεις για μοντέλα σημασιολογικής ομοιότητας (Κεφάλαιο 5.2.6).

Γνώση σε εξηγήσεις όρασης–γλώσσας (Vision-Language)

Στο Κεφάλαιο 6 επεκτείνουμε τις εννοιολογικές CEs σε πολυτροπικά συστήματα και συγκεκριμένα σε συστήματα όρασης-γλώσσας. Συνεισφέρουμε:

- Δημιουργία εικόνας από κείμενο (Text-to-image generation) με εννοιολογική ευθυγράμμιση κειμένου–εικόνας (Κεφάλαιο 6.2).
- Ανίχνευση ψευδαισθήσεων (hallucinations) σε κειμενικές περιγραφές εικόνων, αποδεικνύοντας την ανεπάρκεια των κλασικών μετρικών (Κεφάλαιο 6.3).

Ενίσχυση γνώσης σε πολυτροπικές εργασίες

Στο Κεφάλαιο 7 παρουσιάζουμε επισκόπηση της συνεισφοράς της εκωτερικής γνώσης σε εργασίες όπως απάντηση οπτικών ερωτήσεων (VQA), οπτική συλλογιστική (visual reasoning), παραγωγή οπτικών περιγραφών (captioning), διάλογο με εικόνες (visual dialog), παραγωγή ιστοριών (storytelling) και δημιουργία εικόνων (image generation), εξηγώντας στη συνέχεια γιατί τα LLMs δρουν ως πρακτικές βάσεις γνώσης όταν οι δομημένοι γράφοι είναι ανεπαρκείς.

LLMs ως βάσεις γνώσης

Τέλος, στο Κεφάλαιο 8 εξετάζουμε τα LLMs ως βάσεις γνώσης για εμπλουτισμό δεδομένων με σκοπό την ενίσχυση της επίδοσης των σχετικών συστημάτων. Συνεισφέρουμε στα ακόλουθα:

- Ανασκόπηση χρήσης LLMs σε πολυτροπικές εργασίες (Κεφάλαιο 8.1).
- Βελτίωση της αποσαφήνισης οπτικών εννοιών (Visual Word Sense Disambiguation - VWSD) μέσω γνώσης των LLMs (Κεφάλαιο 8.2).
- Ανάλυση ορίων των LLMs, όπως ψευδαισθήσεις και αποτυχίες συλλογισμού (Κεφάλαια 8.3, 8.4).

Συνολικά, η διατριβή παρουσιάζει πώς εξωτερικές πηγές γνώσης ενισχύουν την εξηγησιμότητα και την απόδοση των μοντέλων τεχνητής νοημοσύνης χωρίς αλλαγές στην αρχιτεκτονική τους, προωθώντας πιο αξιόπιστα, προσαρμόσιμα και ουσιαστικά συστήματα.

1.2 Τύποι γνώσης

Η ενσωμάτωση εξωτερικών πηγών γνώσης σε νευρωνικά μοντέλα αποτελεί ένα ενεργό πεδίο έρευνας, χάρη στα πλεονεκτήματα που προσφέρει ως προς τη βελτίωση της απόδοσης και της επεξηγησιμότητας. Η γνώση διακρίνεται σε διάφορες «λειτουργικές αισθήσεις», δηλαδή κατηγορίες ή διαστάσεις γνώσης που μπορούν να κατέχουν οι άνθρωποι ή τα συστήματα τεχνητής νοημοσύνης, καθεμία από τις οποίες χαρακτηρίζεται από τη φύση, την προέλευση και τον τρόπο αξιοποίησής της. Μεταξύ των πιο συνηθισμένων ειδών πρόσθετης γνώσης που προσφέρουν αυτά τα οφέλη συναντώνται διάφορες μορφές δομημένης, γλωσσικής, πραγματολογικής και κοινής γνώσης. [206]

Η **ιεραρχική γνώση (hierarchical knowledge)** βασίζεται σε σχέσεις τύπου «isA», που σχηματίζουν δενδροειδείς δομές με τη ρίζα να αντιστοιχεί στην πιο γενική έννοια και τα φύλλα στις πιο εξειδικευμένες. Για παράδειγμα, η σχέση «η γάτα είναι θηλαστικό» αποτελεί τυπικό παράδειγμα τέτοιας δομής.

Η **λεξική γνώση (lexical knowledge)** λειτουργεί ως ένα δομημένο λεξικό που παρέχει γλωσσικούς κανόνες και συμβάλλει στην αποσαφήνιση σημασίας, ενώ μπορεί να συνδυαστεί με ιεραρχική γνώση για την απόδοση υπερωνυμιών ή υπωνυμιών. Οι ονομαζόμενες οντότητες καλύπτουν κύρια ονόματα ανθρώπων, τοποθεσιών, οργανισμών και άλλων πραγματικών οντοτήτων· για παράδειγμα, στη φράση «ο Joe Biden είναι ο πρόεδρος των Ηνωμένων Πολιτειών» οι όροι «Joe Biden» και «Ηνωμένες Πολιτείες» αποτελούν ονομαζόμενες οντότητες.

Η **πραγματολογική γνώση (factual knowledge)** περιλαμβάνει εγκυκλοπαιδικές πληροφορίες για τον κόσμο, όπως το ιστορικό γεγονός ότι ο Β΄ Παγκόσμιος Πόλεμος διήρκεσε από το 1939 έως το 1945. Σε αυτήν ανήκουν και πιο εξειδικευμένες επιστημονικές γνώσεις, π.χ. στον χώρο της ιατρικής, της βιολογίας ή της χημείας. Συχνά συνδυάζεται με ονομαζόμενες οντότητες, όπως στη δήλωση «οι ζέβρες ζουν στην Αφρική».

Από την άλλη πλευρά, η **κοινή γνώση (commonsense knowledge)** αντιστοιχεί στην προφανή κατανόηση του κόσμου όπως τον αντιλαμβάνονται οι άνθρωποι. Δηλώσεις όπως «η ζάχαρη είναι γλυκιά» ή «αν βγω στη βροχή θα βραχώ» θεωρούνται αυταπόδεικτες. Η κοινή γνώση περιλαμβάνει μια σειρά από υποκατηγορίες, όπως σχέση **ομοιότητας** ή **αντιδιαστολής** μεταξύ εννοιών, σχέσεις **μέρους-όλου** (εκφρασμένες λεξικά μέσω μερωνυμιών και ολονυμιών), σχέσεις **χρησιμότητας** ή **ικανότητας** («το πιρούνι χρησιμοποιείται για φαγητό»), **χωρικές** πληροφορίες για τη θέση αντικειμένων στον φυσικό κόσμο («τα καράβια βρίσκονται κοντά στο νερό»), **συγκριτικές** σχέσεις («οι λεοπαρδάλεις είναι μεγαλύτερες από τις γάτες»), **αριθμητική** γνώση («οι άνθρωποι έχουν δύο μάτια»), καθώς και γνώσεις **προθέσεων**, επιθυμιών και σχεδίων («οι πεινασμένοι άνθρωποι θέλουν να φάνε»). Επιπλέον, η **συμπεριφορική** γνώση προκύπτει μέσω λογικής από κοινοτοπίες («ένα παιδί δεν μπορεί να πιει 10 λίτρα νερό σε μία μέρα»), ενώ η «γνώση **δημιουργού**» εμπερικλείει σχέσεις όπως «ένα τραγούδι δημιουργείται από μουσικό».

Η **χρονική ή γεγονοτική γνώση (temporal or event knowledge)** συνδυάζει πραγματολογικές και κοινές πληροφορίες, παρέχοντας χρονική σειρά και αιτιακές σχέσεις. Ένα παράδειγμα πραγματολογικής ακολουθίας είναι ότι «η COVID-19 ξεκίνησε το 2019 και τα εμβόλια αναπτύχθηκαν το 2020», ενώ ένα παράδειγμα κοινής ακολουθίας είναι ότι «η άνοιξη έρχεται μετά τον χειμώνα». Οι ακολουθίες μπορεί επίσης να εκφράζουν **αιτιότητα** («το αγόρι έριξε το ποτήρι και το ποτήρι έσπασε»), η οποία μπορεί να αναδιατυπωθεί σε υποθετικές ή ακόμη και σε **αντιπαραδειγματικές (counterfactual)** δηλώσεις.

Η **οπτική γνώση (visual knowledge)** περιλαμβάνει εικόνες και ενδεχομένως επιπλέον σχολιασμό που συνδέει την οπτική αντίληψη με την κοινή γνώση. Ιδιότητες αντικειμένων όπως σχήμα, χρώμα και υφή αποτυπώνονται οπτικά, ενώ χωρικές σχέσεις, λεπτομέρειες αντικειμένων, ενέργειες, σκηνικά κείμενα, παροχές ικανοτήτων και άλλες πληροφορίες μπορούν επίσης να εξαχθούν από εικόνες. Συχνά οι εικόνες επιτρέπουν την εξαγωγή πιο σύνθετων συμπερασμάτων, όπως προθέσεων («ένα άτομο με βαλίτσα και διαβατήριο σκοπεύει να ταξιδέψει»), αιτιών/συνεπειών ή πολιτισμικών και πραγματολογικών πληροφοριών («ναός της αρχαίας Ελλάδας του 5ου αιώνα π.Χ.»).

Στην παρούσα εργασία, η εξωτερική γνώση διακρίνεται σε τρεις κύριες κατηγορίες, ανάλογα με τον βαθμό δομής των πηγών: την άρρητη (implicit) γνώση, η οποία βρίσκεται σε μη συμβολική μορφή, τη ρητή (explicit) γνώση, που συνήθως αποθηκεύεται σε δομημένες βάσεις γνώσης, και τη γνώση αντλημένη από τον ιστό (web-crawled knowledge), η οποία προέρχεται από μη δομημένες διαδικτυακές πηγές. Επιπλέον, αναγνωρίζεται μία ακόμη κατηγορία, η εσωτερική ή αυτο-γνώση, η οποία δεν εξαρτάται από εξωτερικές πηγές, αλλά προκύπτει από την ίδια την υπάρχουσα πληροφορία στα δεδομένα.

Άρρητη γνώση

Η άρρητη γνώση αναφέρεται σε πληροφορίες που είναι αποθηκευμένες με μη συμβολικό τρόπο, όπως στα βάρη των νευρωνικών μοντέλων. Η προεκπαίδευση μοντέλων μετασχηματιστών με μη εποπτευόμενες ή αυτο-εποπτευόμενες μεθόδους επιτρέπει την απόκτηση τέτοιας γνώσης, σύμφωνα με το παράδειγμα χρήσης γλωσσικών μοντέλων ως βάσεις γνώσης (LM-as-KB). Εκεί, η γνώση ενσωματώνεται σε υψηλής διάστασης συναρτήσεις, ως στατιστικές κανονικότητες μεταξύ των λέξεων· οι οντότητες εκπροσωπούνται ως διανύσματα και οι σχέσεις εκφράζονται ως γεωμετρικές μετασχηματίσεις. Παρ' όλα αυτά, η άρρητη γνώση δεν επαρκεί πάντοτε για την απάντηση ερωτήσεων που απαιτούν σπάνια, ακριβή ή καλά δομημένη γνώση, ενώ το μαύρο-κουτί της φύσης της δυσχεραίνει τον εντοπισμό λαθών, προκαταλήψεων ή ασυνεπειών. Επιπλέον, ο βαθμός στον οποίο τα γλωσσικά μοντέλα πράγματι «λογικεύονται» ή απλώς απομνημονεύουν παραμένει ένα ανοιχτό ζήτημα.

Η πρόσφατη άνοδος των Μεγάλων Γλωσσικών Μοντέλων έχει ενισχύσει την αντίληψη ότι μπορούν να λειτουργήσουν ως ισχυρές άρρητες βάσεις γνώσης, χάρη στην κλιμάκωση, την ευελιξία και την ικανότητά τους να γενικεύουν. Παρόλο που είναι επιρρεπή σε αστοχίες, ασυνέπειες και παραισθήσεις, η δυναμική τους είναι σημαντική, ειδικά αν συνδυαστούν με τεχνικές όπως η ανάκτηση πληροφοριών (RAG), η προσαρμογή τομέα και η αντιφατική συλλογιστική.

Ρητή γνώση

Η ρητή γνώση βασίζεται σε σαφή, δομημένα δεδομένα, συνήθως με τη μορφή γράφων γνώσης, και μπορεί να καλύψει κενά που δεν μπορεί να καλύψει η άρρητη γνώση. Οι γράφοι γνώσης επιτρέπουν τη διαφανή αξιολόγηση της συμβολής της γνώσης και την εύκολη αντιμετώπιση σφαλμάτων ή μεροληψιών. Παρά τα οφέλη, η κατασκευή και η συντήρησή τους συχνά απαιτούν σημαντικό ανθρώπινο κόπο και εξειδίκευση, ιδιαίτερα σε εξειδικευμένα πεδία όπως η ιατρική.

Γνώση από τον ιστό

Η γνώση που προκύπτει από διαδικτυακές πηγές ανήκει σε μια ενδιάμεση κατηγορία, χωρίς ανάγκη προεκπαιδευμένων ή επισημασμένων δεδομένων, αλλά με προκλήσεις ως προς την ποιότητα και την αξιοπιστία. Αν και πιο ευέλικτη και κλιμακούμενη, απαιτεί μηχανισμούς φιλτραρίσματος και ποιοτικού ελέγχου.

Εσωτερική γνώση

Τέλος, η εσωτερική ή αυτο-γνώση προκύπτει από τα ίδια τα διαθέσιμα δεδομένα, όπως συμβαίνει με τη δημιουργία γράφων σκηνής από εικόνες. Παρέχει λεπτομερή και πλούσια αναπαράσταση, αλλά δεν επεκτείνει τη γνώση πέραν του αρχικού συνόλου εκπαίδευσης και είναι ευάλωτη σε σφάλματα που παράγονται κατά την εξαγωγή της.

1.3 Αντιπαραδειγματικές εξηγήσεις

Το παρόν κεφάλαιο επικεντρώνεται στις «αντιπαραδειγματικές» ή «αντιφατικές» εξηγήσεις (counterfactual explanations - CEs) και αναλύει τη μετάβασή τους από τη θεωρία της αιτιότητας έως τις σύγχρονες εφαρμογές τους στη μηχανική μάθηση. Η εργασία αυτή στοχεύει κυρίως στην αξιοποίηση εξωτερικής γνώσης με σκοπό την απόκτηση ερμηνειών, είτε για νευρωνικά μοντέλα είτε για αυτόματες μεθόδους αξιολόγησης. Οι Γράφοι Γνώσης (Knowledge Graphs – KGs) αποτελούν ιδανικά εργαλεία για αυτόν τον στόχο, καθώς προσφέρουν δομημένη, ερμηνεύσιμη και εξηγητική αναπαράσταση γνώσης, καθιστώντας τα κατάλληλα όταν απαιτείται ιχνηλασιμότητα, αξιοπιστία και σαφήνεια σημασιολογικού περιεχομένου.

Επειδή το εύρος των πιθανών μεθόδων εξήγησης είναι μεγάλο, εστιάζουμε κυρίως στις αντιπαραδειγματικές εξηγήσεις, οι οποίες επιχειρούν να απαντήσουν στο ερώτημα: «Τι πρέπει να αλλάξει σε μια δεδομένη περίπτωση ώστε το αποτέλεσμα να είναι διαφορετικό;». Το σύνολο των απαραίτητων αλλαγών προς το εναλλακτικό αποτέλεσμα αποκαλύπτει, με αιτιοκρατικό τρόπο, γιατί το σύστημα έλαβε μια συγκεκριμένη απόφαση. Στο πλαίσιο αυτό, τα χαρακτηριστικά εισόδου δρουν ως «αιτία» και η έξοδος του συστήματος ως «αποτέλεσμα». Οι αντιπαραδειγματικές εξηγήσεις ευθυγραμμίζονται με τον τρόπο που σκέφτονται οι άνθρωποι, μέσω υποθετικών σεναρίων (what-if), στα οποία τροποποιούνται ελεγχόμενες μεταβλητές για να προκύψει ένα εναλλακτικό ενδεχόμενο.

Το θεμέλιο της αντιπαραδειγματικής συλλογιστικής βρίσκεται στην αιτιακή συμπερασματολογία, η οποία επιδιώκει όχι απλώς να εντοπίσει συσχετίσεις αλλά να ανακαλύψει αιτιακές εξαρτήσεις. Τα δομικά αιτιακά μοντέλα (Structured Causal Models - SCMs) του Judea Pearl [259] εισήγαγαν την έννοια της «παρέμβασης»: της στοχευμένης μεταβολής μιας μεταβλητής ώστε να παρατηρηθούν οι επακόλουθες αλλαγές. Στην αιτιακή ιεραρχία, η αντιπαραδειγματική σκέψη βρίσκεται στην κορυφή, πάνω από τον συσχετιστικό και τον παρεμβατικό συλλογισμό. Ενώ ο συσχετισμός περιγράφει τι ισχύει και η παρέμβαση τι θα συμβεί αν πράξουμε κάτι, ο αντιπαραδειγματικός συλλογισμός εξετάζει τι θα είχε συμβεί αν κάποια πτυχή της πραγματικότητας ήταν διαφορετική. Έτσι, απαιτεί βαθύτερη κατανόηση των υποκείμενων μηχανισμών. Στα SCMs, η παρέμβαση ισοδυναμεί με την αντικατάσταση μίας εξίσωσης από μια σταθερή τιμή, δηλαδή τη δημιουργία ενός «εναλλακτικού κόσμου». Οι αντιπαραδειγματικές εξηγήσεις στη μηχανική μάθηση υιοθετούν την ίδια λογική: αλλαγές στα χαρακτηριστικά εισόδου ισοδυναμούν με παρεμβάσεις στον μηχανισμό λήψης αποφάσεων του μοντέλου. Έτσι, οι αντιπαραδειγματικές εξηγήσεις γεφυρώνουν τη σκέψη τύπου «τι θα γινόταν αν» με τα προβλεπτικά μοντέλα μηχανικής μάθησης.

Χαρακτηριστικά των αντιπαραδειγματικών εξηγήσεων

Μια αντιπαραδειγματική εξήγηση στο πλαίσιο ενός μοντέλου M ορίζεται ως η ελάχιστη δυνατή αλλαγή στα χαρακτηριστικά εισόδου x που οδηγεί σε διαφορετική πρόβλεψη. Ο βαθμός ελαχιστοποίησης εξαρτάται από τις ανάγκες της εκάστοτε εφαρμογής [100]. Για να είναι μια αντιπαραδειγματική εξήγηση χρήσιμη πρέπει να είναι πειστική, δηλαδή να αντιστοιχεί σε πραγματικό σημείο του χώρου δεδομένων, να είναι ελάχιστη ως προς τις απαραίτητες μεταβολές, να παρέχει ποικιλία όταν υπάρχουν πολλές εναλλακτικές, να είναι εφαρμόσιμη στην πράξη, να είναι αραιή ώστε να μεταβάλλονται λίγα χαρακτηριστικά και να διατηρεί εγγύτητα και εσωτερική συνοχή με το αρχικό δείγμα. Ο συνδυασμός αυτών των ιδιοτήτων δεν είναι πάντοτε εύκολος: η εξασφάλιση ρεαλισμού μπορεί να συγκρούεται με την ελαχιστοποίηση της αλλαγής και αντιστρόφως. Έτσι, οι σύγχρονες μέθοδοι διατυπώνουν τη δημιουργία αντιπαραδειγμάτων ως πρόβλημα βελτιστοποίησης, όπου συνδυάζονται ταυτόχρονα η ελαχιστοποίηση της απόστασης και η μεγιστοποίηση της πειστικότητας.

Ένα ακόμη σημείο αφορά την πρόσβαση στο μοντέλο. Τα αντιπαραδείγματα είναι ιδιαίτερα πρακτικά επειδή είναι κατά κανόνα μοντελο-αγνωστικά (model agnostic): το μοντέλο υπό εξήγηση μπορεί να είναι «μαύρο κουτί», είτε επειδή δεν είναι εκ φύσεως ερμηνεύσιμο (π.χ. νευρωνικά δίκτυα), είτε επειδή δεν έχουμε πρόσβαση στον εσωτερικό τους μηχανισμό (π.χ. API μοντέλων όπως το ChatGPT [244]). Επομένως, τα αντιπαραδείγματα προσφέρουν έναν ενιαίο τρόπο εξήγησης που δεν απαιτεί εξειδικευμένη τεχνική για κάθε διαφορετικό μοντέλο. [204, 78, 62, 203, 320, 154, 67, 69, 208, 316, 202]

Άλλες τεχνικές εξηγησιμότητας

Η εξηγήσιμη τεχνητή νοημοσύνη περιλαμβάνει πολλές τεχνικές, όπως οι μέθοδοι απόδοσης χαρακτηριστικών, τα προσεγγιστικά μοντέλα, οι εξηγήσεις βασισμένες σε πρωτότυπα ή παραδείγματα, οι χάρτες προσοχής και οι συμβολικές μέθοδοι. Οι αντιπαραδειγματικές εξηγήσεις διακρίνονται επειδή δεν περιγράφουν απλώς την επιρροή των χαρακτηριστικών, αλλά προσδιορίζουν συγκεκριμένες αλλαγές που θα οδηγούσαν σε διαφορετική απόφαση. Οι μέθοδοι απόδοσης χαρακτηριστικών, όπως οι LIME [285], SHAP [198] και Integrated Gradients [328], προσδιορίζουν ποια χαρακτηριστικά συνέβαλαν περισσότερο στην πρόβλεψη αλλά όχι πώς θα αλλάξει η απόφαση αν μεταβληθεί κάποιο χαρακτηριστικό. Τα προσεγγιστικά μοντέλα επιχειρούν να «μιμηθούν» το αρχικό μοντέλο με απλούστερες δομές, όμως η πιστότητά τους δεν είναι εγγυημένη. Οι εξηγήσεις βασισμένες σε πρωτότυπα παρουσιάζουν παραδείγματα που μοιάζουν με την είσοδο, αλλά δεν δείχνουν πώς θα επιτυγχανόταν ένα διαφορετικό αποτέλεσμα. Οι μέθοδοι προσοχής και οι χάρτες κλίσεων αναδεικνύουν τις περιοχές που επηρέασαν την απόφαση, χωρίς όμως να έχουν αιτιακό χαρακτήρα. Οι συμβολικές μέθοδοι, όπως κανόνες απόφασης, είναι ερμηνεύσιμες αλλά δυσκολεύονται να συλλάβουν πολύπλοκα όρια απόφασης.

Σε αντίθεση με αυτά, οι αντιπαραδειγματικές εξηγήσεις παρέχουν ένα αιτιακό και εποικοδομητικό εργαλείο κατανόησης: εντοπίζουν συγκεκριμένα σημεία όπου η πρόβλεψη του μοντέλου αλλάζει και αποτυπώνουν καθαρά τι θα έπρεπε να μεταβληθεί για να προκύψει ένα διαφορετικό αποτέλεσμα. Επομένως, συμπληρώνουν όλες τις άλλες κατηγορίες εξηγήσεων και ενισχύουν την κατανόηση του μοντέλου σε επίπεδο άμεσης δράσης.

Αναγκαιότητα της σημασιολογίας στην εξηγησιμότητα

Παρά την πολυετή έρευνα στην εξηγησιμότητα, έχει επισημανθεί ότι πολλές μέθοδοι, ιδίως όταν βασίζονται αποκλειστικά σε χαμηλού επιπέδου μεταβολές όπως οι επιθέσεις (adversarials), αν και αλλάζουν την πρόβλεψη ενός μοντέλου, δεν παράγουν ανθρώπινα κατανοητές εξηγήσεις. Παρότι μια μικροσκοπική αλλαγή ενός pixel μπορεί να αλλάξει την πρόβλεψη ενός ταξινομητή εικόνας, δεν εξηγεί «γιατί» επήλθε η αλλαγή. Αυτό αναδεικνύει την αναγκαιότητα της σημασιολογικής πληροφορίας για ουσιαστικές εξηγήσεις. Οι σημασιολογικές επεμβάσεις, όπως η αλλαγή ενός αντικειμένου στη σκηνή, είναι ευθυγραμμισμένες με την ανθρώπινη κατανόηση και αποτελούν πράγματι ερμηνεύσιμες αντιφατικές παρεμβάσεις.

Η πρόσφατη βιβλιογραφία δείχνει ότι για να είναι ουσιαστικά τα αντιπαραδείγματα, απαιτείται πρόσβαση σε δομημένη γνώση [32]. Οι Γράφοι Γνώσης, οι οντολογίες και άλλες δομημένες μορφές γνώσης επιτρέπουν αντιφατικές επεμβάσεις που είναι σημασιολογικά θεμελιωμένες. Όταν η γνώση αυτή ενσωματώνεται σε διανυσματικές αναπαραστάσεις, είτε μέσω embeddings γραφημάτων είτε μέσω σημασιολογικών κωδικοποιητών, οι παραγόμενες αντιπαραδειγματικές εξηγήσεις συνδυάζουν μαθηματική ευκολία και εννοιολογική σαφήνεια. Αυτό ευθυγραμμίζεται και με την ανθρώπινη νόηση: οι άνθρωποι δεν σκέφτονται σε επίπεδο pixel αλλά σε επίπεδο εννοιών και σχέσεων.

Για τη δημιουργία σημασιολογικών αντιφατικών εξηγήσεων, είναι απαραίτητος ο προσδιορισμός αποστάσεων μεταξύ εννοιών, καθώς η έννοια της ελαχιστότητας εξαρτάται από έναν ουσιαστικό και συνεπή ορισμό εγγύτητας. Αυτές οι αποστάσεις μπορούν να προέλθουν από ταξινομικές δομές (π.χ. μήκος μονοπατιού σε οντολογίες), από κατανεμημένη σημασιολογία (π.χ. ομοιότητα ως προς το διανυσματικό χώρο embeddings) ή από embeddings γραφημάτων που κωδικοποιούν τη δομή των σχέσεων. Με τον καθορισμό τέτοιων αποστάσεων είναι δυνατόν να πραγματοποιηθούν σημασιολογικές παρεμβάσεις που αντιστοιχούν σε ρεαλιστικές αλλαγές στον πραγματικό κόσμο, παράγοντας έτσι αντιφατικά που όχι μόνο αλλάζουν την πρόβλεψη του μοντέλου αλλά και αντικατοπτρίζουν τον τρόπο με τον οποίο οι άνθρωποι νοηματοδοτούν εναλλακτικές καταστάσεις πραγμάτων.

1.4 Ρητές βάσεις γνώσης για εξηγησιμότητα

Τα γράφηματα γνώσης αποτελούν θεμελιώδες εργαλείο για την εξηγησιμότητα, επειδή διαθέτουν ντετερμινιστική δομή, ξεκάθαρη εννοιολογική οργάνωση και καλά ορισμένες σημασιολογικές αποστάσεις μεταξύ εννοιών [320, 207, 204, 67, 69, 203]. Αυτά τα χαρακτηριστικά τα καθιστούν εξαιρετικά κατάλληλα για αντιπαραδειγματική συλλογιστική, καθώς υποδεικνύουν με ακρίβεια ποιες εννοιολογικές μεταβολές είναι ικανές να αλλάξουν το τελικό αποτέλεσμα ενός μοντέλου. Ανάλογα με τον βαθμό εκφραστικότητας που απαιτείται, οι εξηγήσεις μπορούν να βασίζονται μόνο σε σύνολα εννοιών ή να επεκτείνονται σε πλήρη σημασιολογικά γραφήματα που ενσωματώνουν και τις μεταξύ τους σχέσεις, γεγονός που αν και αυξάνει το υπολογιστικό κόστος, επιτρέπει πιο λεπτομερείς και δομημένες επεξηγήσεις.

Η παραγωγή αντιπαραδειγματικών επεξηγήσεων μπορεί να επιτευχθεί μέσω βέλτιστης αντιστοίχισης εννοιών χρησιμοποιώντας διμερή γραφήματα, όπου αλγόριθμοι όπως ο Hungarian [160] ή οι μέθοδοι δημοπρασίας υπολογίζουν το ελάχιστο κόστος υποκατάστασης μεταξύ δύο συνόλων εννοιών. Όταν οι σχέσεις μεταξύ εννοιών είναι κρίσιμες, το πρόβλημα επεκτείνεται στο ταίριασμα πλήρων γραφημάτων. Σε αυτήν την περίπτωση, η ακριβής αντιστοίχιση επιδιώκει δομική ισομορφία, ενώ η προσεγγιστική αντιστοίχιση αναζητά τις βέλτιστες αντιστοιχίες ακόμη και όταν τα γραφήματα διαφέρουν σε μέγεθος, περιέχουν θόρυβο ή παρουσιάζουν μερική επικαλυπτόμενη πληροφορία. Οι μέθοδοι αυτές στηρίζονται σε φασματικές τεχνικές, πιθανοτικές προσεγγίσεις, μετρικές επεξεργασίας γραφημάτων ή ακόμη και νευρωνικά μοντέλα που μαθαίνουν αντιστοιχίες βάσει δεδομένων.

Η διαδικασία ταιριάσματος μπορεί επίσης να διατυπωθεί ως τετραγωνικό ή γραμμικό πρόβλημα ανάθεσης [152], επιτρέποντας την ανάπτυξη πιο αποδοτικών και επεκτάσιμων αλγορίθμων για μεγάλες ή πολύπλοκες δομές. Σε σημασιολογικά περιβάλλοντα, το ταίριασμα πρέπει να συνδυάζει δομική ομοιότητα και σημασιολογική συνοχή, ώστε οι αντιπαραδειγματικές επεξηγήσεις να προκύπτουν ως ουσιαστικές εννοιολογικές τροποποιήσεις που διατηρούν νόημα και μπορούν να ερμηνευθούν από τον άνθρωπο.

Ανεξάρτητα από την προσέγγιση που ακολουθείται, υπάρχουν ορισμένα βασικά ζητούμενα για τη δημιουργία αντιπαραδειγματικών επεξηγήσεων βασισμένων σε γραφήματα. Πρώτον, απαιτείται βέλτιστη επιλογή εννοιολογικών αλλαγών, ώστε οι επεμβάσεις να είναι όσο το δυνατόν μικρότερες αλλά και αποτελεσματικές. Δεύτερον, οι αλλαγές πρέπει να είναι ελέγξιμες, επιτρέποντας στους χρήστες να κατανοήσουν και να αναπαράγουν τις προτεινόμενες τροποποιήσεις. Τρίτον, η υπολογιστική αποδοτικότητα είναι κρίσιμη, δεδομένης της ανάπτυξης ολοένα και πιο μεγάλων μοντέλων και πλούσιων εννοιολογικών χώρων. Τέλος, οι διαδρομές εννοιολογικής υποκατάστασης πρέπει να είναι πλήρως διαφανείς, αποκαλύπτοντας όχι μόνο το αποτέλεσμα αλλά και τη λογική που οδήγησε στην επιλογή κάθε εννοιολογικής αλλαγής. Μέσα από αυτή την οπτική, τα γραφήματα γνώσης αποτελούν μια ιδιαίτερα ισχυρή υποδομή για την ανάπτυξη εύρωστων, ερμηνεύσιμων και αξιοποιήσιμων αντιπαραδειγματικών επεξηγήσεων.

1.4.1 Επιτάχυνση εξαγωγής σημασιολογικών αντιπαραδειγμάτων

Παρότι αλγόριθμοι όπως ο Hungarian και η Απόσταση Επεξεργασίας Γράφου (Graph Edit Distance GED) παρέχουν ντετερμινιστικές εγγυήσεις ορθότητας στην παραγωγή σημασιολογικών αντιπαραδειγμάτων, κλιμακώνονται εξαιρετικά αργά σε πραγματικών διαστάσεων εφαρμογές, όπου τα λεξιλόγια εννοιών αριθμούν χιλιάδες λέξεις και οι γράφοι σκηνής περιέχουν πυκνή σχεσιακή πληροφορία. Για την αντιμετώπιση αυτής της υπολογιστικής πολυπλοκότητας, προτείνουμε ένα ενιαίο πλαίσιο που επιταχύνει την αναζήτηση σημασιολογικών αντιπαραδειγμάτων μέσω Νευρωνικών Δικτύων Γράφων (Graph Neural Networks - GNNs). Η προσέγγισή μας αξιοποιεί τη δύναμη των GNNs να χαρτογραφούν διακριτές δομές σε συνεχείς χώρους, επιτρέποντας προσεγγιστικό ταίριασμα με υψηλή αποδοτικότητα και διατήρηση της σημασιολογικής συνοχής.

Αρχικά ευθυγραμμίζουμε τις εξαγόμενες έννοιες με δομημένες βάσεις γνώσης όπως το WordNet [226], ώστε να υπολογίσουμε σημασιολογικές αποστάσεις μέσω path-based similarity ή εννοιολογικών embedding μοντέλων (π.χ. SBERT [282]). Οι αποστάσεις αυτές λειτουργούν ως βάρη ακμών στο διμερή γράφο που χρησιμοποιούμε για εννοιολογικές αντικαταστάσεις. Αποδεικνύουμε πειραματικά ότι τα GNNs, συμπεριλαμβανομένων των GCN [149], GAT [351], GIN [383], μπορούν να κωδικοποιήσουν αποτελεσματικά τοπική και σχεσιακή πληροφορία, παράγοντας κόμβικες αναπαραστάσεις που προσεγγίζουν βέλτιστες αντιστοιχίες με σημαντικά χαμηλότερη πολυπλοκότητα σε σχέση με τους παραδοσιακούς συνδυαστικούς αλγόριθμους.

Χρήση GNNs για εννοιολογικά αντιπαραδείγματα

Μέσω της ένωσης εννοιών με ρητούς γράφους γνώσης, προτείνουμε ένα μοντέλο βασισμένο σε GNNs για την παραγωγή εννοιολογικών αντιπαραδειγμάτων σε NLP συστήματα. Σε έναν διμερή γράφο όπου οι αρχικές και οι πιθανές εννοιολογικές αντικαταστάσεις σχηματίζουν δύο διακριτά σύνολα κόμβων, το μοντέλο μας εφαρμόζει μια αρχιτεκτονική encoder–convolution–decoder με στόχο την προσεγγιστική επίλυση του προβλήματος τετραγωνικής αντιστοίχησης (RLAP). Αφού διαχυθεί σημασιολογική πληροφορία σε ολόκληρο τον γράφο μέσω κόμβικων και ακμιακών συνελίξεων, ένας beam-search μηχανισμός επιλέγει τις πιο εύλογες αντικαταστάσεις που μπορούν να προκαλέσουν αναστροφή ετικέτας του εκάστοτε κειμένου. Αξιολογούμε τη μέθοδο στα σύνολα δεδομένων IMDB και 20 Newsgroups, συγκρίνοντας με τις προηγούμενες τεχνικές MiCE [290] και Polyjuice [372]. Τα αποτελέσματα δείχνουν ότι το GNN μοντέλο μας επιτυγχάνει καλύτερη φυσικότητα (fluency), μεγαλύτερη σημασιολογική εγγύτητα, ανταγωνιστικούς ρυθμούς αναστροφής ετικέτας και εντυπωσιακή μείωση χρόνου εκτέλεσης—έως και 9–15× σε σχέση με το ντετερμινιστικό RLAP και 40–50× σε σχέση με το MiCE.

Χρήση GNNs για αντιπαραδείγματα γράφων

Επεκτεινόμενοι πέρα από μεμονωμένες έννοιες, εξετάζουμε γράφους σκηνής όπου οι αντιπαραδειγματικές επεξηγήσεις απαιτούν εύρεση κρισίμων σχεσιακών αλλαγών (π.χ. από “safe” σε “not safe”). Εδώ το πρόβλημα αντιστοίχισης γράφων είναι κεντρικό. Μελετάμε επιβλεπόμενα GNNs (GCN, GAT, GIN) σε σιαμαία αρχιτεκτονική ώστε να προσεγγίσουμε τη GED, παρακάμπτοντας την αλγοριθμική της πολυπλοκότητα [67, 69, 39]. Με χρήση των συνόλων δεδομένων CUB και Visual Genome, δείχνουμε ότι τα embeddings γράφων διατηρούν σημασιολογική δομή και επιτρέπουν γρήγορη ανάκτηση αντιπαραδειγμάτων μέσω ομοιότητας συνημιτόνου (cosine similarity), χωρίς επαναλαμβανόμενο υπολογισμό GED.

Συνολικά, τα συμπεράσματά μας αποδεικνύουν ότι τα GNNs μπορούν να αντικαταστήσουν αποτελεσματικά τις παραδοσιακές συνδυαστικές μεθόδους για τη δημιουργία σημασιολογικών αντιπαραδειγμάτων σε κείμενο και εικόνες. Με τη δύναμη της σχεσιακής επαγωγικής μάθησης και της ενσωμάτωσης γνώσης, τα GNNs επιτρέπουν πραγματικού χρόνου και ερμηνεύσιμες αντιπαραδειγματικές εξηγήσεις σε μεγάλης κλίμακας δομές.

1.4.2 Μη ελάχιστες παρεμβάσεις

Η αντιπαραδειγματική συλλογιστική διατυπώνεται παραδοσιακά με βάση την αρχή της ελάχιστης παρέμβασης: μεταβάλλεται το ελάχιστο δυνατό τμήμα της εισόδου ώστε να προκληθεί αλλαγή στο αποτέλεσμα, απομονώνοντας έτσι τη μεταβλητή με αιτιακή σημασία. Παρότι η αρχή αυτή είναι θεωρητικά κομψή, αποδεικνύεται υπερβολικά περιοριστική όταν εφαρμόζεται σε υψηλού επιπέδου σημασιολογικές αναπαραστάσεις όπως αυτές που χρησιμοποιούνται σε γλωσσικά ή πολυτροπικά μοντέλα. Οι ελάχιστες παρεμβάσεις διερευνούν κυρίως τοπικές ευαισθησίες και αδυνατούν να αποκαλύψουν βαθύτερες εννοιολογικές δομές ή αδυναμίες των μοντέλων.

Για τον λόγο αυτό εισάγουμε σημασιολογικά μη-ελάχιστες παρεμβάσεις [204], δηλαδή σκόπιμες αντικαταστάσεις που επιφέρουν κατηγορικές ή αντιθετικές αλλαγές στο νόημα, όπως αντιώνυμα, αναστροφές σχέσεων ή εννοιολογικά διακριτά γνωρίσματα. Τέτοιες παρεμβάσεις λειτουργούν ως διαγνωστικός μηχανισμός εννοιολογικής συνάφειας: αναδεικνύουν πότε ένα μοντέλο ανταποκρίνεται σε ουσιαστικές σημασιολογικές διαφοροποιήσεις και πότε παρουσιάζει «σημασιολογική άγνοια», παραμένοντας αμετάβλητο ακόμη και υπό μεγάλες αλλαγές στο νόημα. Επιπλέον, φανερώνουν έναν κρίσιμο συμβιβασμό μεταξύ ευρωστίας και παθολογίας: ενώ τα μοντέλα πρέπει να είναι ανθεκτικά σε άσχετες ή μικρές τροποποιήσεις, η αμεταβλησία απέναντι σε μεγάλες σημασιολογικές αλλαγές υποδηλώνει αποτυχία διακριτικής κατανόησης, με άμεσες συνέπειες για την αξιοπιστία και την πρακτική τους χρήση.

Έτσι, οι μη-ελάχιστες παρεμβάσεις αποτελούν δοκιμασία πίεσης για τη σημασιολογική θεμελίωση των μοντέλων, εξετάζοντας αν αυτά κωδικοποιούν όχι μόνο επιφανειακές γλωσσικές κανονικότητες αλλά και τη δομή αντίθεσης και κατευθυντικότητας που ενυπάρχει στις έννοιες. Παράλληλα, το πλαίσιο αυτό αμφισβητεί τις παραδοσιακές αντιλήψεις περί ευρωστίας, διαχωρίζοντας την ορθή σταθερότητα από την επιβλαβή σημασιολογική αναλγησία. Σε εργασίες κατάταξης—όπως η εννοιολογική ομοιότητα και η ανάκτηση εικόνων μέσω κειμένου—οι μη-ελάχιστες παρεμβάσεις αποκαλύπτουν τον τρόπο με τον οποίο μεταβάλλεται η σειρά κατάταξης υπό αλλαγές στο νόημα, υποδεικνύοντας την ευαισθησία ή την αδιαφορία του μοντέλου. Φωτίζουν επίσης πτυχές παραγνωρισμένες από τις τυπικές αντιπαραδειγματικές προσεγγίσεις, όπως η μη-ευαισθησία σε μη-αιτιακά γνωρίσματα, η εξάρτηση από στρεβλές συσχετίσεις, η ευθραυστότητα σε σημασιολογικές μεταβολές και οι διαστάσεις δικαιοσύνης που συνδέονται με αμετάβλητες αποφάσεις σε προστατευμένα γνωρίσματα.

Με βάση αυτές τις ιδέες, αναπτύσσουμε ένα ανεξάρτητο από το μοντέλο πλαίσιο ερμηνευσιμότητας για μοντέλα σημασιολογικής ομοιότητας. Στο πλαίσιο πολυτροπικών διαδικασιών ανάκτησης, αντιμετωπίζουμε την ανάκτηση εικόνων από κείμενο ως μονοτροπική διαδικασία ενσωμάτωσης και κατάταξης: κάθε εικόνα εκπροσωπείται από ένα σύνολο περιγραφικών προτάσεων, ενώ μετασχηματιστές σημασιολογικής ομοιότητας παράγουν ενσωματώσεις για τόσο το ερώτημα όσο και τις περιγραφές. Η απόδοση αξιολογείται μέσω της κατάταξης των ενσωματώσεων με βάση την συνημιτονική ομοιότητα, και τα σφάλματα καταγράφονται κάθε φορά που η κορυφαία επιλογή δεν αντιστοιχεί στη σωστή εικόνα.

Συνολικά, η προσέγγισή μας καλείται να καταδείξει εάν οι σημασιολογικά μη-ελάχιστες παρεμβάσεις αποκαλύπτουν βαθύτερα επίπεδα εννοιολογικής ευαισθησίας στα μοντέλα σημασιολογικής ομοιότητας, φωτίζοντας πτυχές της συμπεριφοράς τους που παραμένουν αόρατες υπό την παραδοσιακή, ελάχιστης παρέμβασης αντιπαραδειγματική ανάλυση.

Τεχνικές λεπτομέρειες

Το προτεινόμενο πλαίσιο ενσωματώνει μη-ελάχιστες παρεμβάσεις προς δημιουργία αντιπαραδειγμάτων σε αυτή τη διαδικασία, ώστε να αποκαλυφθεί ποιες σημασιολογικές διαστάσεις επιδρούν ουσιαστικά στη συμπεριφορά κατάταξης. Με χρήση σημασιολογικά πλούσιων συνόλων δεδομένων όπως COCO [183], Flickr [396] και Visual Genome [155], κατασκευάζουμε αμιγώς γλωσσικά ζεύγη ερωτήματος-εγγράφου κειμένου και υλοποιούμε στοχευμένες παρεμβάσεις αξιοποιώντας δομές γνώσης όπως τη ιεραρχία του WordNet και μια προσαρμοσμένη ιεραρχία χρωμάτων. Εξερευνούμε διαφορετικές τεχνικές αναπαράστασης για εκτενείς περιγραφές, από μέσο όρο ενσωματώσεων προτάσεων έως αφαιρετική περίληψη. Η απόδοση αποτιμάται μέσω δεικτών Recall@k, MRR@k και της διάμεσης θέσης κατάταξης (median rank).

Η ανάλυσή μας βασίζεται σε ένα μεγάλο σύνολο δεδομένων 34K εικόνων από το Visual Genome ∩ COCO, το οποίο περιέχει τόσο συνολικές περιγραφές εικόνων όσο και λεπτομερείς επισημειώσεις περιοχών. Δημιουργούμε τρεις κατηγορίες μη-ελάχιστων παρεμβάσεων:

1. Αντικαταστάσεις με αντώνυμα για επίθετα
2. Μεταβολές χρωμάτων είτε εντός του συνόλου χρωμάτων του dataset είτε σε πλήρες φάσμα RGB
3. Μεταβολές μεγέθους μέσω αντωνυμικών υποκαταστάσεων

Παρατηρούμε ότι, ανεξάρτητα από τον τύπο παρέμβασης, το 70% των εικόνων αλλάζει θέση στην κατάταξη, ωστόσο οι καθολικές μετρικές κατάταξης παραμένουν **σχεδόν αμετάβλητες**, παρουσιάζοντας μόνο μικρές μειώσεις στο Recall@k και στο MRR. Η σταθερότητα αυτή υποδηλώνει ότι τα συστήματα ανάκτησης συχνά αγνοούν τα μεταβαλλόμενα γνωρίσματα και διατηρούν τις ταξινομήσεις βάσει της ταυτότητας των αντικειμένων, αποκαλύπτοντας υποκείμενες μεροληψίες αναπαράστασης που δεν ανιχνεύονται από τις κλασικές μετρικές.

Εξηγήσιμες μετρικές αξιολόγησης

Για την αντιμετώπιση της αδιαφάνειας αυτής, προτείνουμε μια νέα οικογένεια ερμηνεύσιμων μετρικών αξιολόγησης κατάταξης που ποσοτικοποιούν τις εννοιολογικές σχέσεις μεταξύ της ground-truth εικόνας και της ανακτηθείσας. Οι μετρικές αυτές περιλαμβάνουν:

1. Concept Agreement (CA), ποσοστό κοινών εντοπισμένων αντικειμένων.
2. Non-common Concept Similarity (NCS), το οποίο μετρά τη σημασιολογική εγγύτητα μη κοινών αντικειμένων μέσω μέγιστης αντιστοίχισης σε διμερές γράφημα βασισμένο σε WordNet synsets.
3. Concept Enumeration (CE), που συλλαμβάνει αποκλίσεις στις πληθικότητες αντικειμένων.
4. Size Disagreement (SD), το οποίο ποσοτικοποιεί διαφορές στα μεγέθη bounding boxes για κοινές κατηγορίες αντικειμένων μέσω ελάχιστης αντιστοίχισης κόστους.

Οι προτεινόμενες μετρικές αποκαλύπτουν με ακρίβεια τις σημασιολογικές διαστάσεις στις οποίες αποτυγχάνουν τα συστήματα ανάκτησης, αναδεικνύοντας μοτίβα σφαλμάτων και μεροληψίες που παραμένουν κρυμμένες κάτω από τις συμβατικές μετρικές ανάκλησης. Επιπλέον, η εκτεταμένη αξιολόγηση πολλών μοντέλων ομοιότητας δείχνει ότι τα paraphrase-oriented μοντέλα—εκπαιδευμένα εντατικά σε λεζάντες εικόνων—υπερτερούν σταθερά, υπογραμμίζοντας τη σημασία της έκθεσης σε οπτικο-γλωσσικά δεδομένα.

Συνολικά, τα ευρήματά μας δείχνουν ότι τα σύγχρονα συστήματα σημασιολογικής αντιστοίχισης κειμένου-σε-εικόνα (text to image retrieval) παρουσιάζουν μεγάλη **αμεταβλητότητα** ως προς ουσιαστικές αλλαγές στα σημασιολογικά γνωρίσματα, γεγονός που εγείρει ερωτήματα σχετικά με την αξιοπιστία τους για λεπτομερή σημασιολογική συλλογιστική. Το προτεινόμενο ερμηνεύσιμο πλαίσιο αξιολόγησης προσφέρει μια συστηματική μεθοδολογία για την κατανόηση των αποτυχιών κατάταξης και τη βελτίωση μελλοντικών, πιο αξιόπιστων μοντέλων ανάκτησης.

1.5 Γνώση σε εξηγήσεις εικόνας-γλώσσας

Η ενσωμάτωση γνώσης σε εργασίες όρασης-γλώσσας (Vision-Language - VL) αποτελεί κρίσιμο βήμα για την επίτευξη πραγματικής πολυτροπικής κατανόησης και συλλογιστικής. Παρόλο που η εκπαίδευση σε μεγάλη κλίμακα έχει βελτιώσει την απόδοση σε ποικίλες εργασίες, αυτά τα μοντέλα συχνά βασίζονται στην αναγνώριση προτύπων αντί για πραγματική κατανόηση, συλλαμβάνοντας στατιστικές συσχετίσεις μεταξύ οπτικών και κειμενικών μορφών αλλά δυσκολεύοντας στη συλλογιστική για έννοιες και κοινή λογική.

Η εισαγωγή γνώσης γεφυρώνει αυτό το κενό, επιτρέποντας στα μοντέλα να αξιοποιούν προηγούμενη κατανόηση του κόσμου για να συναγάγουν πλαίσιο, πρόθεση και νόημα, όπως οι άνθρωποι. Δομημένες και μη δομημένες πηγές γνώσης, όπως γραφήματα γνώσης και αναπαραστάσεις από LLMs αντίστοιχα, επιτρέπουν στα συστήματα VL να ξεπερνούν την επιφανειακή αντιστοίχιση, βελτιώνοντας την ερμηνευσιμότητα, την ανθεκτικότητα και τη γενίκευση. Αυτό υποστηρίζει συλλογιστική πάνω σε άγνωστες έννοιες, συνθετική κατανόηση και μείωση προκαταλήψεων των δεδομένων μέσω σημασιολογικά ουσιαστικών συσχετίσεων.

Απάντηση οπτικών ερωτήσεων

Η απάντηση οπτικών ερωτήσεων (Visual Question Answering - VQA) είναι μια πολυτροπική εργασία που συνδυάζει την υπολογιστική όραση με την κατανόηση φυσικής γλώσσας, όπου ένα μοντέλο πρέπει να απαντήσει σε μια ερώτηση φυσικής γλώσσας σχετικά με μια εικόνα. Οι εργασίες VQA αξιολογούν την ικανότητα του μοντέλου να αναγνωρίζει αντικείμενα και σκηνές, να κατανοεί σχέσεις και ιδιότητες, να εκτελεί συλλογισμούς και να αξιοποιεί γνώση κοινής λογικής.

Συνήθη συστήματα VQA ενσωματώνουν οπτικούς κωδικοποιητές (π.χ. CNNs ή Vision Transformers) με γλωσσικούς κωδικοποιητές (π.χ. BERT ή LLMs), χρησιμοποιώντας μηχανισμούς σύντηξης όπως cross-attention. Οι γλωσσικές προκαταλήψεις εμφανίζονται όταν οι ερωτήσεις μόνες τους προβλέπουν την απάντηση, ενώ οι οπτικές προκαταλήψεις προκύπτουν από τη συσχέτιση συγκεκριμένων οπτικών χαρακτηριστικών με ορισμένες απαντήσεις λόγω επαναλαμβανόμενων μοτίβων στα δεδομένα. Αυτές οι προκαταλήψεις οδηγούν σε «shortcut learning», όπου τα μοντέλα εκμεταλλεύονται τις τακτικότητες των δεδομένων αντί να εκτελούν πραγματική πολυτροπική συλλογιστική.

Ακολουθούμε μια προσέγγιση δημιουργίας αντιπαραδειγμάτων ερωτήσεων με χρήση εξωτερικής γνώσης και παρατηρούμε τις εξαγόμενες απαντήσεις [320]. Μέσω αυτού του πειραματισμού, παρατηρούμε ότι όλες οι εννοιολογικές παραλλαγές οδηγούν σε μείωση της ακρίβειας, με τις αντικαταστάσεις ουσιαστικών να προκαλούν τη μεγαλύτερη πτώση, υπογραμμίζοντας τον κεντρικό τους ρόλο στη σημασία της ερώτησης. Ποιοτικές παρατηρήσεις δείχνουν ότι το VQA μοντέλο χειρίζεται καλά την παραμετροποίηση επιθέτων, όπως η αλλαγή του “hot” στο “hot dog”, αλλά δυσκολεύεται με σπάνια συνώνυμα επιθέτων. Οι παραλλαγές ρημάτων γενικά αντιμετωπίζονται σωστά, ενώ οι αντικαταστάσεις ουσιαστικών δείχνουν περιορισμούς στη συλλογιστική με υπερώνυμα και ουσιαστικά με κοινό πρόγονο. Οι απαντήσεις είναι πιο σταθερές όταν ζωντανά όντα αντικαθίστανται από υπερώνυμα ή υποώνυμα, ενώ για τα ουσιαστικά με κοινό πρόγονο παρατηρούνται μικτά φαινόμενα, μερικές φορές αγνοώντας εντελώς την παραλλαγή.

Εξηγησιμότητα στη δημιουργία εικόνας από κείμενο

Οι αντιπαραδειγματικές εξηγήσεις προσφέρουν μια μεθοδική απάντηση στο ερώτημα: «Τι πρέπει να εισαχθεί, διαγραφεί ή αντικατασταθεί εννοιολογικά για να μεταβούμε από τα παραγόμενα δεδομένα στα πραγματικά;» Με την εφαρμογή ελάχιστων εννοιολογικών τροποποιήσεων, ένα σύνολο εννοιών πηγής S (από παραγόμενη εικόνα) μετασχηματίζεται σε σύνολο στόχο T (από την πραγματικότητα), διορθώνοντας τη σημασιολογία των παραγόμενων εικόνων. [203]

Οι ιεραρχίες εννοιών και οι αποστάσεις εννοιών παρέχουν μετρήσιμη δομή για τη διαφοροποίηση των εννοιών, είτε χρησιμοποιώντας εξωτερικές βάσεις γνώσης όπως το WordNet, είτε χειροποίητες ιεραρχίες. Ορίζονται τρεις βασικές λειτουργίες τροποποίησης: Αντικατάσταση (Replacement - R), Διαγραφή (Deletion - D) και Εισαγωγή (Insertion - I), με ενσωμάτωση αποστάσεων για ελαχιστοποίηση κόστους. Το σύνολο αυτών των επεξεργασιών παράγει την Απόσταση Επεξεργασίας Συνόλου Εννοιών (CSED), η οποία λειτουργεί ως μετρική απόστασης εννοιών εξαγόμενων από την εικόνα και από το κείμενο, ποσοτικοποιώντας το «Τι πρέπει να μεταβληθεί εννοιολογικά».

Η βασική μέθοδος χρησιμοποιεί ένα προεκπαιδευμένο black-box γενετικό μοντέλο, το οποίο λαμβάνει μια περιγραφή c (κείμενο ή σύμβολα) και παράγει εικόνα I. Οι έννοιες που εξάγονται από την εικόνα ορίζουν το σύνολο πηγής S, ενώ οι έννοιες από την περιγραφή ορίζουν το σύνολο στόχο T. Η εξαγωγή εννοιών γίνεται μέσω αυτοματοποιημένων μεθόδων όπως ανίχνευση αντικειμένων και σημασιολογική τμηματοποίηση. Η διαδρομή ελάχιστου κόστους $S \rightarrow T$ παρέχει τοπικές αντιφατικές εξηγήσεις, αναδεικνύοντας ποια εννοιολογικά στοιχεία παράγονται εσφαλμένα ή λείπουν.

Η προσέγγιση εφαρμόζεται σε δύο δύσκολες δημιουργικές εργασίες:

- Οπτικοποίηση Ιστοριών (Story Visualization - SV): Διαδοχική παραγωγή εικόνων βάσει διαδοχικών κειμενικών συνθηκών. Η αξιολόγηση απαιτεί πιστότητα (όλα τα αντικείμενα και χαρακτηριστικά εμφανίζονται) και συνέπεια (τα αντικείμενα διατηρούνται στα πλαίσια). Ορίζουμε τις μετρικές Story Loss (SL) και Consistency Loss (CL). Τα συνολικά metrics (GSL, GCL) συνοψίζουν την απόδοση πολλαπλών ιστοριών και αναδεικνύουν συχνά σφάλματα ή ασυνεπή μοτίβα.
- Δημιουργία Σκηνών (Scene Generation - SG): Σύνθεση σύνθετων εικόνων με βάση την κειμενική περιγραφή c, όπου πολλά αντικείμενα με χαρακτηριστικά αλληλεπιδρούν.

Η ιεραρχική γνώση (WordNet) επιτρέπει την ακριβή σύγκριση εννοιών και υπολογισμό των προαναφερθέντων μετρικών, επιτρέποντας ακριβή αξιολόγηση μοντέρνων diffusion models όπως Stable Diffusion και Protogen.

Συνολικά, το πλαίσιο αντιπαραδειγματικών εννοιολογικών τροποποιήσεων παρέχει ποσοτικά metrics και ερμηνεύσιμες εξηγήσεις, επιτρέποντας τη συστηματική αξιολόγηση της πιστότητας και συνέπειας των παραγόμενων εικόνων, καθώς και την αναγνώριση τοπικών και παγκόσμιων μοτίβων σφάλματος.

Μετά την εκτέλεση των πειραμάτων προκύπτει ότι η αξιολόγηση βασισμένη σε έννοιες (conceptual evaluation) παρέχει πιο εξηγήσιμη εικόνα από τα κλασικά pixel-level metrics. Οι προτεινόμενες μετρικές δείχνουν τι χρειάζεται να αλλάξει σε μια παραγόμενη εικόνα για να ταιριάζει με την πραγματική κειμενική περιγραφή (μέσω εισαγωγής, διαγραφής, αντικατάστασης εννοιών). Μέσω ανάλυσης αυτών των συμπερασμάτων σε διάφορα δείγματα, αποκτάται μια γενική εικόνα για τα προβληματικά concepts που υπεραντιπροσωπεύονται ή υποεκπροσωπεύονται από το μοντέλο. Οι παρατηρούμενες μεροληψίες μπορούν να καθοδηγήσουν βελτιώσεις στα μοντέλα, είτε μέσω επανεκπαίδευσης είτε μέσω τροποποιήσεων στην αρχιτεκτονική. Το προτεινόμενο πλαίσιο δημιουργίας σημασιολογικών παρεμβάσεων είναι εύκολα προσαρμόσιμο σε άλλες αρχιτεκτονικές generative μοντέλων, θεωρώντας τα και ερμηνεύοντάς τα ως μαύρα κουτιά.

Εξηγησιμότητα στη δημιουργία κειμένου από εικόνα

Η αυτόματη περιγραφή εικόνων αποτελεί ένα από τα κεντρικά προβλήματα στη πολυτροπική τεχνητή νοημοσύνη, καθώς απαιτεί από τα μοντέλα να κατανοήσουν οπτικό περιεχόμενο και να το εκφράσουν σε φυσική γλώσσα. Παρά την εντυπωσιακή πρόοδο των σύγχρονων αρχιτεκτονικών, ένα κρίσιμο πρόβλημα παραμένει: οι ψευδαισθήσεις (hallucinations). Οι ψευδαισθήσεις εμφανίζονται όταν το μοντέλο δημιουργεί περιεχόμενο που είναι γλωσσικά πειστικό αλλά δεν υποστηρίζεται από την εικόνα, όπως αντικείμενα που δεν υπάρχουν, λανθασμένα χαρακτηριστικά ή ανύπαρκτες σχέσεις. Αυτό το φαινόμενο μειώνει την αξιοπιστία των συστημάτων, ιδίως σε εφαρμογές υψηλής σημασίας.

Η βιβλιογραφία διαχωρίζει τις ψευδαισθήσεις σε επιπέδου αντικειμένων και επιπέδου προτάσεων, με το CHAIR να αποτελεί το πρώτο μετρικό εργαλείο συστηματικής μέτρησης [288]. Ωστόσο, τα υπάρχοντα εργαλεία βασίζονται σε περιορισμένα λεξιλόγια, δεν ανιχνεύουν φαινόμενα όπως η υπερεξειδίκευση ή η ψευδαισθησιακή περιγραφή ρόλων, και συχνά αξιοποιούν LLMs τα οποία μπορεί και τα ίδια να παράγουν ψευδαισθήσεις. Παράλληλα, οι σημερινές μετρικές στερούνται εξηγησιμότητας: δεν παρέχουν οδηγίες για το πώς πρέπει να τροποποιηθεί μια λεζάντα ώστε να γίνει οπτικά πιστή.

Για να αντιμετωπίσουμε αυτά τα προβλήματα, επεκτείνουμε το πλαίσιο εννοιολογικών αντιπαραδειγματικών εξηγήσεων που περιγράψαμε προηγουμένως στην αξιολόγηση της περιγραφής εικόνων [202]. Μια λεζάντα παράγει ένα σύνολο εννοιών S, ενώ η εικόνα παράγει το σύνολο T (θεωρώντας τα υπάρχοντα annotations ως σύνολο πραγματικών εννοιών). Επιθυμούμε να μετασχηματίσουμε το $S \rightarrow T$ μέσω τριών τύπων διορθωτικών ενεργειών: Αντικατάσταση (Replacement - R), Διαγραφή (Deletion - D) και Εισαγωγή (Insertion - I). Ο μετασχηματισμός βασίζεται στη σημασιολογική ιεραρχία του WordNet, εξασφαλίζοντας μετρήσιμες σημασιολογικές αποστάσεις, νοηματικούς και γλωσσικά ορθούς μετασχηματισμούς και βέλτιστες επεμβάσεις με χρήση συντομότερων μονοπατιών και bipartite matching. Οι εισαγωγές δεν τιμωρούνται, καθώς συχνά οι λεζάντες είναι δικαιολογημένα συνοπτικές και αφαιρούν υπάρχουσες έννοιες προς διατήρηση της συντομίας.

Επιπλέον, εισάγουμε για πρώτη φορά αξιολόγηση ρόλων στο πλαίσιο ανίχνευσης ψευδαισθήσεων. Οι σχέσεις μεταξύ αντικειμένων (όπως χωρικές ή λειτουργικές αλληλεπιδράσεις) συχνά ψευδαισθησιακά περιγράφονται και αγνοούνται από προηγούμενες μεθόδους. Για να αξιολογήσουμε τέτοιες περιπτώσεις, εξετάζουμε τριάδες $(αντικείμενο_1, σχέση, αντικείμενο_2)$, οι οποίες εξάγονται από γράφους σκηνής της εικόνας και γράφους ανάλυσης της λεζάντας. Η απόσταση μεταξύ των δύο γράφων υπολογίζεται με προσεγγιστικό Graph Edit Distance (GED), παρέχοντας το ελάχιστο σύνολο επεμβάσεων που εξηγεί τη διαφορά.

Το προτεινόμενο πλαίσιο HalCECE παρέχει μια πλήρη, εξηγήσιμη ανάλυση ψευδαισθήσεων. Σε επίπεδο αντικειμένων αναγνωρίζει τα φαινόμενα:

- Διαγραφή (D): αντικείμενα παρόντα στη λεζάντα αλλά όχι στην εικόνα.
- Αντικατάσταση (R): λανθασμένα αντικείμενα χωρίς σημασιολογική ιεραρχική σχέση με τα σωστά.
- Υπερεξειδίκευση (O): όταν η λεζάντα χρησιμοποιεί πιο ειδική έννοια από την οπτική πραγματικότητα.

Υπολογίζουμε το συνολικό σφάλμα ως το άθροισμα των τριών αυτών φαινομένων, καθώς και το ποσοστό ψευδαισθήσεων, κανονικοποιώντας το άθροισμα ως προς το πλήθος των εννοιών του συνόλου S.

Επιπλέον, μετράμε σημασιολογική ομοιότητα (Wu–Palmer), υποεξειδίκευση, και γενικότητα περιγραφής. Σε επίπεδο ρόλων, υπολογίζουμε διορθώσεις τριάδων, αποκαλύπτοντας δομικά και σημασιολογικά λάθη που δεν ανιχνεύονται από αντικειμενοκεντρικές προσεγγίσεις.

Ένα ακόμη πρόβλημα που εξετάζει το HalCECE είναι η ανικανότητα των τυπικών μετρικών αξιολόγησης παραγωγής κειμένου από εικόνα να συμπεριλάβουν περιπτώσεις hallucinations. Συγκεκριμένα, οι γλωσσικές μετρικές όπως ROUGE, BLEU, Google BLEU, Mauve και η perplexity (PPL) αποτελούν καθιερωμένα εργαλεία για την αξιολόγηση παραγόμενου κειμένου. Παρότι η εκτεταμένη χρήση τους οφείλεται στην απλότητα και τη μακρόχρονη παρουσία τους στη βιβλιογραφία, συχνά αποτυγχάνουν να αποτυπώσουν την πραγματική σημασιολογική ποιότητα μιας λεζάντας, ιδίως όταν αυτή περιέχει φαινόμενα hallucinations. Οι μετρικές ROUGE και BLEU βασίζονται αποκλειστικά σε επικαλύψεις n-gram, επιβραβεύοντας επιφανειακή ομοιότητα και τιμωρώντας οποιαδήποτε απόκλιση από τα κείμενα αναφοράς, ανεξάρτητα από το αν η απόκλιση διατηρεί τη σημασιολογική ορθότητα. Έτσι, αντιμετωπίζουν το ίδιο μια μικρή σημασιολογική διαφοροποίηση, όπως η αντικατάσταση «γάτα» με «γατάκι», και ένα πλήρως εσφαλμένο υποκατάστατο, όπως η αντικατάσταση «γάτα» με «πλοίο». Η αδυναμία τους να διακρίνουν μεταξύ ελαφράς σημασιολογικής μετατόπισης και σοβαρών εννοιολογικών λαθών τις καθιστά ακατάλληλες για την αποτίμηση hallucinations. Ακόμη και οι μετρικές που δεν εξαρτώνται τόσο από την ακριβή διατύπωση, όπως η Mauve, αδυνατούν να ανιχνεύσουν πραγματολογικές ασυνέπειες μεταξύ κειμένου και εικόνας, αφού λειτουργούν σε επίπεδο κατανομών και όχι νοηματικής γείωσης. Αντίστοιχα, η perplexity μετρά αποκλειστικά τη γλωσσική ευχέρεια και τη βεβαιότητα του μοντέλου για την επόμενη λέξη, αγνοώντας πλήρως τη διασταύρωση με την οπτική πληροφορία.

Οι αδυναμίες αυτές γίνονται ιδιαίτερα εμφανείς όταν εξετάζουμε hallucinations, αφού η αξιολόγησή τους απαιτεί αυξημένη ευαισθησία στη σημασιολογική ακρίβεια και όχι στην επιφανειακή ομοιότητα. Τα αποτελέσματά μας δείχνουν ότι οι γλωσσικές μετρικές παράγουν ασυνεπείς και συχνά αντιφατικές ενδείξεις ανάμεσα στα μοντέλα· ένα μοντέλο μπορεί να επιτυγχάνει υψηλή βαθμολογία σε n-gram-based μετρικές αλλά να εμφανίζει από τις χαμηλότερες επιδόσεις στη Mauve, ή και το αντίστροφο, αποκαλύπτοντας έλλειψη συνοχής στη σημασιολογική τους κρίση. Αντιθέτως, το HalCECE, το επεξηγήσιμο εννοιολογικό πλαίσιο αξιολόγησης hallucinations που προτείνουμε, απομονώνει ρητά την πηγή της σημασιολογικής ασυμφωνίας, προσδιορίζοντας ποιες έννοιες έχουν υποστεί υποκατάσταση, διαγραφή, υπερεξειδίκευση ή υποεξειδίκευση, ενώ ταυτόχρονα προτείνει τι ακριβώς πρέπει να αλλάξει ώστε η λεζάντα να ευθυγραμμιστεί με το οπτικό περιεχόμενο.

Επιπλέον, οι γλωσσικές μετρικές απαιτούν την ύπαρξη λεζάντας αναφοράς, περιορίζοντας σημαντικά τη χρησιμότητά τους σε ρεαλιστικά ή zero-reference περιβάλλοντα, αντίθετα με το HalCECE που χρειάζεται μόνο ένα σύνολο μεμονωμένων εννοιών. Η ανάλυση συσχέτισης δείχνει ότι όλες οι γλωσσικές μετρικές εμφανίζουν εξαιρετικά χαμηλή συσχέτιση τόσο με τις μετρικές hallucinations αντικειμένων όσο και με εκείνες ρόλων που παράγει το HalCECE, επιβεβαιώνοντας εμπειρικά ότι δεν μπορούν να λειτουργήσουν ως αξιόπιστοι δείκτες παρουσίας ή έντασης hallucinations. Σε πολλές περιπτώσεις μάλιστα προσφέρουν παραπλανητική εικόνα: μια λεζάντα με υψηλή ευφράδεια και χαμηλή perplexity ενδέχεται να περιέχει σοβαρά hallucinations, ενώ μια λεζάντα με χαμηλή n-gram συμφωνία μπορεί να είναι απόλυτα πιστή στο περιεχόμενο της εικόνας.

Συνολικά, το HalCECE αποτελεί το πρώτο ντετερμινιστικό, εξηγήσιμο, εννοιολογικά συνεπές και role-aware πλαίσιο αξιολόγησης ψευδαισθήσεων για περιγραφές εικόνων. Προσφέρει όχι μόνο ακριβή μέτρηση αλλά και μια σαφή, βέλτιστη και σημασιολογικά τεκμηριωμένη πορεία διόρθωσης της λεζάντας, ενισχύοντας την αξιοπιστία των πολυτροπικών συστημάτων. Παράλληλα, τα ευρήματα δείχνουν ξεκάθαρα ότι, ενώ οι γλωσσικές μετρικές παραμένουν χρήσιμες για τη μέτρηση της ροής και του ύφους, είναι εκ φύσεως ακατάλληλες για την ανίχνευση hallucinations. Η αδιαφάνειά τους, η εξάρτησή τους από λεξικές επικαλύψεις και η πλήρης αδιαφορία τους για τη σημασιολογική γείωση καταδεικνύουν την ανάγκη για εξηγήσιμα, εννοιοκεντρικά πλαίσια όπως το HalCECE. Τα αποτελέσματα υπογραμμίζουν την ανεπαρκή φύση των παραδοσιακών γλωσσικών μετρικών στην αξιολόγηση δημιουργίας λεζάντας και αναδεικνύουν τη σημασία σημασιολογικά θεμελιωμένων εναλλακτικών.

1.6 Ενίσχυση συστημάτων κειμένου-εικόνας μέσω γνώσης

Το πρόν κεφάλαιο παρουσιάζει μια εκτενή επισκόπηση της ενίσχυσης γνώσης σε οπτικο-γλωσσικές εργασίες (VL). Τα συστήματα VL επωφελούνται σημαντικά από γράφους γνώσης (KGs), τα οποία βελτιώνουν τις δυνατότητες συλλογιστικής. Οι VL εργασίες ταξινομούνται σε διακριτικές (λογισμός σε εικόνες + κείμενο) και γενετικές (παραγωγή μιας μορφής από την άλλη). [206]

Ενίσχυση γνώσης στα συστήματα απάντησης οπτικών ερωτήσεων (K-VQA)

Datasets Παρουσιάζονται πολλά σύνολα δεδομένων που ενσωματώνουν εξωτερική γνώση:

- **Γενικά σύνολα VQA** (VQA, VQA-E, DAQUAR, COCO-QA), όπου η γνώση μπορεί να βοηθήσει αλλά δεν απαιτείται.
- **Εξειδικευμένα σύνολα** (KB-VQA, FVQA, KVQA, OK-VQA, Text-KVQA, Visual7W+KB, S3VQA, ZS-F-VQA, AQUA) σχεδιασμένα ώστε οι ερωτήσεις να απαιτούν εξωτερική γνώση: κοινού νου, πραγματολογική, οντοτήτων, αναγνώρισης κειμένου σε σκηνή ή γνώση τέχνης.

Τα datasets αναδεικνύουν προκλήσεις όπως η ανάγκη χρήσης όλων των modalities, η αποφυγή διαρροής απαντήσεων και η υποστήριξη zero-shot σεναρίων.

Μέθοδοι Οι μέθοδοι K-VQA οργανώνονται σε τρεις μεγάλες κατηγορίες:

- **Ρητή αναζήτηση γνώσης μέσω keywords.** Πρώιμες προσεγγίσεις εξήγαγαν χαρακτηριστικά από εικόνες, παρήγαγαν SPARQL queries προς DBpedia, και προέβλεπαν απαντήσεις μέσω RNNs, δίνοντας έμφαση αργότερα και στην εξηγησιμότητα.
- **Sequential models & embedding-based KG retrieval.** Με την έλευση των LSTMs, οι μέθοδοι βασίστηκαν σε ενσωματωμένους χώρους όπου εικόνα–ερώτηση και γεγονότα του KG χαρτογραφούνται σε κοινό embedding.
- **Μοντέλα Transformers.** Οι προσεγγίσεις χωρίζονται επιμέρους σε περιπτώσεις όπου οι Transformers χρησιμοποιούνται για γλωσσική κωδικοποίηση (BERT, SBERT, GPT-3, XLNet) για κατάταξη facts ή παραγωγή απαντήσεων, ή εναλλακτικά οπτικο-γλωσσικοί Transformers (ViLBERT, LXMERT, multimodal BERT) που ενσωματώνουν απευθείας εικόνα + κείμενο εμπλουτίζονται παράλληλα και στα δύο modalities.

Συνολικά, το πρόβλημα ισορροπεί μεταξύ ευρείας κάλυψης γνώσης και μείωσης θορύβου, ώστε το σύστημα να αντλεί μόνο τις πιο σχετικές πληροφορίες.

Ενίσχυση γνώσης στα συστήματα οπτικής συλλογιστικής (K-VR)

Datasets Στην περίπτωση του K-VR, διάφορα σύνολα δεδομένων υποστηρίζουν ρητό και δομημένο συλλογισμό. Το HVQR εισάγει πολυβηματικό επεξηγήσιμο reasoning, συνδυάζοντας scene graphs με γνώση από WebChild, ConceptNet και DBpedia· κάθε ερώτηση συνοδεύεται από ολόκληρη αλυσίδα reasoning, διακρίνεται σε first- και second-order ερωτήσεις και χωρίζει ερωτήματα που απαιτούν γνώση από αυτά που δεν τη χρειάζονται. Τα CLEVR και CLEVR-CoGenT προσφέρουν ελεγχόμενα, συνθετικά benchmark για συνθετικό λογισμό, με ρητά προγράμματα και ερωτήσεις φυσικής γλώσσας που δημιουργούνται από πρότυπα. Μια ευρύτερη οικογένεια datasets—Kandinsky Patterns, NLVR/NLVR2, Winoground, WinoGAViL—διευρύνει την αξιολόγηση σε γεωμετρικό λογισμό, αντιστοίχιση εικόνας–κείμενου και πολύπλοκες συλλογιστικές διαδικασίες.

Μέθοδοι Οι μέθοδοι K-VR περιλαμβάνουν LSTM-βασισμένα συστήματα που αξιοποιούν εξωτερική γνώση, όπως το KM-net, το οποίο αποσυνθέτει την ερώτηση σε υπο-ερωτήματα και τα δρομολογεί σε οπτικές και γνωσιακές μονάδες reasoning. Από την άλλη, μέθοδοι με εσωτερική γνώση κατασκευάζουν scene graphs και εκτελούν λογικούς τελεστές (“and”, “or”, “not”) πάνω σε περιοχές προσοχής, υποστηρίζοντας σύνθετο συλλογισμό. Η αξιολόγηση βασίζεται σε accuracy και αναλύσεις ως προς τα βήματα reasoning, τον τύπο ερώτησης και την ανάγκη εξωτερικής γνώσης, ενώ η ανάκτηση υποστηρικτικών facts αξιολογείται με ranking metrics όπως average recall.

Ενίσχυση γνώσης στα συστήματα κοινής οπτικής γνώσης (K-VCR)

Datasets Στο K-VCR, εξωτερικές πηγές commonsense γνώσης ενισχύουν τη βασική συλλογιστική των μοντέλων που καλούνται να απαντήσουν σε ανάλογες ερωτήσεις πάνω σε εικόνες. Το βασικό σύνολο δεδομένων που χρησιμοποιείται στη συγκεκριμένη εργασία, το VCR, περιλαμβάνει δύσκολες ερωτήσεις πολλαπλής επιλογής με επεξηγήσεις (rationales), οργανωμένες σε κατηγορίες όπως εξήγηση, δραστηριότητα, χρονική αλληλουχία, νοητικές καταστάσεις και υποθετικά σενάρια. Μία βασική τεχνική αναπαράστασης δεδομένων που χρησιμοποιείται στη βιβλιογραφία είναι το Visual Commonsense Graphs (VCG), η οποία επεκτείνει την κοινή γνώση του βασικού μοντέλου με μεγάλης κλίμακας παραγωγή χρονικών inferenced (πριν/μετά, προθέσεις), συνδεδεμένων με τις εικόνες.

Μέθοδοι Πληθώρα τεχνικών αξιοποιεί αρχιτεκτονικές Transformer για την επίλυση του βασικού προβλήματος συλλογιστικής, η οποία επιδέχεται την είσοδο εξωτερικής γνώσης. Αρκετές προσεγγίσεις βασίζονται στην αρχιτεκτονική του BERT, ενώ πιο εκειδικευμένες επεκτάσεις επιτυγχάνουν την ενίσχυση γνώσης δημιουργώντας ένα κοινό χώρο αναπαράστασης για τη γλώσσα, την εικόνα και την εξωτερική γνώση, καταλήγοντας σε ενιαίες knowledge-vision-language αναπαραστάσεις. Συγκεκριμένα, το KVL-BERT εμπλουτίζει προτάσεις με τριπλέτες από το ConceptNet και τις ενσωματώνει σε έναν BERT-ομοιά VL transformer, χρησιμοποιώντας μάσκες ορατότητας για τον έλεγχο του self-attention. Το ViLaKC ενσωματώνει ξεχωριστά τις τρεις modalities (εικόνα, γλώσσα, γνώση) με ResNet, BERT και GCN και τις συγχωνεύει μέσω multi-head attention, με νέα στάδια προεκπαίδευσης όπως MLMIK και MOCTK. Το CKRM μεταφέρει χρονική γνώση από το SWAG μέσω πολυεπίπεδης χαρτογράφησης (σε επίπεδο κελιού, στρώματος και attention), η οποία συνδυάζεται τελικά με οπτική προσοχή για παραγωγή απάντησης. Η αξιολόγηση ακολουθεί το πρωτόκολλο του VCR, ελέγχοντας την ακρίβεια επιλογής απάντησης, επεξήγησης και του συνδυασμού τους.

Ενίσχυση γνώσης στα συστήματα κειμενικής περιγραφής εικόνων (K-IC)

Datasets Στην εργασία K-IC δεν υπάρχουν ειδικά σύνολα δεδομένων που απαιτούν γνώση. Ως εκ τούτου, τα COCO Captions και Flickr30k αποτελούν τη βασική βάση δεδομένων.

Μέθοδοι Οι μέθοδοι χωρίζονται σε διαδοχικά μοντέλα τύπου LSTM και σε προσεγγίσεις Transformer, ενσωματώνοντας εξωτερική ή εσωτερική γνώση. Οι πρώτες υλοποιήσεις με εξωτερική γνώση εμπλουτίζουν τις λεζάντες αξιοποιώντας πηγές κοινού νου όπως το ConceptNet ή το Visual Genome: τα ανιχνευμένα αντικείμενα αναζητούν σχετική γνώση και οι οντότητες ή σχέσεις που ανακτώνται συγχωνεύονται με τα οπτικά χαρακτηριστικά μέσω LSTM, GCN ή σημασιολογικών γράφων. Με αυτόν τον τρόπο, τα συστήματα μπορούν να παράγουν λέξεις που δεν εμφανίζονται ρητά στην εικόνα, να περιορίζουν σφάλματα ανιχνευτών μέσω σημασιολογικών priors και να προσφέρουν ερμηνεύσιμη καθοδήγηση του περιεχομένου μέσω attention masks.

Οι προσεγγίσεις εσωτερικής γνώσης δημιουργούν δικά τους visiolinguistic priors, συνήθως μέσω παραγωγής scene graphs. Το SGAE μαθαίνει ένα λεξικό που αντιστοιχεί σχέσεις ανάμεσα σε κείμενο και scene graphs, μεταφέροντας γλωσσικά priors στη διαδικασία δημιουργίας λεζάντας, ενώ μη-ρητές δομές προκύπτουν και μέσω latent-topic attention μηχανισμών όπως το CLTA, το οποίο αναγνωρίζει κρυφές θεματικές ενότητες χωρίς κατασκευή scene graph και ανακατανέμει το βάρος οπτικά σημαντικών περιοχών.

Τα Transformer-based συστήματα κλιμακώνουν αυτές τις ιδέες. Η εξωτερική γνώση χρησιμοποιείται σε captioning με οντότητες ή γεγονότα, κατασκευάζοντας διασταυρούμενα υπο-γραφήματα μεταξύ αντικειμένων της εικόνας και ονομαστικών οντοτήτων κειμένου, ενισχυμένα από πολυτροπική γνώση της Wikipedia. Οι ενσωματώσεις από GAT knowledge graphs, RoBERTa και ResNet διοχετεύονται σε Transformer decoders. Το KM-BART ενοποιεί οπτική και γλωσσική πληροφορία με δομημένη γνώση κοινού νου από COMET και ATOMIC/ConceptNet, χρησιμοποιώντας task tokens για να καθορίζει εργασίες οπτικού κοινού νου (π.χ. τι συνέβη πριν/μετά ή ποια είναι η πρόθεση). Οι εσωτερικές γνώσεις σε Transformers αντιμετωπίζουν περιορισμούς της MLE προσθέτοντας KL όρους που βαρύνουν σημασιολογική συγγένεια και ενισχύοντας τις ενσωματώσεις με γράφους βασισμένους σε γλωσσική ομοιότητα.

Ενίσχυση γνώσης στα συστήματα οπτικού διαλόγου (K-VD)

Datasets Στην εργασία K-VD, το VisDial αποτελεί τον κύριο πόρο, ενώ το VisDialCK επεκτείνει το σύνολο δεδομένων με διαλόγους που απαιτούν γνώση και ιστορική συνέπεια.

Μέθοδοι Η μοναδική ενισχυμένη με γνώση μέθοδος χρησιμοποιεί εσωτερική γνώση, εμπλουτίζοντας οπτικά και γλωσσικά γραφήματα με γεγονότα από το ConceptNet. Μετά τον καθαρισμό και τη σύντηξη των γράφων, ένα transformer-based fusion module παράγει την απάντηση.

Ενίσχυση γνώσης στα συστήματα οπτικής αφήγησης (K-VIST)

Datasets Στην εργασία K-VIST, η ανάγκη για χρονικό συλλογισμό είναι έντονη, καθώς οι αφηγήσεις πρέπει να ακολουθούν λογική αλληλουχία εικόνων. Το VIST αποτελεί το βασικό dataset.

Μέθοδοι Οι μέθοδοι με εξωτερική γνώση βασίζονται σε πολυσταδιακές διαδικασίες: αρχικά εντοπίζουν αντικείμενα, ανακτούν σχετικές οντότητες από ConceptNet, επιλέγουν τις πιο συναφείς μέσω attention και στη συνέχεια παράγουν προτάσεις χρησιμοποιώντας visual features και προηγούμενα κείμενα. Άλλες προσεγγίσεις περιορίζουν την επαναληπτικότητα της αφήγησης με λεξιλογικό εμπλουτισμό, ιεράρχηση σχέσεων και παραγωγή κειμένου με Transformers που ενσωματώνουν ποινές επανάληψης, βελτιωμένη αναφορά (anaphora) και ευέλικτες positional encodings. Τρισταδιακά μοντέλα (imagine–reason–write) ενσωματώνουν γράφους κοινού νου, scene graphs και event graphs μέσω GCN για να επιλέξουν τις καταλληλότερες έννοιες.

Transformer-based αφηγήσεις επεκτείνουν αυτές τις διαδικασίες μέσω pipelines εμπλουτισμού εννοιών, GAT reasoning, sequential ή clique-based επιλογής εννοιών και τελικής δημιουργίας ιστοριών με RNN ή BART, με το BART να επιτυγχάνει πιο ποικιλόμορφες αφηγήσεις. Η αξιολόγηση περιλαμβάνει BLEU, ROUGE, METEOR, CIDEr, τα Dist-n για μετρήσεις ποικιλίας και πλούσια ανθρώπινη αξιολόγηση που εξετάζει ευφράδεια, συνάφεια, πληροφορικότητα, συνοχή, λογικότητα και συνολική προτίμηση.

Ενίσχυση γνώσης στα συστήματα παραγωγής εικόνων από κείμενο (K-cIG)

Datasets Η γνώση στη δημιουργία εικόνων εκτείνεται τόσο στη συνθετική δημιουργία υπό συνθήκες όσο και στην οπτική αφήγηση, παρότι τα περισσότερα διαθέσιμα σύνολα δεδομένων, όπως τα ImageNet, CIFAR, FFHQ, Oxford Flowers και CUB, δεν έχουν σχεδιαστεί για απαιτήσεις γνώσης. Για τη διαδοχική δημιουργία χρησιμοποιούνται δεδομένα κινουμένων σχεδίων όπως τα Pororo-SV και FlintstonesSV ή σύνολα λεζαντών βίντεο όπως το DiDeMo-SV, τα οποία παρέχουν χρονικά ευθυγραμμισμένες σκηνές, περιγραφές και αναλυτικές σημειώσεις που ευνοούν την οπτική αφήγηση.

Μέθοδοι Στην παραγωγή εικόνων υπό συνθήκη (K-cIG), τα GAN δυσκολεύονται με άγνωστους συνδυασμούς γνωρισμάτων, οδηγώντας σε απώλεια σημασιολογίας όταν οι κειμενικές συνθήκες αναφέρονται σε μη εμφανιζόμενες καταστάσεις. Το KG-GAN εισάγει εσωτερική γνώση εκπαιδεύοντας έναν δεύτερο γεννήτορα σε γνώση του πεδίου μέσω μιας knowledge loss και μοιράζοντας παραμέτρους με τον γεννήτορα που συνθέτει εικόνες από κείμενο. Ένα δίκτυο παλινδρόμησης επιβάλλει περιορισμούς ευλογοφάνειας για άγνωστους συνδυασμούς, επιτρέποντας στο μοντέλο να παράγει συνεπείς εικόνες πέρα από την κατανομή εκπαίδευσης χωρίς να χρειάζεται εξωτερικά knowledge graphs.

Η γνώση στην Οπτική Αφήγηση (K-SV) αξιοποιεί εξωτερικό κοινό νου και δομημένες γλωσσικές αναπαραστάσεις. Τα συντακτικά δέντρα κωδικοποιούν ιεραρχικά το κείμενο, ενώ το ConceptNet συμπληρώνει λείπουσες σημασιολογικές σχέσεις και φέρνει εννοιολογικά παρόμοιες προτάσεις πιο κοντά στον χώρο ενσωμάτωσης. Η dense captioning δρα ως μορφή εσωτερικής χωρικής γνώσης, παρέχοντας αναλυτική πληροφόρηση για θέσεις αντικειμένων που συχνά απουσιάζει από το αφήγημα. Συστήματα όπως το MARTT συνδυάζουν τη σημασιολογική και χωρική γνώση μέσω tree-transformers και graph transformers ώστε να βελτιώσουν την κειμενοεικονοποιητική συνέπεια, την οπτική πιστότητα και τη χρονική συνέχεια. Επιπλέον, οι επιτυχίες του DALL-E ενέπνευσαν τη χρήση μεγάλων προεκπαιδευμένων μοντέλων ως πηγές μη δομημένης γνώσης, επιτρέποντας ακόμη και zero-shot σύνθεση ιστοριών ή συνέχιση μιας οπτικής ιστορίας με βάση μια αρχική εικόνα.

Ενίσχυση γνώσης σε πολυ-εργασιακές αρχιτεκτονικές

Στις πολυ-εργασιακές αρχιτεκτονικές βασισμένες σε transformers, η εξωτερική ή εσωτερική γνώση επιτρέπει την ενοποίηση πολλών VL εργασιών. Ορισμένες προσεγγίσεις ενσωματώνουν grounded situation recognition και Visual Commonsense Graphs μέσω του VisualCOMET για χρονικό και αιτιώδη συλλογισμό σε εργασίες όπως VQA, VE και VCR. Άλλες χρησιμοποιούν δομημένη γνώση από ConceptNet και Wikidata, ευθυγραμμίζουν embeddings γνώσης με προτάσεις εκπαίδευσης ή αξιοποιούν οντότητες Wikidata μέσω Wikipedia2Vec για να εμπλουτίσουν κείμενο και ετικέτες αντικειμένων. Οι εσωτερικές μέθοδοι, όπως το OSCAR, χρησιμοποιούν ετικέτες αντικειμένων ως διαμεσολαβητές μεταξύ εικόνας και κειμένου, ενώ το ERNIE-ViL μαθαίνει λεπτομερή σημασιολογία μέσω προεκπαίδευσης με scene graphs. Το ROSITA κατασκευάζει εικόνο-κείμενο και δια-τροπικούς γράφους, ευθυγραμμίζοντας οπτικές και γλωσσικές έννοιες και σχηματίζοντας εμπλουτισμένα υπογραφήματα που βελτιώνουν εργασίες όπως VQA, αναφορές σε αντικείμενα και cross-modal retrieval.

1.7 Ενίσχυση γνώσης μέσω Μεγάλων Γλωσσικών Μοντέλων

Τα knowledge graphs, παρότι προσφέρουν διαφάνεια και ακρίβεια, παρουσιάζουν σοβαρούς περιορισμούς στην επεκτασιμότητα, την κάλυψη και την προσαρμοστικότητα, με αποτέλεσμα να αδυνατούν να συμβαδίσουν με τις διαρκώς αναδυόμενες οντότητες και έννοιες που εμφανίζονται σε πραγματικά δεδομένα όρασης. Η συμβολική τους φύση τα καθιστά ανεπαρκή για λεπτές, συμφραζομενικές ή αιτιώδεις μορφές γνώσης, ενώ η αντιστοίχιση οπτικών χαρακτηριστικών σε σταθερά οντολογικά σχήματα δημιουργεί συχνά σημασιολογικά κενά. Αυτοί οι περιορισμοί οδήγησαν στη σταδιακή μετάβαση προς το παράδειγμα των μεγάλων γλωσσικών μοντέλων ως βάσεις γνώσης (LLM-as-KB), όπου τα LLMs λειτουργούν ως άρρηκτες, διανεμημένες βάσεις γνώσης. Τα LLM μπορούν να ενσωματώνουν συμφραζόμενα, να πραγματοποιούν συνθετική και αναλογική συλλογιστική και να κλιμακώνονται φυσικά μέσω εκπαίδευσης σε δεδομένα μεγάλης κλίμακας, καλύπτοντας μακρές ουρές εννοιών και νέα γεγονότα χωρίς χειροκίνητη επιμέλεια. Παράλληλα, η γλωσσική τους φύση τα καθιστά συμβατά με τα κειμενικά outputs των VL μοντέλων, επιτρέποντας ολοκληρωμένη διαφορική συλλογιστική χωρίς την ανάγκη άκαμπτων γραφημάτων και λογικών κανόνων.

Ωστόσο, τα LLM μπορεί να παρεκκλίνουν ή να υπεργενικεύσουν, γεγονός που έχει οδηγήσει στην ανάπτυξη υβριδικών συστημάτων όπου τα KGs προσφέρουν επαληθεύσιμη, ελεγχόμενη γνώση, ενώ τα LLM αναλαμβάνουν την αφαιρετική, συμφραζομενική κατανόηση. Στο πεδίο των vision-language εργασιών, αυτή η προσέγγιση έχει ενισχύσει σημαντικά τις δυνατότητες των μοντέλων σε γνωσιοκεντρικές εργασίες όπως K-VQA, K-VCR και η γνώση-ενισχυμένη περιγραφή εικόνων. Στο K-VQA, οι πρώτες μέθοδοι χρησιμοποίησαν LLMs για παροχή γνώσης μέσω λεζαντών και few-shot prompts, ενώ νεότερες προσεγγίσεις προσαρμόζουν οπτικούς κωδικοποιητές ή συνδυάζουν ρητή και άρρηκτη γνώση για βελτιωμένη ανάκτηση και συλλογιστική. Στο K-VCR, τα LLMs συμβάλλουν στην παραγωγή αιτιολογήσεων, στη χρονική και κοινής λογικής ανάλυση και στην υιοθέτηση τεχνικών chain-of-thought, συχνά αποκαλύπτοντας ενδιάμεσα βήματα λογισμού. Στην περιγραφή εικόνων, η ενσωμάτωση γνώσεων γεγονότων, οντοτήτων και commonsense, αλλά και η αξιοποίηση άρρηκτης γνώσης μέσω μοντέλων όπως GPT-2 και CLIP, ενισχύουν τη γενετική ικανότητα των συστημάτων. Παρά την πρόοδο, η έλλειψη εξειδικευμένων datasets που απαιτούν εξωτερική γνώση περιορίζει την πλήρη αξιολόγηση των K-IC μοντέλων, υπογραμμίζοντας την ανάγκη για πιο απαιτητικές συνθήκες ελέγχου.

Αποσαφήνιση οπτικών εννοιών

Το πρόβλημα της αποσαφήνισης οπτικών εννοιών (Visual Word Sense Disambiguation - VWSD) αποτελεί μια πολυτροπική επέκταση του κλασικού Word Sense Disambiguation, όπου μια ασαφής ή πολυσημική λεκτική φράση πρέπει να αντιστοιχιστεί στη σωστή εικόνα ανάμεσα σε δέκα υποψήφιες. Το πρόβλημα είναι ιδιαίτερα απαιτητικό, καθώς τα σύγχρονα μοντέλα όρασης–γλώσσας συχνά αποτυγχάνουν να αναγνωρίσουν σπάνιες σημασίες ή να στηριχθούν σε ουσιαστική σημασιολογική κατανόηση. Αντίθετα, τείνουν να αξιοποιούν επιφανειακές συσχετίσεις και στατιστικές κανονικότητες του προεκπαιδευτικού τους corpus, παρακάμπτοντας τον πραγματικό εννοιολογικό πυρήνα του προβλήματος.

Η πρώτη μεθοδολογική μας προσέγγιση βασίζεται στον εμπλουτισμό της φράσης εισόδου μέσω LLMs, τα οποία λειτουργούν ως βάσεις γνώσης. Το ζητούμενο είναι να παραχθούν εμπλουτισμένες εκδοχές της αρχικής φράσης με τρόπο που να αποσαφηνίζουν τη σωστή σημασία της. Χρησιμοποιούνται στοχευμένα prompts, όπως “What is <phrase>?” ή “Describe <phrase>”, τα οποία οδηγούν σε περιγραφές που ενημερώνουν τα μοντέλα όρασης–γλώσσας για πιο λεπτές και συχνά παραμελημένες σημασιολογικές πτυχές. Η ποιότητα και η κλίμακα του

LLM αποδεικνύονται κρίσιμες: μεγάλα μοντέλα όπως το GPT-3/GPT-3.5 αποφέρουν σημαντικές βελτιώσεις, ενώ μικρότερα μοντέλα συχνά επιδεινώνουν τα αποτελέσματα. Το κατάλληλο prompt επηρεάζει έντονα την απόδοση, με τα πιο ερμηνευτικά και περιγραφικά prompts να υπερέχουν έναντι εκείνων που ζητούν λέξη προς λέξη ακρίβεια.

Στη δεύτερη μεθοδολογική προσέγγιση, το VWSD μετατρέπεται σε πρόβλημα Ερώτησης–Απάντησης (QA). Κάθε εικόνα μετατρέπεται σε λεζάντα και το LLM καλείται να επιλέξει την καταλληλότερη λεζάντα ανάμεσα σε δέκα υποψήφιες. Χρησιμοποιούνται διαφορετικές στρατηγικές prompting, όπως zero-shot ερωτήσεις τύπου "think" ή "choose", few-shot παραδείγματα και τεχνικές Chain-of-Thought. Τα αποτελέσματα δείχνουν ότι τα μεγαλύτερα LLMs, κυρίως το GPT-3.5, αποδίδουν σαφώς καλύτερα από μικρότερα μοντέλα, αν και η χρήση συλλογιστικής αλυσίδας (Chain of Thought - CoT) δεν προσφέρει πάντα σαφή πλεονεκτήματα. Ωστόσο, το συνολικό πλαίσιο τύπου QA τείνει να αποδίδει χαμηλότερα από τη μέθοδο εμπλουτισμού της φράσης, γεγονός που οφείλεται στο ότι η μετατροπή εικόνας σε κείμενο εισάγει θόρυβο και απώλεια πληροφορίας, καθιστώντας πιο δύσκολη την αποσαφήνιση της σωστής σημασίας της φράσης.

Περιορισμοί των LLMs ως βάσεις γνώσης

Η χρήση των LLMs ως βάσεων γνώσης συνοδεύεται από σημαντικές προκλήσεις, λόγω της εκτενούς προεκπαίδευσής τους σε ποικίλες πηγές δεδομένων και της πολύπλοκης αρχιτεκτονικής τους. Ιδιαίτερα, οι αδυναμίες στη λογική σκέψη (rasoning) και οι «παραισθήσεις» (hallucinations) περιορίζουν την πρακτική τους εφαρμογή σε κρίσιμους τομείς. Πρόσφατες μελέτες έχουν αναδείξει προόδους σε διάφορους τύπους συλλογιστικής, όπως η κοινή λογική, η μαθηματική, η αιτιακή και άλλες υποκατηγορίες. Παρόλα αυτά, η συλλογιστική των LLMs βασίζεται συχνά σε πιθανοτικά μοτίβα παρά σε πραγματική κατανόηση αιτίας-αποτελέσματος ή σε λογικά πλαίσια, με αποτέλεσμα ασυνέπειες ή λανθασμένη λογική σε πολύπλοκα προβλήματα πολλών βημάτων. Τα hallucinations εμφανίζονται όταν το μοντέλο παράγει φαινομενικά λογικές αλλά εσφαλμένες ή μη σχετικές πληροφορίες, συχνά λόγω ελλιπούς συλλογιστικής. Η αντιμετώπιση αυτών των προκλήσεων απαιτεί νέες στρατηγικές για την επεξεργασία της συλλογιστικής και τη μείωση των hallucinations, ώστε να επιτευχθεί πιο αξιόπιστη αναπαράσταση γνώσης [97, 141].

Προς την κατεύθυνση της πιο «ανθρωπόμορφης» συλλογιστικής στα LLMs, είναι αναγκαία η μελέτη δυσκολότερων και πιο πολύπλοκων κατηγοριών συλλογισμού, προκειμένου να διερευνηθεί και να επαυξηθεί η εμπιστοσύνη στα LLMs σε οποιεσδήποτε εργασίες, συμπεριλαμβανομένης και της χρήσης τους ως βάσεις γνώσης [88, 249, 248]. Η αφηρημένη συλλογιστική στα LLMs (abstract) προσομοιάζει ανθρώπινες διαδικασίες σκέψης, χωριζόμενες σε κάθετη και πλευρική σκέψη. Η κάθετη βασίζεται στη γραμμική λογική και στους κανόνες, ενώ η πλευρική απαιτεί δημιουργικότητα και ευελιξία για «εκτός πλαισίου» λύσεις. Τα προβλήματα τύπου γρίφων συνδυάζουν και τις δύο μορφές, απαιτώντας λογική και δημιουργική σκέψη. Η αξιολόγηση μέσω γρίφων επιτρέπει την ανάλυση αδυναμιών των μοντέλων, όπως η υπερβολική εξάρτηση από ψευδείς συσχετίσεις, ενώ η σωστή κατηγοριοποίηση των γρίφων—σε βασισμένους σε κανόνες (deterministic και stochastic) και μη βασισμένους σε κανόνες (riddles, προγραμματιστικούς, με κοινή λογική)—καθορίζει τις απαιτήσεις σε συλλογιστική, γνώση κόσμου και ευελιξία. Εξειδικευμένες τεχνικές προτροπής (prompting) αξιοποιούνται για την ανάδειξη αφηρημένης συλλογιστικής, με ιδιαίτερη έμφαση στη πλευρική σκέψη μέσω ανακατασκευασμένων σεναρίων (contextual reconstructions), που μειώνουν την παραπλανητική επίδραση των σημασιολογικών λεπτομερειών. Η άντληση αφηρημένης γνώσης από τα LLMs αποδεικνύει ότι τα μοντέλα αυτά διαθέτουν τη συγκεκριμένη ικανότητα σε ένα βαθμό, τουλάχιστον όταν αξιοποιείται η κατάλληλη τεχνική prompting, ανοίγοντας προοπτικές για την αξιόπιστη χρήση τους στον εμπλουτισμό γνώσης.

Παρόλο που τα LLMs επιτυγχάνουν κορυφαία αποτελέσματα σε ποικίλες εργασίες [74, 334, 279, 353, 251, 252] δίνοντας παράλληλα κάποιες προοπτικές και εγγυήσεις σχετικά με τις δυνατότητες συλλογιστικής τους, η αξιοπιστία τους μπορεί να αποτύχει, ειδικά σε αντιδιαισθητικά ή μεροληπτικά σενάρια, αναδεικνύοντας ότι η μελέτη και η ενίσχυση της συλλογιστικής αποτελεί ένα μόνο κομμάτι του συνολικού προβλήματος [321, 77].

Ένα χαρακτηριστικό παράδειγμα είναι το φαινόμενο της αντίστροφης κλιμάκωσης (inverse scaling), όπου η απόδοση ενός μοντέλου επιδεινώνεται καθώς αυξάνεται το μέγεθός του, λόγω υπερβολικής γενίκευσης ψευδών συσχετίσεων που βρέθηκαν στα δεδομένα προεκπαίδευσης. Στο πλαίσιο αυτό, προτείνουμε την «εργασία επανακαθορισμού» (redefinition task), στην οποία τα LLMs καλούνται να ακολουθήσουν εναλλακτικές λογικές πορείες αντικαθιστώντας τις προεπιλεγμένες γνώσεις τους με κατάλληλο prompting. Οι φυσικές σταθερές, όπως το π, ο αριθμός του Euler, η ταχύτητα του φωτός, και άλλες, επανακαθορίζονται είτε με αριθμητικές τιμές είτε με αντικατάσταση από άλλες σταθερές, σε τρία επίπεδα δυσκολίας, από κοντινές έως τελείως μη ρεαλιστικές τιμές.

Στη συνέχεια σχεδιάζονται ερωτήσεις τριών επιπέδων δυσκολίας. Η ανάλυση των απαντήσεων δείχνει ότι όλα τα LLMs παρουσιάζουν φαινόμενο «άγκυρας» (anchoring), δηλαδή υπερβολική προσκόλληση στις προηγούμενες γνώσεις τους, ειδικά στις απαντήσεις πολλαπλής επιλογής και σε μεγαλύτερα μοντέλα. Παράλληλα, παρατηρείται ότι τα μεγαλύτερα μοντέλα τείνουν να απαντούν με ψευδή αυτοπεποίθηση ακόμη και όταν δεν μπορούν να επεξεργαστούν σωστά τις επανακαθορισμένες τιμές, αναδεικνύοντας έναν αφανή κίνδυνο οργανικά συνδεδεμένο με τη φύση των LLMs.

Επιπλέον, μελετώνται οι γνωστικές μεροληψίες (cognitive biases) ως φαινόμενο σε LLMs και η δυνατότητα να χρησιμοποιηθούν ως επιθετικές στρατηγικές σε συστάσεις προϊόντων. Οι επιθέσεις αυτές, είτε χειροποίητες είτε παραγόμενες από LLM, επηρεάζουν τη συχνότητα και τη θέση σύστασης των προϊόντων, μετρώντας αλλαγές στην αναλογία σύστασης και στην κατάταξη του προϊόντος. Τα αποτελέσματα δείχνουν ότι διάφορες γνωστικές μεροληψίες όπως η κοινωνική απόδειξη (social proof), η αίσθηση αποκλειστικότητας (exclusivity), η σπανιότητα (scarcity), η πλασματική έκπτωση (discount framing) και η επίκληση στην αυθεντία (authority bias), μπορούν να μεταβάλλουν σημαντικά τη σύσταση και την κατάταξη προϊόντων σε LLMs, είτε θετικά είτε αρνητικά, επιβεβαιώνοντας ότι τα μοντέλα είναι ευάλωτα σε γνωστικές επιρροές που μπορούν να χρησιμοποιηθούν για χειραγώγηση.

Συνολικά, αποκαλύπτεται ότι η εμπιστοσύνη στα LLMs δεν είναι δεδομένη, καθώς οι περιορισμοί τους στην προσαρμογή σε νέα ή επανακαθορισμένα δεδομένα και η ευαισθησία τους σε γνωστικές μεροληψίες μπορεί να οδηγήσουν σε αναξιόπιστες ή παραπλανητικές απαντήσεις.

1.8 Συμπεράσματα

Στο σύνολό της, η παρούσα διατριβή ανέδειξε ότι η γνώση αποτελεί θεμέλιο στοιχείο για την κατανόηση, την εξήγηση και τη γενίκευση σε πολυτροπικά συστήματα. Τα συστήματα όρασης-γλώσσας εξακολουθούν να υστερούν όταν καλούνται να διαχειριστούν λεπτομερή, χρονικά ή αιτιακά δομημένη γνώση, με αποτέλεσμα να εμφανίζουν ελλείψεις τόσο στην παραγωγή εξηγήσεων όσο και στην αξιοπιστία των συλλογισμών τους. Η ανάλυση των τύπων γνώσης, η διερεύνηση των counterfactual επεμβάσεων και η μελέτη ρητών βάσεων γνώσης έδειξαν ότι η ενσωμάτωση εξωτερικών πηγών παραμένει κρίσιμη για επεξηγησιμότητα και σταθερότητα. Παράλληλα, τα αποτελέσματα έδειξαν πως η μετάβαση σε πιο εύπλαστα, κατανεμημένα μοντέλα γνώσης –όπως τα LLMs– επεκτείνει σημαντικά τις δυνατότητες των VL συστημάτων, χωρίς όμως να εξαλείφει πλήρως τις αδυναμίες τους ως προς την αξιοπιστία ή τον έλεγχο των παραγόμενων γνώσεων. Έτσι, προκύπτει ότι η συνδυαστική χρήση ρητών και άρρητων μηχανισμών γνώσης αποτελεί την πιο ρεαλιστική και παραγωγική επιλογή.

Από τα ευρήματα της διατριβής, προκύπτει ότι παρά την αξιοσημείωτη πρόοδο των LLMs, εξακολουθούν να υπάρχουν σημαντικά κενά στη διαχείριση γνώσης, στην αποφυγή hallucinations και στη δυνατότητα σύνδεσης μεταξύ αντιληπτικών και συμβολικών αναπαραστάσεων. Τα LLMs λειτουργούν πιο αποτελεσματικά ως αποθήκες και επεξεργαστές γνώσης, ενώ οι VL μηχανισμοί παραμένουν απαραίτητοι για την οπτική θεμελίωση και την πολυτροπική συλλογιστική. Αυτό το χάσμα υπογραμμίζει την ανάγκη για συστήματα τα οποία δεν ενσωματώνουν απλώς γνώση, αλλά μπορούν να τη διαχειριστούν, να την επαληθεύσουν και να τη χρησιμοποιήσουν με τρόπο που ευθυγραμμίζεται με την οπτική αντίληψη. Οι αναλύσεις της διατριβής καταδεικνύουν ότι η βιώσιμη πρόοδος δεν βρίσκεται ούτε στην πλήρη εξάρτηση από συμβολικά γραφήματα ούτε στον απόλυτο ενστερνισμό της άρρητης γνώσης των LLMs, αλλά σε συνδυαστικές αρχιτεκτονικές που επιτρέπουν έλεγχο, διαφάνεια και ευελιξία.

Οι μελλοντικές κατευθύνσεις συγκλίνουν προς την ανάπτυξη πιο δομημένων, υβριδικών προσεγγίσεων όπου η γνώση (είτε ρητή είτε άρρητη) λειτουργεί ως συντονισμένο σύστημα πόρων. Η ανάγκη για αντικειμενικές μετρικές αξιολόγησης, datasets που απαιτούν πραγματική γνώση και πιο σταθερές στρατηγικές ενσωμάτωσης γνώσης είναι εμφανής. Ειδικά ο ρόλος των LLMs ως βάσεις γνώσης θα ενισχυθεί, είτε ως ανεξάρτητες πηγές πληροφόρησης είτε ως επαληθευτικοί μηχανισμοί που ελέγχουν ή βελτιώνουν τα LVLMs. Ωστόσο, το πιο σημαντικό βήμα είναι η μετάβαση προς πρακτορικά πολυτροπικά συστήματα, όπου ένα LLM και ένα LVLM λειτουργούν ως δύο συνεργαζόμενοι πράκτορες: ο πρώτος ως γνώστης, επιμελητής και ελεγκτής της πληροφορίας, ο δεύτερος ως αντιληπτικό σύστημα που μεταφράζει τη γνώση σε οπτικά θεμελιωμένες πράξεις ή εξηγήσεις. Αυτή η συνεργατική δυναμική ανοίγει τον δρόμο για πολυτροπική νοημοσύνη που δεν περιορίζεται σε στατικές αναπαραστάσεις, αλλά μπορεί να σχεδιάζει, να τεκμηριώνει, να αμφισβητεί, να επαληθεύει και να αναθεωρεί τις ίδιες της τις διαδικασίες.

Συνολικά, οι προτεινόμενες μελλοντικές κατευθύνσεις συγκλίνουν προς μια νέα εποχή όπου τα πολυτροπικά συστήματα δεν θα είναι απλώς μοντέλα που συνδυάζουν εικόνα και κείμενο, αλλά πολυπράκτορες με αυτόνομη διαχείριση γνώσης, ικανοί να συνθέτουν ρητές και άρρητες μορφές πληροφορίας, να εφαρμόζουν counterfactual συλλογισμούς και να αλληλεπιδρούν με τον κόσμο μέσω μιας συνεκτικής και επεξηγήσιμης γνωστικής δομής. Αυτή η μετάβαση προς την πράκτορα πολυτροπικότητα αποτελεί το φυσικό και αναγκαίο βήμα για την ανάπτυξη συστημάτων που είναι όχι μόνο πιο ισχυρά, αλλά και πιο κατανοήσιμα, ελεγχόμενα και διαφανή.

Chapter 2

Introduction

In recent years, Artificial Intelligence (AI) has become a focal point across industries and academia, driving innovation and reshaping numerous sectors. Once confined to theoretical discussions and science fiction, AI now permeates nearly every aspect of modern life—from healthcare and finance to transportation, education, and entertainment. Its ability to learn from data, adapt to new information, and perform complex tasks has not only revolutionized industries but also positioned AI as a critical area of academic inquiry.

As AI systems increasingly influence high-stakes decisions, questions around trust, accountability, and transparency have become more urgent. This has led to the growing importance of Explainable Artificial Intelligence (XAI), a subfield focused on making AI decision-making processes understandable to humans. Unlike traditional “black-box” models, explainable AI seeks to provide insights into how and why a system reaches a particular outcome, thereby enabling users to verify, challenge, or refine its behavior. This transparency is especially crucial in sensitive domains where ethical, legal, and social consequences are significant. Furthermore, explainability is not only vital for building trust with end-users but also for improving model reliability, uncovering biases, and facilitating regulatory compliance. As AI becomes more embedded in decision-making processes, the demand for systems that are not only powerful but also interpretable is rapidly growing.

Among the various approaches to explainable AI, counterfactual explanations (CEs) have gained significant attention for their intuitive appeal and practical usefulness. A counterfactual explanation describes how a minimal change to an input could lead to a different output—effectively answering the question, “What would need to be different for the AI to make another decision?” For example, in a loan application scenario, a counterfactual explanation might state: “Your loan was denied because your annual income was $30,000; if it had been $45,000, the loan would have been approved.” This kind of explanation is not only easy to understand for non-expert users, but also actionable, as it offers concrete guidance on how an outcome could be improved. By focusing on the smallest possible changes necessary to alter a prediction, counterfactuals provide both transparency and empowerment—bridging the gap between model logic and user expectations. As such, they are particularly well-suited for domains where interpretability, fairness, and user trust are essential.

A significant limitation in current literature around CEs is the focus to *low-level* uninterpretable features, such as pixel manipulations in images: changing an RGB value of a certain pixel in an image may alter the classification label of the image overall, even though this manipulation is totally uninformative to humans; therefore, answering "What would need to be different" cannot be interpreted in human-level concepts, nor can it be replicated for other images in a consistent manner. On the other hand, introducing human-interpretable *high-level* concepts for CEs allows for more useful and valid explanations, satisfying the final recipients-the humans themselves. Such concepts involve meaningful units of information, such as visual objects (e.g. "chair", "person", "dog", "car" etc), concept characteristics (e.g. "great movie", "blue sky" etc) and interactions between concepts (e.g. "man rides bike", "cat on roof" etc). Those naturally correspond to how humans perceive the world, therefore they are effortlessly interpretable and replicable.

An upcoming literature branch suggests the exploitation of external knowledge sources to enforce and guarantee conceptuality in CEs and XAI in general, establishing well-defined algorithms that advance explainable systems and validate their proper functionality, especially in high-stakes scenarios. Nevertheless, the multifaceted applicability of knowledge sources is not only restricted to enabling conceptual XAI techniques: the same knowledge sources can be readily applied for enhancing the performance of AI systems, without significant modifications to the original algorithm that connects conceptuality with CEs.

A parallel venture views knowledge enhancement from external knowledge sources as an important step towards more complex AI systems, which can seemingly *reason* and approach human-level performance as much as possible. Knowledge-enhanced systems are able to take existing AI systems to the next level, serving more purposes rather than perceiving what is evident: those systems can possess knowledge about real-world facts, commonsense cognition, temporal associations, complex linguistic terms and many other senses that humans normally acquire through learning, experience and exposure to varying unpredictable situations.

Especially with the advent of Large Language Models (LLMs), knowledge enhancement becomes more effortless, while the knowledge spectrum fused to existing systems grows to unprecedented levels. There is much evidence in literature that LLM-driven knowledge enhancement enables impressive capabilities, taking AI to a new era of efficiently assisting humans in an abundance of tasks, without significant limitations regarding the amount and variety of knowledge that they possess.

However, this is not without a cost: LLMs are undoubtedly knowledgeable but their opaque nature raises several concerns. They often *hallucinate*, producing unfaithful and erroneous outputs, obscured under linguistically correct utterances. At the same time, reasoning shortcomings and failure to adhere to trustworthy suggestions limit their unquestionable integration to several applications. Thus, a significant discussion across stakeholders commences, addressing all these shortcomings naturally accompanying LLMs, with the hope that more reliable LLMs will be promoted in the future.

Several of these issues are addressed in the current thesis: inspired by the evolution of external knowledge sources and their increasing integration in several AI systems, we study them for crafting trustworthy conceptual CEs, as well as for advancing performance of existing models using knowledge enhancement.

2.1 Thesis contributions

In this thesis, we investigate the contribution of knowledge bases in various tasks in Natural Language Processing and Multimodal Learning, either in terms of explainability or performance enhancement. The structure is introduced as following.

2.1.1 Survey and taxonomy of knowledge types

At first, we explore the types of knowledge that can be fused in AI systems, involving both the level of transparency of the knowledge source itself, as well as the different senses that have been incorporated in knowledge sources [206, 205]. The types of knowledge are introduced and analyzed in chapter 3.

The nature and structure level of each knowledge source suggests its optimal usage; for example, knowledge graphs serve as clearly structured and well-defined knowledge sources, perfectly fulfilling the purpose of conceptuality in XAI applications. This characteristic can be attributed to the deterministic semantic distances between concepts, requesting interpretable algorithms (such as graph transversing) to discover conceptual relationships. On the other hand, LLM knowledge is dispersed and cannot be located, rendering the retrieval process of a concept uninterpretable. As a trade-off, the vast amounts of knowledge, which is acquired in an unsupervised manner, allows knowledge enhancement of an abundance of tasks, without explicitly seeking for specific knowledge senses in dedicated knowledge sources.

2.1.2 Background in Counterfactual Explanations

Then, we delve into the intersection of knowledge and explanations, specifically focusing on CEs. A thorough introduction on CEs is presented in chapter 4, underlying the importance of involving knowledge to ensure conceptuality in CEs. In that case, we revisit existing literature that supports our basic claims of utilizing structured knowledge graphs to obtain semantic distances that drive concept substitutions. This way, the question of "What would need to be different" is translated to "What is the optimal conceptual change" that results in an observable change in the outcome. Minimal conceptual changes verify that humans are able to comprehend and replicate them in a controllable manner.

2.1.3 Explicit knowledge bases for explainability

In chapter 5 we delve into the utilization of explicit knowledge bases to enable post-hoc explainability of models in a black-box manner [204, 208, 320]. Our core method is primarily based on *graph matching* between two sets of concepts. By optimally bridging these two sets, we obtain a set of conceptual edits, which ultimately act as explanations of "What is the optimal conceptual change". In other words, defining the lowest cost matching between these two concept sets provides us with theoretical guarantees of optimality, with the assignment cost being defined from the conceptual distances on a knowledge graph. In any case, we focus on model-agnostic interventions, in the sense that we do not optimize our algorithm to explain a specific model at a time. This black-box nature permits the extendability of our method to any model that can be probed by observing input-output relationships. Consequently, we present our research contributions in:

- Finding approximate minimal interventions for NLP classifiers using concept-level counterfactuals (Section 5.2.4), as well as visual classifiers harnessing graph-based counterfactuals (section 5.2.5). In the classic counterfactual setup where minimal changes are requested, we introduce novelties regarding accelerating the matching procedure as driven from distances stemming from knowledge graphs.

- Exploring the novel setup of non-minimal interventions for **semantic similarity** models (section 5.2.6).

2.1.4 Knowledge in vision-language explanations

In chapter 6, we present the utilization of knowledge graphs in explaining multimodal systems combining vision and language. As a basic use-case, we present knowledge enhancement in Visual Question Answering (VQA) leveraging counterfactual interventions as presented previously in section 5.2.6.

Consequently, we extend the basic algorithm for crafting conceptual CEs based on concept *substitutions* to allow for more operations, such as concept *deletions* and *insertions* [203, 69, 67, 39, 202, 316]. In that case, selecting between insertion, deletion and substitution of concepts translates into transversing a knowledge graph to discover the optimal set of such operations to propose counterfactual data instances within given datasets. Based on an existing framework for conceptual CEs, we present our contributions in the following tasks:

- **Text to image generation** (section 6.2). Specifically, we experiment on two challenging tasks: 1) Story visualization involves the sequential generation of frames based on related captions, maintaining continuity from frame to frame while accurately transferring concepts from the textual to the visual modality. 2) Scene generation takes traditional image generation one step further by requesting generating several interacting objects while respecting potential object attributes. The key question of "What is the optimal conceptual change" translates to finding optimal paths between textual and visual concepts, so that concepts extracted from the image should change to concepts present in the text, since text forms the ground-truth modality. Once again, optimally pairing visual and textual concepts is reflected to a graph matching problem between these two concept sets.

- **Hallucination detection in image captioning** (section 6.3). Ultimately, the setup is the inverse of the one leveraged in the text to image generation case: now, we have to discover conceptual edits in order to transit from the textual modality to the visual modality, which serves as the ground truth. Both concept and graph matching are tested in order to incorporate more complex cases, as the ones in which conceptual relationships are definitive. To conclude the value of the semantic edits framework proposed, we showcase that typical image captioning measures are fully uninformative in capturing conceptual discrepancies between the two modalities, strengthening the value of conceptuality in image captioning evaluation strategies.

2.1.5 Knowledge enhancement in vision-language tasks

In chapter 7 we focus on multimodal applications in conjunction to knowledge enhancement via external sources. By continuing the survey and taxonomy introduced in chapter 3, we present a thorough analysis of the vision-language tasks that have received knowledge enhancement applications in recent literature [206], involving tasks such as Visual Question Answering, Visual Reasoning, Visual Commonsense Reasoning, Image Captioning, Visual Dialog, Visual Storytelling, Image Generation. We then justify the need for considering LLMs as knowledge bases for multimodal tasks, serving scenarios where incorporating structured knowledge graphs is inefficient or even impossible, as in the case where knowledge retrieval and multi-step reasoning processes have to be employed [205].

2.1.6 LLMs as knowledge bases

Finally, in chapter 8 we focus on the usage of LLMs as Knowledge Bases in place of the traditionally employed structured Knowledge Graphs. LLMs are leveraged to enrich existing data harnessing the rich knowledge they have acquired during pre-training, thus boosting the final performance of pre-trained models for downstream tasks. We therefore present our contributions in the following:

- Initially, we review the usage of LLMs as knowledge bases for Vision-Language tasks in section 8.1, demonstrating the value of the approach and the current state of research.
- We enhance performance on the **Visual Word Sense Disambiguation (VWSD)** task using LLMs as knowledge bases (section 8.2). In this application, prompting LLMs to fuse knowledge in order to disambiguate terms in context improves the performance of knowledge-free approaches [157, 156].
- We delve into challenges concerning the **LLM capabilities** themselves; these are mostly tied to open research questions regarding hallucinations and LLM reasoning that may hinder the effortless adoption of LLMs for knowledge enrichment, as well as for other tasks in general [88, 249, 97, 141] (section 8.3). Finally, tasks where LLMs collapse are analyzed in section 8.4 [321, 77].

Overall, this thesis outlines and analyzes the varying contributions that external knowledge sources can have to advance expainability and performance of AI models without intruding to their particular architecture and internal mechanisms. This is a significant step towards reusable, efficient and adaptable algorithms that can serve several purposes with zero or minimal modification in their functionality to address significant gaps in research literature with the ultimate goal of advancing AI models to be more trustworthy, accountable, meaningful and competent.

Chapter 3

Types of knowledge

The incorporation of external knowledge sources in neural models has been an area of active research due to the advantage it offers in terms of performance enhancement and explainability. Knowledge is divided in several *senses*, referring to distinct categories or dimensions of knowledge that humans or AI systems can possess, each characterized by its source, nature, and application. Most common senses of additional knowledge offering those benefits are analyzed below.

Hierarchical knowledge refers to *is-a* relationships forming a tree structure, with the root serving as the most generic concept and parent node of all the rest, while leaves constitute the most specific concepts. For example, *cat is-a mammal* is an instance that represents such hierarchical relationships.

Lexical knowledge serves as a structured dictionary, offering linguistic rules, while being able to resolve issues such as word sense disambiguation. Lexical knowledge can be combined with hierarchical knowledge, providing hypernym/hyponym relationships.

Named entities cover a variety of proper names as instances of entities, and include names of people, locations, companies, organizations etc [98]. For example, the sentence *Joe Biden is the president of the United States* contains the named entities *Joe Biden* and *United States*.

Factual knowledge includes encyclopedic information of the world, such as the historical fact *WW2 lasted from 1939 to 1945*. Such knowledge can also refer to more specific scientific facts, including knowledge in medical, biological, chemistry domains and many more. Facts can also be combined with named entities, forming statements such as *Zebras live in Africa* (*Africa* is a named entity).

Commonsense knowledge is the self-evident perception of the world according to humans; *sugar is sweet* and *if I go out in the rain I'll get wet* are obvious commonsense statements. We can identify several discrete senses of commonsense knowledge, affecting aspects of the world a human experiences. Such subcategories refer to **similarity/dissimilarity** concepts, such as synonyms and antonyms of words. Another commonsense variant includes **part-whole** (part-of) relationships representing concepts belonging to more generic ones or consisting members of a group, for example *the bark is a part of a tree, the tree is part of the forest*. Part-whole in terms of lexical knowledge is expressed via **meronyms** (part) and **holonyms** (whole). **Utility** relationships describe usage scenarios, such as *a fork is used for eating* or capability (*wheels can rotate*). **Spatial** information offers knowledge about usual locations of objects in the physical world, for example *boats are situated near water*, or even geographic information, such as *Italy is located at Europe*, which sits on the intersection with **factual knowledge** and **named entities**. **Comparative knowledge** provides rules of comparison between objects, for example *leopards are larger than cats*. Such statements are crucial towards learning logical reasoning scenarios. **Numerical knowledge** addresses common enumerations in real life, providing facts such as *humans have two eyes*. **Intents**, **desires** and **plans** constitute another sense of commonsense knowledge, including facts such as *hungry people want to eat* and *a hungry person cooks to eat*. **Behavioral** knowledge results from logical reasoning over commonsense facts, forming rules e.g. *a child cannot drink 10 liters of water in one day*. **Creator knowledge** contains statements such as *a song is created by a musician* or *bread is made from wheat*.

Event/temporal knowledge contains chronological information and order of events, blending factual and commonsense knowledge. Events can refer to a large number of chronologically distinct time periods from widely known events such as *world wars, significant political events, sports, social/scientific movements* and many more, to more specialized events known to smaller audiences. Temporal sequences can contain chronologically ordered events: for example, *COVID-19 started in 2019. Vaccines for COVID-19 were developed during 2020* is a **factual sequence** of events. A **commonsense sequence** of events could contain information such as *Spring comes after winter.* Sequences may also refer to **causal** relationships with the cause preceding the event, such as *the boy dropped a glass of water and then the glass broke*, which can also be transformed to **hypothetical if/then** statements, for example *if a boy drops the glass of water, the glass will break*, or even **counterfactual** statements expressing what would have happened if an alternative scenario occurred, e.g. *if the boy had not dropped the glass of water, the glass would not have been broken.*

Visual knowledge contains images and possibly additional annotations to connect **visual perception** with **commonsense**. Attributes of objects, such as shape, color, texture and others can be connected with their visual counterpart, visualizing commonsense situations such as *tomatoes are red and round.* Visual knowledge is ideal for learning instances of the world involving object relationships and attributes, paving the way for more complex reasoning required in several multimodal tasks. **Spatial** relationships are naturally combined with images; for example *apples placed inside a bowl, bowl placed on a table.* More types of relationships can be further visualized, including **actions** between visual entities (*a girl is holding a tennis racket*), object **details** (*black and white stripped hat*), **part-whole** has-a relationships (*woman has long blonde hair*) or **scene text** (*a truck with Coca-Cola logo*). Those rather obvious statements can be extended to commonsense **assumptions** (*the temperature is low*), when an image of an icy landscape is provided. More complex visual instances can provide information about **intents** (*a customer enters a restaurant to eat, a person holding a suitcase and a passport plans to travel*), **causes/effects** (*a biker cycling out in the rain will get wet*), **factual** instantiations (*an ancient Greek temple of the 5th century BC, girls with Japanese kimono dresses*), **similarity reasoning** (*the dog's toy looks similar to a plate*), similarity including **named entities** (*a man looking similar to Brad Pitt*), **creator knowledge** (*the painting was created by a person holding a paintbrush*), **capability** (*a cat can jump on the tree branch*).

In our work, we divide external knowledge in three main categories, based on the degree of structure of their sources and its consequent advantages or limitations: **implicit knowledge**, present in a non-symbolic form, **explicit knowledge**, typically stored in structured knowledge bases, and **web-crawled knowledge**, acquired from various online sources, usually in unstructured format. Moreover, we can recognize the category of *internal knowledge* or *self-knowledge*, which does not rely on external sources, but rather obtains extra knowledge from the existing data. [206]

An overview of the available types of knowledge sources is provided in Figure 3.0.1.

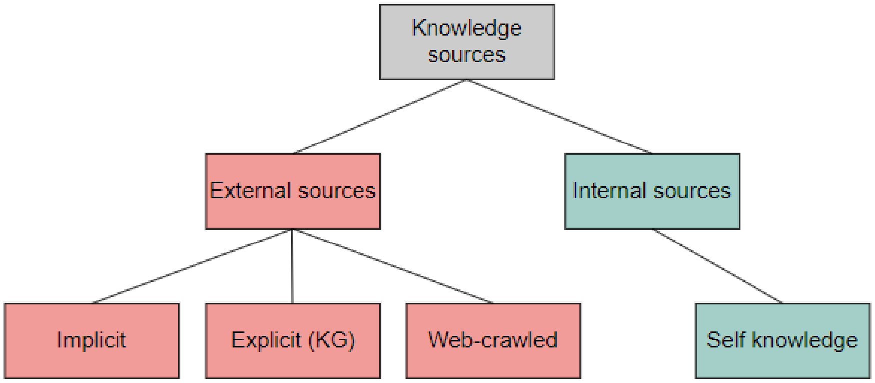


Figure 3.0.1: Overview of knowledge sources. [206]

3.1 Implicit knowledge

Implicit knowledge refers to information stored in a non-symbolic form, such as neural network weights. The indisputable popularity of neural architectures in recent deep learning literature has led to numerous relevant contributions, even if their primary goal deviates from knowledge representation. Unsupervised or self-supervised pre-training of transformer models (without any fine-tuning) is able to provide implicit knowledge in several downstream tasks, as introduced in the LM-as-KB paradigm [262]. In this case, the knowledge in an LM is embedded in high-dimensional weight matrices as statistical regularities among tokens. Conceptual relations (such as synonymy, hypernymy, or causal connections) are not stored symbolically but as geometric relationships in the embedding space; to this end, entities are represented as contextualized vectors, relations emerge as transformations or directions in the vector space, while queries are resolved through probabilistic inference rather than discrete lookup. This distributed encoding allows flexible generalization e.g., inferring unseen relations from analogical reasoning. In the multimodal case, incorporating large-scale linguistic and visual data in the pre-training stage can seemingly form unstructured knowledge bases, following the LM-as-KB paradigm.

Nevertheless, implicit knowledge is not always sufficient to answer questions requiring general, factual and commonsense knowledge, especially when rare information is requested. At the same time, the requested knowledge may be inconsistent, potentially returning varying responses to a given query. Additionally, its black-box nature raises concerns about how and what a pre-trained model has learned; for example, biased, outdated or erroneous data received during pre-training will be reflected in all later stages, resulting in decreased performance of the downstream model. Tracing back the source of such a problem is not possible due to the lack of interpretability tied to implicit knowledge bases, as well as the distributed nature of implicit knowledge. This distributed nature also contributes to the difficulty of querying an LM, shifting the weight towards exploring solutions such as fill-in-the-blank probing, fine-tuning, and more recently prompting. [7, 292] Another shortcoming of the LM-as-KB paradigm is the intractable reasoning process [7]. In fact, it is an open question to which extend LMs can reason or merely memorize as much as possible [363, 399, 118], while performing poorly on certain reasoning types [399, 144].

3.1.1 Large Language Models as Knowledge Bases

The surge of Large Language Models (LLMs) [31, 244, 245, 338, 339, 48, 8, 132, 133, 333] suggests that they can serve as ideal implicit knowledge bases [105] thanks to the massive information they have stored and their emergent abilities [363]. This novel implicit knowledge paradigm positions LLMs as dynamic, flexible, and ever-evolving systems that can adapt to diverse informational needs without the constraints of predefined schema or rigid structures, as in the case of knowledge graphs. In our work, we prove that LLMs are successful in enhancing a variety of tasks such as multimodal retrieval [157, 156] and image generation [253, 12, 13].

On the one hand, the LLMs' capacity to generalize across data and synthesize new insights from disparate pieces of information allows for unparalleled versatility in handling complex queries. Ón the other hand, some shortcomings such as errors, inconsistencies and hallucinations remain pertinent when LLMs are employed as knowledge bases [119, 416, 97], not to mention reasoning-related restrictions [118, 88, 249], which question their wide applicability in practice, especially in high-stakes decisions.

Nevertheless, the LLM-as-KB paradigm holds immense promise overall. The scalability, flexibility and the human-friendly interaction the LLMs offer tend to shift the knowledge representation paradigm to their favor. Their integration with domain-specific fine-tuning, retrieval-augmented generation (RAG), and counterfactual reasoning methods can mitigate inherent limitations and further enhance its capacity for reliable knowledge representation. As research advances, the LLM-as-KB paradigm could redefine how knowledge is stored, accessed, and interpreted, paving the way for AI systems that are not only informationally rich but also trustworthy and explainable.

3.2 Explicit knowledge

Explicit knowledge is based on clear, structured facts in the form of a knowledge graph and it is able to explicitly fill some aforementioned gaps that cannot be covered via the LM-as-KB paradigm. Even though several contemporary neural models have acquired a certain understanding of their participating modalities, they cannot effectively handle concepts and relationships they have never seen during training [124]. The same discrepancy may apply even when an implicit knowledge source is used, if the implicit distribution remains rather distant from the desired one. For example, a model trained on pairs of generic images and corresponding captions will inevitably present much lower metrics when asked to infer from medical images accompanied by relevant captions with scientific vocabulary. The same limitation is prevalent when there is a shortage of training data [124]. Although an intuitive scenario would suggest to repeat the pre-training procedure, so that this extra information will be reflected via updated neural weights, the pre-training cost is in reality computationally unaffordable [303] for the majority of research institutions. Even in that case, repeated occurrences of out-of-distribution data would demand from scratch pre-training or at least fine tuning each time, preventing the scalability of related tasks.

On the contrary, in the case of explicit knowledge bases, the contribution of the knowledge source can be measured and evaluated, offering valuable transparent insights regarding the role of knowledge. Such out-of-distribution information is well represented in structured knowledge graphs. Large scale knowledge can complement pre-trained models by extending their understanding to previously unseen concepts, either by substituting the need for extra training, or by enriching existing datasets to achieve more informative, fair and high quality representations, if (re-)training is necessary. Even in that case, pre-training demands can be reduced, achieving similar representation capabilities to larger models pre-trained without additional explicit knowledge. The quality of such representations is somehow controllable, a benefit which can be attributed to the explicit and transparent nature of KGs: issues regarding biases, errors, concept drifts and inconsistencies can be captured and resolved easily, exploiting automatic techniques or manual interventions. In any case, KGs should contain relevant information to the downstream task in order to be beneficial [124].

There are some downsides regarding the usage of explicit knowledge in the form of KGs. First, many KGs may require manual labor for data collection and curation. The same disadvantage also applies on the construction and maintenance of the graph itself. In certain cases, such as in the medical domain, experts are necessary in order to design and construct dedicated KGs. Moreover, there are difficulties regarding alignment and co-operation between different KGs [124], thus sometimes limiting in practice the improvements they offer.

Combining implicit and explicit information can offer advanced capabilities to downstream tasks, as implicit sources can fuse large-scale general knowledge to a model, while explicit sources can fix errors, enrich existing knowledge and increase a model's transparency. [206]

3.3 Web-crawled knowledge

Web-crawled knowledge refers to unstructured knowledge obtained from the web, which is able to combine the benefits present in implicit and explicit knowledge bases. There is no need for labelled data, but also no need for expensive pre-training, which is one major limitation of implicit knowledge. Online sources are readily accessible, while the amount and the content of retrieved knowledge is easily controlled and customized to the task's needs. A questionable part of web-scrapped knowledge is its quality, as it is hard to validate each available web source. Low-quality data may deteriorate the final performance of the model, therefore time and effort has to be invested in techniques that automatically ensure high-quality data. Web knowledge can offer some amount of transparency, as a sentence leading to the final prediction can be tracked, even if reasoning is not as fully explicit, as in cases of structured graphs.

3.4 Internal knowledge

Internal or **self-knowledge** is a knowledge type that does not rely on any external source, as it is obtained from existing textual and visual data themselves. For example, producing a scene graph enables learning more fine-grained representations compared to merely utilizing VL data in their original format [397]. Self-knowledge has demonstrated improvements in downstream model performance, especially when detailed disambiguation is necessary, as it enables better associations between existing data. However, self-knowledge does not extend the knowledge a model has already acquired from the data it has been trained on. Furthermore, it is prone to errors associated with the knowledge acquisition process, such as scene graph generation errors. [206]

Chapter 4

Counterfactual explanations

This thesis primarily focuses on harnessing external knowledge to obtain *explanations* with the goal of interpreting either neural models or automatic evaluation methods. Knowledge graphs (KGs) excel in tasks requiring structured, interpretable, and explainable representations of knowledge, making them ideal candidates for explainability ventures where traceability, trustworthiness and semantic clarity are essential.

Since there is a broad spectrum of possible explanation methods, we mainly focus on *counterfactual explanations*: through them we attempt to answer the question of what needs to be changed in a given instance to observe an alternative outcome. The set of changes towards the alternative outcome answer why a system made a decision in a *cause-effect* manner. Under this interpretation, the input features present in data act as the *cause*, while the output of a system acts as the *effect*. Counterfactuals closely represent the way humans are performing reasoning, allowing what-if scenarios to understand situations, which are ultimately based on altering controllable variables to transit to the alternative scenario. [259, 100]

4.1 From Causality to Counterfactuals

The foundation of counterfactual reasoning lies in causal inference, which aims to uncover not just correlations but causal dependencies between variables. Judea Pearl's structural causal models (SCMs) formalized this reasoning by introducing the concept of *interventions*—the act of forcibly changing a variable to observe downstream effects [259]. In the causal hierarchy, counterfactual reasoning resides at the top, above associational and interventional reasoning. While associational reasoning describes what *is* (i.e., observed data) and interventional reasoning describes what *will happen if we do* (i.e., outcome of manipulation), counterfactual reasoning explores what *would have happened if things were different*. This capability distinguishes explanation from mere prediction, as counterfactuals demand an understanding of the underlying causal mechanisms.

Formally, an SCM comprises a set of structural equations that define how each variable in a system is generated from others and an exogenous noise component. Intervening on a variable corresponds to replacing one of these equations with a fixed value, effectively simulating an alternative world. Within explainable AI, counterfactuals operationalize this same principle: changing the input features simulates an intervention in the model's decision-making process. Thus, counterfactual explanations bridge human reasoning (*what-if* thinking) and machine-level prediction models.

4.2 Desiderata for Counterfactual Explanations

In terms of machine learning models, a counterfactual explanation targeting a predictive model M comprises the *minimal* possible change in the input features to alter the predicted outcome: given a model M and a corresponding prediction $y = M(x)$, where x refers to an input instance, a counterfactual explanation x' is a minimally different instance to x that changes the prediction $M(x') \neq y$. The notion of minimality between instances may be different between applications and datasets. [100]

Several properties determine whether a counterfactual explanation is meaningful, useful, and reliable. Following [355, 100], these properties can be summarized as:

Plausibility: The counterfactual instance x' should correspond to a realistic point in the data distribution. Implausible or non-existent configurations (e.g., "a person aged -3 years") diminish the value of an explanation.

Minimality: The perturbation $\Delta x = x' - x$ must be as small as possible while still changing the prediction. This ensures that only the most influential factors are identified.

Diversity: When multiple counterfactuals exist, providing a diverse set enhances the user's understanding of multiple causal pathways.

Actionability: Counterfactuals should be grounded in changes the user can realistically act upon. For example, "increase annual income by $5,000" is actionable, whereas "change gender" is not.

Sparsity: Explanations should ideally involve modifying only a small number of features. Sparse explanations are easier for humans to comprehend and act upon.

Proximity and Coherence: The counterfactual should be close to the original instance not only numerically but semantically, preserving the internal consistency of features (e.g., age, education, and occupation relationships).

These properties collectively ensure that counterfactual explanations are not only technically valid but also human-aligned, facilitating trust and interpretability. However, balancing these properties is non-trivial: for example, minimizing the distance between x and x' may yield counterfactuals that are unrealistic, while enforcing realism might require larger changes. Consequently, modern approaches often formulate counterfactual generation as an optimization problem that jointly minimizes distance while maximizing plausibility, often under the constraints of a trained model or data distribution.

Another aspect of counterfactual explanations in machine learning concerns the accessibility an explanation can have within the predictive model. Typically, the model M under explanation is *black-box*, i.e. the model is not interpretable by nature (for example, the model can be a neural network[1]) and/or there is no access to its internal workings (irrespectively of the nature of the model, we are unable to know its architecture or access it externally[2]). [100] This characteristic makes counterfactual explanations highly favorable in practice, as they allow explaining a wide array of neural models, without the need to tailor explanation techniques specific to the respective model. Therefore, when employing counterfactuals, we are able to produce *model-agnostic explanations* [204, 78, 62, 203, 320, 154, 67, 69, 208, 316, 202], allowing for transferability and reusability of the explanation algorithm.

4.3 Counterfactual Explanations vs. Other XAI Methods

Explainable AI encompasses a broad range of techniques that aim to make machine learning models transparent, interpretable, and trustworthy. These methods differ fundamentally in their goals, assumptions, and outputs. While *counterfactual explanations* focus on describing what minimal changes to the input would alter the model's decision, other XAI methods such as *feature attribution*, *surrogate modeling*, *prototype-based explanations*, and *example-based explanations* aim to describe model behavior from complementary perspectives. Understanding how counterfactual explanations relate to these approaches provides important insights into their strengths, limitations, and appropriate use cases.

4.3.1 Feature attribution methods

Feature attribution methods are among the most widely used XAI techniques. Their goal is to assign importance scores to input features to indicate how much each feature contributed to a particular model

---

[1] Models that are based on explicit rules, such as decision trees or linear regression are interpretable by nature. On the other hand, models performing complex transformations of input features, such as neural networks, lead to untractable decision-making steps.

[2] An example could be ChatGPT, which can only be accessed via its API. Even if ChatGPT was a fully interpretable model or not, it is impossible to access it.

prediction. Methods such as *LIME* (Local Interpretable Model-Agnostic Explanations) [285], *SHAP* (SHapley Additive exPlanations) [198], and *Integrated Gradients* [328] fall within this category.

Feature attribution provides a snapshot of the internal reasoning process of a model by identifying which features were influential in producing a decision. For example, in a loan approval model, such methods might indicate that "income" and "credit score" were the strongest contributors to the model's output. However, they do not indicate how the decision might change if a feature were different. In contrast, counterfactual explanations explicitly answer such "what-if" questions: *"If the applicant's income were increased by $5,000, the loan would be approved."*

This distinction has important implications. Feature attribution is descriptive (it summarizes *which features mattered* in the decision) whereas counterfactual explanations are prescriptive (they indicate how the decision *could be changed*). The latter thus offers a more actionable and user-centric form of interpretability, particularly in high-stakes settings such as finance, law, or healthcare, where understanding what actions can lead to a different outcome is essential.

4.3.2 Surrogate model explanations

Another major category of XAI methods is surrogate modeling, which involves training an interpretable model to approximate the behavior of a complex, black-box model. Examples include global surrogates (e.g., decision trees or rule lists trained on the black-box model's predictions) and local surrogates (e.g., LIME's local linear approximations).

The advantage of surrogate models is that they provide an interpretable representation — such as a set of rules or a decision boundary — that can be analyzed and visualized. However, they depend heavily on how well the surrogate approximates the original model. If the surrogate's fidelity is low, the resulting explanations can be misleading.

Counterfactual explanations differ from surrogate models in that they do not attempt to approximate the global or local decision surface of the model. Instead, they probe it directly by identifying boundary-crossing points — specific input variations that lead to a change in prediction. Thus, counterfactuals offer localized but highly faithful explanations of the model's behavior in the immediate neighborhood of a data instance. Whereas surrogate models explain *how the model behaves*, counterfactuals explain *what would make the model behave differently*.

4.3.3 Prototype and example-based explanations

Prototype-based explanations operate on the intuition that humans understand complex categories by reference to representative examples. Methods such as *case-based reasoning*, *prototypical networks*, or *nearest-neighbor explanations* provide users with instances from the dataset that are similar to the input under consideration or representative of a particular class.

For instance, a prototype-based explanation for an image classifier might state: "This image was classified as a *dog* because it resembles this prototypical image of a dog." Such explanations leverage the human ability to reason analogically. However, they lack explicit causal directionality: they describe similarity but not change.

Counterfactuals, in contrast, provide contrastive examples rather than prototypical ones: "This image would have been classified as a *cat* if the ears were pointier and the nose smaller." Humans naturally think in contrastive terms — we rarely ask "Why X?" but rather "Why X instead of Y?" — making counterfactuals more psychologically aligned with how people seek understanding. Empirical studies in cognitive psychology confirm that humans tend to generate explanations by mentally simulating alternative realities — precisely the kind of reasoning counterfactuals formalize.

4.3.4 Saliency and gradient-based methods

Saliency-based methods, primarily used in computer vision and deep learning, aim to identify which regions of an input (e.g., pixels in an image or tokens in a text) most influenced a model's output. Techniques such as *Grad-CAM* [298] or *SmoothGrad* [313] visualize the gradient of the output with respect to the input, effectively highlighting where the model "looked."

While saliency maps are useful for interpreting deep neural networks, they face several limitations. First, they often produce noisy or unstable visualizations that can vary significantly with minor input perturbations. Second, they are descriptive but not explanatory — they show which regions are correlated with a decision but not what changes would alter it.

Counterfactual explanations overcome these limitations by introducing causal direction: instead of highlighting where the model focuses, they simulate how the model's decision changes when the input is systematically modified. For instance, instead of merely visualizing the pixels responsible for a "dog" classification, a counterfactual method might generate an alternative image showing that removing the snout or changing the fur pattern would cause the classifier to label it as a "cat." This causal and generative nature makes counterfactual explanations more suitable for evaluating robustness and reasoning pathways in vision models.

4.3.5 Rule-based and symbolic explanations

Some XAI approaches construct symbolic or rule-based explanations by extracting human-readable logical rules that describe model behavior. Decision rules of the form $\textit{IF } feature_1 > \textit{threshold AND } feature_2 < \textit{threshold THEN class} = A$ provide explicit, interpretable reasoning structures. These are intuitive but often limited in expressiveness and applicability, especially for models with complex, nonlinear decision boundaries.

Counterfactual explanations can complement or extend rule-based reasoning by showing how transitions between rule regions occur. For example, given two rules defining different output classes, a counterfactual explanation identifies the minimal change in feature space that moves an instance from satisfying one rule to satisfying another. Thus, counterfactuals bridge the gap between symbolic interpretability and numerical optimization-based explanations.

4.3.6 Complementarity and integration of methods

In practice, counterfactual explanations are not necessarily competitors to other XAI methods but rather complementary tools that provide a distinct type of insight. Feature attribution methods answer *"Which features matter?"*, surrogate models answer *"How does the model behave overall?"*, and counterfactuals answer *"What needs to change to obtain a different result?"*.

Integrating these perspectives can yield richer and more trustworthy explanations. For example, one could use feature importance scores to identify relevant features and then use counterfactual reasoning to determine actionable modifications among them. Similarly, counterfactual generation can be constrained by knowledge derived from surrogate or causal models to ensure plausibility and coherence. In multimodal or high-dimensional domains, combining saliency-based localization with counterfactual editing (e.g., generating semantic image edits where the saliency map is most active) can provide both visual intuition and causal reasoning.

4.4 Knowledge and semantics in counterfactual generation

Even though the field of eXplainabile AI (XAI) has enjoyed several implementations capable of peeking inside otherwise opaque models, there have been some concerns regarding the field's fundamentals [5]. For example, a parallel line of research targets adversarial attacks to neural models, which are not primarily focused on explaining the decision-making process of a model, even though they reveal the model's vulnerabilities and instabilities via perturbations similar to the ones applied for counterfactuals. For example, an adversarial attack could be the modification of a certain pixel in an image that changes the prediction of an image classifier from class A to class B, indicating lack of robustness on respective classifiers. However, this change does not really explain the causal trajectory that the classifier followed to perform this classification; in other words, in terms of human intuition, we would be unable to answer *why* this change occurs, since a change in a pixel's value would be imperceptible from our perspective. This fact shifts our attention to high-level characteristics that are meaningful to humans: it showcases that *semantics* are crucial towards understanding the reasons behind a decision-making process.

Therefore, since counterfactuals hold a causal role in XAI, semantics should be crucial for them. This assumption has been validated in prior literature, concluding to the statement that *there can be no explanation without semantics*. At the same time, computationally speaking, a vector holding semantic information has the same mathematical representation to a vector that holds pixel values. Therefore, while the input to the explainer retains the same format with or without semantics, the informativeness of the resulting explanation is highly diverging between cases that semantics are employed or not, concluding the fundamental role of semantics in XAI. [32] As an example, controlled interventions applied on raw data, such as pixel-level edits do not align with human perception (change a pixel from red to blue and explain the classifier's change in decision) while semantic-based edits do (change a person holding a flower with a person holding a gun and explain the classifier's change in decision).

A key insight of recent research is that generating meaningful counterfactuals often requires access to *structured knowledge* [32, 69, 316]. Purely data-driven perturbations (e.g., pixel-level or feature-wise changes) can lead to counterfactuals that, although effective in changing model predictions, fail to provide insight into the model's reasoning. For instance, changing a few pixel intensities in an image might switch a classifier's output from "cat" to "dog," yet this change is imperceptible to humans and devoid of semantic interpretability.

Knowledge graphs, ontologies, and other structured representations of world knowledge offer a solution to this limitation. By encoding relationships between entities (such as "dog" is-a "animal" or "car" has-part "wheel"), they allow for counterfactual interventions that are semantically grounded. When knowledge is embedded in vector space representations, such as through graph embeddings or semantic encoders, the resulting counterfactuals can be both mathematically tractable and conceptually meaningful. For example, substituting "dog" with "cat" in a scene description preserves semantic coherence while altering the model's prediction in a human-understandable way.

Integrating semantics into counterfactual reasoning also aligns with cognitive theories of human reasoning [342, 34, 259, 227]. Humans do not typically reason in terms of raw sensory inputs, but rather in terms of abstract, conceptual entities and their relationships. When people imagine counterfactuals — "What if the person had smiled instead of frowned?" — they manipulate meaningful, high-level attributes rather than low-level data. Thus, embedding semantic knowledge into counterfactual frameworks enhances both their interpretability and their alignment with human cognitive processes.

4.5 Semantic distances and conceptual interventions

Finally, in the cases of semantic-based counterfactuals we need to consider the notion of minimality [78, 208]: minimal changes denote finding the *semantically closest instance* to the one to be changed. The minimality needs to be *guaranteed* in order to still remain a counterfactual, requesting a rather deterministic way of assigning distances to semantics. These distances should correspond to real-world scenarios in order to be meaningful: for example, the semantic *dog* can be semantically close to the semantic *cat*, as they are both *domestic animals*. Under the same schema of semantic relationships, the semantic *dog* is not as close to the semantic *ocean*, at least in comparison to the semantic *cat*. Therefore, crafting semantically rich counterfactuals is transformed to finding a way to deterministically imbue meaningful distances to concepts present in data, and then perform interventions based on them.

To operationalize semantic counterfactuals, one must define a way to measure distances between concepts [78, 62, 69]. This step is crucial because the notion of minimality in counterfactual explanations depends on a meaningful metric of closeness between features. Several approaches exist to define semantic distances:

- **Taxonomic distances:** In ontological structures such as WordNet [226], distance can be measured as the length of the shortest path between two concepts in the hierarchy (e.g., "dog" and "cat" are closer than "dog" and "ocean").
- **Distributional semantics:** Using embedding-based representations (e.g., Word2Vec [225], BERT [63], or CLIP [273] embeddings), distance can be quantified as the cosine similarity between vectors, capturing semantic proximity learned from large corpora.
- **Knowledge graph embeddings:** When concepts are embedded through graph-based learning methods (e.g., TransE [28], RotatE [327]), the relational structure of the graph informs the distance measure,

enabling counterfactual generation that respects relational semantics.

By assigning deterministic and interpretable distances to semantic entities, one can perform conceptual interventions that correspond to realistic world changes. This allows for the construction of counterfactuals that not only modify model predictions but also mirror how humans would imagine alternative realities — a fundamental requirement for truly explainable artificial intelligence.

Chapter 5

Explicit knowledge bases for explainability

One of the advantages of harnessing knowledge graphs for explainability is their deterministic nature, allowing highly desirable explainability pipelines, where each step of the explanation algorithm is tractable and well-justified. Moreover, knowledge graphs provide curated and grounded world knowledge in conceptual units, enabling a clear definition of conceptual distances. All those characteristics can perfectly serve the counterfactual question of *"What (concepts) should be changed to alter the observed outcome?"* by suggesting conceptual edits driven by knowledge graph concepts.

Knowledge graphs provide meaning and relatedness to concepts that constitute our counterfactual explanations. According to the degree of expressivity we wish to include in our explanations, we decide on whether we work using concepts on their own, or whether we also include their in-between semantic relationships, in which case we obtain *semantic graphs*. The expressivity gains are counterbalanced by computational constraints naturally accompanying graph structures, a factor that further complicates the accuracy of derived explanations.

In the following section 5.1, we are covering existing techniques in concept-based and graph-based matching, thus highlighting powerful directions that can be effortlessly leveraged to provide semantically rich counterfactual explanations.

5.1 Matching algorithms for counterfactual explanations

5.1.1 Concept Matching using bipartite graph structures

The first scenario we explore is to optimally match sets of concepts, given a notion of conceptual distance, as imposed by a knowledge graph. Seeking for the *minimum total distance* between two given concept sets in order to respond on *which concepts should be minimally changed* so that the two sets become identical, we ultimately request an optimization algorithm that finds the *minimum cost* conceptual combination of intra-set concepts.

If we place such concepts on a graph structure, the source concepts will form a set S, while the target ones form a set T. This is actually a *bipartite graph* $G = (V, E)$: this graph structure allows us to perform several assignment operations, matching concepts from $s \in S$ to concepts $t \in T$ ($S \cup T = V$), thus defining what concepts should be changed. On this bipartite graph, edge weights $w_e > 0 \in W$ of the edge set E directly reflect the conceptual distance of possible $S - T$ pairs, providing an accurate description of how expensive (i.e. heavy) an assignment is. An optimization algorithm can find the cheapest $S - T$ matching $M \subseteq E$ considering all possible unique pairs by minimizing the overall weight $\sum w_e$ of the assignment. The *minimum weight bipartite matching* covers all the nodes of the $min(|S|, |T|)$ set of G. Fortunately, graph matching is a well-studied optimization problem in graph literature, offering an abundance of deterministic as well as approximate solutions.

A purely deterministic solution is able to showcase all optimal pathways between S and T to perform as many conceptual substitutions as possible. This solution is highly explainable, offering tractability of operations, while preserving optimality, since deterministic minimum weight bipartite matching algorithms are guaranteed to offer the optimal solution. A guarantee that all possible substitutions for S will be performed On the other hand, if the given bipartite graph is large enough, the assignment may be inefficient and computationally heavy, thus possible approximations may be needed. In that case, the fully interpretable and optimal assignments can be partially sacrificed for the sake of efficiency by employing faster neural methods for graph matching.

5.1.2 Bipartite matching as an assignment problem

Linear Assignment Problem (LAP)

The *assignment problem* is a fundamental combinatorial optimization problem that consists of finding an optimal one-to-one assignment between two sets of elements of equal cardinality n, given a cost or weight for each possible pair. Formally, regard sets S and T:

$$S = \{1, \ldots, n\}, \quad T = \{1, \ldots, n\} \tag{5.1.1}$$

Let $W \in \mathbb{R}^{n \times n}$ be a weight matrix where W_{ij} represents the benefit (or cost) of assigning element $i \in S$ to element $j \in T$.

The *Linear Assignment Problem* (LAP) can then be formulated as:

$$\begin{aligned} \max_{P \in \mathcal{P}} \sum_{i=1}^{n} \sum_{j=1}^{n} W_{ij} P_{ij}, \\ \textit{subject to}: \\ \sum_{j=1}^{n} P_{ij} = 1, \quad \forall i = 1, \ldots, n, \\ \sum_{i=1}^{n} P_{ij} = 1, \quad \forall j = 1, \ldots, n, \\ P_{ij} \in \{0, 1\}. \end{aligned} \tag{5.1.2}$$

Here, P is a permutation matrix representing the assignment, and $\mathcal{P}$ denotes the set of all valid permutation matrices. The objective is to maximize the total weight (or minimize total cost, depending on the application).

Rectangular Linear Assignment Problem (RLAP)

In many practical applications such as concept matching, the two sets of nodes to be matched may *not* have the same cardinality. Let:

$$|S| = n, \quad |T| = m, \quad n \leq m, \tag{5.1.3}$$

where S and T are the sets of nodes from the two conceptual sets. The goal is then to find a one-to-one assignment from each node in the smaller set S to a distinct node in T that maximizes total similarity, while some nodes in T may remain unmatched.

Formally, given a weight (or similarity) matrix $W \in \mathbb{R}^{n \times m}$, the RLAP can be formulated as:

$$\max_{P \in \mathcal{P}} \sum_{i=1}^{n} \sum_{j=1}^{m} W_{ij} P_{ij}, \tag{5.1.4}$$

where $\mathcal{P}$ is the set of *partial permutation matrices* satisfying:

$$\sum_{j=1}^{m} P_{ij} = 1 \quad \forall i = 1, \ldots, n, \qquad \sum_{i=1}^{n} P_{ij} \leq 1 \quad \forall j = 1, \ldots, m, \qquad P_{ij} \in \{0, 1\}. \tag{5.1.5}$$

5.1.3 Algorithms for Bipartite Matching

Several algorithms have been developed to solve the maximum-weight bipartite matching problem efficiently.

a) Hungarian Algorithm. The Hungarian algorithm [160] is a combinatorial optimization method designed to find a perfect matching in a weighted bipartite graph that maximizes (or minimizes) the total assignment weight. Given two sets of nodes, S and T, and a weight matrix $W \in \mathbb{R}^{|S|\times|T|}$ representing the value of matching each node pair, the goal is to determine a permutation matrix P such that the overall weight is maximized using equation 5.1.5.

The Hungarian algorithm provides an exact solution with cubic time complexity, $O(n^3)$, where $n = \max(|S|, |T|)$. It operates by constructing a cost matrix $C = -W$ and finding a minimum-cost perfect matching through combinatorial optimization. Specifically, it executes the following steps:

1. **Row and column reduction:** Subtract the minimum value in each row and column from all elements in that row or column, ensuring that every row and column contains at least one zero.
2. **Zero assignment and cover:** Attempt to assign zeros such that each row and column has at most one assignment. If a complete matching is possible, the solution is optimal.
3. **Adjustment step:** If a complete matching is not yet possible, cover all zeros with a minimum number of horizontal and vertical lines, identify the smallest uncovered element, subtract it from uncovered elements, and add it to elements at intersections of lines. Repeat until a perfect assignment is found.

For large or sparse graphs, the auction algorithm is often preferred.

b) Auction Algorithm. The auction algorithm, originally proposed by Bertsekas [22], is an iterative, primal-dual method for solving the Linear Assignment Problem (LAP). Its core idea is to interpret the assignment problem as a market, where agents compete to acquire items based on their valuations, which correspond to assignment weights.

Consider a bipartite graph with sets of agents $S = \{1, \ldots, n\}$ and items $T = \{1, \ldots, n\}$, and let w_{ij} denote the weight or value of assigning agent i to item j. The objective is to find a one-to-one assignment P maximizing the total weight, under the optimization imposed in equation 5.1.5.

In the auction framework, each agent acts as a bidder, and each item has a price p_j reflecting its current value in the auction. Each bidder i evaluates the net benefit of acquiring item j as:

$$b_{ij} = w_{ij} - p_j. \tag{5.1.6}$$

During each round, unassigned agents bid for the items that maximize their net benefit. The price of an item is increased according to the difference between the highest and second-highest bids, plus a small increment $\epsilon > 0$ to ensure convergence:

$$p_{j^*} \leftarrow p_{j^*} + (b_{ij^*} - b_{ij'}) + \epsilon, \tag{5.1.7}$$

where j^* is the agent's top choice and j' is its second choice. The item is then temporarily assigned to the highest bidder, potentially displacing a previous owner.

The auction algorithm converges to an ϵ-optimal solutions, meaning the total assignment weight is within $n\epsilon$ of the optimal weight. The computational complexity is $O(n^2 \log(nC/\epsilon))$, where C bounds the maximum weight magnitude. Empirically, it is often faster than the Hungarian algorithm for sparse or structured graphs. Bidders act independently, making the method highly parallelizable and scalable for large graphs.

c) Linear Programming Relaxation. Another approach involves relaxing the binary constraints on the assignment matrix P from $\{0,1\}^{|S|\times|T|}$ to the continuous interval $[0,1]^{|S|\times|T|}$. This linear programming relaxation allows the problem to be solved in polynomial time using standard LP solvers or via Sinkhorn normalization, providing an efficient approximate solution that can be further refined if needed.

5.1.4 Graph matching

Graphs provide a powerful and expressive data structure for representing entities (nodes) and their pairwise relationships (edges). They naturally encode complex, structured information that cannot be easily captured by vector-based or tabular representations; the same applies to concept-based explanations, which can be limited in case their in-between relationships are important. One fundamental computational problem involving graphs is *graph matching* –extending the previously analyzed concept matching– which concerns finding correspondences between the nodes (and edges) of two or more graphs that preserve structural or semantic similarity.

At a high level, graph matching aims to answer the question:

Given two graphs G_1, G_2:

$$G_1 = (V_1, E_1) \quad \text{and} \quad G_2 = (V_2, E_2),$$

which nodes of G_1 correspond to which nodes of G_2? Depending on the context, this correspondence may be exact (isomorphism) or approximate (similarity-based). The problem plays a central role in numerous applications, including image and pattern recognition [51], molecule alignment in cheminformatics [280], ontology alignment and knowledge integration [309], and, more recently, XAI algorithms involving semantic scene graphs or causal structures [67, 69].

5.1.5 Exact versus Inexact Graph Matching

Two principal categories of graph matching are commonly distinguished:

1. **Exact graph matching**, where the goal is to determine whether two graphs are identical in structure up to a relabeling of their nodes, i.e., whether they are isomorphic.
2. **Inexact (or approximate) graph matching**, where the objective is to find the best possible correspondence between graphs that may differ in size, connectivity, or attributes.

Exact graph matching corresponds to the classical *graph isomorphism problem*, which seeks a bijective mapping:

$$f : V_1 \to V_2 \tag{5.1.8}$$

such that

$$(u, v) \in E_1 \Leftrightarrow (f(u), f(v)) \in E_2. \tag{5.1.9}$$

This ensures perfect structural equivalence. However, real-world data are rarely noise-free or perfectly aligned; thus, exact isomorphism is too rigid for most practical purposes.

Inexact graph matching, on the other hand, relaxes the isomorphism constraints to allow partial correspondences and structural deviations. The goal becomes finding a mapping f that minimizes a cost function $C(f)$, which quantifies the dissimilarity between the structures or node/edge attributes of the two graphs. This formulation transforms the problem into an optimization task and allows meaningful comparison between graphs of different sizes, topologies, or semantic content.

5.1.6 Formulations and Optimization Objectives

In general, the inexact graph matching problem can be formulated as an optimization over a correspondence matrix:

$$X \in \{0,1\}^{|V_1|\times|V_2|} \tag{5.1.10}$$

where $X_{ij} = 1$ indicates that node i in G_1 is matched with node j in G_2. The objective can take multiple forms, but a common formulation seeks to minimize the structural difference:

$$\min_X \|A_1 - XA_2X^\top\|_F^2, \tag{5.1.11}$$

where A_1 and A_2 denote the adjacency matrices of the two graphs, and $\|\cdot\|_F$ is the Frobenius norm. This expression measures how well the edge structures align under the proposed correspondence X. Constraints such as one-to-one matching are typically enforced:

$$X\mathbf{1} = \mathbf{1}, \quad X^\top\mathbf{1} = \mathbf{1}. \tag{5.1.12}$$

However, exact optimization of this combinatorial problem is NP-hard [51]. As a result, numerous approximate or heuristic methods have been proposed, ranging from spectral relaxations and probabilistic formulations to graph neural network (GNN)-based embeddings.

5.1.7 Major Approaches to Graph Matching

Several algorithmic families have emerged for tackling the graph matching problem:

(a) Combinatorial and Exact Algorithms. Early methods such as the Ullmann algorithm [344] and the VF2 algorithm [52] rely on depth-first search (DFS) and constraint propagation to efficiently test for subgraph isomorphism. Regarding computational constraints, the worst-case time complexity for the Ullmann algorithm is exponential, typically denoted as $O(|V_1|!)$, since it may explore all possible permutations of node correspondences. The VF2 algorithm improves on Ullmann's method by introducing effective feasibility rules and state-space pruning. Nevertheless, its theoretical complexity remains exponential in the number of nodes, with complexity of $O(b^d)$, where b is the branching factor and d the depth of the search tree. In practice, however, the VF2 algorithm performs efficiently on sparse and labeled graphs due to pruning. These approaches are suitable for small to medium-sized graphs but scale poorly with increasing graph size due to their prohibiting computational complexity.

(b) Spectral Methods. Spectral graph matching leverages the eigen-decomposition of the graph Laplacian or adjacency matrices, transforming the discrete matching problem into a continuous one. By comparing eigenvectors (which encode global structural information), these methods relax the discrete optimization problem into a continuous one. Examples include the Umeyama algorithm [345], which aligns graphs based on their spectral embeddings. This algorithm computes an approximate permutation matrix P by aligning the eigenvectors of the Laplacians:

$$L_1 = U_1\Lambda_1U_1^\top, \qquad L_2 = U_2\Lambda_2U_2^\top, \tag{5.1.13}$$

and setting $P = U_1U_2^\top$. This relaxation allows matching to be solved in polynomial time, dominated by the eigen-decomposition cost of $O(|V|^3)$, where $|V|$ is the number of nodes in the graph. This makes spectral methods significantly more scalable than combinatorial ones. However, they are sensitive to noise and structural perturbations, and they cannot guarantee integer-valued (i.e., discrete) node correspondences, often requiring post-processing to restore permutation matrices.

(c) Probabilistic and Relaxation-Based Approaches. Probabilistic methods, such as the graduated assignment algorithm [92], treat matching as a continuous optimization over soft correspondence probabilities. Similarly, relaxation labeling and expectation–maximization schemes provide flexible frameworks for approximate matching by iteratively refining node correspondences. Specifically, these algorithms typically exhibit a per-iteration complexity of $O(|V_1|^2|V_2| + |V_1||V_2|^2)$, and converge in a number of iterations that is problem-dependent. While not polynomial in strict terms, their empirical runtime is often acceptable for graphs with up to several thousand nodes. However, convergence to the global optimum is not guaranteed.

(d) Graph Edit Distance. Another prominent approach conceptualizes matching as an edit operation problem: transforming one graph into another through a sequence of edits (node/edge insertion, deletion, or substitution). The *graph edit distance* (GED) quantifies the cost of these operations [82], capturing structural and attribute differences simultaneously. Formally, if Σ denotes the set of edit operations and $c : \Sigma \to \mathbb{R}_{\geq 0}$ assigns a cost to each operation, then the GED is:

$$\mathrm{GED}(G_1, G_2) = \min_{\pi \in \Sigma^*} \sum_{o \in \pi} c(o). \tag{5.1.14}$$

Computing the exact GED requires exploring an exponentially large search space of $O(|V_1|!)$, and is therefore NP-hard. Approximation algorithms, such as beam search, bipartite matching relaxations, or A* heuristics, are used in practice to reduce runtime to sub-exponential complexity, often around $O(|V_1|^2|V_2|^2)$ for approximate GED solvers. Despite their high complexity, GED-based approaches remain attractive due to their flexibility and their ability to incorporate both structural and attribute-based costs, proving particularly useful when graphs differ significantly in size or topology.

5.1.7.0.1 (e) Learning-Based and Neural Approaches. Recent work leverages machine learning, particularly GNNs, to learn embeddings that capture structural and semantic properties of nodes and edges. These embeddings are then compared using distance metrics to infer correspondences. Notable examples include neural graph matching networks [180], differentiable matching layers [361], and permutation-invariant architectures that relax discrete constraints into differentiable forms. Such methods achieve high scalability and adaptability, particularly in domains like visual correspondence or scene graph alignment.

5.1.8 Graph Matching as an Assignment Problem

A significant line of research reformulates the graph matching problem as a *quadratic assignment problem* (QAP), one of the most general and well-studied formulations in combinatorial optimization. This perspective provides both theoretical insight and a foundation for developing scalable approximate algorithms.

5.1.8.1 Quadratic Assignment Problem (QAP) Formulation

Given two graphs

$$G_1 = (V_1, E_1), \qquad G_2 = (V_2, E_2),$$

with corresponding adjacency matrices $A_1 \in \mathbb{R}^{|V_1|\times|V_1|}$ and $A_2 \in \mathbb{R}^{|V_2|\times|V_2|}$, the objective of graph matching can be written as minimizing the structural discrepancy between them under a permutation matrix $P \in \{0,1\}^{|V_1|\times|V_2|}$:

$$\min_{P \in \mathcal{P}} \|A_1 - PA_2P^\top\|_F^2, \tag{5.1.15}$$

where $\mathcal{P}$ denotes the set of all valid permutation matrices, satisfying $P\mathbf{1} = \mathbf{1}$ and $P^\top\mathbf{1} = \mathbf{1}$.

Expanding the objective and omitting constant terms leads to the equivalent quadratic form:

$$\max_{P \in \mathcal{P}} \operatorname{trace}(A_1^\top P A_2 P^\top), \tag{5.1.16}$$

which is recognized as a *Quadratic Assignment Problem* (QAP) [152]. The QAP is a generalization of the linear assignment problem (LAP) and is NP-hard. Indeed, exact graph matching (isomorphism) is a special case of the QAP, making it among the most challenging combinatorial optimization problems known [33].

5.1.8.2 Linear Assignment Problem (LAP) Relaxations

To improve tractability, a common strategy is to linearize or relax the quadratic cost function. This can be achieved by assuming that node correspondences are independent of one another, transforming the QAP into a *Linear Assignment Problem* (LAP):

$$\min_{P \in \mathcal{P}} \sum_{i \in V_1} \sum_{j \in V_2} C_{ij} P_{ij}, \tag{5.1.17}$$

where $C \in \mathbb{R}^{|V_1|\times|V_2|}$ is a cost matrix representing pairwise dissimilarities between node attributes or local graph features. The LAP can be efficiently solved in polynomial time using the Hungarian (Kuhn–Munkres) algorithm [160] with time complexity $O(|V|^3)$.

This linear relaxation is particularly effective when edge structures are weakly informative or when graphs are large and attribute-rich (e.g., in knowledge or scene graphs). It provides a tractable first approximation to the matching problem, often followed by local refinement steps.

5.1.8.3 Continuous Relaxations and Probabilistic Assignments

Inexact matching can also be cast as a *continuous relaxation* of the assignment matrix. Instead of restricting P to binary values, one allows $P \in [0,1]^{|V_1|\times|V_2|}$ with doubly-stochastic constraints:

$$P\mathbf{1} = \mathbf{1}, \quad P^\top\mathbf{1} = \mathbf{1}.$$

This defines the Birkhoff polytope — the convex hull of all permutation matrices. Optimization over this set can be efficiently approximated via methods such as the Sinkhorn–Knopp algorithm [312], which iteratively normalizes rows and columns of P to enforce the constraints.

The relaxed version of the quadratic objective becomes:

$$\max_{P\in\mathcal{B}} \operatorname{trace}(A_1^\top P A_2 P^\top), \tag{5.1.18}$$

where $\mathcal{B}$ denotes the Birkhoff polytope. Although the relaxation loses the strict combinatorial guarantees, it enables differentiable optimization and gradient-based methods, making it particularly suitable for neural or probabilistic models.

5.1.8.4 Connections to Graph Edit Distance and Neural Matching

The assignment formulation also underlies approximate solutions to the Graph Edit Distance (GED) problem [82], where node substitutions correspond to assignment costs C_{ij}, and insertions/deletions are handled by augmenting the cost matrix with dummy nodes. This transforms the GED computation into a form of the *linear sum assignment problem (LSAP)*, solvable by Hungarian [160] or auction algorithms.

More recently, neural graph matching approaches [361, 180] adopt differentiable assignment layers based on the Sinkhorn operator to compute soft permutation matrices during training. This allows end-to-end learning of matching functions that jointly optimize both the node embeddings and their alignment.

5.1.8.5 Computational Considerations

The QAP, being NP-hard, has exponential worst-case complexity of $O(|V|!)$, while its linear relaxation (LAP) is solvable in polynomial time $O(|V|^3)$ via the Hungarian algorithm. Continuous relaxations further reduce computational cost, with iterative Sinkhorn updates scaling as $O(|V|^2)$ per iteration. Thus, assignment-based formulations provide a principled and computationally controllable trade-off between exactness and scalability.

5.1.9 Graph Matching and Semantics

In semantic domains, such as knowledge graphs or scene graphs, matching involves not only structural similarity but also semantic coherence between node and edge labels. For example, two nodes labeled “dog” and “cat” may be considered close due to their shared ontological category (“domestic animal”). Hence, semantic graph matching incorporates external knowledge sources, such as ontologies or word embeddings, to compute similarity between labels. The objective function is then augmented to include both structural and semantic terms:

$$\min_X \|A_1 - XA_2X^\top\|_F^2 + \lambda\, D(S_1, S_2, X), \tag{5.1.19}$$

where $D(S_1, S_2, X)$ measures semantic dissimilarity between corresponding nodes or edges (based on, for example, embedding distances), and λ balances structural versus semantic alignment.

This semantic perspective is essential for explainability tasks, as it allows matching to be performed in the space of interpretable concepts rather than raw data. In counterfactual reasoning on structured domains, for instance, one can generate or evaluate alternative scenarios by aligning or substituting semantically similar subgraphs, a process fundamentally reliant on graph matching.

In all cases, we introduce some fundamental desiderata of graph-based counterfactual explanations, applicable in any domain, as long as conceptual units can be extracted [204, 208]. Specifically, we request the following:

- **Optimality**: proposed substitutions should be optimal, under a well defined notion of semantic-driven optimality. Approximations are acceptable and even desirable to increase execution speed.
- **Controllability**: if an edit can be performed it should be performed, as long as it satisfies its purpose (e.g. label flipping of a classifier). This means that there is no randomization over the perturbation process; some valid reasons to not perform an edit it because the goal is achieved (e.g. the targeted classifier changed its label) or as a trade-off with optimality (perform less edits, even if more can be realized, but keep the total number of edits as low as possible).
- **Efficiency**: an optimal solution should be reached within non-exhaustive time constrains. This may involve deterministic algoritmical optimizations, heuristics or even neural approximations in case of heavy compute. Ideally, the search space of alternatives should be explored as much as possible, should the execution time and resources allow; to this point, there may be a trade-off between optimality and efficiency.
- **Explainability**: the paths defining the interventions should be tractable. In the graph matching scenario, the matching M occurring under the *minimum weight bipartite matching* reveals the edits (paths from S to T), eliminating any fuzziness regarding the substitution process. Approximations on the matching definition results in decreased explainability, even though the theoretical guarantees of optimality (i.e. optimality of bipartite matching) are still valid. Explainability may be sacrificed to increase efficiency, indicating another potential inherent trade-off.

5.2 Accelerating semantic counterfactual computations

5.2.1 Graph Neural Networks

Traditional graph algorithms, such as graph matching or bipartite assignment, often become computationally expensive as the size of the graph grows. To address these challenges, recent advances in deep learning have introduced *Graph Neural Networks (GNNs)*, a family of models designed to operate directly on graph-structured data.

The key idea behind GNNs is to learn vector representations (embeddings) for nodes, edges, or entire graphs by recursively aggregating information from local neighborhoods. This allows each node to capture both its own attributes and the structural context of its neighbors in a continuous latent space. Formally, for a node v in a graph $G = (V, E)$, a GNN layer updates its embedding $\mathbf{h}_v$ as:

$$\mathbf{h}_v^{(k+1)} = \text{UPDATE}\Big(\mathbf{h}_v^{(k)}, \text{AGGREGATE}\big(\{\mathbf{h}_u^{(k)} : u \in \mathcal{N}(v)\}\big), \mathbf{f}_v\Big), \tag{5.2.1}$$

where $\mathbf{h}_v^{(k)}$ is the node's embedding at layer k, $\mathcal{N}(v)$ denotes its neighbors, and $\mathbf{f}_v$ represents the node's initial features. After multiple layers, the embedding encodes both the local structure around the node and its intrinsic properties.

One of the main advantages of GNNs in graph problems is their ability to map discrete combinatorial structures into a continuous space. This enables the use of efficient linear algebra operations for downstream tasks, including node/graph classification, link prediction, as well as graph matching and assignment.

5.2.2 GNN Architectures

GNNs constitute a diverse family of architectures that differ in how they aggregate and transform information across nodes and edges. Despite this diversity, most models follow a common paradigm of *message passing* [89], where each node updates its embedding by combining its own features with information aggregated from its neighbors. Below, we describe the most widely used GNN architectures, their mathematical underpinnings, and their distinctive characteristics.

5.2.2.1 Graph Convolutional Networks (GCN)

The *Graph Convolutional Network* (GCN) [149] generalizes the concept of convolution from Euclidean domains (e.g., images) to graph-structured data. Instead of using spatial filters, GCNs perform spectral filtering using the graph Laplacian. In its popular simplified form, the layer update rule is:

$$\mathbf{H}^{(k+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}\mathbf{H}^{(k)}\mathbf{W}^{(k)}\right), \tag{5.2.2}$$

where $\tilde{A} = A + I$ is the adjacency matrix with added self-loops, $\tilde{D}$ is its degree matrix, $\mathbf{W}^{(k)}$ is a trainable weight matrix, and $\sigma(\cdot)$ denotes a non-linear activation (e.g., ReLU). GCNs perform a weighted average of neighbor embeddings followed by a linear transformation, allowing nodes to gather information from their local graph structure. They are effective for semi-supervised classification and node representation learning.

5.2.2.2 GraphSAGE

GraphSAGE (Graph Sample and Aggregate) [102] extends GCNs to inductive learning scenarios, where unseen nodes or graphs may appear at test time. Instead of using a fixed propagation rule, GraphSAGE learns an aggregation function from data. For each node v, the update rule is:

$$\mathbf{h}_v^{(k+1)} = \sigma\left(\mathbf{W}^{(k)} \cdot \mathrm{CONCAT}\left(\mathbf{h}_v^{(k)}, \mathrm{AGGREGATE}\big(\{\mathbf{h}_u^{(k)} : u \in \mathcal{N}(v)\}\big)\right)\right), \tag{5.2.3}$$

where the aggregation function can take various forms such as mean, LSTM-based, or pooling-based. GraphSAGE efficiently handles large-scale graphs by sampling a subset of neighbors during training, making it particularly suitable for industrial applications and web-scale knowledge graphs.

5.2.2.3 Graph Attention Networks (GAT)

Graph Attention Networks (GATs) [351] introduce an attention mechanism to weigh the importance of different neighbors during aggregation. Instead of treating all neighboring nodes equally, GATs compute attention coefficients that quantify the relevance of each neighbor:

$$\alpha_{vu} = \frac{\exp\left(\mathrm{LeakyReLU}(\mathbf{a}^\top[\mathbf{W}\mathbf{h}_v \,\|\, \mathbf{W}\mathbf{h}_u])\right)}{\sum_{k\in\mathcal{N}(v)} \exp\left(\mathrm{LeakyReLU}(\mathbf{a}^\top[\mathbf{W}\mathbf{h}_v \,\|\, \mathbf{W}\mathbf{h}_k])\right)}, \tag{5.2.4}$$

$$\mathbf{h}_v^{(k+1)} = \sigma\left(\sum_{u\in\mathcal{N}(v)} \alpha_{vu}\mathbf{W}\mathbf{h}_u\right), \tag{5.2.5}$$

where $\mathbf{a}$ and $\mathbf{W}$ are trainable parameters. By dynamically focusing on the most relevant neighbors, GATs improve representation quality, especially in heterogeneous or noisy graphs. Multi-head attention variants further stabilize the learning process and enhance expressivity.

5.2.2.4 Graph Isomorphism Networks (GIN)

The *Graph Isomorphism Network* (GIN) [383] is motivated by the desire to achieve maximum discriminative power, equivalent to the Weisfeiler-Lehman (WL) graph isomorphism test. The update rule is:

$$\mathbf{h}_v^{(k+1)} = \mathrm{MLP}^{(k)}\Big((1+\epsilon)\cdot\mathbf{h}_v^{(k)} + \sum_{u\in\mathcal{N}(v)}\mathbf{h}_u^{(k)}\Big), \tag{5.2.6}$$

where ϵ is a learnable or fixed scalar, and MLP denotes a multi-layer perceptron. GINs use simple summation instead of averaging to ensure that structurally different graphs map to different embeddings. They are often employed in molecular property prediction and graph classification tasks due to their high expressivity.

5.2.2.5 Relational Graph Convolutional Networks (R-GCN)

The *Relational GCN* (R-GCN) [297] extends GCNs to handle multi-relational graphs, such as knowledge graphs, where edges are labeled with relation types. Each relation type $r\in\mathcal{R}$ has a dedicated weight matrix $\mathbf{W}_r$, and the update rule becomes:

$$\mathbf{h}_v^{(k+1)} = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{u\in\mathcal{N}_r(v)}\frac{1}{c_{v,r}}\mathbf{W}_r\mathbf{h}_u^{(k)} + \mathbf{W}_0\mathbf{h}_v^{(k)}\right), \tag{5.2.7}$$

where $c_{v,r}$ is a normalization constant. R-GCNs are particularly effective in reasoning over heterogeneous knowledge graphs such as WordNet [226], Freebase [27], or ConceptNet [317].

5.2.3 Connection with knowledge

In the context of concept and graph matching, GNNs offer a powerful acceleration mechanism: by embedding nodes into a continuous vector space, classical combinatorial problems, such as maximum-weight bipartite matching or full graph matching, can operate on similarity scores between embeddings rather than the original discrete graphs. This not only reduces computational complexity but also allows incorporation of semantic and structural information in a flexible way.

Knowledge-aligned concepts

In technical terms, we commence by aligning concepts extracted by some modality (e.g. objects in an image, as denoted by object labels) to a structured knowledge base, such as WordNet [226]: for example, the concept “dog” is mapped onto a WordNet *synset* “dog.n.01” using an automated function. After mapping all concepts of interest, semantic relationships can be quantified by computing distance or similarity scores between synsets within the Wordnet hierarchy; for example, the path-based similarity algorithm provided by calculates the shortest path length between two synsets in WordNet.

Conceptual distances

Edge weights between concepts can then be defined as a function of the similarity score:

$$w_{ij} = \mathrm{path\_similarity}(s_i, s_j), \tag{5.2.8}$$

where s_i, s_j are the synsets corresponding to concepts c_i, c_j. A score of 1 represents identity i.e. comparing a concept with itself will return 1.

Another way of defining semantic distances is to embed concepts in a low-dimensional space using a pre-trained models for semantic similarity [282]. Then, conceptual distances within this semantic space can be extracted using cosine similarity, with higher values (closer to 1) denoting more similar concepts.

Therefore, in both cases, when searching for counterfactuals, i.e. the most similar concept compared to a given one, their in-between path similarity has to be as high as possible. Scaling up to all concepts between two sets, the weights of their in-between relationships should be as large as possible, which directly translates to finding the *heaviest bipartite match* on the constructed bipartite graph. This process can be solved in polynomial time using algorithms referenced in Section 5.1.1, while approximate accelerations can be offered using GNNs: in such cases, computational complexity can be reduced to $O(mn\log n)$, or even further.

Knowledge-aligned graphs

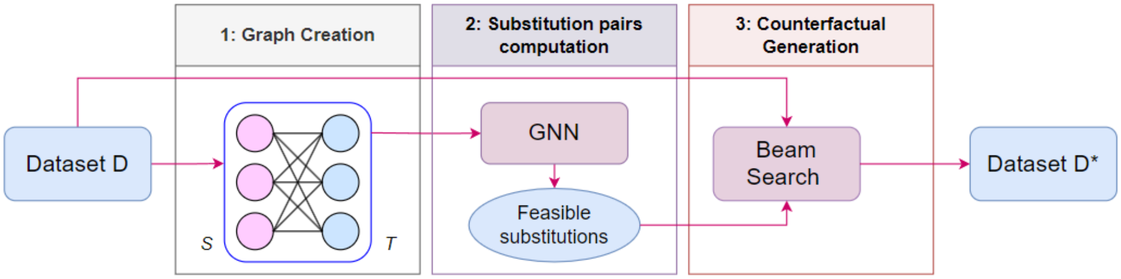


Figure 5.2.1: The pipeline of our method. In the first stage, we construct a bipartite graph using words as nodes, and in the second stage we utilize a GNN to get feasible substitutions that approximately solve the RLAP. In the final stage, we use beam search to change appropriate words of the original dataset, thus getting a new counterfactual dataset. [208]

When we employ scene graphs instead of concepts, we face the more complex problem of graph matching, where the semantic distance between two graphs should be minimized. Once again, given two graphs G_1 and G_2 to be matched, both their nodes and edges are mapped onto WordNet synsets. The distances between graph concepts and edges are again calculated using WordNet's path similarity, again with the goal of findings *maximum weight* pairings, in order for the graph distance to be *minimized*. The baseline algorithm in these cases is GED, due to its deterministic guarantees; however, its NP-hard complexity strictly permits usage in small and sparse graphs, even when heuristics are employed. In that case, leveraging GNNs becomes even more pertinent, allowing scalability and efficiency.

5.2.4 Concept-based counterfactuals using GNNs

We demonstrate our contributions in explaining NLP classifiers in conceptual level by placing concepts on a bipartite graph $G = (V, E)$, $S \cup T = V$. The concepts are extracted from a textual dataset D and separated by their part-of-speech (POS) tags (noun, adjective etc). Eligible counterfactual pairs suggesting *what needs to be changed in order for the text's label to flip* are returned after GNN inference, with the final pair (s_i, t_j) decided using a beam search module. A counterfactual dataset D^* is constructed based on target concepts $t' \in T' \subset T$ formulated from those pairs, as demonstrated in Figure 5.2.1.

Ultimately, the GNN module is tasked to approximate a solution to RLAP, with n source nodes from S optimally assigned to $m \geq n$ target nodes from T. Specifically, we harness a Graph Convolutional Network (GCN) [149] consisting of three modules: the encoder, the convolution module and the decoder. The **encoder** applies a Multi-Layer Perceptron (MLP) to each edge to transform its attributes into latent representations, producing embedding features. Initially, each edge attribute is set to its weight, i.e., $e_{ij} = w_{ij}$, where e_{ij} denotes the attribute of the edge connecting nodes i and j, and w_{ij} is the corresponding weight. Node attributes are initialized as zero-valued vectors. The transformed graph is then passed into a convolutional module to update the graph state.

The **convolution module** consists of a **node** convolution layer and an **edge** convolution layer. For the ith node in the graph, the node convolution layer aggregates information from its adjacent edges and 1st-order neighboring nodes using adaptive aggregation weights, and updates the node's attribute. For each edge, the edge convolution layer aggregates the attribute vectors of the two connected nodes and updates the edge attribute vector accordingly. Although the receptive field of the convolution module is limited to 1st-order neighborhoods, the bipartite structure of the graph—composed of two disjoint node sets where each node in one set is connected to all nodes in the other—ensures that messages from any node can reach all other nodes after two convolution iterations. Therefore, the receptive field effectively covers the entire graph after the 2nd iteration.

The **decoder**, coupled with the encoder, reads the edge attributes from the output graph and predicts edge labels via an update function. This update function is also implemented as an MLP and is applied to each edge, producing edge labels through a sigmoid activation.

Experimental setup

Datasets We evaluate our framework against existing counterfactual editors using two English-language datasets: the IMDB dataset for binary sentiment classification [210] and a 6-class subset of the 20 Newsgroups dataset for topic classification [164].

Editors We benchmark our approach against two state-of-the-art counterfactual generators: MiCE [290], which produces task-specific minimal edits to induce label changes, and Polyjuice [372], a general-purpose editor not constrained to any specific task. Within our own framework, we include both the deterministic RLAP-based substitution baseline and our proposed GNN-based RLAP optimization. To evaluate the flexibility of our method, we further test two substitution settings: POS-restricted and POS-unrestricted.

Predictors For consistency with prior work, we adopt the same predictor models used in MiCE [290] for each dataset. Both are fine-tuned RoBERTa$_{\text{LARGE}}$ models [190], achieving test accuracies of 95.9% on IMDB and 85.3% on Newsgroups.

Metrics Following the evaluation setup of MiCE, we assess performance using four core metrics:

1. **Flip-rate:** The percentage of edited instances that result in a different model prediction, measuring the effectiveness of label flipping.
2. **Minimality:** The degree of textual change, computed via the normalized word-level Levenshtein distance between the original and edited inputs. This is given by:

$$\text{minimality} = \frac{\text{Levenshtein}(D, \hat{D})}{|D|} \tag{5.2.9}$$

   where D is the original input, $\hat{D}$ the edited version, and $|D|$ the number of words in D.
3. **Closeness:** The semantic similarity between original and edited text, assessed using BERTScore [414].
4. **Fluency:** A measure of distributional similarity between original and edited inputs. We compute this using the loss values from a pretrained T5-BASE model [274], calculating the loss ratio as:

$$\text{loss\_ratio} = \frac{\text{Loss}(\hat{D})}{\text{Loss}(D)} \tag{5.2.10}$$

   and define the fluency metric as:

$$\text{fluency} = |1 - \text{loss\_ratio}| \tag{5.2.11}$$

   A lower fluency score (closer to 0) indicates that the edited text remains distributionally close to the original.

All experiments were conducted on a system with a *16 GB GPU*, *Intel i7 CPU*, and *16 GB RAM.*

Quantitative results are presented in Table 5.1.

5.2.5 Graph-based counterfactuals using GNNs

Transitioning from concepts to graphs, we first have to research problems in which interconnections are meaningful by nature. To ensure this, we initially deviate from extracting concepts and relationships (as in the case of NLP counterfactuals in Section 5.2.4) and stick to existing annotations. To this end, scene graphs serve as the ideal testbed to experiment with semantic graph matching, since graph edges correspond to highly-valued relationships between objects, or actions definitive for the meaning of a scene. This can be easily demonstrated in Figure 5.2.2.

5.2.5.1 Supervised GNNs

In the supervised setup [69], a ground-truth set of graph pairs is needed to guide GNN training, with the goal of learning to place semantically similar scene graphs closer in the embedding space. Since the supervision signal needs to be of high-quality, GED calculation is unavoidable, at least for a subset of $N' \subset N$ graphs, where N denotes the sample size of the dataset. A siamese component is tasked to learn graph similarities

	IMDB					
	Editor	**Fluency↓**	**Closeness↑**	**Flip Rate↑**	**Minimality↓**	**Runtime↓**
WordNet	Deterministic w. fluency	0.14	0.969	0.892	0.08	4:09:41
	GNN w. fluency	0.07	0.986	0.861	0.12	3:17:51
	GNN w. fluency & dynamic thresh	0.057	0.986	0.851	0.146	4:18:34
	GNN w. fluency & POS_filter	0.08	0.992	0.862	0.123	**0:32:05**
	GNN w. fluency & edge filter	0.105	0.993	0.845	0.149	3:00:38
	GNN w. fluency_contrastive	0.112	**0.999**	0.914	0.014	2:12:06
	GNN w. contrastive	0.048	0.996	0.927	textbf0.01	2:00:15
Embeddings	GNN w. AnglE & contrastive	0.063	0.995	0.944	0.011	0:45:38
	GNN w. GIST & contrastive	0.037	0.995	0.882	0.016	0:58:14
	GNN w. JinaAI & contrastive	0.047	0.995	0.928	0.017	1:00:56
	GNN w. MUG & contrastive	**0.036**	0.996	0.889	0.013	0:52:19
	Polyjuice	0.394	0.787	0.782	0.705	5:01:58
	MiCE	0.201	0.949	**1.000**	0.173	48:37:56
	Newsgroups					
	Editor	**Fluency↓**	**Closeness↑**	**Flip Rate↑**	**Minimality↓**	**Runtime↓**
WordNet	Deterministic w. fluency	0.182	0.951	0.870	0.135	4:20:52
	GNN w. fluency	0.074	0.985	0.826	0.151	3:48:37
	GNN w. fluency & dynamic thresh	0.043	0.984	0.823	0.148	4:47:14
	GNN w. fluency & POS filter	0.044	0.989	0.841	0.143	1:19:57
	GNN w. fluency & edge filter	0.12	0.989	0.834	0.151	3:05:08
	GNN w. fluency_contrastive	0.088	0.979	0.875	0.033	2:45:31
	GNN w. contrastive	0.033	0.989	0.920	0.033	2:02:34
Embeddings	GNN w. AnglE & contrastive	0.005	0.995	0.904	0.027	1:09:13
	GNN w. GIST & contrastive	**0.001**	0.995	0.898	0.02	1:02:55
	GNN w. JinaAI & contrastive	0.013	0.993	0.882	0.025	0:57:31
	GNN w. MUG & contrastive	0.005	**0.996**	0.900	**0.016**	**0:53:04**
	Polyjuice	1.153	0.667	0.8	0.997	6:00:10
	MiCE	0.152	0.922	**0.992**	0.261	47:23:35

Table 5.1: Experimental results of counterfactual generation. We evaluate different versions of our framework using the metrics described above, and we compare it with MiCE and Polyjuice. For each metric (column) the best value is highlighted in **bold**. Reported runtimes refer to inference.

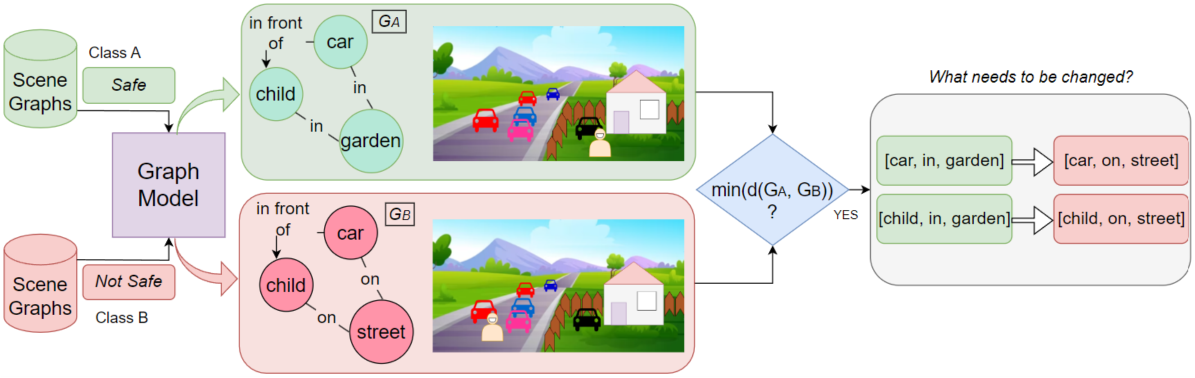


Figure 5.2.2: This image illustrates a real-world scenario where scene graphs meet counterfactual explanations: what delineates the class "safe" from "not safe" can be attributed to certain graph edits [68].

employing a GCN backbone and GloVe [260] initialization of features. Further experiments substitute GCN with either GAT or GIN. The graph embedding calculation is provided by the following equation:

$$h_G = \frac{1}{n}\sum_{i=1}^{n}(u_i^{K-1} + \sum_{j\in\mathcal{N}(i)} u_j^{K-1}) \tag{5.2.12}$$

where u_i is the representation of node i, $\mathcal{N}(i)$ is the neighborhood of i, n is the number of nodes for G and K is the number of GCN layers. The similarity of $(h_{G_{C_x}}, h_{G'_{C_y}})$ (C_x, C_y denote any class, as long as $x \neq y$) embeddings is preserved using the dimensionality reduction technique of Multi-Dimensional Scaling [368]. The model is then trained in a transductive fashion to minimize the objective function $\mathcal{L}$:

$$\mathcal{L} = \mathbb{E}(\left\|(h_{G_{C_x}} - h_{G'_{C_y}}\right\|_2^2 - GED(G_{C_x}, G'_{C_y})) \tag{5.2.13}$$

During ranking, given a query graph G_A from class A the closest graph in terms of cosine similarity G_B from a class $B \neq A$ serves as its counterfactual.

$$G_B = G_B^i,\ \arg\max_i(\frac{h_{G_B^i} \cdot h_{G_A}}{\left\|h_{G_B^i}\right\| \left\|h_{G_A}\right\|})\ if\ B \neq A \tag{5.2.14}$$

This process requires searching within at most $N-1$ graph pairs, demonstrating the need for substituting further GED calculations with efficient approximate techniques.

Experimental setup

Datasets harnessed in this work in order to cover diverse image classification cases are the following:

- **Caltech-UCSD Birds (CUB)**: contains more than 11K images of birds belonging to 200 fine-grained subcategories, with annotations that indicate bird parts and attributes such as colors and textures. [106] Graphs are automatically constructed in a star-like structure, with a central "bird" node, connected with "parts" via "has" edges.
- **Visual Genome (VG)**: comprises more than 108K images of varying scenes accompanied by human-annotated scene graphs that include objects, relationships and attributes. [155] We regard two subsets: a randomnly sampled one containing 500 scene graphs called VG-RANDOM, as well as a subset promoting dense graphs with fewer isolated nodes called VG-DENSE.

Comparison includes a white-box and a black-box counterfactual method, referenced as CVE [346] and SC [62] respectively. None of these employs scene graph, while CVE is not explicitly based on semantics at all.

Evaluation of retrieved counterfactual graphs is primarily based on ranking metrics, including:

- **Precision@k**: returns the number of top-k items in terms of GED found within the top-k items as retrieved by the GNN.
- **Binary Precision@k**: This variant serves the counterfactual retrieval case, where only the top-1 retrieved item is considered to be the desired counterfactual; therefore, *binary Precision@k* denotes the number of top-1 items in terms of GED found in the top-k items as retrieved by the GNN.
- **Binary NDCG@k**: it also considers the position of the counterfactual by comparing the ideal rank, as suggested by GED, placing the counterfactual item in the top-1 position, with the top-k GNN-retrieved rank. The relevance of each item is instructed by their position in the ground truth rank, with items placed in a higher position being more relevant, and each position being equally weighted.

Apart from ranking metrics, we further report metrics related to the semantic edits performed, as suggested from post-hoc GED calculation between the top-1 ground truth item and the top-1 GNN- retrieved item:

- **Number of node *insertions*, *deletions*, *substitutions***.
- **Cost of edits**: sum of weights to reach from source to target for each edit performed.

Qualitative results are showcased in Figure 5.2.3: we report counterfactuals between the source class A (Rusty Blackbird) and the target class B (Brewer Blackbird). These class pairs were selected based on the idea of the *most confused classes* as analyzed in [346]. Overall, our approach produces the **fewest and cheapest concept edits**, which are reported per image, together with the resulting GED.

We can spot some conceptual fallacies of the other frameworks, such as the appearance of two Brewer Blackbirds in the third row (left), a counterfactual image proposed by [62]. Another sub-optimal counterfactual

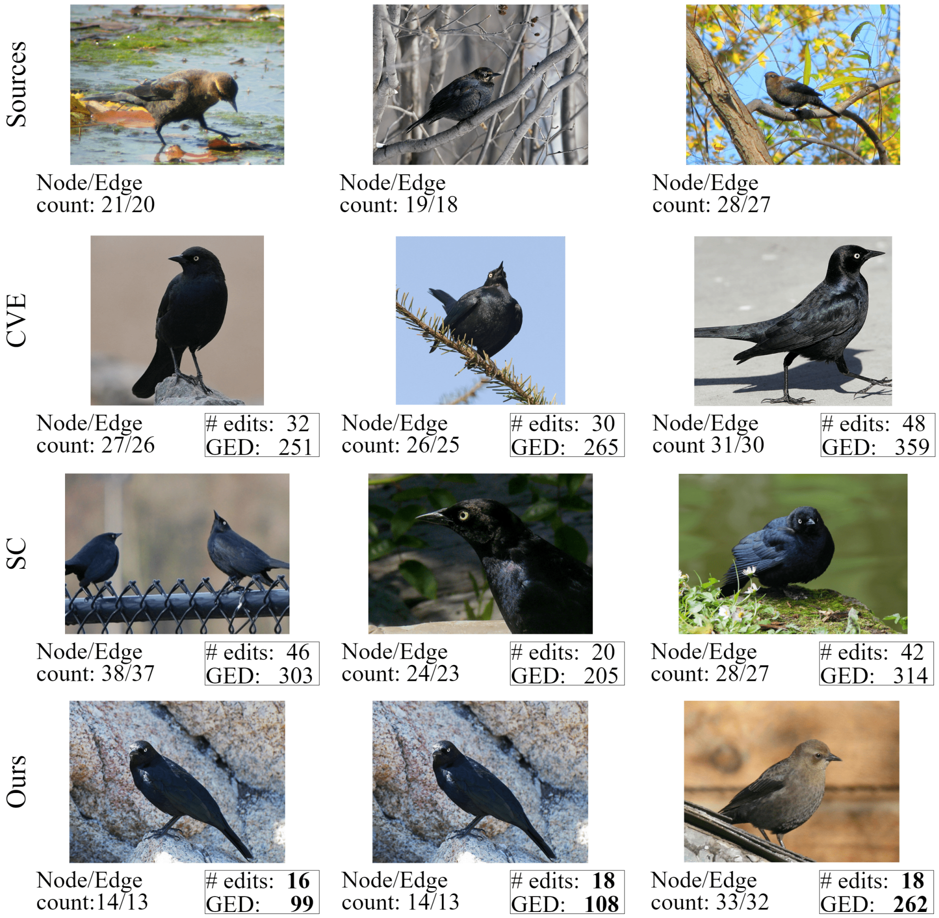


Figure 5.2.3: Results for Rusty → Brewer Blackbird. **Bold** denotes best results (lowest number of edits and GED scores). [69]

retrieved by [62] is presented in the middle of the third row, with only one part of the bird appearing in the frame. The retrieval of such instances is directly associated with elevated number of edits, since a semantic graph containing all bird parts and attributes needs to be inserted in the case of two birds, while several parts and attributes need to be deleted in the case where only a part of the bird appears. The graph structures leveraged in our approach impose a stricter interpretation of related concepts via their interconnections, penalizing erroneous edits in a more refined fashion compared to edits between sets of sets. The method of [346] takes into account some visual cues (e.g. zoom), thus avoiding the aforementioned fallacies; nevertheless, despite the seemingly successful counterfactuals retrieved by [346], the number of edits needed as well as the GED values between the source and the counterfactuals are elevated in comparison to ours. This is due to the lack of semantic guarantees associated with their framework, even though they characterize as "semantic" edits the changes performed within specific image areas, which however still operate in pixel-level.

Human Evaluation Some critique arising from Figure 5.2.3 could be that the counterfactuals returned by [346] are still visually appealing according to the counterfactual formulation, where a minimally altered bird has to be returned, especially since a human is only provided with the image but no other specific characteristics associated with each sample. We acknowledge that both techniques (ours and [346]) perform an optimization in a different level, ultimately shifting the weight towards defining which ground truth is more interpretable, thus which ground truth should be used for optimization.

Apart from consulting prior work defending the interconnected nature of semantics and explanation we prove the power and expressiveness of semantic-based counterfactuals based on human perception. To this end, we present results of the human survey in the following Tables. Specifically, Table 5.2 demonstrates results of the comparative human survey, indicating a clear preference over both SC and CVE according to human evaluators. Therefore, we can safely deduct that our counterfactual images are more interpretable, justifying the graph-based advancements accompanying our method. Despite the conceptual proximity with SC, annotators selected the GNN-powered version almost twice, confirming the meaningful addition of linking concepts within a graph.

Ours	Win%	Lose%	Tie%
SC	**48.86**	19.32	31.82
CVE	**48.42**	26.27	25.31

Table 5.2: Human preference; Win%=% times our method was preferred, Lose% for vice-versa, Tie% when equally preferred. **Bold** denotes higher human preference per method.

Quantitative results are presented in Table 5.3 for both CUB and VG. For computational efficiency, GNNs on VG are trained using only $N/2 = 70k$ graph pairs. Based on the results, it becomes evident that GNN-based variants (and specifically GCN ones) consistently perform better in retrieving appropriate counterfactual instances, proving that both semantics and structure are important. On the other hand, approaches that ignore structure (CVE, SC) or semantics (graph kernels) are unable to incorporate all aspects needed for successfully providing counterfactual scene graphs. Overall, the incorporation of semantic graphs for retrieving counterfactual images approaches ground truth GED-driven ranks, accelerating a brute force scenario in which no GNN is leveraged, and all computation is supported by GED. The computational overhead of these methods is demonstrated in Table 5.4, verifying the merits of employing GNNs for efficiency.

5.2.5.2 Unsupervised GNNs

Previously, we demonstrated the efficiency of supervised GNNs on approximating the NP-hard GED problem. Still, there are points that could potentially be improved, such as the computational time needed for GNN training (Table 5.4). However, reducing the number of training pairs below $N/2$ is not a viable choice, since it leads to decreased performance, as proven in Figure 5.2.4, even though it would reduce the time needed to learn similarities between given scene graph pairs.

Due to these limitations, a totally diverging line of work needs to be considered in order to mitigate the training time needed without relying on examining graph pairs to produce proper graph embedding representations. To this end, an appropriate approach is to leverage Graph Autoencoders (GAEs) [150] to obtain graph embeddings. GAEs trespass the need for labeled data –in our case being a ground truth GED value between each graph pair of the training data. Since this supervision signal is now eliminated, GAEs may struggle in learning all the similarity characteristics between input graphs, thus often ranking lower in comparison to supervised GNNs. [68]

GAE preliminaries

GAEs employ an encoder-decoder structure, where the encoder is trained to map input graphs on a common embedding space. The networks comprising the encoder and the decoder can be any aforementioned GNN module, as long as the encoder's and the decoder's architecture remains the same from model to model. Distance between these embeddings, for example using cosine similarity, reveals the actual graph similarity: embeddings lying closer in the embedding space correspond to more similar graphs. This representation is implicitly learned during the training phase of a GAE, therefore avoiding the need for ground-truth GED values at all in this phase[2]. Overall for N graphs, instead of $O(N^2)$ operations requested in the supervised case, unsupervised GAEs only consume $O(N)$ computational time to learn similarities between input graphs.

Experimentation with unsupervised GNNs includes the vanilla GAE, Variational GAE (VGAE) [150], and Adversarially Regularized Variational Graph Autoencoder (ARVGA) [247], together with a variation of Graph Feature Autoencoder (GFA) [103]. GAE and VGAE are built upon similar properties and only differ in terms of the selected loss function. ARVGA extends the VGAE idea drawing inspiration from Generative Adversarial Networks (GANs) [93] to further regularize the produced embeddings. GFA utilizes the Feature

---

[2] GED is still needed in order to form the ground truth rank that compares the GAE rank with the more deterministic GED similarity perception during the evaluation phase.

	P@k↑		**P@k** (binary)↑		**NCDG@k** (binary)↑		**Node** ↓	**Edge** ↓	**Total** ↓
	k=1	k=4	k=1	k=4	k=1	k=4			
CUB									
CVE [346]	0.02	0.10	0.02	0.11	0.11	0.26	8.43	4.70	13.13
SC [62]*[1]	-	-	-	-	-	-	8.07	**3.66**	11.73
Ours	**0.19**	**0.34**	**0.19**	**0.49**	**0.23**	**0.36**	**6.16**	4.34	**10.5**
VG-DENSE									
Kernels									
WL kernel	0.076	0.164	0.076	0.186	0.138	0.287	5.24	11.67	16.908
SP kernel	0.122	0.154	0.122	0.232	0.194	0.334	4.96	12.08	17.04
RW kernel	0.001	0.0	0.0	0.0	0.0	0.0	9.16	20.01	29.17
NH kernel	0.05	0.05	0.124	0.088	0.131	0.282	5.35	11.26	16.61
GS kernel	0.008	0.022	0.008	0.018	0.103	0.259	6.44	16.75	23.19
Semantic Counterfactuals (SC)									
SC [62]*	-	-	-	-	-	-	**4.91**	7.29	12.2
GNNs									
GIN-70K	0.16	0.27	0.16	0.38	0.20	0.34	5.11	10.77	15.87
GAT-70K	0.18	0.32	0.18	0.44	0.22	0.35	5.28	10.89	16.17
GCN-70K	**0.25**	**0.37**	**0.25**	**0.49**	**0.28**	**0.41**	4.95	**7.15**	**12.11**
VG-RANDOM									
Kernels									
WL kernel	0.096	0.13	0.096	0.166	0.16	0.306	13.44	11.95	25.39
SP kernel	0.064	0.079	0.064	0.11	0.144	0.292	14.14	12.39	26.53
RW kernel	0.0	0.007	0.0	0.0	0.0	0.0	17.81	22.36	40.17
NH kernel	0.092	0.135	0.092	0.168	0.157	0.303	12.57	11.71	24.28
GS kernel	0.0	0.002	0.0	0.05	0.0	0.0	17.22	15.34	32.56
Semantic Counterfactuals (SC)									
SC [62]*	-	-	-	-	-	-	**12.15**	**7.52**	**19.67**
GNNs									
GIN-70K	0.03	0.07	0.03	0.07	0.22	0.38	12.79	11.39	24.19
GAT-70K	0.18	0.29	0.18	0.38	0.11	0.27	12.81	11.98	24.79
GCN-70K	**0.21**	**0.30**	**0.21**	**0.42**	**0.25**	**0.38**	12.18	7.54	19.72

Table 5.3: Comparison of counterfactual retrieval results with ground truth GED rankings. **Bold** denotes best results per dataset, while green cells indicate best results within model family per dataset.

	GED ↓	GCN-N/2 (train)↓	GCN-N/2 (retrieval)↓	GCN-N/2 (inference)↓
CUB	46220	32691	0.03	0.06
VG-DENSE	13982	12059	0.03	0.06
VG-RANDOM	18787	16271	0.03	0.10

Table 5.4: Time (sec) for counterfactual calculation.

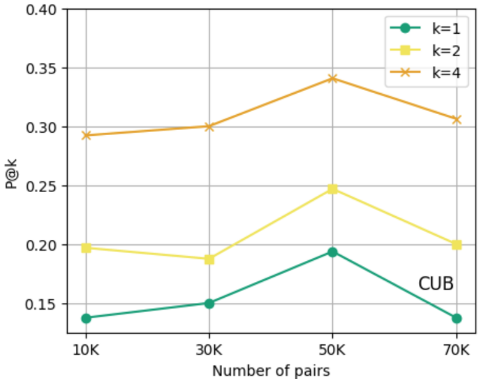
Decoder, alongside the original VGAE Inner-Product Decoder. GNNs comprising the encoder and the decoder remain single-layered.

Experimental setup

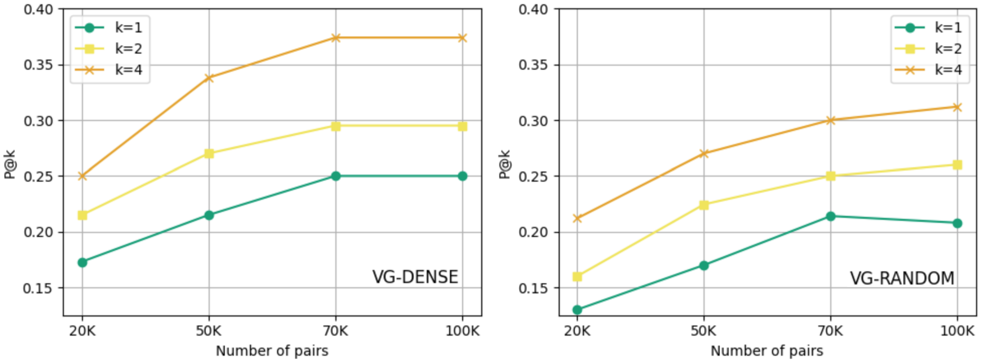
Throughout experimentation, the setup proposed for the supervised case remains as it is, maintaining the same graph splits for VG-DENSE and VG-RANDOM, while also considering the same evaluation setting, where GED forms the ground truth rank for scene graphs.

Quantitative results for unsupervised GNNs are provided in Table 5.5, together with the best performing supervised GNN for comparison. It becomes evident that unsupervised GNNs lack in performance compared to the best performing supervised variants, and overall there is no clear winner in the unsupervised regime. Overall, existing unsupervised architectures are not capable enough of encoding the necessary discriminative characteristics and drive appropriate counterfactual retrieval.

Advanced retrieval can be on par with efficiency



(a) CUB



(b) VG variants

Figure 5.2.4: P@k of GCN variant for different training pairs p on CUB and VG. [69]

As demonstrated previously, unsupervised GNNs based on GAE architectures pose some merits in terms of counterfactual retrieval even though they often stay behind best supervised counterparts. To this end, possible optimizations over off-the-self GAE choices may lead to upgraded performance, thus proposing a universal efficient graph-based solution for counterfactual retrieval.

This solution is embodied in our proposed SCENIR framework [39], primarily designed as a general-purpose image retrieval method. Designed upon the typical GAE architecture, SCENIR introduces certain architectural advancements for advanced scene graph retrieval results. Specifically, SCENIR acknowledges the source of limited scene graph retrieval capability, attributing it to the inherent complexity of properly representing node-edge relationships tied to design choices such as the usage of inner-product decoders, resulting in limited expressiveness, as well as the limited discriminative power of such architectures as a result of oversmoothing.

We first need to consider a scene graph $G = (V, E)$, formulated as a *feature matrix* $\mathbf{X} \in \mathbb{R}^{n \times d}$ (information about objects in image I), and an *adjacency matrix* $\mathbf{A} \in \mathbb{R}^{n \times n}$ (information about relations in image I), where $n = |V|$ is the number of nodes, and d the feature vector dimensionality per node. These two matrices are received by two GNN modules (GNN_μ and GNN_σ) comprising the **encoder**, whose goal is to encode the input graph into a learned latent space, outputting a latent node embeddings matrix $\mathbf{Z} \in \mathbb{R}^{n \times d_l}$. Instead of leveraging a shared GNN layer followed by separate GNN_μ and GNN_σ for variational training [150], SCENIR maintains them as two independent 3-layered streams, imposing separately learning structural features (mean embeddings) and uncertainty (variance embeddings). In place of the GNN components, any related architecture, such as GCN, GIN, GAT, can be selected.

Regarding the **decoder**, instead of employing a simple inner-product function, two parallel streams involving two 2-layered MLPs are inserted, each of which constituting an Edge Decoder and a Feature Decoder. By introducing this more complex decoding scheme, SCENIR aspires to capture more intricate and challenging relationships present in scene graphs. However, GNNs are not preferred over MLPs as decoding functions, as evidently, multiple graph convolutional layers stacked one after another lead to undesirable oversmoothing

Models	P@k ↑		P@k (binary)↑		NDCG@k (binary)↑		Node ↓	Edge ↓	Total ↓
	k=1	k=4	k=1	k=4	k=1	k=4			
VG-DENSE									
Supervised GNNs									
GCN-70K	**0.25**	**0.37**	**0.25**	**0.49**	**0.28**	**0.41**	4.95	**7.15**	**12.11**
Unsupervised GNNs									
GAE GAT	0.05	0.08	0.05	0.11	0.13	0.28	5.50	11.85	17.35
GAE GIN	0.05	0.08	0.05	0.13	0.13	0.28	5.4	12.002	17.402
GAE GCN	0.05	0.07	0.05	0.10	0.14	0.29	5.178	11.27	16.448
VGAE GAT	0.08	0.14	0.08	0.20	0.15	0.30	5.03	11.4	16.43
VGAE GIN	0.09	0.14	0.09	0.20	0.16	0.30	4.918	10.97	15.90
VGAE GCN	0.07	0.12	0.07	0.17	0.14	0.29	5.114	11.702	16.816
GFA GAT	0.09	0.12	0.09	0.18	0.16	0.30	5.07	11.56	16.63
GFA GIN	0.10	0.14	0.10	0.21	0.16	0.31	**4.86**	10.79	15.64
GFA GCN	0.068	0.12	0.14	0.29	0.158	0.16	5.03	11.432	16.46
ARVGA GAT	0.08	0.13	0.08	0.19	0.15	0.3	5.06	11.48	16.54
ARVGA GIN	0.11	0.14	0.11	0.22	0.17	0.317	4.89	10.91	15.81
ARVGA GCN	0.08	0.12	0.08	0.18	0.15	0.30	5.048	11.316	16.36
VG-RANDOM									
Supervised GNNs									
GCN-70K	**0.21**	**0.30**	**0.21**	**0.42**	**0.25**	**0.38**	12.18	**7.54**	**19.72**
Unsupervised GNNs									
GAE GAT	0.07	0.092	0.07	0.13	0.146	0.29	**11.95**	11.43	23.38
GAE GIN	0.07	0.08	0.07	0.11	0.15	0.30	12.55	11.74	24.29
GAE GCN	0.066	0.08	0.07	0.11	0.14	0.29	12.13	11.38	23.52
VGAE GAT	0.09	0.10	0.09	0.16	0.16	0.31	12.88	12.38	25.26
VGAE GIN	0.09	0.11	0.09	0.15	0.16	0.30	12.69	12.04	24.73
VGAE GCN	0.09	0.10	0.09	0.16	0.16	0.30	12.91	12.39	25.30
GFA GAT	0.076	0.11	0.08	0.14	0.14	0.29	12.99	12.41	25.40
GFA GIN	0.08	0.11	0.08	0.14	0.15	0.30	12.92	12.22	25.13
GFA GCN	0.08	0.11	0.08	0.14	0.15	0.30	12.70	12.34	25.03
ARVGA GAT	0.09	0.11	0.09	0.156	0.16	0.31	12.51	12.19	24.70
ARVGA GIN	0.08	0.12	0.08	0.15	0.15	0.30	12.23	11.79	24.03
ARVGA GCN	0.09	0.11	0.09	0.15	0.16	0.31	12.46	12.09	24.55

Table 5.5: Ranking results on the two VG variants for various unsupervised GNNs in comparison to the best supervised one (Table 5.3). **Bold** numbers indicate best ranking metrics overall, while **green cells** indicate best results within model family per dataset.

[174]. The output of the decoder comprises $\mathbf{A}_p \in \mathbb{R}^{n \times n}$, with $\mathbf{A}_{p,ij} \in [0,1]$, is the predicted adjacency matrix and $\mathbf{Z}_f \in \mathbb{R}^{n \times d}$ is the predicted feature matrix.

To further boost discriminative capacity of the main autoencoding module, a **discriminator** is incorporated [247], enhancing differentiation between real samples, as drawn from a prior function, and fake ones as generated by the encoder. The discriminator is designed as a 2-layered MLP with binary classification output.

The model is trained end-to-end with a loss function that combines multiple objectives from all the aforementioned modules:

$$\mathcal{L} = \lambda_1(\mathcal{L}_{feat\_recon} + \mathcal{L}_{edge\_recon}) + \lambda_2\mathcal{L}_{adv} + \lambda_3\mathcal{L}_{KL} \tag{5.2.15}$$

where $\mathcal{L}_{feat\_recon}$ is the Mean Squared Error (MSE) loss between the original $\mathbf{X}$ and predicted $\mathbf{Z}_f$ feature matrices, while $\mathcal{L}_{edge\_recon}$ is the reconstruction loss between the original $\mathbf{A}$ and the predicted $\mathbf{A}_p$ adjacency matrices defined as $\mathbb{E}_{q(\mathbf{X}|\mathbf{Z}_e,\mathbf{A})}\left[log\ p\left(\mathbf{A}|\mathbf{Z}_e\right)\right]$. $\mathcal{L}_{adv}$ is the adversarial loss implemented as Binary Cross Entropy Loss for the discriminator prediction, and $\mathcal{L}_{KL}$ is the Kullback-Leibler Divergence of the latent embeddings to the prior Gaussian distribution $\mathcal{N}(0, I)$. At inference time, global scene graph representations are generated by applying sum-pooling to the latent node embeddings $\mathbf{Z}_\mu$ from the trained GNN_μ encoder.

Experimental setup

Dataset Experiments are conducted on an improved version of VG called PSG [386], containing $\sim 49K$ annotated image, caption and scene graph samples; we consider $11K$ scene graphs for training, and $1K$ for testing. We utilize PyTorch Geometric [76] for SCENIR training, only requesting a single P100 GPU.

Comparisons To put SCENIR into perspective regarding its counterfactual retrieval capabilities, we compare with unsupervised GNN techniques, including VGAE [150] and ARVGA [247], as well as will the supervised GNN-based SC framework [69]. Furthermore, we compare with IRSGS [395], which utilizes image captions instead of GED as the supervision signal to perform ground truth graph matching.

Evaluation As an evaluation measure, GED is again proposed as the golden standard, thanks to its semantic and structural properties over image similarity metrics. In the interest of evaluating counterfactual retrieval, where only the top-1 instance is relevant, we focus on the previously proposed *binary* variants of ranking metrics [69].

Quantitative results are illustrated in Table 5.6. Apparently, the introduction of a more refined unsupervised GNN architecture introduced in SCENIR is capable of advancing counterfactual retrieval, ultimately achieving higher retrieval metrics even in comparison to its supervised counterpart [69].

Model	NDCG(Binary)@1↑	MAP(Binary)@3↑	MRR(Binary)↑
VGAE	8.79	11.09	14.23
ARVGA	8.2	10.63	13.79
IRSGS	8.9	11.63	14.96
SC	7.0	8.7	10.67
SCENIR	**9.7**	**11.83**	**14.99**

Table 5.6: Counterfactual scene graph retrieval on PSG.

Efficiency is exhibited in Table 5.7. The counterfactual setup of SCENIR is capable of retrieving counterfactual instances in just 8 minutes in comparison to its supervised competitor [69] which requires 3 hours for training on the same dataset size.

Model	Preprocessing	Training	Inference	Total Time
SC	$\mathcal{O}(n^2)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n)$	~3 hr.
IRSGS	$\mathcal{O}(n^2)$	$\mathcal{O}(n)$	$\mathcal{O}(n)$	~50 min.
SCENIR	$\mathcal{O}(n)$	$\mathcal{O}(n)$	$\mathcal{O}(n)$	**~8 min.**

Table 5.7: Complexities of scene graph retrieval frameworks, with respect to the dataset size (PSG dataset, 11k/1k train/test graphs).

All in all, the introduction of appropriate unsupervised GNN architectures for counterfactual image retrieval is capable of achieving competitive performance in ranking metrics, while significantly accelerating the time needed for retrieving the counterfactual images.

5.2.6 Non-minimal interventions

Traditionally, counterfactual reasoning is framed around the idea of minimal intervention by changing as little as possible in the input while altering the target outcome. This principle stems from causal inference [259]: when we ask "What if X had been different?", we want to isolate the effect of X and nothing else. Minimal interventions ensure that only the causal variable of interest changes, keeping all else equal.

This principle, while analytically elegant, assumes that the counterfactual world should remain as close as possible to the factual one, differing only in the variable of interest. However, when dealing with high-level semantic representations, particularly in vision-language models, minimal interventions can be overly restrictive and may fail to expose how the model truly encodes and utilizes conceptual distinctions.

In contrast, semantically non-minimal interventions, i.e. those that involve substantial or categorical shifts in meaning, such as substituting antonyms (dark → bright, open → closed, man → woman) or opposing relational attributes (above → below), offer a complementary perspective on model understanding and robustness. Rather than probing the local sensitivity of the model to small, synonymous perturbations, these interventions test the model's ability to differentiate between semantically opposite or conceptually distinct scenarios.

The intuition behind employing such non-minimal interventions is twofold.

First, they serve as a diagnostic for *conceptual relevance*. If a model remains invariant to a semantically non-minimal change (e.g., producing the same caption or retrieving the same image when a concept "dark" is replaced by a contrasting concept, such as "bright"), it indicates that the altered concept **does not play a meaningful role** in the model's internal representation. In other words, invariance to non-minimal interventions reveals *semantic blindness*: the model's inability to distinguish between conceptually opposing states. Conversely, a change in model output in response to such an intervention implies that the model indeed encodes and conditions on that concept, demonstrating semantic sensitivity and appropriate contextual grounding.

Second, semantically non-minimal interventions reveal robustness-pathology trade-offs in internal model reasoning. Conventional robustness evaluation often rewards models that produce stable outputs under small perturbations (e.g., synonym replacements). However, robustness to large semantic changes, such as flipping antonyms or negations can be undesirable, as it implies insensitivity to critical meaning shifts. For instance, in a text-to-image retrieval system, the captions "a dark room" and "a bright room" should yield entirely different retrieval results. If a model's retrieval remains invariant under this semantically non-minimal intervention, it reflects a fundamental **failure of discriminative understanding**: the model conflates distinct conceptual states, undermining its reliability in downstream or real-world tasks.

Thus, semantically non-minimal interventions act as a **stress test for semantic grounding**. They examine whether models capture not just surface-level linguistic regularities, but also the directionality and oppositional structure of meaning embedded in language and visual concepts. This is particularly crucial for linguistic and multimodal reasoning, where subtle semantic contrasts (e.g., full vs. empty glass, happy vs. sad person) correspond to significant visual and contextual differences.

Moreover, this paradigm challenges the narrow notion of "faithful robustness" in linguistic and multimodal systems. True robustness should not imply invariance to all perturbations, but rather context-sensitive sensitivity—stability to irrelevant changes (e.g., rephrasing or paraphrasing) and sensitivity to meaning-altering ones. Semantically non-minimal interventions, therefore, provide a principled mechanism to disentangle these two dimensions of model behavior.

By introducing interventions that intentionally violate minimality at the semantic level, we gain deeper insight into whether a model's predictions are grounded in meaningful conceptual structure or driven by shallow correlations. In this sense, semantically non-minimal counterfactuals operate not as adversarial attacks, but as diagnostic instruments—revealing when a model's invariance ceases to reflect robustness and instead signals a lack of semantic discrimination.

Non-minimal interventions are adjacent to **invariant outcomes**: the outcome of a model before and after intervention should remain almost identical, or demonstrate a minimal change which however does not significantly shift the overall interpretation of this outcome. To this end, non-minimal interventions leading to invariant outcomes is not meaningful in discrete classification settings, but rather in more continuous outcomes, such as in ranking tasks, where an ordered set of items can be significantly or minimally altered under the presence of a non-minimal intervention, indicating the degree of sensitivity of semantic ordering. Overall, this direction can help uncover model properties ignored by typical counterfactual approaches which include:

- **Model insensitivity** (which features are *not* driving decisions). Going one step further, from a causal perspective, if an intervention on a non-causal variable leaves the output unchanged, that supports the causal sufficiency of the model.
- **Spurious correlations** (model does not change output even with meaningful changes from human perspective). Such an observation may even suggests that a model may operate in a different semantic

level in comparison to a human. In practical terms, the presented invariance after semantically non-minimal interventions may be proven problematic in real-world systems, where discrimination driven from key semantics is crucial for the operationality of the underlying system.

- Robustness to irrelevant or adversarially-crafted changes: If a model predicts the same label despite large changes to irrelevant features, this shows good robustness. Conversely, lack of invariance can reveal brittle behavior.
- Fairness auditing: When changes in protected attributes do not change the decision. For example, changing race, gender, or ZIP code shouldn't change the outcome in a hiring model.

The above frame the crucial contribution of non-minimal interventions, which are going to be introduced on the task of semantic similarity, as thoroughly analyzed in the following section.

Explaining semantic similarity models

We propose a model-agnostic explainability framework that enforces non-minimal interventions over key semantics in sentences, evaluating retrieval outcomes produced by semantic similarity models.

The first use-case of our proposed explainability framework targets semantic similarity models on visual vocabularies [204], exemplifying the value of non-minimal interventions. Generally, a semantic similarity model M is tasked to produce embedding representations for sentences. In a dataset comprised of N such natural language sentences, given a query sentence q, a ranking R for the rest $N-1$ sentences is obtained based on the proximity of their embeddings.

5.2.6.1 Problem overview

In our study, we utilize semantic similarity models for text-image retrieval, i.e. retrieving the most appropriate image given a textual query. Despite being a multimodal problem, image-text retrieval is appropriately converted in a unimodal one. As input, we consider a dataset of size N that contains complex scene images $I_i \in \mathcal{I}$, accompanied by query-corpus pairs $(q_i, c_i), q_i \in \mathcal{Q}, c_i \in \mathcal{C}, i = 1, 2, ..., N$ with each corpus c_i consisting of an arbitrary number of sentences $s_j, j = 1, 2, ..., l_c$. Queries $q_i \in \mathcal{Q}$ correspond to high-level descriptions of images, similar to the descriptions a user would input to a search engine for images. On the other hand, corpora $c_i \in \mathcal{C}$ provide a more detailed description of an image, incorporating several details within a longer visual passage. This passage can be constructed by aggregating individual sentences from areas of the scene.

By excluding images $I_i \in \mathcal{I}$ and only considering purely linguistic (q_i, c_i) pairs as representatives for each data sample, we obtain embedding ***representations*** of $\mathcal{Q}$,$\mathcal{C}$ instances in a common vector space U by leveraging pre-trained sentence similarity transformers M from SBERT [282]. Cosine similarity scores between query-corpus embedding pairs in U are sorted to provide a rank R_i per query q_i in the consequent ***ranking*** stage, with R_i either lead to *success*, if the ground truth image I_{g_i} with corpus c_i is returned at the top of the rank, or *failure* otherwise. Finally, in the ***evaluation*** stage we assess the ranking quality via appropriate metrics. Moreover, all failures per model M, i.e. image pairs (I_g, I_r) for which $I_g \neq I_r$ are stored in a set $\mathcal{F}$, which is further passed to the ***evaluation*** stage. More specifically, we count as failure $f_i = (I_g, I_r)_i \in \mathcal{F}$ any instance of a ground truth image I_{g_i} with corpus c_i that was not ranked in the first position ($\mathsf{rank}_i \neq 1$) given q_i; instead another image $I_{r_i} \neq I_{g_i}$ with $c_r \neq c_i$ achieved $\mathsf{rank}_r = 1$. Following the 'blind' evaluation strategy of traditional ranking metrics, we provide a measure of retrieval failures as the cardinality of the failure set: $F = |\mathcal{F}|$ for each M. The overview of the aforementioned procedure is exhibited in Figure 5.2.5.

Visual concepts in language

Visual vocabularies contain descriptions about real life scenes, including objects, relationships and attributes. Datasets that connect visual vocabularies paired with images, such as Visual Genome [155], COCO [183] and Flickr [396] set our sources to construct purely textual query-corpus pairs, assuming that necessary visual information is contained within the high quality annotations of those datasets. In particular, the annotation diversity allows either shorter, global descriptions, as in Flickr and COCO captions, or detailed descriptions in local level, as in Visual Genome region descriptions, concatenated in a corpus c_i per image I_i.

Choice of explicit knowledge base

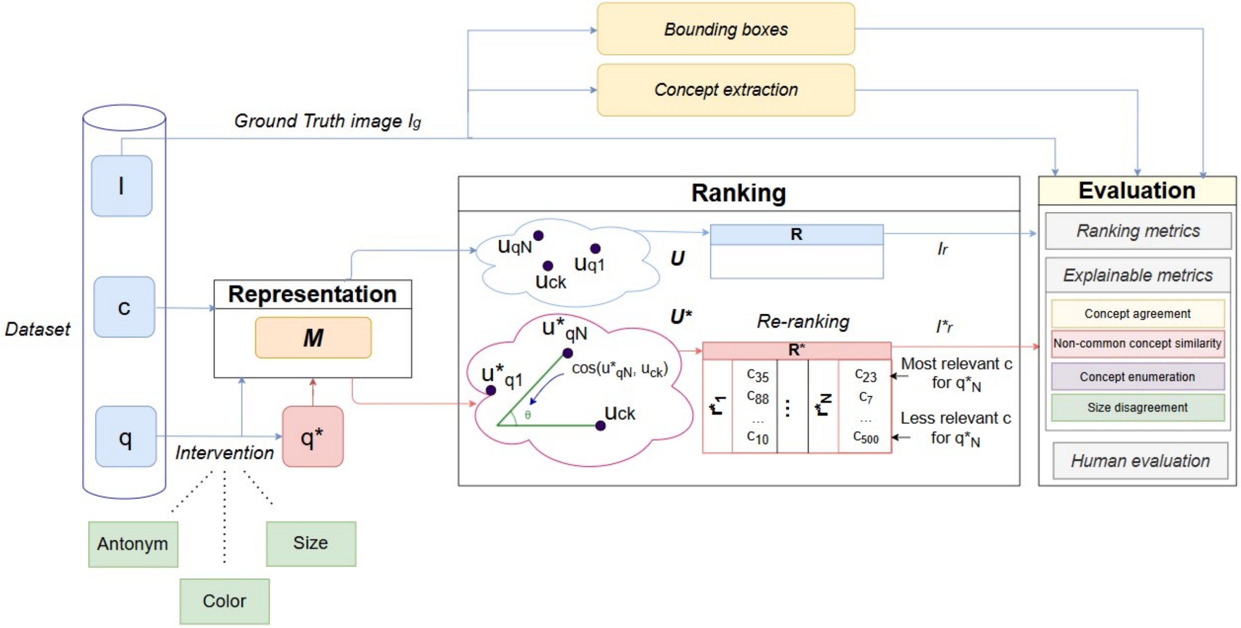


Figure 5.2.5: An overview of the *representation-ranking-evaluation* procedure to explain semantic similarity models. Both the default stream (blue) and the intervened one (pink) are provided.

In order to define distances for perturbed semantics, we explore two options. The first one includes the existing **Wordnet** hierarchy [226], providing an explicit tree structure to describe semantic interrelationships between concepts. As a second option, we construct a **Color hierarchy** that provides the relativeness of colors present in common sentences, such as the ones in COCO, Flickr or Visual Genome. Color distance is provided via the RGB values of Matplotlib colors[3].

Optimal embedding representation

Obtaining an overall representation of a corpus c_i is not trivial, as existing transformers can handle up to a certain number of input tokens per sentence. To resolve this, we can independently embed each corpus sentence $s_j \in c_i, j = 1, 2, ..., l_c$ using a model M, and then calculate the average of all vectors v_j^c. Therefore, $u_i^c = \frac{1}{l_c} \sum_{j=1}^{l_c} v_j^c \in U$ serves as the averaged representation for c_i. Another approach is to leverage state-of-the-art abstractive summarizers [407, 274] to obtain a meaningful shorter version of c_i while maintaining semantics as much as possible, and then apply M only once per c_i. Query representations $u_i^c \in U$ are produced by inserting each $q_i \in \mathcal{Q}$ in a model M, or by averaging over representations when q_i comprises from more than one sentences. The quality of each representation can be evaluated based on their efficiency upon a downstream task targeting semantic similarity.

Ranking

Given a model M, each query representation $u_i^q \in U$ is paired with all corpus representations $\{u_1^c, u_2^c, ..., u_N^c\} \in U$, and cosine similarity scores are calculated for each pair. Higher cosine similarity scores yield more similar representations, therefore sorting from higher to lower scores provides the ranking R_i per q_i. The process is repeated for all N images resulting in N^2 calculations. A closer look at the ranking procedure is illustrated in Figure 5.2.6.

Traditional ranking metrics evaluate the ranking success, coarsely indicating the representation quality of each M. Recall@k returns the proportion of ground truth images found in top-k ranked instances for all queries $q_1, q_2, ...q_N$, given that each q_i has only one ground truth c_i. Mean Reciprocal Rank (MRR) is the averaged of the inverse of the ground truth rank position $\texttt{rank}_i$ for each c_i given q_i, considering the top-k

---

[3] Matplotlib colors

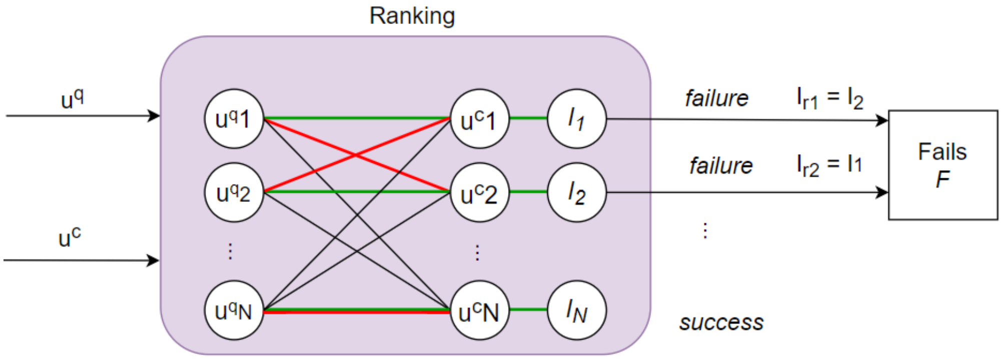


Figure 5.2.6: A closer look at ranking procedure. Green lines denote ground truth matchings, while red lines indicate matchings selected from maximum cosine similarity scores between query and corpus embeddings [204].

items:

$$MRR@k = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\mathsf{rank}_i} for\ each\ \mathsf{rank}_i \geq k. \tag{5.2.16}$$

We calculate Recall@k and MRR@k for k=5, 10, N. Also, we calculate the median rank position for all c_i.

5.2.6.2 Interventions on salient attributes

We utilize counterfactual interventions based on explicit knowledge to obtain counterfactual queries $q \rightarrow q^*$ targeting key attributes and produce respective representations in U, upon which counterfactual rankings R^* per q^* are extracted. Interventions upon attributes are defined based on the WordNet [226] knowledge graph, which allows for deterministically trasversing the path between two WordNet concepts, which correspond to sentence attributes. More specifically, we perturb salient semantics in queries $q_i \in \mathcal{Q}$, producing $q_i^* \in \mathcal{Q}^*$, and evaluate the changes occurring in the rank. Figure 5.2.7 provides the causal graph of interventions for any $q_i \in \mathcal{Q}$.

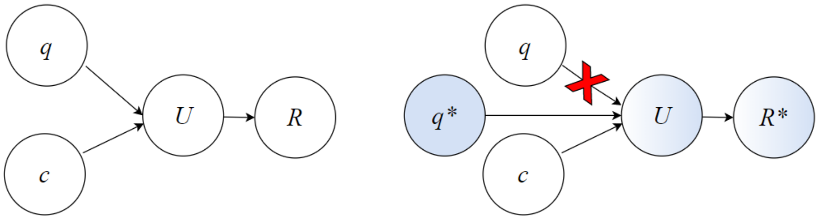


Figure 5.2.7: Conventional causal graph (left) and counterfactual intervention causal graph (right) when $q \rightarrow q^*$.

Let's consider a real-world scenario in which we are searching for images corresponding to the query *a man walking in a dark corridor* and images corresponding to *a man walking in a bright corridor*. Intuitively, we are expecting to receive totally different images from an appropriate retrieval system. To mimic this contrastive behavior, we exploit non-minimal interventions.

An appropriate non-minimal counterfactual perturbation must conceptually reverse salient semantics, be focused on an individual semantic each time, and the resulting query q_i^* should be linguistically correct. With respect to those requirements and in order to restrict the search space of counterfactuals, we target substituting ***object attributes***. An example regarding a graph of such substitutions is presented in Figure 5.2.8. Initially, generic counterfactual queries include replacing attributes with their antonyms. More refined subsequent counterfactuals focus on replacing object colors and sizes; such substitutions are discrete, fast and controllable.

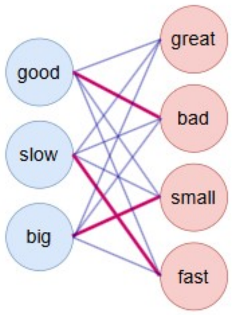


Figure 5.2.8: Example of semantically non-minimal interventions on attributes. With blue color we denote the initial attributes, while the pink nodes correspond to the attributes that substitute the initial ones, so that the intervention is performed. The bold edges between nodes denote the matched concepts-which actually consist the best matches available on this graph.

Antonyms are extracted via relevant WordNet functions for any adjective present in a query. If more than one antonyms are returned, one is randomly picked to substitute the actual word.

Color substitution refers to changing colors present in the sentence with another distant color, based on our constructed *color hierarchy*. We set a proximity threshold to ensure perceptually non-negligible color changes. Two possible substitutions are attempted: either considering all RGB colors (*color-all)*, or colors only mentioned in the dataset (*color-in)*.

Size substitution is an antonym substitution specialized in sizes. Words such as *large, big, enormous, huge* are substituted with a random choice among *small, little, minor, tiny* and vice versa.

5.2.6.3 Re-ranking evaluation

Perturbed query representations $u_i^{q*} \in U$ of $q^* \neq q$ with $u_i^c \in U$ of corpus c_i may directly influence the final ranking R^* when $\mathsf{rank}_i^* < \mathsf{rank}_i$ or inversely $\mathsf{rank}_i^* > \mathsf{rank}_i$. Intuitively, any non-negligible perturbation of q should result in worse position $\mathsf{rank}_i^* > \mathsf{rank}_i$, as the counterfactual query representation u_i^{q*} would diverge from c_i comparing to u_i^q, due to the substitution of the actual semantic with a conceptually different one. However, given the relative nature of ranking, some instances may stay in the same position $\mathsf{rank}_i^*{=}\mathsf{rank}_i$, or even go higher. Ascending in the rank does not imply a better q_i^*, c_i matching, except if their in between cosine similarity increases; instead, the distorted representations push lower previously higher-ranked instances, virtually improving some rank_i^*. In any case, we expect all ranking metrics to perceptibly drop, as we pull apart ground truth matchings in U.

The following evaluation concerns N =34K images from the VG∩COCO subset, where both queries and corpora (corresponding to COCO caption and aggregated VG region captions respectively) can be collected. This large dataset size allows drawing meaningful conclusion regarding the contribution of non-minimal interventions over salient visual attributes on related descriptions.

5.2.6.4 Quantitative results under intervention

In our experimentation, **antonym-based interventions** concerning adjectives are applicable on 10523 (30.93%) queries, producing $q^* \neq q$ which result in updated embedding representations: the cosine similarity $cos(u_i^q, u_i^{q*}) < 1$. By exclusively considering intervened instances with updated representations $u_i^q \neq u_i^{q*}$, we observe that 4346 instances (41.30%) are ranked lower than the original ones, 2918 instances (27.73%) are ranked higher, and 3259 instances (30.97%) remain in the same position. Moreover, for **size-based interventions**, 10471 (30.77%) queries of the dataset are perturbed with respect to size-oriented words. Consequently, 3902 instances (37.26%) are ranked lower than the original ones, 2961 instances (28.28%) are ranked higher, and 3608 instances (34.46%) remain in the same position. As for **Color-based interventions**, 16007 (47.04%) queries that contain colors are perturbed using non-minimal interventions driven by color distance. Regarding the *color-in*) experiment and by considering intervened instances with updated representations, 7960 instances (49.73%) are ranked lower, 3578 instances (22.35%) are ranked higher, and 4469 instances (27.92%) remain in the same position. In the *color-all* experiment, we observe that 7729 instances (48.29%) are ranked lower than the original ones, 3848 instances (24.04%) are ranked higher, and 4430 instances (27.68%) remain in the same position.

Updated query representations result in re-ranking of instances; specifically, on average almost 70% of instances changed position in R^* comparing to R as presented in Table 5.8. The $\#q^*$ column refers to number of perturbed queries, while *Lower, Higher, Same* columns refer to the position change.

Intervention	$\#q^*$		rank*	%
		Lower	Higher	Same
Antonym	10523	41.30	27.73	30.97
Color-all	16007	48.29	24.04	27.68
Color-in	16007	49.73	22.35	27.92
Size	10471	37.26	28.28	34.46

Table 5.8: Changes observed for all counterfactual interventions on VG∩COCO subset.

Overall, despite the re-arrangements of individual instances, R^* was only *marginally altered* in global level for any of the counterfactual perturbations according to all *query-agnostic* ranking metrics, as presented in Table 5.9.

Intervention	Recall	(%)	MRR	(%)	Fails
	@1	@10	@10	@*all*	(%)
Original	15.31	39.48	22.11	23.42	84.69
Antonym	15.13	38.82	21.76	23.07	84.87
Color-all	14.52	38.16	21.15	22.48	85.48
Color-in	14.52	38.05	21.12	22.46	85.48
Size	15.19	39.32	21.97	23.29	84.81

Table 5.9: Ranking results on counterfactual queries.

Therefore, either by providing meaningful and relevant queries or conceptually divergent ones, the response of a semantic similarity system is virtually the same. This invariance over non-minimal interventions generally questions the trustworthiness of opaque ranking metrics, highlighting the need for *explainable evaluation.*

By qualitatively assessing retrieval failures post-intervention, we observe that perturbed semantics are rather bypassed in favor of preserving object class. Even if this could imply representation robustness, on the other hand it can be attributed to language model biases towards object identities. In any case, existing ranking metrics cannot indicate potential biases, patterns and rules in the linguistic representations due to their opaque nature.

5.2.6.5 Explainable evaluation

Another aspect concerning explainability issues of semantic similarity models is the lack of transparent evaluation [204]. Apart from the observed invariance under non-minimal interventions, we get one step back, acknowledging that existing metrics are not informative enough by design, at least in the text-image problem as formulated in the previous sections: traditional ranking metrics [218] provide either a binary answer (item found in top-k items or not), or position-informed variants (item found in the k-th position). However, such measures cannot provide detailed insights regarding the contribution of the scene constituents of a scene image to the rank position. For example, if an instance is ranked in the k-th position, items in previous k-1 positions may be highly relevant to the ground truth one or on the contrary, highly irrelevant. To this end, we propose novel explainable ranking evaluation metrics that decompose and quantify the conceptual differences between ground truth and retrieved instances in local and global level.

Our purpose is to expand the existing ranking evaluation strategy within the *image similarity* setting: since our dataset is comprised of image-text pairs, while also bounding boxes for objects, together with the corresponding object classes are provided, we design evaluation algorithms to measure the agreement between ground truth and top-1 retrieved instances.

5.2.6.6 Towards explainable evaluation metrics

We design four evaluation stages for all failures $f_i \in \mathcal{F}$, starting from more influential concepts and moving towards less prevalent details. Visual concepts are focused on scene ***objects***. For fair comparison with traditional ranking metrics, we demonstrate a *query-agnostic* evaluation approach: we compare concepts between retrieved and ground truth images without considering query semantics. In the next paragraphs we drop i subscript for simplicity.

Concept agreement - CA Considering $\mathcal{V}$ as a set of visual concepts, **concept agreement** measures the percentage of ground truth concepts $\mathcal{V}(I_g)$ contained in the retrieved concept set $\mathcal{V}(I_r)$ over all $\mathcal{V}(I_g)$ concepts for each f_i. Let $V_{(g,r)} = \mathcal{V}(I_g) \cap \mathcal{V}(I_r)$ the set of common concepts:

$$CA_f = \frac{|V_{(g,r)}|}{|\mathcal{V}(I_g)|},\ f = (I_g, I_r)$$

Higher ***CA*** indicates higher concept similarity. For example, if $\mathcal{V}(I_r) = \{\mathsf{Dog}, \mathsf{Frisbee}, \mathsf{Park}\}$ and $\mathcal{V}(I_g) = \{\mathsf{Dog}, \mathsf{Ball}, \mathsf{Park}\}$, then the $\boldsymbol{CA}=\frac{2}{3}$. On the other hand, if $\mathcal{V}'(I_r) = \{\mathsf{Cat}, \mathsf{Fish}\}$, then $\boldsymbol{CA}=0$, as no overlap exists. This way we can confidently conclude that the first retrieved image is conceptually closer to the ground truth than the second, and by extension the model used to retrieve the first image is better with respect to ***CA***.

Non-common concept similarity - NCS aims to provide a distance measure between concepts present exclusively in either $\mathcal{V}(I_g)$ or $\mathcal{V}(I_r)$. For example, we would expect the set $\{\mathsf{Dog}, \mathsf{Frisbee}, \mathsf{Park}\}$ to be more similar to $\{\mathsf{Dog}, \mathsf{Ball}, \mathsf{Park}\}$ than $\{\mathsf{Dog}, \mathsf{Cat}, \mathsf{Park}\}$, since the non-common concept Frisbee is conceptually closer to Ball than Cat. Mathematically, let $D_g = V_{(g,r)} - \mathcal{V}(I_r)$ and $D_r = V_{(g,r)} - \mathcal{V}(I_g)$, with both $D_g, D_r \neq \emptyset$. Other than that, D_g and D_r may contain different number of concepts. Then, a measure of concept distance can be provided by calculating the path similarity score ps of corresponding WordNet [226] synset pairs, based on the shortest available path between those two concepts. Path similarity ps ranges between 0 and 1.

An *optimistic* ***NCS*** metric returns the maximum possible cumulative ps averaged over the number of pairs, by appropriately selecting concept pairs between non-empty D_g and D_r. The maximization of ***NCS*** requires a dynamic programming solution, as naive strategies taking into account all possible D_g and D_r pairs would yield a factorial amount of combinations. To trespass this prohibitive complexity, we create a bipartite graph $G = (D_g, D_r, E)$ from D_g and D_r: all concept nodes from the one set are matched with all the nodes of the other via edges $e_y \in E, y = 1, 2, ..., |D_g| \times |D_r|$, while no edges are allowed within the same set. Edge weights w_{e_y} correspond to WordNet ps scores between synsets of connected nodes.

Consequently, the *maximum weight bipartite matching* on G refers to pairing D_g and D_r concepts so that the cumulative edge weight is maximized. An optimized version of the Hungarian algorithm [**hungarian_algo**, **max_weight_matching**] implemented by NetworkX[4] reduces the computational complexity of finding the maximum ps to $O(|V|^3)$, where $|V| = max(|D_g|, |D_r|)$.

Therefore, ***NCS*** can be written as:

$$NCS_f = avg(max\_weight\_match(G)),\ G = (V_{(g,r)} - \mathcal{V}(I_r), V_{(g,r)} - \mathcal{V}(I_g), E))$$

Higher ***NCS*** scores reveal more similar concepts.

Concept enumeration - CE Real world scenes may contain repeated instances of same-class concepts, forming concept multisets $\mathcal{V}_m = \{(\mathcal{V}_1, |\mathcal{V}_1|), (\mathcal{V}_2, |\mathcal{V}_2|), ..., (\mathcal{V}_x, |\mathcal{V}_x|)\}$, where $\mathcal{V}_1, \mathcal{V}_2, ..., \mathcal{V}_x$ denote concept categories, and $|\mathcal{V}_1|, |\mathcal{V}_2|, ..., |\mathcal{V}_x|$ cardinalities per category. The cardinality per concept category is called concept multiplicity in the multiset. ***CE*** penalizes differences in multiplicities between common concepts of I_g and I_r for each f_i:

$$CE_f = \sum_{j=1}^{x} ||\mathcal{V}_j(I_g)| - |\mathcal{V}_j(I_r)||_{\mathcal{V}_j(I_g)=\mathcal{V}_j(I_r)}$$

---

[4] NetworkX max weight matching

Higher $\boldsymbol{CE}$ scores demonstrate higher enumeration disagreement, deeming lower $\boldsymbol{CE}$ values more favorable. For example, if I_g contained 10 dogs and 1 frisbee $\{(\mathsf{Dog}, 10), (\mathsf{Frisbee}, 1)\}$, a retrieved I_r with 1 dog and 1 frisbee would have $\boldsymbol{CE}$=9, while an I'_r with 10 dogs and 1 ball would have a $\boldsymbol{CE}$=0. Therefore, the first image yields a worse CE score than the second, even though the second would have worse $\boldsymbol{CA}$ and $\boldsymbol{NCS}$ scores than the first one.

Size disagreement - SD Even in cases where there is a high agreement of objects and multiplicities between I_g and I_r, disagreement in object sizes may correspond to semantically divergent scenes. For example an image with a dog in the foreground (large bounding box) is different than an image of a dog in the background (small bounding box). To capture this difference, we design an *optimistic* $\boldsymbol{SD}$ metric which returns the area differences of bounding boxes $\mathcal{D}_A = |A_g - A_r|$ for all available object matchings. Such matchings occur by pairing concepts of the same category u between $\mathcal{I}_g$ and $\mathcal{I}_r$ up to the point that no more unique pairs can be constructed. This is equivalent of creating bipartite graph $G_u = (\mathcal{V}_u(I_g), \mathcal{V}_u(I_r), E)$, where $\mathcal{V}_u(I_g), \mathcal{V}_u(I_r)$ belong in the same u and edge weights $w_{e_y}, e_y \in E, y = 1, 2, ..., |\mathcal{V}(I_g)| \times |\mathcal{V}(I_r)|$ denote the area difference $\mathcal{D}_A$ between concept nodes. Pairing concepts with similar bounding box areas can be considered as the optimal choice, therefore node pairs connected by lower edge weights w_{e_y} are preferred. Finding the *minimum weight matching* provides the most similar pairs size-wise, and can be solved in polynomial time using the NetworkX[5] implementation of Karp algorithm [**min_weight_matching**]. The matching process is repeated for all concept categories in the multiset $\mathcal{V}_m$, resulting in a set of graphs G_m:

$$SD_f = \sum^{\mathcal{V}_u \in \mathcal{V}_m} avg(min\_weight\_match(G_u)),$$

$$G_u = (\mathcal{V}_u(I_g), \mathcal{V}_u(I_r), E), G_u \in G_m$$

A simplified *binary* version of SD increases a sum if area differences of paired concepts are above a predefined threshold T_D.

5.2.6.7 Experimental details

We focus on presenting experiments on our construted VG∩COCO of N=34k images: the dataset size N itself, as well as the more detailed region descriptions of Visual Genome which comprise a larger corpus set $\mathcal{C}$, require accurate linguistic representations in order to retrieve more relevant images.

Selected language models are designed for semantic similarity, and according to the datasets they have been pre-trained on, they can be divided in: *all-* models pretrained and fine-tuned on 1B sentence pairs from multiple sources; *multi-qa-* trained on 215M diverse question-answer pairs, learning to map queries to passages; *stsb-* models trained on the STSbenchmark, which contains sentence pairs annotated with similarity scores; *paraphrase-* models with more than 86M paraphrase sentence pairs containing more challenging and uncurated characteristics comparing to STSb; *nq-* models trained with 100k real Google search queries mapped to Wikipedia passages; *nli-* models incorporate natural language inference data pairs (premise/hypothesis), included in AllNLI dataset.

By conducting a large number of experiments to estimate the performance of such models on visual vocabularies, we observe certain patterns in ranking results. In all experiments, most paraphrase models consistently outperform the rest. Paraphrasers have been pre-trained on image captions (COCO, Flickr), which actually serve as paraphrasing data: during the construction of these datasets, annotators have independently produced varying descriptions for the same concepts. Query-corpus pairs can be viewed as the one being a paraphrase of the other, thus paraphrasers have learned a suitable representation for this matching, together with their exposure to visual vocabularies.

Quantitative Results Table 5.10 presents ranking results on VG∩COCO.

Qualitative results In all following experiments, we consider results from the best performing model MiniLM-L3 on VG∩COCO. In total, F=28817 queries failed to retrieve their corresponding ground truth I_g.

---

[5] NetworkX min weight full matching

Model Name	Recall(%)↑			MRR(%)↑			Median Rank ↓	Fails (%) ↓
	@1	@5	@10	@5	@10	@*all*		
all MiniLM L12	13.31	26.93	34.31	18.18	19.16	20.45	34	86.69
all MiniLM L6	13.07	26.80	34.26	18.02	19.02	20.29	35	86.93
all distilroberta	12.20	26.10	33.58	17.22	18.22	19.55	35	87.80
all mpnet base	12.64	26.12	33.82	17.45	18.47	19.78	35	87.36
all roberta large	12.01	25.63	33.42	16.87	17.91	19.24	35	87.99
multi qa MiniLM L6 cos	9.29	20.10	26.77	13.14	14.03	15.29	60	90.71
multi qa distilbert cos	9.22	20.35	27.04	13.16	14.04	15.32	57	90.78
multi qa mpnet base cos	9.53	21.03	27.81	13.63	14.52	15.79	55	90.47
nli distilroberta base	11.73	24.94	32.33	16.48	17.46	18.75	39	88.27
nq distilbert base	8.03	17.95	23.95	11.53	12.32	13.54	76	91.97
paraphrase MiniLM L12	14.24	29.89	38.34	19.89	21.02	22.37	24	85.76
paraphrase MiniLM L3	**15.31**	**30.92**	**39.48**	**20.97**	**22.11**	**23.42**	23	**84.69**
paraphrase MiniLM L6	14.39	30.01	38.51	19.96	21.10	22.46	24	85.61
paraphrase TinyBERT L6	14.55	30.38	39.12	20.27	21.43	22.82	**22**	85.45
paraphrase albert base	13.12	27.83	36.11	18.43	19.54	20.91	28	86.88
paraphrase albert small	14.56	30.25	39.04	20.23	21.40	22.75	23	85.44
paraphrase distilroberta	14.51	30.39	38.91	20.23	21.36	22.74	**22**	85.49
paraphrase mpnet	13.99	28.99	37.40	19.39	20.50	21.84	26	86.01
xlm distilroberta paraphrase	11.59	24.62	31.87	16.26	17.23	18.52	40	88.41
stsb distilroberta base	13.41	27.04	34.28	18.32	19.28	20.55	35	86.59
stsb mpnet base	14.05	28.26	35.93	19.23	20.24	21.49	32	85.95
stsb roberta base	13.69	27.38	34.79	18.67	19.65	20.92	34	86.31
stsb roberta large	10.32	22.13	28.71	14.60	15.47	16.73	53	89.68

Table 5.10: Rank results on the VG ∩ COCO dataset for our semantic similarity models.

5.2.6.7.1 Local evaluation

The real power of our proposed metrics lies in local level. We present an example from the *color* and *details* failure category below. Given a query q_i (caption), Figure 5.2.9 shows the ground truth I_{g_i} (left) and the retrieved I_{r_i} (right).

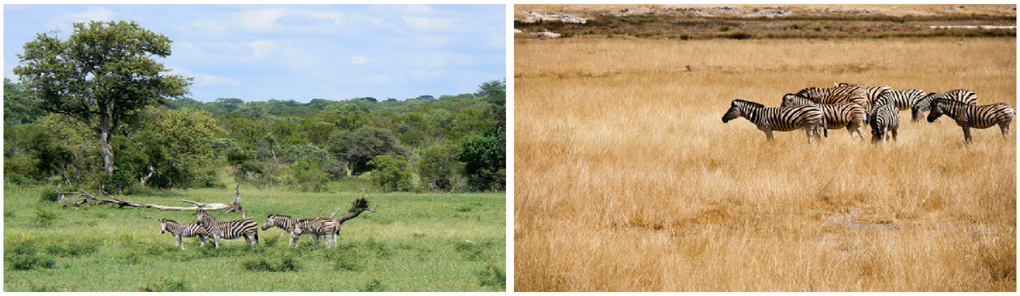


Figure 5.2.9: A herd of zebras grazing in a lush green field

The set of ground truth object synsets is {trunk.n.01, hill.n.01, tree.n.01, sky.n.01, field.n.01, branch.n.01, head.n.01, leg.n.01, leaf.n.01, zebra.n.01, mane.n.01} of cardinality 11, and the set of retrieved ones is {grassland.n.01, field.n.01, zebra.n.01, mane.n.01, grass.n.01} of cardinality 5. Common synsets are {zebra.n.01, mane.n.01, field.n.01} of cardinality=3, resulting in $\boldsymbol{CA}_i$=27.28%. Regarding $\boldsymbol{NCS}$, the constructed bipartite graph G contains $|V|$=10, $|E|$=16, and the best matched synset pairs according to the *maximum weight matching* are {hill.n.01, grassland.n.01} with ps=0.111, and {tree.n.01, grass.n.01} with ps=0.167. The average ps for all matched pairs leads to $\boldsymbol{NCS}_i$=0.139. Common object enumeration provides the following multisets: $\mathcal{V}_m^g$ = {zebra.n.01, 5, field.n.01, 1, mane.n.01, 1} and $\mathcal{V}_m^r$ ={zebra.n.01, 7, field.n.01, 1, mane.n.01, 5}. Therefore, $\boldsymbol{CE}_i$= 6. As for $\boldsymbol{SD}$ for T_D=1, 3 bipartite graphs are created for the 3 common synsets. The first graph G_{mane} contains $|V|$=6 and $|E|$=5, resulting in 1 *minimum weight matching* of weight D_A=2.30≥ T_D. Therefore $\boldsymbol{SD}_{mane}$=1=$\boldsymbol{SD}_i$. The second graph G_{zebra} consists of $|V|$=11 and $|E|$=28, resulting in 4 *minimum weight matchings*, from which none trespassed the threshold T_D, resulting in $\boldsymbol{SD}_{zebra}$=0, thus maintaining $\boldsymbol{SD}_i$=1. Finally, the G_{field} graph of $|V|$=2 and $|E|$=1, leads to 1 *minimum weight matching* of weight

D_A=1.369$\geq T_D$, resulting in $\boldsymbol{SD}_{field}$=1, which increases the total sum $\boldsymbol{SD}_i$=2. Having in total 6 matches for all three graphs, the ***averaged*** $\boldsymbol{SD}_i$=33.3% for this f_i.

Perceptually, a major I_{g_i}, I_{r_i} disagreement can be attributed to not satisfying *lush, green* attributes rather than semantics addressed by our metrics. As for traditional ranking metrics, I_{g_i} was placed in rank_{g_i}=294 with reciprocal rank score of 0.0034 and R@k=0, k=1,5,10. Obviously, we cannot extract much information in local level about how much I_{g_i} and I_{r_i} conceptually deviate and what we should potentially regard and request from retrieved instances (*colors* such as *green* instead of *yellow*, *details* such as *lush* instead of *arid*) to ascend in the rank. To this end, we conclude that traditional ranking metrics are only helpful in a very abstract level.

The following Figures 5.2.10, 5.2.11, 5.2.12, 5.2.13 demonstrate some interesting results regarding retrieved images and their ground truth matchings with respect to a given query. Images to the left correspond to the retrieved image I_{r_i}, while images to the right denote the ground truth image I_{g_i} with respect to a query q_i appearing in the caption. The caption also mentions the failure category the images belong, according to human evaluation results.

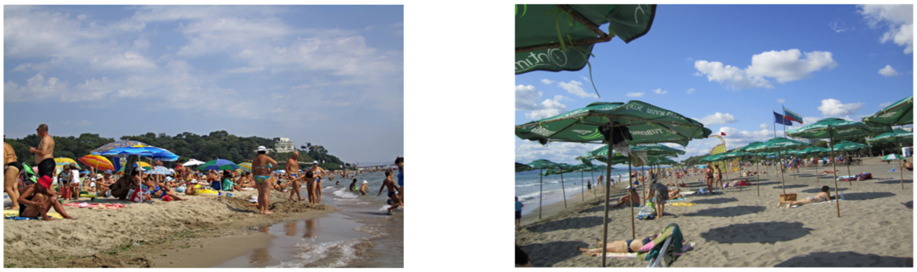


Figure 5.2.10: ***Successful alternative.***Many people are relaxing under their umbrellas on the beach.

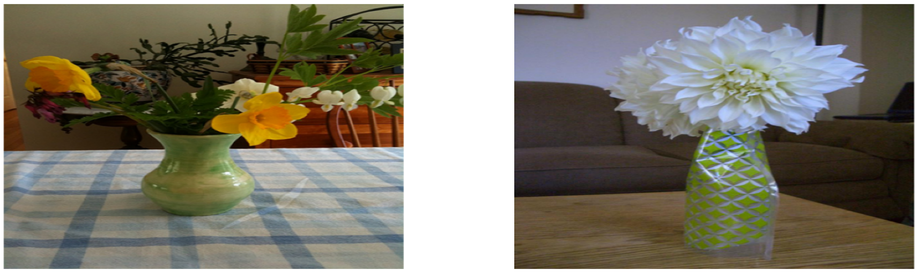


Figure 5.2.11: ***Object color.***A vase sitting on a table with white flowers in it.

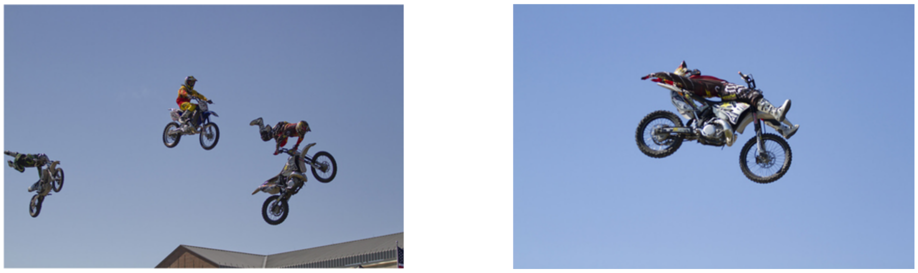


Figure 5.2.12: ***Object enumeration.*** A dirt bike rider performing a stunt while in the air.

Rules in failures

The frequent class of *successful alternatives* indicates that even when automatic metrics consider an I_{r_i} as failure, it may actually be a conceptually correct answer to q_i. Qualitative analysis over *successful alternatives* further demonstrated that almost all (I_{g_i}, I_{r_i}) pairs of this class were visually divergent, even though conceptually equivalent. Also, *details* and *object color* failure classes appeared often enough, indicating that

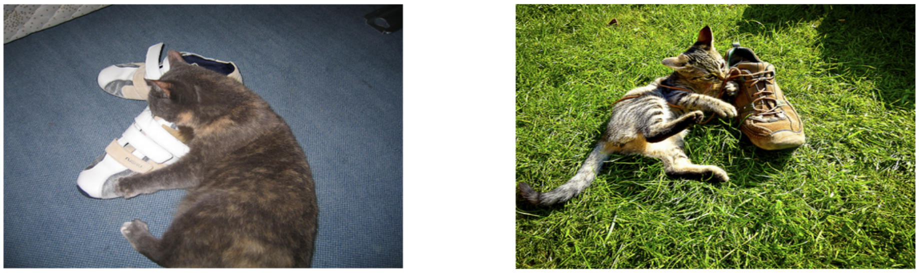


Figure 5.2.13: ***Detailed semantics.***A cat playing with a shoe in a grassy field.

those semantics are rather bypassed in order for others to be preserved. Combinations of semantics did not present any significant pattern; all semantics co-occurrences appeared in less than 10% of the evaluated instances. However, we did observe some frequent rules, which can be translated as: *if semantic A disagrees, then semantic B will disagree as well.* The rule action→details (*if action appears then details will appear*) is observed in 54.37% of the instances containing *action*; object enumeration →object color covers 17.48% of the instances containing *color*; finally, the reverse rule object color→object enumeration was observed in 11.65% of the instances containing *enumeration.*

Global evaluation

We present global *query-agnostic* results for our metrics. Despite our metrics being more meaningful in local level, global evaluation is useful for model benchmarking.

With 134630 common concepts between all $(I_g, I_r)_i \in \mathcal{F}$, the ***average concept agreement (CA)*** value is 22.29%, meaning that on average almost the 1/4th of I_g concepts appear in I_r.

With 903987 non-common concepts between all (I_g, I_r), and 134630 common ones, we retrieve 627833 and 110839 WordNet synsets respectively. The *maximum weight matching* between non-common synsets results in 184747 maximum weight matchings, equivalent to the 29.43% of all non-common synsets. Averaging over matchings (WordNet path similarities) for all (I_g, I_r), provides the ***average non-common concept similarity (NCS)*** score of 0.122.

With 41244 concept sets of same multiplicity and 69595 of different multiplicities regarding matched concept categories for all (I_g, I_r), ***most common concept enumeration (CE)***=1 and ***average CE***=8.638 instances for concepts of the same category reveals that in most cases there are not major enumeration differences.

Focusing on the binary ***SD***, we set the area difference threshold T_D=100%, increasing ***size disagreement (SD)*** by 1 iff $\mathcal{D}_A \geq 1$ between two concept bounding box areas. Thus, ***average SD***=20.35% for all (I_g, I_r), indicating that around 1/5th of common objects have non-negligible size differences.

Our metrics in global level reveal some extra capabilities. Most lower-ranked instances contained erroneous annotations, allowing a *post-hoc dataset cleaning* step that could not have been automatically realized otherwise.

Global results regarding our proposed metrics for all the models are presented in Table 5.11. Moreover, we offer some additional insights:

- Object hit: total number of common objects found between ground truth - wrongly retrieved images (I_g, I_r) at top-1 position.
- Object miss: total number of ground truth objects not found in top-1 retrieved images.
- Matched % synsets: Percentage of ground truth synsets found in top-1 retrieved images out of all ground truth synsets.
- Average % object enumeration disagreement: percentage of objects having the wrong number of instances between ground truth and top-1 retrieved over all ground truth objects (both having right or wrong number of instances).

As observed, the various explainable metrics indicate different models as best/worst performers, revealing that fine-grained evaluation may disagree with traditional coarse evaluation, while providing some useful insights.

Model Name	CA ↑	NCS ↑	CE ↓	SD ↓	obj hit↑	obj miss↓	matched % synsets ↑	avg % enum disagr.↓
all distilroberta	21.75	0.12	9.14	20.02	136025	932261	29.43	2.40
all MiniLM L12	21.72	0.12	8.88	19.88	134188	920214	29.23	2.35
all MiniLM L6	21.90	0.12	8.91	19.53	135687	926717	29.27	2.39
all mpnet base	21.74	0.12	9.14	19.24	135278	926724	29.44	2.39
all roberta large	21.85	0.12	9.13	19.60	136720	935614	29.42	2.39
multi qa distilbert cos	21.84	**0.13**	9.06	16.97	140754	1007800	**30.60**	2.49
multi qa MiniLM L6 cos	21.93	**0.13**	8.69	18.86	**141399**	1008889	30.21	2.45
multi qa mpnet base cos	21.87	**0.13**	9.24	**16.69**	140455	1004153	30.54	2.52
nli distilroberta base	22.12	0.12	8.89	19.68	138882	943947	29.75	2.45
nq distilbert base	21.39	0.12	9.03	17.53	139608	1020776	30.27	2.38
paraphrase albert base	22.13	0.12	8.92	19.71	136867	927060	29.52	2.43
paraphrase albert small	**22.40**	0.12	8.89	19.48	136319	909759	29.69	2.47
paraphrase distilroberta	22.28	0.12	8.63	19.49	135679	911325	29.32	2.39
paraphrase MiniLM L12	22.28	0.12	8.69	20.05	136102	910631	29.23	2.39
paraphrase MiniLM L3	22.29	0.12	8.64	20.35	134630	**903987**	29.43	2.42
paraphrase MiniLM L6	22.07	0.12	8.61	19.37	134623	914535	29.46	2.37
paraphrase mpnet	22.26	0.12	8.71	19.46	136482	911716	29.08	2.39
paraphrase TinyBERT L6	22.15	0.12	8.55	19.67	134808	916573	29.17	2.36
xlm distilroberta paraphrase	22.08	0.12	9.05	19.35	138934	951708	29.60	2.44
stsb distilroberta base	22.09	0.12	8.91	19.99	136181	928388	29.81	2.45
stsb mpnet base	21.84	0.12	9.10	19.95	133670	916167	29.58	2.41
stsb roberta base	22.04	0.12	8.90	19.68	135544	925948	29.75	2.44
stsb roberta large	21.10	0.12	**8.29**	20.11	134706	961836	29.18	**2.24**

Table 5.11: Results from our proposed metrics plus some additional information occurring from our metrics per model

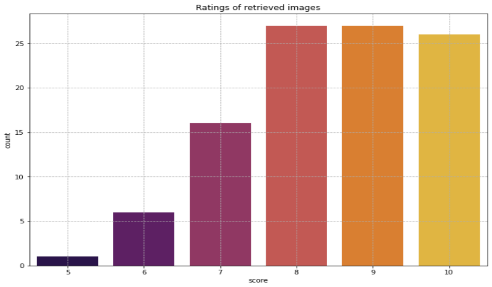
Human evaluation

Query-agnostic evaluation regards all scene semantics, even if in fact they are not present in the query. On the other hand, incorporating query information at evaluation stage conditions concept importance upon the presence of a concept in the query, forming a *query-informed* evaluation strategy. We conducted *query-informed* human evaluation experiments considering all failures in $\mathcal{F}$ and penalizing semantic disagreements only if those semantics are mentioned in q_i. Evaluators were primarily asked to mark which salient semantics were clearly misinterpreted in retrieved images with respect to the given query among the options: *object class, object color, object enumeration, action, size, details*. Otherwise, if I_{r_i} can be considered as conceptually similar to I_{g_i}, it is marked as *successful alternative*. Additionally, the overall retrieval quality is cross checked via qualitative ratings, assessing the conceptual similarity between I_{g_i}, I_{r_i} given q_i. Despite being unfair to compare with the -stricter- automated metrics, we expect lower values for *object enumeration* and *size* failure classes comparing to ***CE, SD*** metrics.

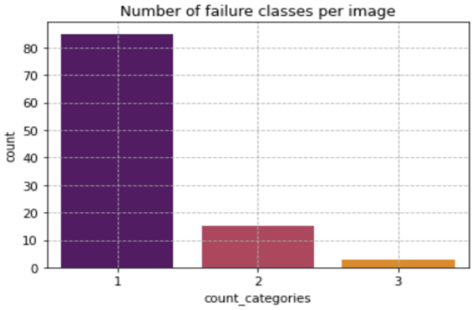
We present some distributions regarding human evaluation experiments. Figure 5.2.14a regards the rating distribution according to our evaluators' perception of ground truth-retrieved image relevance with respect to the given query. Figure 5.2.14b presents the number of failed semantics categories per image.

The crowdsourcing experiment reveals the most frequently misinterpreted attributes or combinations of attributes. Loss of conceptual information can be either attributed to dataset quality, i.e. salient query semantics not present in corpus, or on the capacity of the linguistic representations. Keyword matching between q_i and c_i excludes cases where the ground truth query-corpus pair contains very few common concepts, enabling the remaining samples to reveal patterns within the learned representations.

Results regarding misperceived semantics classes are presented in Table 5.12. The 82.52% of evaluated image pairs resulted in one semantic class disagreement, while the remaining 14.56% and 2.91% contained two and



(a) Human evaluation ratings (1-10).



(b) Number of marked disagreeing semantics per I_{g_i}, I_{r_i} pair for all evaluated image pairs in $\mathcal{F}$.

three semantic class disagreements respectively. The average rating over all classes was 8.47/10.

Regarding our example Figure 5.2.9, human evaluators rated I_{r_i}-q_i relevance with 6/10 on average and all of them marked *details* and *color* as the failure categories.

First, human evaluation experiments can indicate the degree of strictness of our automated metrics, as any *query-agnostic* metric may over-penalize semantics present in the I_{g_i} and c_i but not in q_i. Indeed, *query-informed* variants of our metrics are more relaxed. Moreover, patterns in reported failures also indicate patterns imprinted in the learned linguistic representations. Traditional ranking metrics cannot derive such fine-grained observations.

Alternatives %	Object %				Action %	Detail %
	class	color	enum.	size		
23.30	5.83	17.48	11.65	6.14	7.77	54.37

Table 5.12: Semantics disagreement percentage per class

5.2.6.8 Key findings on non-minimal interventions

The previous section highlighted the shortcomings of widely used ranking metrics, which can only provide a generic sense of the retrieval success of related systems. First of all, counterfactual interventions indicate that several state-of-the-art semantic similarity models are invariant against non-minimal semantic interventions focused on salient query attributes, at least when a averaged ranking metrics are considered. By qualitatively assessing counterfactual failures, we observe that perturbed semantics are rather bypassed in favor of preserving object class. Even if this could imply representation robustness, on the other hand it can be attributed to language model biases towards object identities. In any case, existing ranking metrics cannot indicate potential biases, patterns and rules in the linguistic representations due to their opaque nature. As a parallel venture in terms of knowledge-driven evaluation, explainable metrics are capable of shedding some light in the opaque ranking evaluation strategies commonly utilized, quantifying the degree of failure in cases that the ground truth is not retrieved by the semantic similarity model. A wide array of implemented models showcase the capabilities demonstrated by our proposed technique, which does not require any specific knowledge regarding the model under evaluation.

Another main take-away of this chapter poses some extra questions regarding ranking evaluation: Even in the case of explainable evaluation metrics, query semantics are not taken into account, generally exposing query-agnostic evaluation against any type of perturbation: even if the best possible answer to a query is returned based on similarity measures, how can we ensure that it is good enough in terms of actual relevance? The query-corpus relevance can be easily and explicitly measured via their common concepts, an approach followed in *query-informed* evaluation. To this end, we conclude that *explainable* and, even better, *query-informed* metrics are necessary to ensure evaluation robustness.

Chapter 6

Knowledge in vision-language explanations

The integration of knowledge into vision-language (VL) tasks represents a crucial step toward achieving genuine multimodal understanding and reasoning. While recent advances in large-scale pre-training have significantly improved performance across a variety of VL benchmarks, these models largely rely on pattern recognition rather than true comprehension. They excel at capturing statistical regularities between visual and textual modalities but often fail to reason about concepts, relationships, and commonsense knowledge that underpin human cognition.

Knowledge plays a fundamental role in bridging this gap. Humans do not interpret visual scenes or language in isolation; they rely on prior world knowledge to infer meaning, context, and intent. For example, when answering the question "Why is the umbrella open?", humans draw upon their understanding of weather and social conventions, linking the visual presence of rain clouds to the purpose of the umbrella. In contrast, standard VL models may answer correctly only when such co-occurrences were frequently observed during training, revealing their dependence on data-driven correlations rather than causal or semantic reasoning.

Incorporating structured and unstructured forms of knowledge—ranging from *knowledge graphs* such as ConceptNet and WordNet, to implicit representations captured by large language models (LLMs)—enables VL systems to go beyond surface-level matching. Knowledge enhances interpretability, robustness, and generalization by grounding visual and textual signals within a semantic space that encodes relations, hierarchies, and real-world constraints. This grounding supports reasoning over unseen or rare concepts, enables compositional understanding, and helps mitigate dataset biases by introducing contextually meaningful associations that are not purely data-dependent.

From a broader perspective, the study of knowledge in VL tasks marks a shift from perception-based intelligence to *reasoning-based intelligence.* It aims to endow models with the ability not only to recognize objects or words, but also to understand how these entities interact and why certain configurations make sense in a given context. This paradigm aligns with the long-term vision of explainable and trustworthy AI, where knowledge serves as the substrate for both interpretability and causal explanation.

In the context of the present thesis, we first aim to discover biases imbued in pre-trained models using conceptualization from explicit knowledge bases. We then proceed with two generative tasks: image generation from text and text generation from images (captioning). Knowledge helps addressing the key question of *what needs to be changed* in order for the generated modality to reach its ground truth conditioning. Based on this general framework, we study the following:

- *What concepts need to change in a generated **image**, for it to reach its **textual** conditioning?* The suggested changes act as a means of conceptually-rich evaluation of conditional image generation, enhancing and extending pixel-based image evaluation metrics, which can only offer limited evaluation details.

- *What concepts need to change in a generated **caption**, for it to reach its **visual** conditioning?* Inversely, we evaluate captioning quality in conceptual terms, often addressed in literature as a widely reported *hallucination* type; image captioners often hypothesize the presence of inexistent objects or relationships due to their over-reliance on statistical priors they have encountered during pre-training.

This analysis proves the value of conceptual counterfactuals in diverging multimodal tasks, instructing a new direction for bias discovery, hallucination detection and evaluation for potentially more tasks to come, as long as they can be represented in a conceptualized format and connected with explicit knowledge.

6.1 Explainability in Visual Question Answering

6.1.1 Introduction to Visual Question Answering (VQA)

Visual Question Answering (VQA) is a multimodal task that combines computer vision and natural language understanding. The goal is to enable a model to answer a natural language question about a given image.

Formally, given an image I and a question Q, the model must generate an answer A:

$$A = f(I, Q) \tag{6.1.1}$$

where $f(.)$ denotes a multimodal reasoning function that jointly interprets visual and textual information.

VQA tasks test a model's ability to:

- Recognize objects and scenes (e.g., "What animal is in the picture?")
- Understand relationships and attributes (e.g., "What color is the car?")
- Perform reasoning (e.g., "Is the person likely to be cooking?")
- Leverage commonsense knowledge (e.g., "Why is the umbrella open?")

Typical VQA systems integrate visual encoders (e.g., CNNs or Vision Transformers) with language encoders (e.g., BERT or LLMs), using a fusion mechanism (such as cross-attention or multimodal transformers) to combine the two modalities.

Biases in VQA systems

Existing systems addressing VQA as susceptible to several types of biases. First of all, language bias arises when the question alone is predictive of the answer, without needing the image. For example, if the question begins with "How many...," the most frequent answer might be "2." Early VQA models heavily overfit to these priors, achieving deceptively high accuracy without true visual grounding. Several works, such as [2] and [94], highlighted this issue, motivating the creation of balanced datasets (e.g., VQA v2.0 [359]) where each question is paired with image pairs leading to different answers (e.g., "yes" vs. "no").

Moreover, visual biases occur when models learn to associate specific visual features or regions with certain answers. For instance, the presence of snow may bias a model to answer "skiing" or "cold," even when irrelevant. This happens because datasets often include co-occurrence patterns—certain objects or scenes are overrepresented with specific questions or answers.

Ultimately, these biases result in shortcut learning: models optimize for answer prediction by exploiting dataset regularities instead of performing multimodal reasoning. For example, models may focus more on textual cues than on actual visual grounding. As discussed in [35] and [299], models can achieve competitive accuracy even when the visual input is masked or randomized, indicating limited true visual understanding.

A counterfactual approach

VQA explainability literature is often dominated by model-specific approaches that often need access to the model internals [6, 117, 134, 295, 250]. On the other hand, we can exclusively work on a fully black-box setting by altering conceptual units over a single modality and let the model reason on this alternative reality. By focusing on perturbing concepts in questions Q in a counterfactual manner, we pose the following research question: [320] *What is the response of the VQA model if we substitute word X with word Y in*

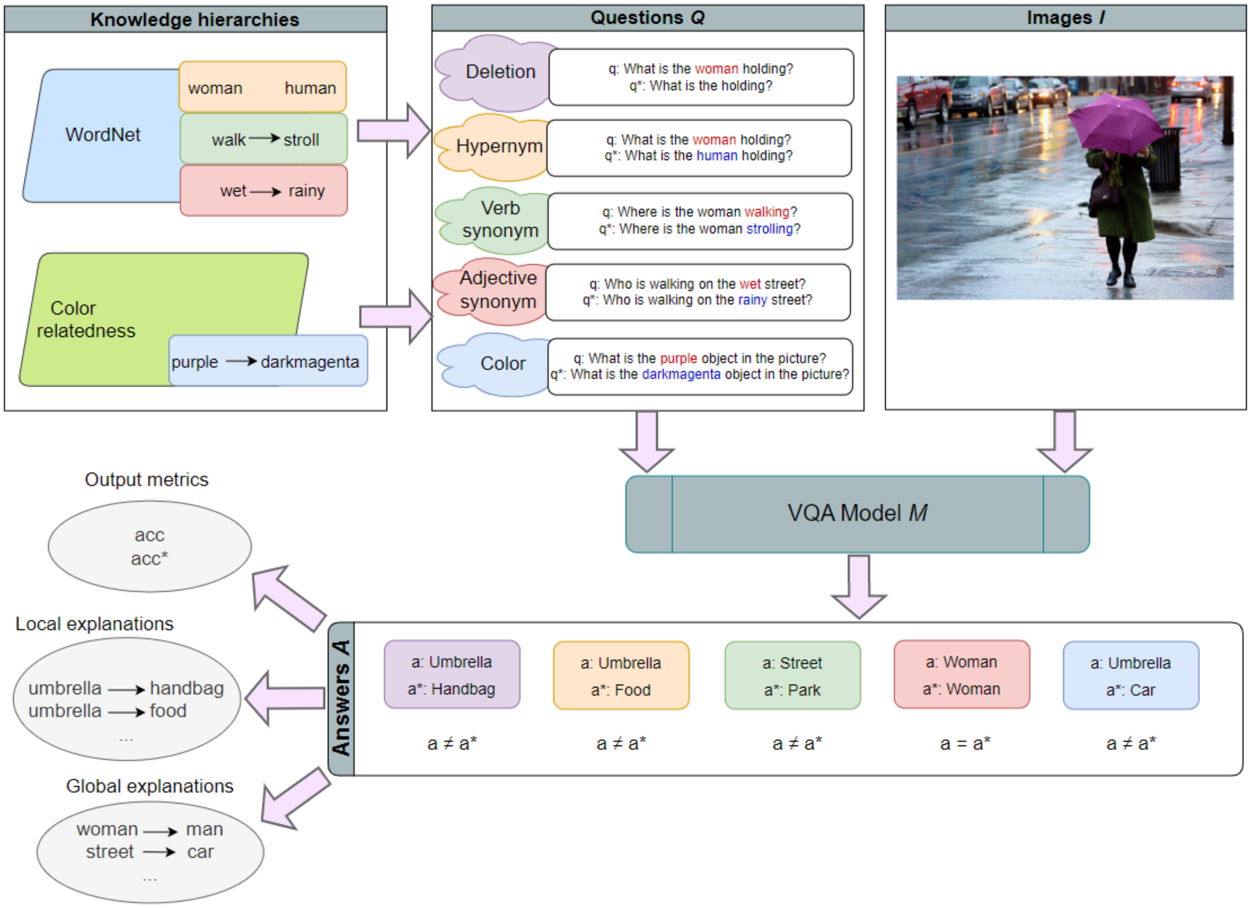


Figure 6.1.1: Overview of our proposed knowledge-based counterfactual VQA framework. [320]

question Q? Substitutions can be effectively driven by a hierarchical knowledge base, such as WordNet, ensuring constraint distances between concepts. Similarly, by exploiting color distances, which correspond to deterministic RGB values, we obtain minimal and non-minimal color substitutions [204]. The perturbations performed via knowledge-driven conceptualization are illustrated in Figure 6.1.1.

Experimental setup

Datasets VQA-v2 [359] is selected as an in-domain evaluation dataset from the perspective of the VQA model ViLT [148], pre-trained on VQA-v2. TO also assess the out-of-domain capabilities of ViLT, we further experiment on Visual Genome (VG) [155].

Evaluation involves the accuracy metric. We compare the counterfactual accuracy acc^* of answers A^* stemming from perturbed questions Q^* in comparison to the initial accuracy acc.

Quantitative results As expected, all substitutions lead to ViLT accuracy degradation for both datasets. The accuracy reduction is more prevalent in *noun substitutions*, underlining the central role of nouns to the meaning of a question. By qualitatively observing response changes, we first highlight unexpected attachments on responses connected with specific color changes, such as *silver* or *gray*, while perturbing more intense colors, such as *green* or *red* leads to redefinition of the derived counterfactual response, denoting the model's capability of handling such perturbations. Furthermore, regarding adjective perturbation, the model has an appropriate understanding of common adjective noun pairs, such as "hot dog", appropriately altering its response when the respective adjective ("hot") is altered. Even though ViLT demonstrates a satisfactory ability of handling adjective contextualization, it fails when encountering rare synonyms of adjectives, thus failing to proper reason over the counterfactual setup. In the case of verbs, ViLT captures the general sense of verbs in the context of the question, presenting a few failure cases. However, as evidenced by accuracy drops for noun substitution, ViLT presents more moderate capability in capturing noun relationships, exhibiting instability in hypernym substitutions that are very broad, inclusive, and polysemous. Robust responses are observed in cases when living creatures are substituted by their hypernyms or hyponyms. In cases when nouns are substituted with a *sibling* noun (i.e. both nouns have the same parent node), mixed behaviors arise: in several cases, ViLT ignores the substitution altogether, providing an identical response in comparison to the non-counterfactual case. However, in other cases, ViLT properly reasons over the updated context.

6.2 Explainability in Text to Image generation

6.2.1 Background in Conditional Image Generation

Text-to-image generation refers to the task of synthesizing realistic or stylized images directly from natural language descriptions. It represents a significant advancement in multimodal learning, where models jointly understand and generate across different data modalities—text and vision.

The objective is to map a linguistic input (a caption, prompt, or description) to a corresponding visual representation, such that the generated image accurately captures both semantic content (what is described) and visual style (how it appears).

Early approaches to text-to-image generation were primarily grounded in **Generative Adversarial Networks (GANs)** [93], which introduced an adversarial learning paradigm consisting of two neural networks, a generator $G(z;\theta_g)$ and a discriminator The generator G learns to map noise vectors z sampled from a prior distribution $p_z(z)$ (e.g., a standard normal distribution) to the data space:

$$x_{\text{fake}} = G(z), \quad \text{where } z \sim p_z(z) \tag{6.2.1}$$

The discriminator D, on the other hand, outputs a probability p_i that a given sample x_i is real (i.e., from the true data distribution p_{data}) rather than generated:

$$p_i = D(x_i) = \text{probability that } x_i \sim p_{\text{data}} \tag{6.2.2}$$

The two networks play a two-player minimax game with the following objective function:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))] \tag{6.2.3}$$

Where:

- G aims to generate samples that fool D.
- D aims to correctly distinguish real from generated samples.
- The generator minimizes the second term, trying to make $D(G(z))$ close to 1.
- The discriminator maximizes both terms, improving its classification ability.

In the case of conditional GANs (cGANs) [228], G is fed not only with random noise z, but also with an additional conditioning vector y, which helps guide the generation of samples from specific sub-regions of the target distribution. Therefore, the generator G of a cGAN is expressed as:

$$x_{\text{fake}} = G(z, y) \tag{6.2.4}$$

The discriminator D receives both the sample and the condition:

$$D(x, y) = \text{probability that } x \sim p_{\text{data}}(\cdot \mid y) \tag{6.2.5}$$

The objective function becomes:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x,y)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z,y),y))] \tag{6.2.6}$$

Where:

- y guides generation and discrimination.
- The model learns to generate data samples conditioned on specific attributes.

Several image generation cGANs [242, 230] perform well when it comes to generating images with distinct textures and colors. However, they tend to struggle with generating coherent overall object structures and other long-range dependencies, due to the limited nature of convolutional filters. The Self-Attention GAN (SAGAN) [405] was proposed as a solution to this problem; it utilizes a self-attention module in both G and D, as well as modern stabilization techniques such as Spectral Normalization of weights [229], while it leverages the two-timescale update rule [110] to impose different learning rates for G and D. In one of the pioneering works, Reed et al. [281] demonstrated that a conditional GAN could produce low-resolution images conditioned on textual embeddings derived from captions. Subsequent developments such as StackGAN [406] extended this framework by adopting a multi-stage architecture, where a first-stage generator produced a coarse image and a second-stage network refined it to higher resolution, thereby enhancing detail and color fidelity. AttnGAN [384] further advanced the field by incorporating attention mechanisms that allowed the model to focus selectively on different words or phrases when rendering corresponding image regions. This innovation improved semantic alignment between textual descriptions and generated imagery. However, GAN-based models were often limited by issues such as training instability, mode collapse, and restricted diversity in generated samples, which constrained their scalability and robustness. StoryGAN [179] is a generative model that synthesizes images based on sequential input (Story Visualization), using an RNN structure to encode the input text and provide context information to the generator. The generator is trained adversarially against two discriminators: the image discriminator, which evaluates image quality and text-image relevance, and the story discriminator, which ensures consistency across images given the entire story context. Recent work has focused on improving the baseline StoryGAN model [169] and exploring alternative story encoding methods, such as using Transformer architectures [213, 212, 340, 253].

A significant paradigm shift occurred with the emergence of **diffusion-based models**, which replaced adversarial learning with a generative process rooted in probabilistic denoising. Diffusion models, introduced through the Denoising Diffusion Probabilistic Models (DDPMs) [111], generate images by progressively removing noise from random Gaussian samples until coherent visual structures emerge.

These models were initially inspired by nonequilibrium thermodynamics, specifically by the process of diffusion, where particles spread out over time from areas of high concentration to low concentration. The generative idea is to simulate a Markovian forward process that gradually corrupts the data with Gaussian noise over many time steps until it becomes nearly pure noise. The model then learns a reverse denoising process—effectively teaching a neural network how to recover the original data from noise, step by step. Some key concepts of diffusion models are the following:

1. Forward Process (Diffusion)

The forward process is a fixed Markov chain that gradually adds Gaussian noise to an initial data sample x_0 over T time steps. At each time step t, the distribution is defined as:

$$q(x_t \mid x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t \mathbf{I}) \tag{6.2.7}$$

Where:

- x_0 is the original data sample (e.g., an image).
- x_t is the noised version of x_0 at time step t.
- β_t is a small positive scalar controlling the variance of added noise at time t.
- $\mathbf{I}$ is the identity matrix, indicating isotropic Gaussian noise.
- $q(x_t \mid x_{t-1})$ is the conditional probability of x_t given x_{t-1} in the forward process.

The forward process produces a sequence:

$$x_0 \rightarrow x_1 \rightarrow x_2 \rightarrow \cdots \rightarrow x_T \tag{6.2.8}$$

such that x_T is nearly pure Gaussian noise.

2. Reverse Process (Generation)

The model learns to reverse the noising process via a parameterized neural network (often a U-Net) that approximates the posterior:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{6.2.9}$$

Where:

- p_θ is the learned reverse distribution.
- $\mu_\theta(x_t, t)$ is the predicted mean of x_{t-1} given x_t.
- $\Sigma_\theta(x_t, t)$ is the predicted variance (often fixed or simplified in practice).
- The process begins at $x_T \sim \mathcal{N}(0, \mathbf{I})$ and recursively samples $x_{t-1} \sim p_\theta(x_{t-1} \mid x_t)$.

This allows for sampling starting from pure noise $x_T \sim \mathcal{N}(0, \mathbf{I})$ and progressively generating cleaner data.

3. Training Objective

The model is trained by minimizing a variational lower bound or a simplified denoising score matching loss. A common training objective is:

$$\mathbb{E}_{x_0,\epsilon,t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right] \tag{6.2.10}$$

Where:

- $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is the Gaussian noise added to the sample.
- $x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1 - \bar{\alpha}_t}\, \epsilon$ is a closed-form expression of the forward process at time t.
- $\bar{\alpha}_t = \prod_{s=1}^{t}(1 - \beta_s)$ is the cumulative product of noise scheduling terms.
- $\epsilon_\theta(x_t, t)$ is the neural network's estimate of the noise ϵ at time t.

Minimizing this loss trains the network to denoise effectively, enabling sampling from pure noise.

Subsequent works, such as the Guided Diffusion model [65], introduced classifier-based guidance to steer the denoising trajectory towards desired attributes, achieving unprecedented image fidelity and controllability. Among these, the introduction of Stable Diffusion [289] marked a major leap in accessibility and performance by conducting diffusion in a latent space rather than pixel space. This latent diffusion approach reduced computational cost while maintaining high-resolution output quality, allowing image synthesis on consumer-grade hardware. These models often leverage pretrained multimodal encoders, most notably CLIP [273], to align textual and visual semantics effectively.

The role of multimodal embeddings has been central in improving text-image correspondence. CLIP established a joint embedding space where textual and visual representations are directly comparable. This representation enabled models to interpret nuanced language and match it to appropriate visual concepts. Building on this, systems such as DALLE [277, 276] and Imagen [291] achieved remarkable success in generating high-quality, semantically aligned images from complex or abstract textual prompts. These models demonstrated compositional reasoning, allowing them to combine multiple concepts (such as "a cat wearing sunglasses under a palm tree") in a single coherent image. The integration of language encoders and vision decoders thus lies at the core of current text-to-image architectures.

Despite substantial progress, evaluating text-to-image generation remains a complex and multifaceted problem. Quantitative metrics such as the Fréchet Inception Distance (FID) [109] measure visual realism by comparing feature distributions of generated and real images, while CLIPScore [108] assesses semantic alignment between the input text and generated output. More pixel-level metrics, such as Inception Score (IS) [294], Learned Perceptual Image Patch Similarity (LPIPS) [410] and other variants are widely preferred in generative evaluation, as they are able to quantify the quality of the generated samples in terms of distributional similarity. Yet, concerns have been raised that their brittleness is leading to inaccurate results [258]. Even though more recent metrics, such as CleanFID [258] can resolve some issues regarding visual artifacts, they still cannot address major issues such as the evaluation of complex images, compositionality, logic, and fairness of generation [29]. Moreover, when it comes to conditional generation, we further require a measure of whether objects and attributes mentioned in the conditioning are successfully depicted on the generated samples. Current attempts in conditional synthesis evaluation remain limited [314, 21] while still facing the shortcomings of their unconditional counterparts, which they are built upon. Overall, many image generation metrics are imperfect and often supplemented by human evaluations, particularly for assessing aesthetic quality, creativity, and coherence. As a parallel venture, explainability in generative modeling can deliver interesting insights, though current efforts either remain model-specific [19, 236, 87, 146] or require discovering interpretable latent directions [266, 305, 38, 366], which is a non-trivial task.

We argue that resolving generative evaluation challenges calls for a conceptual approach to the evaluation process, diverging from the pixel-level track. Relying on concepts instead of pixels offers the advantage of enhanced interpretability of the evaluation process, paving the way for explainable post-hoc evaluation of generative models: identifying concepts (objects or attributes) that can or cannot be generated reveals capabilities and biases of the model at hand, thus driving potential architectural modifications. For this reason, we adopt the approach of [78] in order to discover the edits needed to approach the ground truth images commencing from a corresponding generated instance. In this case, by again viewing images as sets of concepts, we pose the question of: *what needs to be conceptually inserted, deleted or substituted to step from generated concepts to the ground truth ones?* [203]

6.2.2 Conceptual edits for generative evaluation

Counterfactual explanations are capable of addressing the aforementioned "*what needs to be conceptually inserted, deleted or substituted to step from generated concepts to the ground truth ones?*" question by providing the minimum number of conceptual edits to achieve the $S \rightarrow T$ transition, where S serves as the source (generated) image and T as the target (ground truth) one. The ultimate goal is to "approach" the ground truth concepts commencing from the generated ones; in other words, we correct generated semantics via edits that lead to the correct semantics based on ground truth.

Concept distances instruct the shortest path that connects two specific concepts. Concept hierarchies are employed, which deterministically define the transition cost between concepts. We both explore the options to use external hierarchical knowledge sources, such as WordNet [226], mapping extracted concepts to WordNet synsets, or alternatively handcraft specific hierarchies to allow highly controlled semantic distance definition. In any case, we denote as $d(s,t)$ the distance between concepts s and t.

There are three available concept **edit operations** to realize transitions:

- **Replacement (R)** $e_{s \rightarrow t}(S)$: A concept $s \in S$ is replaced with a concept $t \notin S$.
- **Deletion (D)** $e_{s-}(S)$: A concept $s \in S$ is deleted from S.
- **Insertion (I)** $e_{s+}(S)$: A concept $s \in S$ is inserted in S.

Each edit operation inherits the concept distances imposed by the selected hierarchy. Therefore, **R** operation considers the path between s and t so that $min(d(s,t))$ is ensured. **D** and **I** operations regard the *root node* of the hierarchy as t and s respectively; in the case of WordNet, `entity.n.01` serves as the root. **Concept Set Edit Distance (CSED)** $D(S \rightarrow T)$ is obtained by aggregating all possible minimum cost edit operations so that $S \rightarrow T$ is finally achieved:

$$CSED = D(S \rightarrow T) = min \sum_{s \neq t}^{S,T} \sum^{R,D,I} d(s,t) \tag{6.2.11}$$

6.2.3 Method

The heart of our method consists of a pre-trained black-box generative model M which receives a semantic description c (in natural language or in symbolic format) as conditioning and produces an image I corresponding to c. Concepts extracted from I form the *generated* or *source* concept set S, containing discrete features or objects that belong to the generated images as concepts. Their extraction is based on off-the-self automatic methods such as object detection, semantic segmentation and others. Similarly, concepts extracted

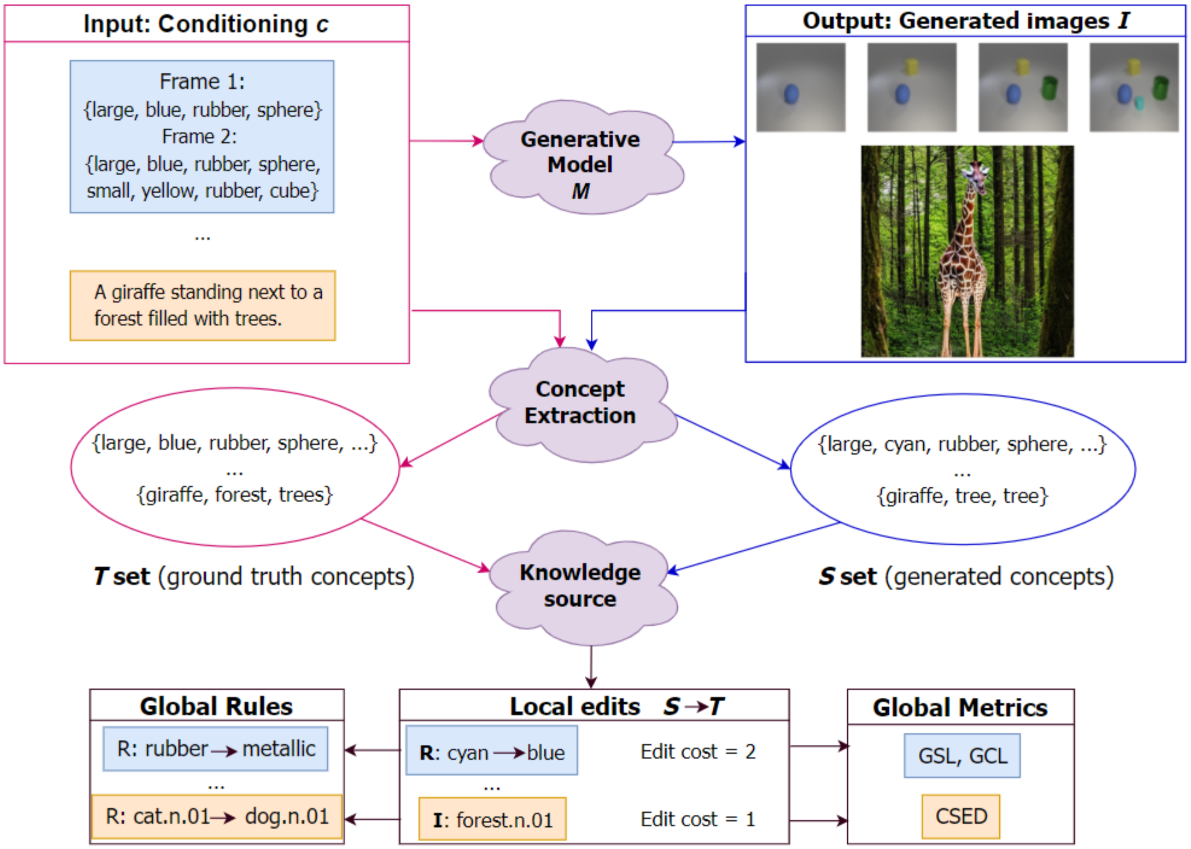


Figure 6.2.1: Outline of the proposed semantic edits framework for image generation evaluation. [203]

from c contribute to the *real* or *target* concept set T. The format of c defines the concept extraction technique to be followed, ranging from linguistic concept extraction, if c is a textual sentence, to simple preprocessing, if c is already in set format. Ultimately, we aspire to answer the following: "What has to *minimally* change in order to transit from S to T?" The outline of our method is presented in Figure 6.2.1.

6.2.4 Generative evaluation

The counterfactual backbone described in Section 6.2.2 highlights the steps to be taken for generative evaluation. We therefore present its adaptation on two difficult tasks of the generative literature: **Story Visualization** and **Scene Generation**.

Story Visualization (SV) targets the sequential creation of images $I_1, I_2, ..., I_L$ that correspond to a given sequential conditioning $c_1, c_2, ..., c_L$ one to one. The generated images need not only to remain *faithful* to their conditioning, but also maintain serial *consistency*. We therefore define the two desiderata applicable to SV:

- **Faithfulness**: objects and attributes mentioned in c_k should also appear in frame I_k.
- **Consistency**: objects or attributes appearing in frame I_k cannot disappear or change in later frames $I_{k+1}, ..., I_L$.

Since well-defined semantics are tied to counterfactual explanations [32], we regard CLEVR-SV [139] as the ideal dataset to demonstrate our approach, as it provides a set of concepts $\mathcal{C}$: shape *(cube, sphere, cylinder)*, size *(small, large)*, material *(rubber, metal)* and one of 8 colors *(blue, cyan, brown, yellow, red, green, purple, gray)*. Each CLEVR-SV object contains $|\mathcal{C}|$=4 concepts that describe its shape, size, material and color. We handcraft a simple hierarchy to group object semantics to generic concept classes, demonstrating the following inclusion relationships:

$$\begin{aligned}\textit{(large, small)} &\subset \textit{Size}\\ \textit{(blue, yellow, brown, grey, green, purple, cyan, red)} &\subset \textit{Color}\\ \textit{(metallic, rubber)} &\subset \textit{Material}\\ \textit{(sphere, cube, cylinder)} &\subset \textit{Shape}\end{aligned} \tag{6.2.12}$$

CLEVR-SV contains stories of length L=4, with the k-th frame strictly containing k objects. Any of the three available edit operations can be relevant per frame: **D** of a concept, when a generated frame contains more objects than its ground truth match; **I** of a concept in the opposite case; **R** equals to a **D** followed by an **I**, and can be applied on frames with proper number of objects when semantics differ. In the default case, we assign equal costs of 1 for all semantics, as well as for **D** and **I** operations (**R** cost is the sum of **D** and **I** costs).

To measure story **faithfulness** we propose the ***Story Loss (SL)*** metric, which sums up the per-frame Concept Set Edit Distance ($CSED_k$) for $k = 1, 2, ..., L$ frames of the story. Generated CLEVR-SV semantics for shape, size, material and color for the k-th frame form the concepts set S_k, while the semantics of the conditioning form T_k, with their $CSED_k$ denoted as $D(S_k, T_k)$. Thus, the cost for the transition $\{S_1, S_2, S_3, S_4\} \rightarrow \{T_1, T_2, T_3, T_4\}$ corresponding to the minimum cost **R D I** edits needed to transform the semantics of the generated sequence $\{I_1, I_2, I_3, I_4\}$ to the semantics of its conditioning $\{c_1, c_2, c_3, c_4\}$ can be expressed as:

$$SL = \sum_{k=0}^{k=L} CSED_k = \sum_{k=0}^{k=L} D(S_k, T_k), \qquad L = 4 \tag{6.2.13}$$

By scaling up the calculation of SL for a dataset containing N stories, we obtain the ***Global Story Loss (GSL)*** metric:

$$GSL = \sum_{i=0}^{N} SL_i = \sum_{i=0}^{N} \sum_{k=1}^{k=L} D(S_k, T_k)_i \tag{6.2.14}$$

As for story **consistency**, we propose the metric of ***Consistency Loss (CL)***: the frame I_k is compared with $I_{k-1}, k = 2, 3, 4$ frames of the generated sequences to capture changes of semantics. A challenging aspect of CL is that there does not actually exist a ground truth concept set. However, since it is known by task definition that the k-th frame contains k objects, and the cardinality $|\mathcal{C}|$ of dataset concepts is predefined ($|\mathcal{C}|$=4 in the case of CLEVR-SV) we can assume that every previous frame constitutes the 'ground truth' corresponding to the concept set T. By commencing from the k=1 frame, we expect the cardinality of T to be equal with $|\mathcal{C}| \cdot k = |\mathcal{C}|$. Any discrepancy results in a penalty $p_k = |T| - |\mathcal{C}| \cdot k = CL_k$ for k=1. For later frames, we define as S the concept set corresponding to the k-th frame, and as T the 'ground truth' set comprised of the k-1 frame concepts. Mathematically, CL can be written as:

$$CL = p_{k=1} + \sum_{k=2}^{k=L} D(S_k, T_k), \qquad T_k = S_{k-1}, L = 4 \tag{6.2.15}$$

In the ideal case, when the k-th frame contains k objects with $\mathcal{C}$ semantics, we expect that $p_{k=1} = 0$ and $CL_{k>1} = |\mathcal{C}| \cdot (k - 1)$. By extending CL to N stories, ***Global Consistency Loss (GSL)*** evaluates the consistency capabilities of a generative model M in total:

$$GCL = \sum_{i=0}^{N} CL_i = \sum_{i=0}^{N} \{p_{k=1,i} + \sum_{k=2}^{k=L} D(S_k, T_k)_i, \;\; T_{k,i} = S_{k-1,i}\} \tag{6.2.16}$$

Average values can be obtained for both local (SL/CL) and global (GSL/GCL) metrics:

$$Avg\ SL = \frac{1}{k} SL, \qquad Avg\ GSL = \frac{1}{N}[Avg\ SL] = \frac{1}{N} GSL \tag{6.2.17}$$

For consistency, instead of exporting an average value over $\sum CL_k$, it is more meaningful to count how many times the $p_{k=1} = 0, CL_{k>1} = |\mathcal{C}| \cdot (k - 1)$ requirement was not respected, averaged for $L = k$ frames:

$$Avg\ CL = \frac{p_{k=1}}{k} + \frac{1}{k} \sum_{k=1}^{k=L} [CL_{k>1} \neq |\mathcal{C}| \cdot (k - 1)], \quad Avg\ GCL = \frac{1}{N}[Avg\ CL] \tag{6.2.18}$$

SL and CL are by nature *explanaible*, as they do not only provide a measure of quality, but also reveal the $S_k \rightarrow T_k$ edit paths. Those paths serve as *local counterfactual explanations*, highlighting the erroneously generated semantics for this particular story, either in terms of **faithfulness** or **consistency**. Overall, higher SL/GSL and CL/GSL values denote lower conceptual generation quality. GSL/GCL edit paths correspond to *global counterfactual explanations*: rule extraction techniques provide frequent patterns, summarizing the behavior of the generative model M under investigation. Frequent GSL edit paths in fact contain *common misconceptions*, i.e. conditioning concepts that M cannot easily generate. In a similar sense, GCL edit paths reveal frequent *inconsistency patterns*, showcasing concepts that arbitrarily change within the story frames. Therefore, by researching the question "What has to *minimally* change in order to transit from S to T?", we eventually answer a more generic one: "Which concepts cannot be generated or preserved by M?"

Scene Generation (SG) aims to synthesize a visual scene I based on a conditioning c. The synthesized image comprises multiple objects which interact with each other. Scene objects are also accompanied by attributes. The given conditioning c is more complex compared to conditionings provided for SV, since the concepts to be generated are numerous and not predefined; this yields a concepts set $\mathcal{C}$ of unknown but comparatively large cardinality.

COCO dataset [183] provides the ideal setting for evaluating generative **faithfulness** for SG, providing textual captions c that can serve as conditioning. We focus our endeavors on on state-of-the-art open source diffusion models [289] from Huggingface[1], and specifically on Stable Diffusion v1.4 & v2 [319, 318] and Protogen x3.4 & 5.8 [268, 269]. These models produce *realistic* images - an important aspect for the concept extraction (object detection) stage-. We omit older SG architectures [257, 402, 209, 177, 326] due to their inferior visual quality and their reliance on scene graphs and layouts for ensuring proper composition.

In the concept extraction stage, YOLO-v8 [138] and YOLOS [75] object detectors are leveraged to construct the generated concept set S. Since c is in textual format, spaCy [127] is used to extract *ground truth concepts* from captions that form the target concept set T. The semantically complex nature of concept distances related to COCO concepts requires a rich knowledge scheme, such as WordNet. For example, if c refers to concepts such as 'food' or 'animal', a diffusion model may generate more refined 'food' or 'animal' instances, for example 'pasta' and 'dog' respectively. The object detectors will then return these refined classes, inducing some noise in the transformation process. Hierarchical knowledge can eliminate such issues: even though $T = \{food, animal\} \neq S = \{pasta, dog\}$, the two sets are **semantically equivalent** if we consider the hierarchical relationships $pasta - isA - food$ and $dog - isA - animal$ provided by mapping S and T concepts on WordNet [226] synsets. In this case, no $S \rightarrow T$ transformation needs to be performed. Therefore, the usage of external knowledge allows more conceptually accurate transitions. Moreover, WordNet provides concept distances necessary for edit operations, precisely reflecting semantic relationships between concepts. Then, CSED can be applied to provide the total cost of the $S \rightarrow T$ transformations.

6.2.5 Experiments on Story Visualization

Since all semantics and **D, I** edit operations have an equal cost, we assign d=1 for all semantics, as well as for **D, I**. For example, deleting a color yields an edit cost of 1. Alternatively, by substituting a color with another one induces an edit cost of 2, equal to deleting the source color and then inserting the target color. The same logic applies to shape, size and material of objects.

6.2.5.1 Quantitative results

In Table 6.1, we present results using typical image generation metrics (FID, Clean-FID, LPIPS, SSIM) over the best variants of the selected SV models [179, 213, 212, 340] pre-trained on CLEVR-SV. Table 6.2 reports conceptual metrics

In general, we observe an agreement between pixel-level and conceptual metrics. This is somehow expected, since the concept extraction stage depends on pixel-level image quality, with better generated objects or semantics being more easily identifiable. Nevertheless, conceptual evaluation offers more *explainable insights*: percentages of losses per concept (Material, Size, Shape, Color) are provided, highlighting strengths and shortcomings of investigated models over different semantics. For example, higher Shape loss for all models

---

[1] https://huggingface.co/models?pipeline_tag=text-to-image&sort=downloads

($>$ 50%), indicates that they synthesize objects of ambiguous shapes in most cases. On the other hand, relatively lower Size losses reveal the models' capability to generate objects having the right size.

M	FID ↓	Clean-FID ↓	LPIPS ↓	SSIM ↑
[340]	41.54 ± 8.55	**115.46**	**0.21** ± 0.05	**0.71**
[179]	**41.45** ± 6.25	123.40	0.25 ± 0.03	0.65
[212]	41.96 ± 9.66	124.97	0.25 ± 0.08	0.67
[213]	41.80 ± 8.81	122.62	0.25 ± 0.05	0.68

Table 6.1: Averaged existing evaluation metrics for all L=4 stories per M.

M	GCL ↓	GSL ↓	Material ↓	Size ↓	Shape ↓	Color ↓
[340]	**4.97**	**7.01**	**20.83%**	**14.55%**	**56.62%**	**33.10%**
[179]	11.44	15.33	30.89%	21.12%	62.34%	37.44%
[212]	10.95	8.06	21.45%	16.02%	56.78%	35.10%
[213]	8.32	11.51	25.34%	16.71%	56.83%	35.14%

Table 6.2: Averaged proposed evaluation metrics for all L=4 stories per M.

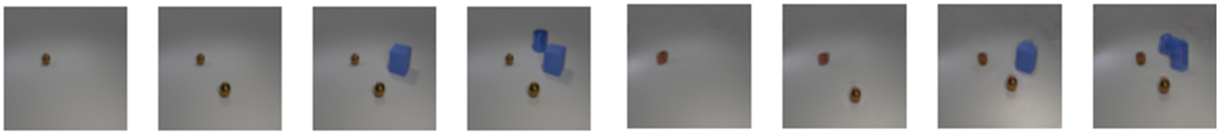
We further investigate our findings by focusing on the best performing SV model of [340] according to the conceptual metrics reported in Table 6.2. Specifically, in Table 6.3 we present results of per frame GSL, GCL and losses per semantic.

Frame	GSC ↓	GSL ↓	Material ↓	Size ↓	Shape ↓	Color ↓
1st	0.00	2.25	40.00%	6.20%	58.75%	7.50%
2nd	4.35	5.66	20.00%	11.88%	57.5%	32.50%
3rd	7.12	8.25	13.33%	16.67%	57.08%	43.33%
4th	8.42	11.49	10.00%	23.44%	53.13%	49.06%

Table 6.3: Average conceptual evaluation metrics per frame for [340].

6.2.5.2 Local explanations

The transparency of the proposed SL/CL metrics is verified by obtaining *local explanations* for [340]. Specifically, we examine edit paths for the sequences of Figure 6.2.2: the 4 leftmost images (Figure **??**) correspond to the ground truth sequence, while the 4 rightmost images (Figure **??**) denote the generated frames. Consequently, S contains concepts of **??** and T contains concepts of **??**. As presented in Table 6.4, a standard **R** operation for all frames is observed, suggesting transforming the material of the small brown sphere from *'rubber'* to *'metallic'* in order to match the ground truth. In the last frame, one more **R** operation is added, suggesting also transforming the shape of the new object from *'sphere'* to *'cylinder'*. The cost for each **R** operation equals to 2, equivalent for one step to remove the wrong semantic and one more step to add the right semantic. However, this cost weight can be tuned appropriately, if needed. SL for this story equals to 10, as a summary of all operation costs per frame. By observing for CL, we realize that the correct number of objects is added in every consequent frame, so that $\text{CL}_{k>1} = |\mathcal{C}| \cdot (k-1), |\mathcal{C}| = 4$ is maintained: starting from $\text{CL}_1 = p_{k=1} = 0$ for the k=1 frame, we verify that only one object is added, respecting that frame number should be equal to the number of objects present in it. CL_2=4 is expected, since the object added in the k=2 frame contains 4 semantics. Any number lower or grater than that would indicate an abnormal behavior: $\text{CL}_{k>1} < |\mathcal{C}| \cdot (k-1)$ marks one (or more) missing objects, while $\text{CL}_{k>1} > |\mathcal{C}| \cdot (k-1)$ indicates one (or more) extra object generated. The desired pattern repeats for the 3rd and 4th frame. Throughout this analysis, the shortcomings of [340] concerning this specific image are revealed, producing a *local* explanation: The semantic *Material* needs to be examined more, as in all story frames of this example the small brown sphere is generated with the attribute *'rubber'* instead of *'metallic'*. In order to obtain insights regarding the model's synthesis capabilities of discrete semantics, global metrics and explanations need to be derived.



(a) Ground truth story frames (b) Generated story frames of [340]

Figure 6.2.2: Ground truth vs generated story frames using [340] for L=4. [203]

Frame	Min edit path	Operation	Edit cost	Semantic	CL
1st	’rubber’ →’metallic’	R	2	Material	0
2nd	’rubber’ →’metallic’	R	2	Material	4
3rd	’rubber’ →’metallic’	R	2	Material	8
4th	{’rubber’, ’sphere’} → {’metallic’, ’cylinder’}	R, R	4	Material, Shape	12

Table 6.4: Interpretable local edits for Figure. 6.2.2

Rules (edits)	Semantic	Support %	Antecedent support%	Consequent support%
’metallic’ →’rubber’	Material	26.77	26.77	26.77
’rubber’ →’metallic’	Material	22.05	22.05	22.83
’cylinder’ →’cube’	Shape	18.11	33.07	31.50
’cylinder’ →’sphere’	Shape	14.96	33.07	18.90

Table 6.5: Interpretable global edits on test set images of CLEVR-SV generated using [340].

6.2.5.3 Global explanations

In order to assess our model’s shortcomings in total, we measure GSL for all test images of CLEVR-SV. Therefore, we can obtain a measure of the model’s inability to capture certain -discrete- semantics, either per frame or in total (Table 6.2). We observe that in later frames, *Material loss* decreases, even though we would expect that the problem gets harder and harder as more objects are added, resulting in higher losses. This expected pattern is followed by *Size loss* and *Color loss*, while no certain pattern can be extracted from *Shape loss*. The high *Shape loss* verifies the need for attention mechanisms within the used GANs, so that long-range relationships can be captured. We can also attribute the rapid rise of *Size* and *Color* losses to consistency deficiencies within the story sequence.

GSL can also reveal patterns in the form of rules for the whole test set. We leverage apriori algorithm [4] to extract frequent semantic combinations and rules. The 4 most common semantic edits are provided in Table 6.5, together with each rule’s frequency (support). The concept category (as occurring from equation 6.2.12), antecedent support (source semantic frequency) and consequent support (target semantic frequency) are also provided.

We observe that *Material* is the most common concept misconception, with both ’rubber’ and ’metallic’ semantics being frequently confused. *Shape* is the second most prominent misconception, with ’cylinder’ appearing in the generated frames more often compared to the ’cylinder’ occurrence in the conditioning; ’cube’ and ’sphere’ shapes are sacrificed for ’cylinder’ to be generated. Since the rule support is not significantly high, with 26.77% being the maximum value, we can safely assume that the SV model of [340] is not heavily biased towards certain semantics. Nevertheless, we spot some tendency to generate wrong material and shape, an observation that can be valuable towards architectural improvements of the model.

6.2.6 Experiments on Scene Generation

We select the first 10K samples from COCO to reduce the inference time needed to extract visual concepts using YOLO-v8 and YOLOS object detectors. COCO provides 5 descriptive sentences per sample, which are paraphrases of each other. For this reason, we only regard the 1st out of the 5 sentences as the conditioning c. We follow two separate processes for SG: actual *generation* conditioned on c and *retrieval* of caption-image pairs based on captions similar to c.

6.2.6.1 Conditional generation on COCO captions

For the generation experiment, we employ pre-trained diffusion models without any further tuning, as mentioned in 6.2.4, which are all tested on the same conditionings c. Each of the four diffusion models required about 15 hours to synthesize 10K images using 2 T4 GPUs, therefore around 60 hours in total.

6.2.6.2 Retrieval of COCO-related captions

In order to obtain considerably more images conditioned on COCO-related queries without having to spend the time and resources to run many more thousand iterations of the diffusion model, we utilized a Stable Diffusion search engine (Lexica.art)[2]. The exact process we used was the following: we use c of the first 10K COCO samples as the 'query' caption. The search engine returned, for each of the 10k captions, 10 images that have been already generated by online communities with the closest input queries to our captions. This technique supplied us with 100.000 more Stable Diffusion images, accompanied by their input queries. We then compare results between web-retrieved and generated images.

6.2.6.3 Object detection

We select a default threshold of T_d =0.6 for detection; objects detected with confidence$\geq$0.6 are added in the generated concept set S. This threshold is experimentally defined to maintain a valid trade-off between false positive and false negative objects; in fact, since no ground truth exists, even defining false predictions is untractable without human inspection. However, our approach can provide relevant hints regarding the probability of false detection, as a higher number of **D** operations may infer higher false positive rates (irrelevant objects being detected, if T_d is too low), while more **I** operations can be correlated with higher false negative rates (relevant objects not being detected, if T_d is too strict).

6.2.6.4 Quantitative results

For comparative reasons we present results for T_d =0.5, 0.6, 0.7 in Tables 6.6 (YOLO-v8) & 6.7 (YOLOS) for generated images, and in Table 6.8 for web images, reporting object extraction from both object detectors. Instances colored in blue denote the lowest scores, which are more desirable, while the highest scores are highlighted with red. We present number of edits (# **I**, # **D**, # **R**), as well as the total cost for each **I**, **D**, **R** operation for all images. Mean CSED is reported as an overall metric regardless of which operation was performed more often.

T_d	M	# **I**	Cost **I**	# **D**	Cost **D**	# **R**	Cost **R**	Mean CSED
0.5	stable diffusion	37651	16762	1196	5655	126004	14323	35.75
	stable diffusion 2	36878	16067	1243	6301	129315	14839	36.32
	protogen base	37072	16208	1233	5944	129290	14744	35.95
	protogen 5.8	38581	17715	1195	4702	117708	13411	34.66
0.6	stable diffusion	39070	18386	1157	4042	110260	12964	34.22
	stable diffusion 2	38678	17782	1200	4514	112499	13397	34.55
	protogen base	38548	17794	1184	4270	114762	13427	34.35
	protogen 5.8	39766	19210	1134	3419	103579	12135	33.38
0.7	stable diffusion	40814	20391	1086	2681	93390	11337	32.96
	stable diffusion 2	40677	19806	1107	2938	95477	11756	33.08
	protogen base	40397	19801	1101	2820	97314	11787	32.94
	protogen 5.8	41295	20944	1039	2308	89850	10726	32.39

Table 6.6: Metric results using YOLO-v8 for object detection on generated images from COCO queries.

Regarding the selected threshold T_d, our initial hypothesis is proven to be correct: more **I** operations are realized for higher threshold T_d=0.7, suggesting that objects from the conditioning where not detected, while fewer **I** were performed for T_d=0.5. Similarly, there are more **D** operations for the lowest T_d=0.5, as spurious objects can be detected more easily. Additionally, more **R** operations are needed for lower thresholds, which is also expected, since more objects are extracted and added to the S set. As for object detectors, results using

---

[2] https://lexica.art/

T_d	M	# **I**	Cost **I**	# **D**	Cost **D**	# **R**	Cost **R**	Mean CSED
0.5	stable diffusion	26302	9032	1382	44189	197623	21097	68.25
	stable diffusion 2	26684	8832	1403	43459	192198	21082	68.05
	protogen base	26887	8966	1404	44406	193327	21035	68.81
	protogen 5.8	28880	10367	1373	34996	189677	19858	60.45
0.6	stable diffusion	27963	9920	1373	33891	188395	20286	60.10
	stable diffusion 2	28145	9662	1394	33933	182767	20322	60.36
	protogen base	28499	9845	1394	34167	185217	20224	60.63
	protogen 5.8	30545	11330	1364	27218	179947	18963	54.13
0.7	stable diffusion	29998	10985	1357	24956	177213	19319	52.51
	stable diffusion 2	29831	10657	1347	25492	172860	19409	53.14
	protogen base	29866	10790	1346	25255	175495	19350	52.98
	protogen 5.8	28880	10367	1373	34996	189677	19858	60.45

Table 6.7: Metric results using YOLOS for object detection on generated images from COCO queries.

T_d	Obj. detector	# **I**	Cost **I**	# **D**	Cost **D**	# **R**	Cost **R**	Mean CSED
0.5	YOLO-v8	186775	857448	1343	52247	1353479	224350	75.87
	YOLOS	163628	605321	1487	421525	2469635	473331	106.41
0.6	YOLO-v8	190047	891454	1317	37418	1174012	190928	73.74
	YOLOS	167576	646112	1467	308346	2303966	432851	98.06
0.7	YOLO-v8	193663	929183	1236	25388	982259	154063	71.81
	YOLOS	171778	688942	1449	214928	2115779	390304	90.56

Table 6.8: Metric results for web-retrieved Stable Diffusion images on similar queries to COCO.

YOLO-v8 are very homogeneous, indicating that the models under investigation follow a rather predictable behavior irrespectively of T_d. Protogen 5.8 consistently yields the lowest mean CSED score, denoting cheaper transitions for all thresholds. This observation slightly changes for T_d=0.7 and YOLOS object detector (Table 6.7), for which, surprisingly, protogen 5.8 produces the more expensive transitions. By comparing Tables 6.6 & 6.7, YOLOS results in higher mean CSED, less **I** operations, significantly more expensive **D** operations (even though the number of **D** operations is not substantially larger), as well as more and expensive **R** operations. Therefore, we can safely assume that YOLOS is comparatively more sensitive in detecting more objects, which may induce some noise in the detection process. All these results will become more interpretable should we delve into the explanations accompanying the evaluation. The patterns arising from evaluating generated images are also supported in Table 6.8 findings, verifying the threshold hypothesis, as well as the increased sensitivity of YOLOS. Nevertheless, web-retrieved images seem to miss objects mentioned in the query, as proven by the large number of **I** and **R** operations.

6.2.6.5 Local explanations

provide edit paths based on the **I**, **D**, **R** operations realized for a specific generated image. For this reason, we employ a scene depicted in Figure 6.2.3.

According to YOLO-v8 with the default threshold T_d=0.6, the generated concepts are S={'car', 'car', 'traffic light', 'car', 'stop sign'}, and ground truth concepts are T={'light', 'buildings'}. The edit operations of total minimum cost 59.00 for this $S \rightarrow T$ transformation are:

I: { }

D: {'car', 'car', 'car'}

R: {'traffic light'→'light', 'stop sign'→'buildings'}

When using YOLOS, the generated concepts are S={'car', 'traffic light', 'car', 'stop sign', 'traffic light', 'car', 'traffic light', 'traffic light', 'traffic light', 'traffic light', 'traffic light', 'traffic light', 'car', 'traffic light', 'traffic light', 'traffic light', 'traffic light', 'traffic light', 'car', 'traffic light', 'traffic light', 'traffic light', 'traffic light', 'car', 'car', 'traffic light', 'traffic light'}, and the ground truth ones are T={'light', 'buildings'}. By

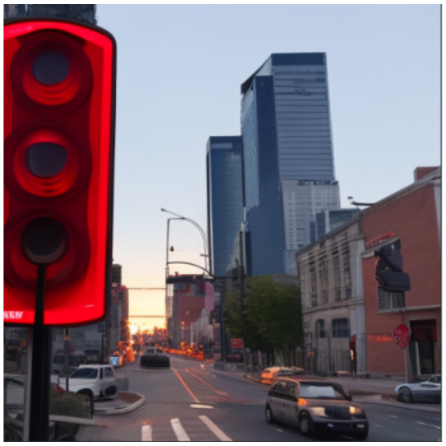


Figure 6.2.3: An image sample generated by Stable Diffusion 2 to extract local explanations. [203]

visually inspecting the image, YOLOS clearly overestimates the actual objects present, inducing noise in the generated concept set S. Nevertheless, our evaluation strategy successfully captures this overestimation, by suggesting the deletion of multiple concepts. Specifically, we obtain the following transformations of total cost 104.04:

I: { }

D: {'car', 'traffic light', 'car', 'traffic light', 'car', 'traffic light', 'traffic light', 'traffic light', 'traffic light', 'traffic light', 'traffic light', 'car', 'traffic light', 'traffic light', 'traffic light', 'traffic light', 'car', 'traffic light', 'traffic light', 'traffic light', 'traffic light', 'car', 'car', 'traffic light', 'traffic light'}

R: {'stop sign'→'light', 'traffic light'→'buildings'}

6.2.6.6 Global explanations

for all images are presented in Table 6.9 for **I**, **D** edits and Table 6.10 for **R** edits. Results only involve YOLO-v8 extracted concepts, as YOLOS results in an overwhelming number of detected instances. Top-3 results are demonstrated, i.e. the 3 most frequent insertion, deletions and replacements. **I** and **D** refers to concepts inserted or deleted respectively, while Freq **I**, **D** denotes how many times a specific concepts was inserted or deleted within all images. **I**, **D** support indicates the frequency a specific edit happens among all **I**, **D** edits respectively. As for **R**, support denotes the frequency of a transformation rule among all produced rules.

We can observe an obvious agreement between models; **I** edits include 'street', 'tennis' and 'table' concepts. It seems that the selected M cannot efficiently generate the **I** concepts, or generated concepts are of low visual quality, so that their detection is not feasible with T_d=0.5, 0.6, 0.7. **D** edits mainly contain 'person', 'sheep', 'car', 'umbrella', 'donut' concepts, indicating some bias towards generating spurious instances of those concept categories. Finally, **R** edits refer to transforming 'person' to 'people', 'man' or 'woman'. Since 'person' is a YOLO category incorporating both genders, such transformations are somehow expected.

6.2.7 Key takeaways on explainable image generation

The aforementioned findings verify the initial intuition of utilizing conceptual approaches for generative evaluation over pixel-level metrics, extending the idea of explainable evaluation initially proposed for semantic similarity [204]. The explainable metrics proposed are highly capable of breaking down the problem of performance evaluation in terms that are meaningful to humans. Therefore, humans acquire a solid understanding of "*What needs to be changed*" in order to transit from a potentially erroneous generated instance to the desired ground truth one, which can be further decomposed to what concepts need to be inserted, deleted or replaced to achieve this transition. By aggregating such edits, we also get a global view of the model in terms of what are the problematic concepts from the model's viewpoint; thus, apart from evaluation, we better comprehend the model's internal perception of semantics, which of them are not generated as often

T_d	M	**I**	Freq **I**	**I** support	**D**	Freq **D**	**D** support
0.5	stable diffusion	street	264	1.57%	person	2075	36.69%
		table	250	1.49%	sheep	363	6.42%
		tennis	247	1.47%	car	252	4.46%
	stable diffusion 2	tennis	253	1.57%	person	2177	34.55%
		street	242	1.51%	sheep	466	7.40%
		table	237	1.48%	car	313	4.97%
	protogen base	tennis	247	1.52%	person	2281	38.37%
		street	244	1.51%	sheep	317	5.33%
		table	229	1.41%	car	311	5.23%
	protogen 5.8	table	270	1.52%	person	1564	33.26%
		tennis	265	1.50%	car	261	5.55%
		street	241	1.36%	umbrella	251	5.34%
0.6	stable diffusion	street	290	1.58%	person	1572	38.89%
		table	281	1.53%	sheep	311	7.69%
		tennis	259	1.41%	car	158	3.91%
	stable diffusion 2	table	274	1.54%	person	1656	36.69%
		street	269	1.51%	sheep	376	8.33%
		tennis	264	1.48%	car	203	4.50%
	protogen base	street	268	1.51%	person	1717	40.21%
		table	261	1.47%	sheep	254	5.95%
		tennis	255	1.43%	car	197	4.61%
	protogen 5.8	table	303	1.58%	person	1220	35.68%
		tennis	278	1.45%	sheep	198	5.79%
		street	274	1.43%	umbrella	176	5.15%
0.7	stable diffusion	table	322	1.58%	person	1075	40.10%
		street	316	1.55%	sheep	254	9.47%
		tennis	268	1.31%	donut	122	4.55%
	stable diffusion 2	table	313	1.58%	person	1134	38.60%
		street	301	1.52%	sheep	291	9.90%
		tennis	267	1.35%	donut	111	3.78%
	protogen base	street	300	1.52%	person	1189	42.16%
		table	289	1.46%	sheep	188	6.67%
		tennis	262	1.32%	umbrella	143	5.07%
	protogen 5.8	table	330	1.58%	person	884	38.30%
		street	299	1.43%	sheep	152	6.59%
		tennis	287	1.37%	umbrella	130	5.63%

Table 6.9: Global explanations (**I** and **D** edits) for YOLO-v8 extracted concepts.

are they should, or on the contrary, which concepts are overgenerated. Such biases-either from the side of the pre-training dataset or from the model itself-pinpoint the main direction of improvement to transit towards more unbiased conditional image generators, potentially be re-training or by employing certain architectural modification. The plug-and-play nature of the semantic edits framework, suited for generative evaluation is able to serve future generative models in a black-box, post-hoc manner, rendering this approach generalizable and easy to adopt in practice.

6.3 Explainability in image captioning hallucinations

6.3.1 Background in Image Captioning

Image captioning is a fundamental task in multimodal artificial intelligence that involves generating natural language descriptions for images. It lies at the intersection of computer vision and natural language processing, requiring models to both perceive visual content and express it linguistically in a coherent and semantically accurate manner. The task poses a dual challenge: accurately identifying objects, attributes, and spatial relationships within an image, and articulating these observations as fluent, contextually appropriate text. As such, image captioning serves as a benchmark for assessing a model's capacity for visual

T_d	M	**R**	Freq **R**	**R** support	M	**R**	Freq **R**	**R** support
0.5	stable diffusion	person → man	1090	7.61%	stable diffusion 2	person → man	1115	7.51%
		person → people	520	3.63%		person → people	551	3.71%
		person → woman	499	3.48%		person → woman	511	3.44%
	protogen base	person → man	1101	7.47%	protogen 5.8	person → man	1061	7.91%
		person → people	507	3.44%		person → woman	476	3.55%
		person → woman	500	3.39%		person → people	441	3.29%
0.6	stable diffusion	person → man	1065	8.22%	stable diffusion 2	person → man	1087	8.11%
		person → people	503	3.88%		person →people	536	4.00%
		person → woman	481	3.71%		person → woman	482	3.60%
	protogen base	person → man	1080	8.04%	protogen 5.8	person → man	1035	8.53%
		person → people	494	3.68%		person → woman	449	3.70%
		person → woman	485	3.61%		person → people	431	3.55%
0.7	stable diffusion	person → man	1022	9.01%	stable diffusion 2	person → man	1033	8.79%
		person → people	473	4.17%		person → people	508	4.32%
		person → woman	458	4.04%		person → woman	441	3.75%
	protogen base	person → man	1054	8.94%	protogen 5.8	person → man	989	9.22%
		person → woman	461	3.91%		person → woman	419	3.91%
		person → people	446	3.78%		person → people	408	3.80%

Table 6.10: Global explanations (**R** edits) for YOLO-v8 extracted concepts.

understanding, language generation, and cross-modal reasoning.

Early approaches to image captioning were dominated by template-based and retrieval-based methods. Template-based systems constructed captions by filling pre-defined sentence templates with detected objects or actions [161]. While these approaches ensured grammatical correctness, they lacked flexibility and creativity, often producing rigid or repetitive sentences. Retrieval-based methods, in contrast, searched large annotated datasets for images visually similar to the query and reused their corresponding captions [246]. Although such models produced fluent sentences, they failed to generalize to novel scenes or unseen compositions. The field underwent a paradigm shift with the advent of deep learning, which enabled end-to-end training of vision-language models. The introduction of the encoder–decoder architecture marked a major milestone, comprising a CNN and an RNN as an encoder and decoder respectively. This model, often referred to as "Show and Tell," significantly improved caption fluency and relevance by learning a joint visual-linguistic representation rather than relying on handcrafted templates [352].

Building upon this foundation, the "Show, Attend and Tell" model [382] introduced an attention mechanism that allowed the captioning system to focus on different image regions when generating each word. This innovation enhanced interpretability and improved descriptive accuracy, particularly for complex scenes containing multiple objects. Attention-based architectures laid the groundwork for more sophisticated models capable of dynamic visual reasoning. Later developments incorporated bottom-up and top-down attention mechanisms [10], combining object-level feature extraction with sentence-level contextual reasoning to generate more fine-grained and contextually appropriate captions.

The recent emergence of transformer-based architectures has further transformed image captioning. Models such as OSCAR [176], VinVL [409], and BLIP [173] leverage pre-trained multimodal transformers that integrate visual and textual information through large-scale self-supervised learning. These models benefit from vision-language pretraining (VLP), where aligned image–text pairs are used to learn joint embeddings capable of capturing rich semantic associations. Unlike earlier models that relied on limited supervised datasets, transformer-based systems scale effectively across large corpora, achieving state-of-the-art results in both caption quality and generalization. Additionally, contrastive learning frameworks such as CLIP [273] have influenced captioning research by providing robust multimodal representations that align visual and linguistic modalities in a shared semantic space.

Evaluating image captioning systems poses unique challenges due to the inherently subjective nature of language. Automatic metrics such as BLEU [254], METEOR [15], ROUGE [181], CIDEr [350], and SPICE [9] are commonly used to measure n-gram overlap or semantic similarity between generated captions and human references. However, these metrics often fail to capture nuanced aspects such as factual accuracy, creativity, or contextual appropriateness, prompting the continued reliance on human evaluations for qualitative assessment.

Despite significant progress, several challenges persist. Models still struggle with compositional generalization, often producing plausible but inaccurate descriptions (a phenomenon known as "hallucination"). Moreover, captioning systems can inherit biases from training data, such as gender or cultural stereotypes, leading to ethically problematic outputs. Current research efforts are addressing these limitations through grounded captioning (anchoring words to visual evidence), explainable attention visualization, and counterfactual data augmentation aimed at improving robustness and fairness.

Hallucinations in Image Captioning

Despite the recent advancements in image captioning techniques, a persistent and well-documented issue remains: **hallucination**. In the context of image captioning, hallucination refers to the generation of textual content that is *semantically plausible* yet visually *unsupported*; that is, words, attributes, or relationships described in the caption that do not actually appear in the corresponding image [288]. Hallucinations undermine the reliability and trustworthiness of vision-language systems, posing challenges for their deployment in real-world applications such as assistive technologies, autonomous systems, and medical image interpretation.

Hallucination in image captioning can be categorized broadly into two types: *object-level* and *sentence-level* hallucination [288]. Object-level hallucination occurs when the model mentions objects that are not present in the image (e.g., describing "a dog" when none exists), whereas sentence-level hallucination involves generating semantically consistent yet globally incorrect statements, such as describing interactions or scenes that differ from the visual context. These errors reveal a fundamental issue in how captioning models integrate and balance visual grounding and linguistic priors. Many deep learning captioning systems, especially those trained end-to-end on paired image–text datasets, tend to over-rely on language models that capture statistical co-occurrences rather than actual visual evidence. For instance, because "a person riding a horse" frequently co-occurs in training data, a model may describe any person–animal pair as a riding scene even when the image shows otherwise.

To quantify and analyze hallucination, Rohrbach et al. [288] introduced the CHAIR metric (Caption Hallucination Assessment with Image Relevance), which evaluates the degree to which generated captions mention objects absent from the image. It consists of two complementary measures: $CHAIR_s$, which reports the proportion of captions containing at least one hallucinated object, and $CHAIR_i$, which measures the proportion of hallucinated object mentions across all generated object tokens. To operationalize these metrics, objects are identified by matching words in captions to a predefined vocabulary aligned with the COCO dataset [183], and visual presence is determined through either ground-truth annotations or automatic object detectors. By separating image-based and text-based verification, CHAIR offers an interpretable quantitative estimate of how frequently captioning models refer to non-existent entities.

The per instance $CHAIR_i$ is defined as:

$$CHAIR_i = \frac{|Hallucinated\ Objects|}{|All\ Predicted\ Objects|} \tag{6.3.1}$$

Furthermore, the per sentence $CHAIR_s$ is formed as:

$$CHAIR_s = \frac{|Sentences\ with\ hallucinated\ objects|}{|All\ sentences|} \tag{6.3.2}$$

Despite its simplicity, CHAIR acts as a first, immediate measure for object hallucinations. The analysis stemming from [288] revealed that models with stronger language modeling components (such as those fine-tuned on textual corpora) tend to hallucinate more frequently, whereas models with stronger visual grounding components exhibit lower hallucination rates. Moreover, CHAIR revealed that hallucination was not merely a rare occurrence but a systematic behavior across state-of-the-art models: even architectures with visual attention, such as Show, Attend and Tell [382] and Bottom-Up and Top-Down Attention [10], frequently generated hallucinated objects. These hallucinations correlated strongly with the models' reliance on language priors—statistical regularities in training data—rather than actual visual evidence. Importantly, CHAIR demonstrated that traditional caption evaluation metrics, such as BLEU or CIDEr, were largely insensitive to hallucination, since they measure lexical overlap rather than visual accuracy. As a result, a caption could achieve a high CIDEr or BLEU score despite containing hallucinated content, underscoring the need for metrics specifically designed to assess factual consistency.

Following CHAIR, researchers proposed several extensions and alternatives to better capture visual grounding fidelity. Some works incorporated object detection-based verification, where captions were cross-checked against outputs from pretrained vision models (e.g., Faster R-CNN [284]), enhancing precision in detecting missing or false visual entities [56]. Others introduced semantic similarity-based metrics, which leveraged pretrained multimodal encoders such as CLIP [273] to compare the alignment between image–caption pairs. These approaches measure hallucination indirectly by quantifying the cosine similarity between visual and textual embeddings; lower alignment indicates greater hallucination likelihood. For instance, CLIPScore [108] computes the similarity between CLIP's image and text embeddings to evaluate caption-image consistency. While such embedding-based metrics offer scalability and generalizability beyond fixed object vocabularies, they can obscure fine-grained grounding errors because they evaluate holistic semantic alignment rather than discrete object presence.

Recent research in vision-language hallucination detection extends beyond object-level errors to encompass attribute and relationship-level hallucinations, where models inaccurately describe object properties or interactions. Metrics such as FaithScore [137] have been proposed to capture these subtler inconsistencies by integrating structured scene graphs or grounding annotations. Some studies also combine human evaluation protocols with automatic metrics to assess perceptual faithfulness, recognizing that hallucination is often context-dependent and not always reducible to binary object presence. Similarly, ALOHa [263] leverages an LLM to identify groundable objects within a candidate caption, assess their semantic similarity to reference objects from both captions and object detections, and utilize Hungarian matching to compute the final hallucination score. Nonetheless, the subcaption process leverages LLMs, which also hallucinate themselves.

The dialog-based evaluation process of POPE [178] suggests answering "yes/no" to questions regarding the existence of an object in an image. Objects are extracted from images based on ground truth annotations or segmentation tools, filling question templates, while an equally sized set of non-existent objects provides negative samples to measure the confidence of prompted models against "yes/no" answer bias. Then, the agreement between answers with ground truth objects is measured. Also using a question-answering pipeline to evaluate object hallucinations, NOPE [193] regards LLM-constructed questions with negative indefinite pronouns (e.g. nowhere, none etc) as ground truth answers.

Involving LLMs in the hallucination detection pipeline, [357] are the first to recognize VL hallucination patterns, driving the construction of prompts for ChatGPT to generate relevant hallucinated instances. Fine-tuning LLama [338] on such hallucinations provides a proficient module for capturing VL hallucinations.

Model performance on standard text generation metrics may be negatively correlated with hallucination occurrence, while the choice of image encoding techniques and training objectives employed in the pre-training stage can be definitive [57]. Statistical factors accompanying object hallucinations were analyzed in [423], examining frequent object co-occurrences, uncertainty during the generation process, and correlations between hallucinations and object positioning within the generated text.

Across these developments, a key challenge persists: balancing interpretability with coverage. Metrics like CHAIR offer transparent, object-level interpretability but depend on exhaustive annotation and limited vocabularies, making them dataset-specific. Embedding-based metrics such as CLIPScore, by contrast, generalize across domains but sacrifice interpretability and may misclassify creative or metaphorical descriptions as hallucinations. Consequently, a hybrid evaluation paradigm is emerging, integrating explicit object-based grounding checks (as in CHAIR) with learned multimodal similarity measures to provide a more comprehensive assessment of visual fidelity.

6.3.2 Explainable hallucinations evaluation

Many of the contributions analyzed above harness LLMs at some point of the hallucination evaluation process. These approaches inevitably induce uncertainty related to the prompt used, while simultaneously facing the possibility of LLMs also hallucinating and ultimately hindering the robustness and trustworthiness of the affected module, and thus the evaluation framework itself. In our framework, we deviate from the usage of LLMs, sacrificing the simplicity they provide in order to enhance the determinism and reliability of the evaluation process.

Other than that, both metrics evaluating linguistic quality, as well as metrics for VL hallucinations lack explainability aspects, since they do not suggest the *direction of change* towards dehallucinated generation. This direction of change should primarily be **measurable** and **meaningful**, while its optimal usage prescribes notions of **optimality**, translated to semantically ***minimal changes***, as well as the ***fewest possible number of edits*** leading to the desired outcome. We will analyze these desiderata:

Measurable change refers to assigning a well defined numerical value for comparative reasons. This requirement demands the connection of concepts to be changed with similarity features within a unified structure, such as their distance on a semantic space or within a semantic graph.

Meaningful change refers to performing operations that are sensible in the real world, such as substituting an object with another object and not with meaningless sequences of characters. For example, swapping the concept "cat" with the character sequence "hfushbfb" does not hold a useful meaning. Moreover, even substituting objects with actions breaks meaningfulness, e.g. replacing the concept "cat" with the concept "swimming" within the same sentence violates the well-defined rules of linguistic syntax.

Optimal change refers to employing a strategy which guarantees that valid and measurable changes are the best ones to be found among a possibly infinite set of valid and measurable changes. For example, replacing "cat" with "person" is meaningful for a human, while also being measurable if we place the concepts "cat" and "person" in a semantic graph structure. However, an alternative suggestion could be replacing "cat" with "dog", as they are both animals, or even "cat" with "tiger" since they are both felines. In this case, optimal edits require finding the most *semantically similar concept* to the source one. Such similarity requirements can be imposed by structured knowledge bases, deterministically ensuring ***semantically minimal edits***. Furthermore, the number of such edits should be controlled, since infinitely performing minimal changes should be naturally excluded from the proposed framework. For example, the transition "cat"→"dog" should not consider extraneous changes, if not required to approach the ground truth sample. Therefore, the set of all proposed changes should be minimized in terms of overall semantic cost, ultimately resulting in ***fewest possible semantically minimal edits***.

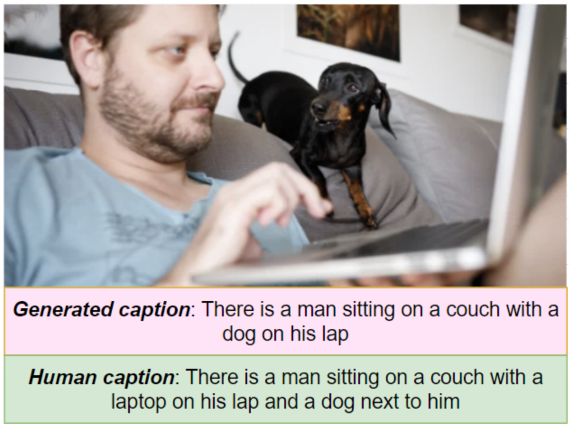
To address these challenges, we leverage the framework first proposed in [78], where counterfactual explanations are provided via edits satisfying our desiderata. This framework, described in Section 6.2, adopted for the evaluation of image generation models [203], where a source set S contains the ground truth concepts as extracted from the generated modality (in our use case being textual captions) and a target set T contains ground truth concepts as extracted from the input modality (in our case being annotated images provided to the captioner).

We wish to perform the $S \rightarrow T$ transition using the fewest possible semantically minimal and meaningful edits, which is achieved via the guarantees offered by the WordNet hierarchy [226]: concepts from S, T are mapped on WordNet synsets, which correspond to sets of cognitive synonyms. Distances of synsets within the hierarchy translate to semantic differences in actual meaning. Finding the minimum path between two synsets entails semantically minimal differences between corresponding concepts. WordNet is a crucial component of this implementation, since it guarantees **measurable** (WordNet distance is a numerical value), **meaningful** (WordNet synsets correspond to lexical entities of the English language) and **semantically minimal** (shortest WordNet distance between two concepts is found using pathfinding algorithms [66]) concept edits. The algorithm of [78] uses bipartite matching to minimize the overall cost of assignment between S and T concepts, ensuring the optimal $S \rightarrow T$ transition.

By breaking down the $S \rightarrow T$ transition, the following three edit operations e are allowed for any source $s \in S$ and target concept $t \in T$ [78, 203]:

- **Replacement (R)** $e_{s \rightarrow t}(S)$: A concept $s \in S$ is replaced with $t \notin S$.
- **Deletion (D)** $e_{s-}(S)$: A concept $s \in S$ is deleted from S.
- **Insertion (I)** $e_{t+}(S)$: A concept $t \in T$ is inserted in S.

Especially in the case of image captioning, we impose higher importance in **D** and **R** edits; the rationale behind this decision is that since hallucinations refer to the presence of irrelevant or extraneous concepts, they should be deleted or replaced to match the ground truth ones. Moreover, in many cases, captions purposefully provide a higher-level description of an image, therefore several visual concepts are omitted, sacrificing coverage for brevity. In that case, **I** suggests the addition of visual concepts to the caption, which may not be always necessary. In our framework, we also include **I** calculations, but we do not consider them in the overall transformation cost; instead, we provide them as *suggestion* for the user to choose whether they may be incorporated in more verbose captions.



Generated caption: There is a man sitting on a couch with a dog on his lap

Human caption: There is a man sitting on a couch with a laptop on his lap and a dog next to him

Figure 6.3.1: Example of hallucination on image captioning. The generated caption c misses an accurate relationship between the man and the dog. The concept "laptop" should replace the concept "dog" in the generated caption, while the relationship "next to" should be added to connect the concepts "dog" and "man". [202]

The role of roles

In Figure 6.3.1, the captioning model (BLIP) [173] confuses the spatial relationship between the man and the dog, showcasing the importance of *role hallucinations*, which were not widely addressed in prior work, since object hallucinations were their primary concern. Additionally, roles should be addressed *in conjunction to objects*, and not on their own, since this more simplistic approach would result in under-detection of hallucinations. For example, if we apply the counterfactual explanations algorithm of [78] on sets of roles, the proposed edits for Figure 6.3.1 would be {**I**("next to")}, referring to the addition of the role "next to" that connects the dog and the man. However, if we consider *triples* of two objects connected with a role, the resulting edits would be: {**R**(["dog", "on", "lap], ["laptop", "on", "lap"]), **I**(["dog", "next to", "man"])}, which is a more valid set of edits, if we view the human-written ground truth caption and the image itself.

To perform the transition to editing triples instead of standalone concepts, we require scene graphs instead of objects to acquire a conceptual representation of the image [**structure-your-data**]. Regarding the caption, we also parse the sentence in a graph structure. Given two graphs G_T representing the image and G_S corresponding to a possibly hallucinated generated caption c, we search for the minimum cost set of **R**, **D**, **I** edits (applied on objects and roles) that transform $G_S \to G_T$, i.e. convert a -possibly- hallucinated graph to a non-hallucinated one.

This cost of transformation is calculated using Graph Edit Distance (GED) between G_S, G_T. Denoting as $c(e_i)$ the cost of an operation $e_i \in \{\mathbf{R}, \mathbf{D}, \mathbf{I}\}$ and $P(G_S, G_T)$ the set of n edit paths to transform $G_S \to G_T$, GED is formed as:

$$GED(G_S,\, G_T) = \min_{(e_1,\ldots,e_n)\in P(G_S,G_T)} \sum_{i=1}^{n} c(e_i) \tag{6.3.3}$$

The shortest paths $P(G_S, G_T)$ are calculated using deterministic pathfinding algorithms, such as Dijkstra [66], ensuring optimality of edits e_i.

However, GED is an NP-hard algorithm as analyzed in Section 5.1, meaning that it cannot be calculated efficiently in its basic, brute-force format. For this reason, we employ some approximations, such as the Volgenant-Jonker (VJ) algorithm [140], which allows for GED calculation in polynomial time.

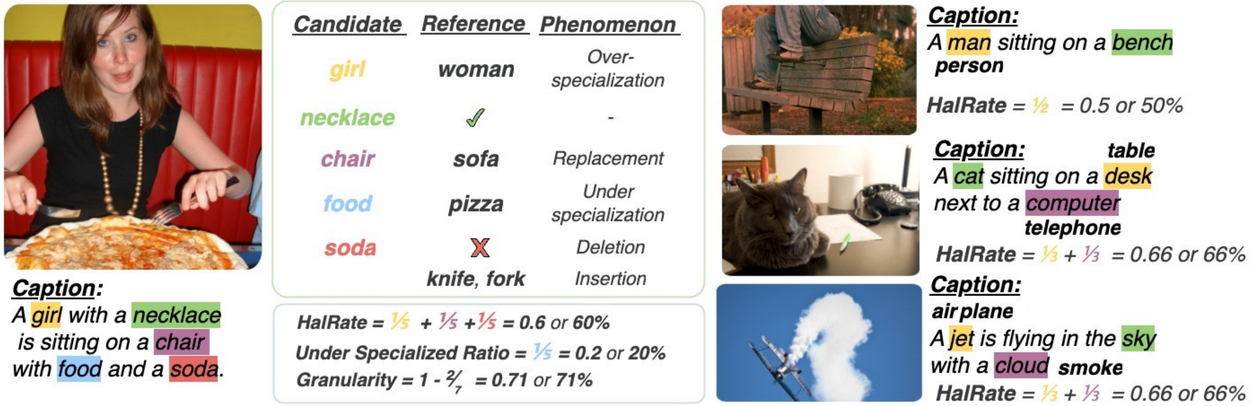

Candidate
Reference
Phenomenon
girl
woman
Over-specialization
necklace
✓
-
chair
sofa
Replacement
food
pizza
Under specialization
soda
✗
Deletion
knife, fork
Insertion
HalRate = 1/5 + 1/5 + 1/5 = 0.6 or 60%
Under Specialized Ratio = 1/5 = 0.2 or 20%
Granularity = 1 - 2/7 = 0.71 or 71%
Caption:
A girl with a necklace is sitting on a chair with food and a soda.
Caption:
A man sitting on a bench
person
HalRate = 1/2 = 0.5 or 50%
Caption:
table
A cat sitting on a desk next to a computer
telephone
HalRate = 1/3 + 1/3 = 0.66 or 66%
Caption:
airplane
A jet is flying in the sky with a cloud smoke
HalRate = 1/3 + 1/3 = 0.66 or 66%

Figure 6.3.2: An example of detected hallucination of objects in image captioning from our framework is presented, depicting each phenomenon along with the proposed metrics. Objects in yellow represent an overspecialized phenomenon, in purple a replacement, and in red a removal. Those in green are correct objects, and those in blue are the underspecialized objects (which do not constitute hallucinations, as the caption contains a more generic concept to the ground truth one). As shown, the hallucination rate is calculated as the sum of the rate of each hallucination phenomenon independently. [202]

6.3.3 Hallucination detection framework

Object hallucinations

In Figure 6.3.2 we illustrate the hallucination analysis provided by our proposed HalCECE framework [202]. We formulate the object hallucination detection problem as follows: each caption c has generated objects $S = \{s_1, s_2, \ldots, s_n\}$, and each image contains the ground truth objects $T = \{t_1, t_2, \ldots, t_m\}$. We find the **R**, **D**, **I** sets of object edits to perform the transition $S \rightarrow T$, as analyzed in Section 6.3.2.

In order to evaluate different granularities of hallucinations, i.e. presence of more generic or more specific concepts compared to the ground truth one, we utilize the Least Common Ancestor (LCA) within the WordNet hierarchy. Specifically, LCA denotes the closest ancestor synset between two synsets in WordNet; we closely examine the case where the LCA between two synsets contains one of the synsets itself: for example, given two synsets v and w, if LCA(v, w)=v, then v is a hypernym (more generic concept) of w.

Based on these, we analyze the following *hallucination phenomena*:

- **Deletion (D)**: When an object $s_i \in S$ must be deleted; e.g., in Figure 6.3.2, the concept "soda" is in the generated caption c but not in the image.
- **Replacement (R)**: When an object $s_i \in S$, is replaced with a different object t_j, where $LCA(s_i, t_j) \neq s_i$, and $LCA(s_i, t_j) \neq t_j$ (meaning that no object is a hypernym of the other). For instance, the caption references a "chair", but the image contains a "sofa".
- **Over-specialization (O)**: When an object $s_i \in S$ is replaced with a different object t_j, where $LCA(s_i, t_j) = t_j$, i.e. t_j is a more general concept than s_i in the hierarchy. For example, the caption states that the image contains a "girl", but the image depicts a "woman"; in this case, the caption erroneously overspecified this term, since "girl" is subcategory of "woman".

Based on these phenomena, we measure the degree of hallucination for a caption c as the number of objects that exhibit at least one of the aforementioned phenomena. Thus, the metric for counting hallucinations in captioning, denoted as $Hal(S, T)$, is defined as the sum of the cardinalities of the sets of **D**, **R**, **O**:

$$Hal(S,T) = |\mathbf{D}(S,T)| + |\mathbf{R}(S,T)| + |\mathbf{O}(S,T)| \tag{6.3.4}$$

The hallucination rate *HalRate* reveals the percentage of hallucinated objects over the total number of objects $|S|$ in c, and it is mathematically expressed as:

$$HalRate(S,T) = \frac{Hal(S,T)}{|S|} \tag{6.3.5}$$

We incorporate additional semantic metrics on these properties, such as quantifying the semantic distance between hallucinatory and ground truth concepts.

- **Similarity of Replacements**: We employ Wu-Palmer similarity [374] to measure the semantic similarity of replacements based on the position of synsets in WordNet. This way, we measure how close the replaced terms are in order to gain further understanding of the behavior of the captioner. For example, semantically related replacements receive a higher Wu-Palmer similarity score, denoting more "justified" hallucination occurrences.

An additional facet of HalCECE lies in its capacity to explore phenomena beyond hallucination. This is exemplified through the following measures:

- **Granularity**: Defined as 1 minus the ratio of **Insertions (I)** over the number of ground truth image objects. In essence, it represents the percentage of objects that c attempts to encapsulate compared to the image objects:
$$Granularity(S, T) = 1 - \frac{|\mathbf{I}(S)|}{|T|} \tag{6.3.6}$$
- **Under-Specialization (U)**: Quantifies the instances of underspecialized objects, where an object $s_i \in S$ from c is replaced with a different object t_j, and $LCA(s_i, t_j) = s_i$, meaning that the caption object is more generic than the corresponding image object. For instance, if the caption indicates the presence of "food", but the image portrays a "pizza", the caption is not incorrect (because a "pizza" is a sub category of "food") but could benefit from greater specificity. The ratio is computed as the division of the number of under-specialized objects by the total number of objects in c, reflecting the proportion of objects in the generated captions that are underspecialized.

In our analysis, we incorporate both the **average number of objects per caption** and the **average number of WordNet ancestors (hypernyms)** associated with each of these objects for all data instances. This approach provides a comprehensive perspective on the content of each caption c.

Role hallucinations

Our framework is directly extended to incorporate edge-level hallucinations. On top of objects included in T and S, images and captions also describe object interactions. As explained, role hallucination is measured using *triples* and not simply relations which would disregard adjacent objects and their transformations. To this end, we denote the sets of triples corresponding to captions and image annotations respectively as $S^r = \{(s_i, r^s_j, s_k), \ldots\}$ and $T^r = \{(t_i, r^t_j, t_k), \ldots\}$. A visual representation of roles within captions can be found in Figure 6.3.3. Examples of caption triples are "horse *over* obstacle" and "people *sitting at* table". To measure role hallucinations, i.e. the transition from $S^r \rightarrow T^r$, we employ an adjusted version of aforementioned equations. Edit sets **D** and **R** are calculated by considering triples instead of objects as following:

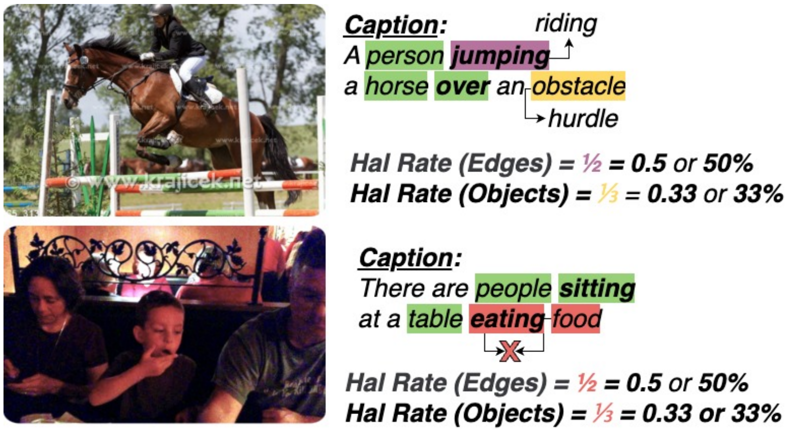

Caption:
riding
A person jumping
a horse over an obstacle
hurdle
Hal Rate (Edges) = ½ = 0.5 or 50%
Hal Rate (Objects) = ⅓ = 0.33 or 33%
Caption:
There are people sitting
at a table eating food
Hal Rate (Edges) = ½ = 0.5 or 50%
Hal Rate (Objects) = ⅓ = 0.33 or 33%

Figure 6.3.3: An demonstration of the edge integration into HalCECE. The edges are highlighted in **bold**, and the different colors correspond to those of Figure 6.3.2.

- **Deletions (D)**: When an edge r_j^s between two objects s_i, s_k must be deleted. Notably, this edit set includes deletions induced as "collateral damage" due to object deletions or replacements, as well as hallucinated relations between correctly detected objects in c. In Figure 6.3.3, the role "eating" between "people" and "food" is deleted because "food" is hallucinated by the captioner.
- **Replacement (R)**: When an edge r_j^s between two objects s_i, s_k is replaced with another edge r_w^t. For example, in Figure 6.3.3 the role "jumping" between "person" and "horse" is hallucinated and needs to be replaced with "riding". Despite "jumping" being a valid relation between "horse" and "obstacle", or even "person" and "obstacle", it is definitely not correct in the presented configuration, placing a great focus on leveraging roles as part of a triple.

It is noteworthy that the definition of over-/under- specialization is not applicable for roles, as edges describe actions, topology or "part of" relations, steering away from hierarchies. To combat this, we leverage the annotation information provided by humans to correctly match caption relations to ground truth ones and map them to WordNet. When captioners produce previously unseen relations (in terms of ground truth), we weight them accordingly, so that they can be easily inserted or deleted during GED computation; they are not likely to be replaced with other roles though, since we lack semantic content. To detect if they are part of **R**, we deploy an extra post-hoc reasoning step and check if a relation r_j^s between the same two objects has been deleted and another r_w^t has been added. Given the previous analysis, role hallucinations are measured as:

$$Hal(S^r, T^r) = |\mathbf{D}(S^r, T^r)| + |\mathbf{R}(S^r, T^r)| \tag{6.3.7}$$

while *HalRate* and *Granularity* are simply adjusted to be:

$$HalRate(S^r, T^r) = \frac{Hal(S^r, T^r)}{|S^r|} \tag{6.3.8}$$

$$Granularity(S^r, T^r) = 1 - \frac{|\mathbf{I}(S^r)|}{|T^r|} \tag{6.3.9}$$

6.3.4 Experiments

Datasets

To evaluate HalCECE on images connected with both captions and scene graphs, we experiment on the intersection of Visual Genome (VG) [155] and COCO [183]. VG contains handcrafted scene graph annotations incorporating objects, attributes and roles. On the other hand, COCO scenes are connected with 5 captions per image, provided by humans. We restrict our experimentation on the COCO validation set (splits are provided by the dataset creators), which demonstrates 2170 common instances with VG; a few of those are eliminated, if the corresponding objects cannot be aligned with WordNet.

Models

We initially experiment with non-proprietary captioners, evaluating both smaller and larger models, since smaller ones can be more easily deployed by every researcher. Specifically, we apply our method on variants of GiT [356] and BLIP [173, 172], namely *GiT-base* (trained on 10 million image-text pairs), *GiT-large* (trained on 20 million image-text pairs) and *GiT-base/large-coco* (fine-tuned on COCO captions); also *BLIP-base* (using ViT [72] base encoder), *BLIP-large* (ViT large encoder), *BLIP2-flan-t5-xl* (Flan-T5 [50] is used as the language decoder) and *BLIP2-opt-2* (using OPT [412] 2.7B as the language decoder). We attempt unconditional and conditional image captioning (related experiments will be denoted as *unc/cond*), where captioners are fine-tuned to estimate conditional and unconditional distributions over captions respectively [153]. Moreover, we experiment with *ViT-GPT2* [162], which leverages ViT as the encoder and GPT2 [272] as the decoder. Finally, we provide results on two proprietary foundational models of the Claude family [11] prompted for captioning, namely *Claude-sonnet*[3] and *Claude-haiku*[4]. This way, we prove the real power of HalCECE on closed-source models where our white-box competitors are not applicable.

Since prompting LVLMs can define the length of the generated captions, we attempt to generate both longer captions (20-30 words), as well as shorter ones (10 words max), which are comparable to the captions produced

---

[3] anthropic.claude-3-5-sonnet-20241022-v2:0

[4] anthropic.claude-3-haiku-20240307-v1:0

from the rest of the captioners. This way, we get the opportunity to explore HalCECE on longer descriptions, something that is not available in smaller VL models. We name the respective experiments using L for *long* generations and S for *short* ones.

Concept sets construction

We construct the linguistic S, S^r and visual concept sets T, T^r, corresponding to source and target concept sets respectively with the goal of transforming $S \rightarrow T$ and $S^r \rightarrow T^r$. Linguistic sets are formed by extracting graphs from text via the Scene Graph Parser tool[5], while visual sets are constructed using ground truth annotations from COCO and VG.

Experimental setup

Non-proprietary pre-trained captioners are loaded from Huggingface[6] using their respective model cards and their inference is executed on a 12GB NVIDIA TITAN Xp GPU. No further training is performed. Proprietary Claude models are accessed via Amazon Web Services (AWS) using API calls (Bedrock service).

6.3.4.1 HalCECE Results

Based on the hallucination detection framework analyzed in the previous section, we present our findings as following: Tables 6.11, 6.12, 6.13 contain averaged results per captioner involving the hallucination phenomena introduced above. In addition, Figures 6.3.4, 6.3.5 demonstrate the distributions of values per hallucination phenomenon in our dataset, addressing object and role hallucinations respectively. These plots refer to GiT-base as a proof-of-concept, since it is one of the best-performing captioners according to our reported explainable metrics.

Model	#objects	#ancestors	HalRate (#hal. objects)↓	Granul.	**U↓**
GiT-base-coco	3.13	27.93	35.56% (1.13)	17.0%	4.06% (0.13)
GiT-large-coco	3.15	27.97	33.93% (1.1)	17.0%	3.92% (0.12)
GiT-base	1.76	16.57	26.41% (0.48)	9.0%	3.27% (0.06)
GiT-large	1.74	16.28	25.38% (0.46)	9.0%	3.31% (0.06)
BLIP-base-unc	2.53	22.55	34.28% (0.91)	13.0%	4.48% (0.12)
BLIP-base-cond	3.23	29.5	58.48% (1.87)	17.0%	2.96% (0.1)
BLIP-large-unc	3.63	32.73	39.2% (1.45)	19.0%	3.47% (0.13)
BLIP-large-cond	4.22	37.5	53.04% (2.24)	22.0%	2.84% (0.12)
BLIP2-flan-t5-xl	2.57	23.16	33.13% (0.89)	14%	4.05% (0.11)
BLIP2-opt-2	2.78	24.89	33.28% (0.96)	15.0%	4.19% (0.12)
ViT-GPT2	2.95	26.51	38.76% (1.18)	16.0%	4.47% (0.14)
Claude sonnet-L	6.85	58.94	58.91% (4.05)	36.0%	4.71% (0.33)
Claude haiku-L	7.12	58.66	64.31% (4.64)	39.0%	5.4% (0.39)
Claude sonnet-S	3.35	30.48	47.16% (1.6)	17.0%	4.67% (0.16)
Claude haiku-S	2.95	25.49	54.36% (1.62)	16.0%	6.74% (0.19)

Table 6.11: Object hallucinations (mean values) on the $VG \cap COCO$ validation subset. **Best** and **worst** results are denoted. Numbers in parenthesis denote absolute #objects.

In all cases, a rather high percentage of hallucinations for all semantics (objects, roles and derived phenomena) is observed; almost $1/3^{rd}$ of the caption objects present some form of hallucinations, while for roles several occurrences contain some hallucinatory inaccuracy, even exceeding 90% in HalRate.

Replacements (**R**) represent the most common type of object edit e, indicating that captioners often generate hallucinated objects that are still related to the source (e.g., "daybed" instead of "couch"), rather than entirely unrelated hallucinated objects. Replacements of relevant objects are preferable to deletions for HalCECE, as the paths leading to another concept within WordNet are often shorter than the path to the root node (*entity.n.01*), which corresponds to the Deletion (**D**) edit. However, for unrelated objects, where the distance between these two is greater than the cost of first deleting the one and then inserting the other, replacement is not preferable. This is because concepts appearing in captions usually lie lower in the hierarchy, being specific

---

[5] https://github.com/vacancy/SceneGraphParser

[6] https://huggingface.co/models?pipeline_tag=image-to-text

Model	**D**↓	**O**↓	**R**↓	Similarity of **R**↑
GiT-base-coco	4.38% (0.15)	3.01% (0.09)	28.18% (0.89)	0.56
GiT-large-coco	4.4% (0.16)	2.46% (0.08)	27.06% (0.87)	0.55
GiT-base	2.11% (0.05)	2.17% (0.04)	22.12% (0.4)	0.61
GiT-large	2.46% (0.05)	2.41% (0.04)	20.51% (0.36)	0.6
BLIP-base-unc	3.78% (0.11)	2.65% (0.07)	27.86% (0.73)	0.57
BLIP-base-cond	23.07% (0.72)	2.76% (0.09)	32.66% (1.05)	0.52
BLIP-large-unc	6.13% (0.24)	3.48% (0.13)	29.59% (1.08)	0.56
BLIP-large-cond	19.27% (0.81)	2.46% (0.11)	31.3% (1.32)	0.52
BLIP2-flan-t5-xl	4.27% (0.12)	3.16% (0.08)	25.7% (0.69)	0.56
BLIP2-opt-2	3.64% (0.11)	2.8% (0.08)	26.84% (0.77)	0.57
ViT-GPT2	3.45% (0.11)	3.16% (0.09)	32.14% (0.97)	0.6
Claude sonnet-L	15.79% (1.05)	2.51% (0.19)	40.61% (2.81)	0.52
Claude haiku-L	17.3% (1.28)	2.69% (0.2)	44.33% (3.15)	0.49
Claude sonnet-S	7.1% (0.25)	5.42% (0.18)	34.63% (1.16)	0.57
Claude haiku-S	7.78% (0.24)	4.59% (0.13)	41.99% (1.26)	0.52

Table 6.12: Continuation of Tab. 6.11. More object hallucination phenomena on $VG \cap COCO$ validation subset. Numbers in parenthesis denote absolute #objects.

Model	#roles	**D**↓	**R**↓	HalRate (#hal. roles)↓	Granul.
GiT-base-coco	1.92	65.32% (1.37)	14.06% (0.29)	79.38% (1.66)	3.93%
GiT-large-coco	1.94	65.33% (1.36)	13.75% (0.29)	79.09% (1.65)	4.08%
GiT-base	0.73	44.05% (0.47)	11.98% (0.13)	56.03% (0.59)	1.8%
GiT-large	0.69	39.15% (0.42)	11.58% (0.12)	50.63% (0.54)	1.89%
BLIP-base-unc	1.44	61.2% (1.01)	13.04% (0.2)	74.23% (1.22)	3.01%
BLIP-base-cond	2.14	90.96% (1.93)	4.22% (0.1)	95.18 (2.03)	1.48%
BLIP-large-unc	2.28	68.32% (1.67)	13.2% (0.31)	81.52% (1.98)	4.38%
BLIP-large-cond	2.98	86.6% (2.54)	6.68% (0.22)	93.28% (2.77)	2.99%
BLIP2-flan-t5-xl	1.62	69.26% (1.16)	14.2% (0.22)	83.47% (1.38)	3.25%
BLIP2-opt-2	1.79	68.87% (1.25)	14.37% (0.25)	83.24% (1.51)	3.65%
ViT-GPT2	1.86	71.05% (1.36)	16.46% (0.28)	87.5% (1.64)	3.42%
Claude sonnet-L	3.9	80.71% (3.17)	9.8% (0.39)	90.51% (3.56)	7.1%
Claude haiku-L	3.99	80.25% (3.29)	10.31% (0.38)	90.56% (3.67)	6.28%
Claude sonnet-S	2.1	75.19% (1.62)	11.85% (0.25)	87.04% (1.87)	5.24 %
Claude haiku-S	1.85	74.31% (1.39)	13.71% (0.14)	88.02% (1.53)	4.99%

Table 6.13: Role hallucinations (mean values per image) on the $VG \cap COCO$ validation subset. Numbers in parenthesis denote absolute #roles.

enough to describe depicted objects. This level of conceptual granularity is imposed during the pre-training of captioners, which utilize descriptive captions, such as the ones of COCO or similar datasets comprising image-text pairs. On the contrary, role hierarchy is much shallower, justifying the higher number of **D** edits in comparison to **R** edits (Table 6.13). This finding is further reinforced by the fact that objects connected in c often are not immediate neighbors in the ground truth, meaning that a completely new edge will need to be inserted.

A comparison between model families regarding object hallucinations (Tables 6.11, 6.12) reveals interesting insights: GiT variants consistently hallucinate less, achieving best results across most metrics compared to other model families; note the colored cells of respective Tables. On the other hand, Claude variants are accompanied with more hallucinations (note the colored cells). This may occur due to the fact that Claude models are not explicitly pre-trained on image captioning using COCO captions or similarly distributed image-text pairs, therefore they tend to hypothesize the existence of out-of-distribution concepts.

This elevated hallucination tendency is more expected on longer captions (20-30 words), since the model is forced to be more verbose, possibly adding extraneous concepts to meet the length requirements; this is verified by the reported results, even though shorter captions are not devoid of object hallucinations as well.

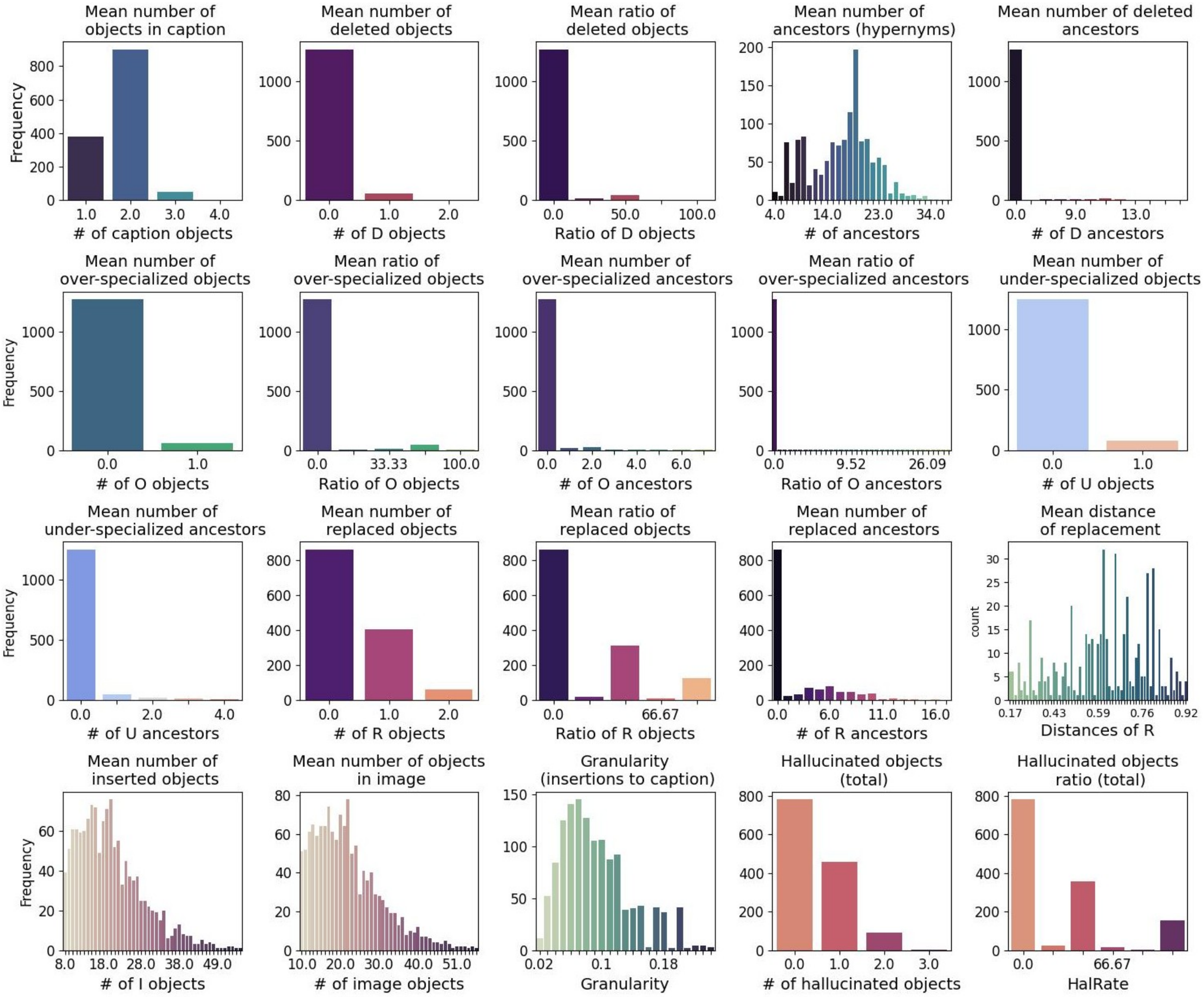


Figure 6.3.4: Statistics of our proposed explainable metrics on ***object hallucinations*** by GiT-base on the $VG \cap COCO$ validation set.

Additionally, as expected, longer descriptions demonstrate roughly twice the Granularity, indicating greater object coverage. Furthermore, shorter generations are accompanied by higher over-specialization (**O**) rates, indicating that Claude models become excessively specific when attempting to condense visual information within a restricted word budget. Another notable observation is that Haiku variants (either prompted for short or longer captions) tend to be more generic, as denoted by the inflated under-specialization (**U**) percentages in comparison to Sonnet variants, despite being prompted with the same instructions. Other than that, Haiku variants require more conceptual replacements (**R**) to assimilate the ground truth captions compared to Sonnet ones; it is possible that those **R** edits can be attributed to substitutions with concept *hyponyms*, so that the **U** rates are also reduced.

A comparatively worse performance in terms of hallucinations is observed when conditional generation is employed over unconditional one in BLIP variants. This can be attributed to over-reliance over linguistic priors [377], amplifying possible biases or noise. HalCECE is able to highlight such discrepancies regarding the generation strategy selected, suggesting straightforward mitigation strategies (in that case being the usage of unconditional caption generation). It also breaks down the source of hallucinations, as indicated in Tables 6.11, 6.12: the rate of hallucinations (HalRate) is significantly higher than their unconditional counterparts, even though the **U** percentages are the lowest, meaning that specificity is not the culprit of hallucinations. On the contrary, the higher percentage of **D** and **R** edits denotes the presence of extraneous objects that have to be removed and substituted accordingly.

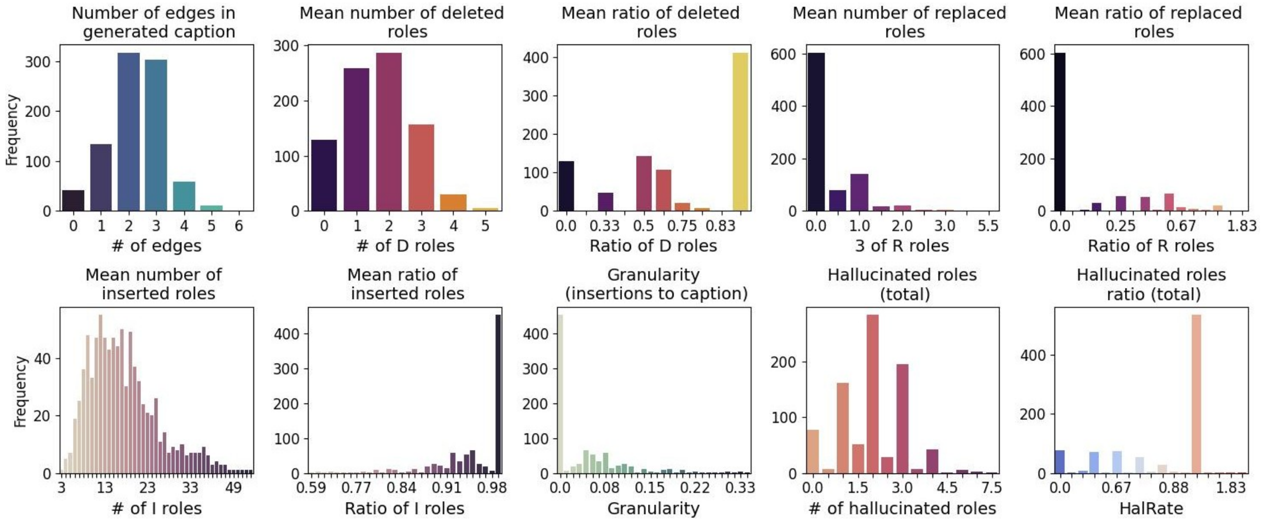


Figure 6.3.5: Statistics of our proposed explainable metrics on ***role hallucinations*** by GiT-base on the $VG \cap COCO$ validation set.

Regarding role hallucinations and comparison between models, similar trends emerge. Larger models exhibit more hallucinations overall, while GiT variants consistently produce fewer, performing best across most metrics. The primary differences across model families emerge in deletions rather than transformations, exemplified by BLIP-base-cond, which has the lowest **R** but the highest overall hallucination rate. This suggests that some models are more prone to omitting role-related information rather than altering it. Notably, Claude models demonstrate greater Granularity in role assignments, which may contribute to their higher hallucination rates. These findings align with object hallucination trends, reinforcing the idea that pre-training differences and generation strategies significantly impact hallucination tendencies across models.

Overall, it is evident that larger models cannot guarantee reduced hallucination rates. On the contrary, lower rates and fewer conceptual edits are observed in smaller captioners, such as the ones from GiT family. Even though this may sound surprising at a first glance, the source of hallucinations can be the data annotations rather than the capacity of the model itself. This means that when the visual input is ambiguous or the vision-language grounding is weak, larger models might rely more on their strong language priors, extensively stored during pre-training, thus producing fluent but unfaithful details. In the interest of image captioning, this can manifest as hallucinations -objects or actions that are statistically likely in language but *not* actually present in the image. Additionally, larger models may overfit to noisy or spurious correlations in the training data, further amplifying hallucinated content. For example, larger vision-language models may generate more detailed captions that sound plausible yet include elements unsupported by the visual evidence [377, 59]. This suggests that the balance between visual grounding and language fluency can be more challenging to maintain as model size increases. In the following section we delve into possible discrepancies between linguistic capacity and hallucinations.

6.3.5 Linguistic metrics may be misleading

Apart from our proposed hallucination evaluation metrics, we report language generation metrics, and specifically *ROUGE* [182], *BLEU* [255], *Google BLEU*[7], *Mauve* [264] and perplexity (*PPL*) [130] to reveal agreements and disagreements.

ROUGE metrics measure recall and structural overlap between ground truth and generated captions. Specifically, *ROUGE1* compares individual words (unigrams), *ROUGE2* evaluates agreement between two-word sequences (bigrams), while *ROUGEL* considers the longest common subsequence (LCS) between ground truth and generated text to decide upon their agreement. All these metrics are extracted by comparing the generated caption with each one of the 5 COCO captions at a time, and then obtaining their average score.

---

[7] https://huggingface.co/spaces/evaluate-metric/google_bleu

Finally, the *ROUGELsum* variant regards LCS scores across multiple ground truth references (in our case being all 5 COCO captions per image), offering similar results to ROUGEL.

BLEU and *Google BLEU* assess unigram precision between the ground truth and the generated caption, once again considering averaged values.

Mauve provides a broader perspective on the text quality and naturalness, measuring the distributional differences between the ground truth and the generated text embeddings. It is less sensitive to exact wording and better reflects semantic similarity and stylistic variability. In technical terms, we opt for GPT2 as the decoder to obtain embedding representations, following the default setup[8].

All those metrics range between [0,1] with higher values being better.

Perplexity (PPL) quantifies how "surprised" a language model is when it sees the next word in a sequence, providing a measure of confidence in accurately predicting the next word. This higher confidence is associated with more predictable, fluent and coherent textual generations, reflected in lower PPL scores. A perfect PPL score equals to 1, while no upper bound exists.

6.3.5.1 What is the issue with language generation metrics?

While these widely used metrics provide useful signals—primarily around fluency, style, and surface-level similarity—they can be misleading indicators of overall quality in text generation, often failing to capture semantic accuracy, contextual appropriateness, and hallucinations, as expressed via factual inconsistencies [79, 159].

For example, n-gram overlaps reward surface-level similarity, totally excluding semantically equivalent expressions or even word ordering variability. For example, if an image contains the concept "cat", n-gram metrics will assign the *same penalty* over captions that contain either the concepts "kitten" or "ship" in place of "cat". On the contrary, HalCECE will provide a significantly higher **R** cost to the "cat"→"ship" edit in comparison to the "cat"→"kitten" one. Even semantically adaptive metrics, such as *Mauve*, are not oriented towards factual inconsistencies, as reflected on disagreements between the visual and the linguistic modalities. This means that a caption can be perfectly natural and well-written, achieving high *Mauve* scores, while also containing several objects or roles not existing in the corresponding image. Similarly, *PPL* penalizes inarticulate generations but totally ignores semantic disagreements between modalities. Overall, apart from the n-gram overlap metrics, the rest are by design *not explainable*; their reliance on linguistic distributions sacrifices senses of semantic interpretability, leading to obscure and dispersed evaluation practices in the first place. Finally, in all cases, linguistic metrics require ground truth captions in order to function, contrary to HalCECE which only requests standalone concepts.

Based on the above, the motivation behind our explainable and conceptual hallucination detection framework is further verified by the unsuitability and opaqueness of common text generation evaluation practices. Therefore, the language generation metrics are incapable of providing proper hallucination signals on their own, and in several cases -e.g. when n-grams are employed to measure agreement- they can even be *misleading*. These arguments will be analyzed with the support of language generation metric results, as presented in Table 6.14.

Analysis

The results presented in Table 6.14 reveal interesting patterns, notably indicating that linguistic metrics are unsatisfactory overall, primarily because exact agreements with ground truth captions are not achieved in most cases. Specifically, even though n-gram-based metrics (i.e. *ROUGE* and *BLEU* variants) can explain their reported low scores, they lead to over-penalization of generations, since they do not respect semantical equivalence between concepts, contrary to HalCECE. On the other hand, *Mauve* and *PPL* are unable to explain themselves, despite being more semantically consistent, a gap that HalCECE is able to fill be breaking down the source of semantic disagreements.

Interestingly, linguistic metrics across models present some unexplainable variability. For example, BLIP2-opt-2 is one of the top-scorers regarding n-gram metrics, though it significantly fails according to *Mauve*. This is somehow contradictory, since the same model presents a higher exact match capability over the rest,

---

[8] https://huggingface.co/spaces/evaluate-metric/mauve

Models	**ROUGE1↑**	**ROUGE2↑**	**ROUGEL↑**	**ROUGELsum↑**
GiT-base-coco	0.152	0.021	0.145	0.145
GiT-large-coco	0.152	0.022	0.146	0.146
GiT-base	0.139	0.01	0.134	0.134
GiT-large	0.127	0.01	0.122	0.122
BLIP-base-unc	0.16	0.021	0.153	0.154
BLIP-base-cond	0.352	0.116	0.317	0.317
BLIP-large-unc	0.134	0.017	0.126	0.126
BLIP-large-cond	0.402	0.163	0.361	0.361
BLIP2-flan-t5-xl	0.435	0.179	0.402	0.402
BLIP2-opt-2	0.44	0.187	0.404	0.404
ViT-GPT2	0.406	0.153	0.370	0.370
Claude sonnet-L	0.133	0.008	0.117	0.117
Claude haiku-L	0.141	0.011	0.125	0.125
Claude sonnet-S	0.062	0.002	0.058	0.058
Claude haiku-S	0.123	0.009	0.114	0.114
	BLEU ↑	**Google BLEU↑**	**Mauve↑**	**PPL↓**
GiT-base-coco	0.0005	0.051	0.186	68.305
GiT-large-coco	0.0005	0.051	0.192	63.629
GiT-base	0.0001	0.027	0.131	1541.317
GiT-large	0.0001	0.025	0.13	1475.033
BLIP-base-unc	0.0004	0.037	0.141	461.076
BLIP-base-cond	0.024	0.099	0.058	506.732
BLIP-large-unc	0.0003	0.033	0.132	67.632
BLIP-large-cond	0.056	0.133	0.064	127.578
BLIP2-flan-t5-xl	0.046	0.132	0.067	211.738
BLIP2-opt-2	0.055	0.139	0.009	130.29
ViT-GPT2	0.051	0.131	0.068	69.605
Claude sonnet-L	0.0001	0.029	0.174	71.307
Claude haiku-L	0.0002	0.029	0.174	42.032
Claude sonnet-S	0.0	0.032	0.174	358.33
Claude haiku-S	0.0004	0.047	0.174	170.585

Table 6.14: Language generation evaluation metrics on the $VG \cap COCO$ validation subset.

but also the lowest semantic agreement at the same time. This confusion is resolved via HalCECE, which places the hallucination performance of BLIP2-opt-2 somewhere in the middle in comparison to the other captioners, as demonstrated in Tables 6.11, 6.12.

Comparisons between model families indicate that BLIP variants score higher in n-gram-related metrics (i.e. *ROUGE* and *BLEU* variants), revealing a comparatively increased adherence to ground truth captions. On the contrary, Claude models present the lowest scores regarding most n-gram metrics, revealing their reduced tendency to follow ground truth distributions. This fact was also reported in HalCECE results and related analysis of Tables 6.11, 6.12, attributing the source of disagreeing semantics to their generic pre-training. Nevertheless, the percentages occurring from HalCECE are less strict, thanks to its semantic-driven foundations: hallucination rate (assimilating a recall-related scenario, where the ratio of generated concepts over all relevant concepts is measured) reaches a maximum of 64.31% (Claude haiku-L at Table 6.11), while *ROUGE* variants, expressing recall-related agreement as well, reach up to 12.5% of conceptual agreement according to $ROUGEL/ROUGELsum$ scores of Table 6.14 for the same model, which equals to a minimum of 87.5% hallucination rate. At the same time, *ROUGE* scores, despite being able to highlight which concepts are responsible for the reported disagreements, they cannot suggest *what needs to be changed*, in order to reach a dehallucinated state; conversely, a lookup in HalCECE recommendation prescribes that the 5.4% of Haiku caption concepts are too generic, the 2.69% are erroneously specific, while the 17.3% and 44.33% of caption concepts should be deleted and replaced respectively (Tables 6.11, 6.12). Finally, *PPL* is highly uninformative when it comes to hallucinations: the high *PPL* scores corresponding to GiT-base captions denote significantly uncertain generations, even though the same captioner is associated with a low HalRate. On the contrary, Claude Haiku L presents the lowest *PPL*, despite being one of the models associated with the highest HalRate. This inverse trend indicates that *PPL* is a completely unsuitable evaluation measure with regard to hallucination detection, rendering any hallucination-related insights driven by *PPL* severely misleading.

	#obj.	#ancest.	HalRate	Granul.	**U**	**D**	**O**	**R**	Sim. **R**
ROUGE1	-0.15	-0.06	-0.04	-0.05	-0.05	-0.03	-0.03	-0.03	-0.03
ROUGE2	0.05	-0.04	-0.15	-0.15	-0.09	-0.06	-0.02	-0.06	-0.02
ROUGEL	-0.04	-0.03	0.0	0.0	-0.01	-0.02	-0.03	-0.02	-0.03
ROUGELsum	-0.04	-0.03	-0.01	-0.02	-0.02	-0.02	-0.03	-0.03	-0.03
BLEU	-0.01	-0.03	-0.08	-0.08	-0.06	-0.05	-0.03	-0.05	-0.03
Google BLEU	-0.02	-0.03	-0.04	-0.05	-0.04	-0.03	-0.02	-0.03	-0.02
Mauve	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
PPL	-0.06	-0.02	0.03	0.01	-0.01	-0.03	-0.05	-0.03	-0.05

Table 6.15: Correlation between the linguistic metrics and the object hallucination metrics provided by HalCECE.

	#roles	D	R	HalRate (#hal. roles)	Granul.
ROUGE1	-0.1	0.05	0.03	0.04	0.03
ROUGE2	0.18	0.01	0.07	-0.03	0.02
ROUGEL	-0.01	0.03	0.01	0.05	0.03
ROUGELsum	0.01	0.03	0.02	0.04	0.03
BLEU	0.07	0.02	0.05	-0.01	0.02
Google BLEU	0.04	0.02	0.03	0.01	0.02
Mauve	0.02	0.02	0.02	0.02	0.02
PPL	-0.09	-0.01	-0.05	0.01	-0.03

Table 6.16: Correlation between the linguistic metrics and the role hallucination metrics provided by HalCECE.

To sum up, we calculate the correlation between the linguistic metrics and object/role hallucination metrics as calculated from HalCECE. Related results are presented in Table 6.15 for object hallucination metrics, and Table 6.16 for role hallucination metrics, denoting weak correlations (close to 0) between the two metric categories in both cases. Ultimately, we conclude that linguistic metrics cannot provide any useful information regarding the presence of hallucinations in image captioning, as detected from HalCECE.

Chapter 7

Knowledge enhancement in vision-language tasks

In the previous years, knowledge enhancement in multimodal tasks has been performed via the extensive exploitation of knowledge structures such as Knowledge Graphs (KGs). Many VL tasks have been benefited in terms of performance boosting, paving the way for hybrid domains such as Knowledge-enhanced Vision-Language (KVL) learning. Tasks previously benefited by knowledge enhancement are presented in Figure 7.0.1. We divide our VL tasks in Discriminative and Generative based on whether the model has to reason over images and text, or generate each modality from the other.

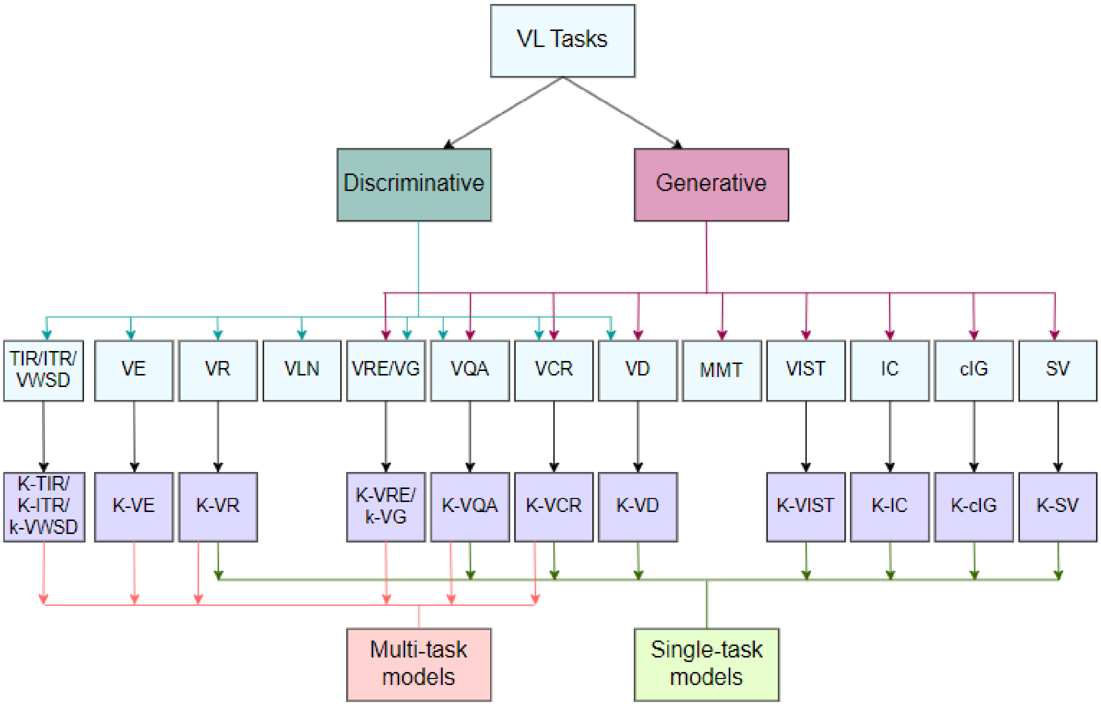


Figure 7.0.1: Taxonomy of VL tasks on which knowledge enhancement via KGs has been performed. [206]

Apparently, since most of the VL tasks have explored knowledge enhancement via KGs, we present a thorough analysis of the previously studied knowledge contribution in each of them.

7.1 Knowledge in Visual Question Answering (K-VQA)

7.1.1 Datasets

The following datasets have been used in implementations that have integrated external knowledge sources in order to provide an answer. Many of those, including VQA [3, 95], VQA-E [175], COCO-QA [283] are also used in knowledge-free versions of VQA, and external knowledge is not required in order for their questions to be answered.

VQA [3] is the first dataset introducing the task of Visual Question Answering. It contains approximately 204k images with diverse and complex scenes, with at least 3 open-ended free-form questions per image and a total of 760k questions in the whole dataset. Many of those questions are commonsense related, such as *Is this a vegeterian pizza?*. For each question 10 ground truth answers are suggested from different annotators, with a total of almost 10 million answers in the dataset. Most answers are short, with the majority of them consisting of a single word. The answers can be evaluated either in open-ended or in multiple-choice settings. Open-ended answers should be validated by 3 annotators agreeing on exactly the same answer for a given question. The multiple-choice scenario regards 18 unique candidate answers per question. Such candidates can be *correct* answers, obtained from the 10 matched answers per question, *plausible* answers, i.e. possibly incorrect answers provided by annotators without viewing the image, *popular* answers, i.e. the most frequently appearing answers in the dataset, and *random* answers sampled from other random questions within the dataset.

VQA with Explanations (VQA-E) [175] pursues the tractability of the reasoning process leading to an answer. In total, it contains around 108k images, and more than 269k explanations assigned to an equal number of QA pairs. Based on VQAv2 [95], it automatically constructs explanations with the help of COCO captions [183], as they are connected with VQAv2 images. Caption embeddings and question/answer embeddings are coupled, forming pairs of highest cosine similarities, thus assigning captions to images. Resulting explanations are highly diverse, with more than 171k unique instances, although they cannot cover images with subjective questions, such as emotional (***Do you** think this pony is cute?*), commonsense (***Can you** cross the street?*) or behavioral (***Could you** eat all these bananas by yourself?*) ones. Human evaluators assess the quality of explanations, measuring if they are fluent, correct, relevant and complementary to the answer.

DAQUAR [217] is a dataset of real world indoor scenes containing fine-grained object categories. Questions and answers related to the images are very rich regarding the objects they express: 573 unique nouns are mentioned within the whole corpus of questions and answers. Questions requiring commonsense knowledge such as *Which object on the table is used for cutting?* are included in DAQUAR, while even spatial questions such as *What is above the desk in front of scissors?* can be benefited from additional knowledge.

COCO-QA [283] addresses shortcomings of the DAQUAR [217] datasets, such as its small size in terms of train/test samples and the limited number of object classes. COCO-QA contains 123,287 images, together with 78,736 train and 38,948 test questions obtained from COCO image descriptions [183]. Questions are divided in 4 types with varying numbers of questions in each of them: *Object*, *Number*, *Color* and *Location* questions.

KB-VQA [371] has been constructed in order to evaluate VQA models on questions that need visual information as well as external knowledge to explicitly infer the right answer. It includes images from COCO [183] containing approximately 150 object classes and 100 scene classes, question-answer pairs following pre-defined templates and question labels. The questions involved in KB-VQA are divided in three categories: visual questions can be answered by extracting information from the image (such as *Is there a dog in this image?*); common-sense questions rely on external knowledge contained in commonsense knowledge bases (*How many road vehicles are in this image?*); finally, KB-knowledge questions require information form Wikipedia or similar sources (*When was the home appliance in this image invented?*).

Factual VQA (FVQA) [360] is a dataset addressing factual VQA, based on images sampled from COCO [183] and ImageNet [61] which form three types of visual content (object, scene and action classes), together with structured visual-related knowledge extracted from DBpedia [14], ConceptNet [317] and WebChild [330]. All this information is stored in a graph of RDF triplets. Annotators construct questions and answers which require both selected visual content and associated facts. In total, FVQA contains 2,190 images of 326 object classes and 221 scene classes, 5,826 questions of 32 categories, which correspond to 4,216 unique facts.

Knowledge-aware VQA (KVQA) [301] targets world knowledge-aware VQA by filling the gap of named entities knowledge. It relies on knowledge present in Wikidata [354] KG, resulting in 183k question-answer pairs which involve more than 18k named entities and 24k images.

Outside-knowledge VQA (OK-VQA) [221] contains more than 14k diverse and difficult questions of 10 mutually exclusive categories which cannot be answered without external knowledge. More than 14k images were sampled and filtered from COCO [183]. In contrast with previous related works, OK-VQA does not consult a fixed knowledge graph to guide answer prediction, but dynamically recognizes what knowledge is needed, either structured or unstructured.

Text-KVQA [311] is a very large dataset addressing scene-text recognition for knowledge-enabled VQA. It contains images from book covers [128] and movie posters [233], as well as Google scraped images of 1000 business brands. All images are evaluated to ensure they contain scene text relevant to the content. Knowledge bases corresponding to each of those 3 scene types were constructed based on Wikidata [354] for business scenes, IMDb [125] for movie posters and [128] for book covers. The train/validation/test splits enable zero-shot capabilities, as there is no entity overlap between them. The supporting facts are not tied with their corresponding entities, but instead are dynamically mined from the knowledge bases.

Visual7W+KB [398] is an extension of the Visual7W [424] test split which further contains knowledge-based visual questions guided from ConceptNet [317]. However, the dataset is not tied with a specific knowledge graph, even though ConceptNet is indeed preferred in practice. In total, it consists of 16,850 open-domain question-answer pairs and 8,425 images from Visual Genome [155]. The questions belong to 7 categories (*what*, *where*, *when*, *who*, *why*, *which* and *how*), while the answers are provided in multiple-choice format.

One of the major challenges in knowledge-enhanced VQA is that questions should encourage exploitation of all participating modalities, therefore data-related weaknesses arise in existing benchmarks. In the meanwhile, information leakage between train and test set answers often promotes guessing rather than reasoning. **S3VQA** [129] is a dataset designed to address these issues by including questions that require the use of a knowledge graph, along with visual and textual information from the image.

Zero-shot Fact VQA (ZS-F-VQA) [47] extends F-VQA [360] for zero-shot learning settings. It considers the image-question-answer triples whose answers belong among the 500 most frequent ones. The filtered dataset is split in train (seen) and test (unseen) triples which contain non-overlapping answers. In total, 5 splits of the original F-VQA dataset are performed, yielding on average 2,732 train and 2,760 test triples.

Art QUestion Answering (AQUA) [84] is a visual reasoning dataset for the art domain. There are many challenges tied with analyzing and reasoning over artworks. First, there are different levels of abstraction regarding common objects and entities, as many paintings deviate from realism. Therefore, recognizing objects and reasoning about them is much harder compared to scenes existing in most datasets. Moreover, domain knowledge regarding artists, art movements, historical periods and other cultural influences can only be recognized with the help of a knowledge source. This information also affects the interpretation of a painting. QA pairs are generated automatically based on paintings and descriptions of the SemArt [83] dataset, which form the knowledge source. In total, AQUA contains more than 69k QA training pairs after cleansing, from which around 29k pairs are visual and 40k pairs are knowledge oriented.

7.1.2 Methods

7.1.2.1 Keyword-based explicit KG querying

First attempts target the construction of a scalable multimodal knowledge base which aims to answer visual queries that require real-world knowledge. Image classes, attributes and actions are extracted from the images, forming logical rules. The knowledge base built upon those rules contains nodes of visual and textual entities, as well as edges of diverse types between the entities. [425] However, this constructed knowledge base remains limited to the visual information present specifically in the SUN [376] dataset. Most subsequent methods utilize already constructed large knowledge bases, targeting a broader range of concepts, commonsense knowledge and more complex questions to be answered.

Towards this direction, early approaches focus on handling open ended questions regarding contents of a scene with the assistance of provided external knowledge. Attributes extracted from images using a fine-tuned VGG-16 [310] model act as SPARQL queries to knowledge bases such as DBPedia [14], and contribute to caption generation. Retrieved knowledge embedded via Doc2Vec [165], together with attributes and LSTM-based caption representations are fed in another LSTM model which generates the final answer. [371] An improvement of this version followed in [370], extending the framework to two more datasets, namely DAQURA-ALL [217] and its reduced version DAQUAR-REDUCED.

Even from early works in knowledge-enhanced VQA, explainability is recognized as an important aspect in understanding how a model learns from visual content and external knowledge to reach a conclusion. Therefore, Wang et al [359] developed a knowledge-enhanced VQA framework that provides the reasoning path from which the answer is inferred. Objects are detected using Fast-RCNN [90] object detectors trained on ImageNet [61] and MS-COCO [183], scene classes are extracted from a VGG-16 [310] pre-trained on MIT-Places [**mit_places**], and scene attributes are captured via a VGG-16 pre-trained on ImageNet and fine-tuned on MS-COCO. All those visual concepts form RDF triples and are linked with corresponding DBPedia [14] entities. Questions are parsed so that key-phrases are extracted and mapped to the knowledge base entities. The same work introduced the KB-VQA dataset.

Consequent works further proceed towards avoiding SPARQL queries on the knowledge graph, but rather fully utilize embedding representations for fact selection and reasoning to provide an answer.

7.1.2.2 Sequential language models for question encoding

Rather than relying on SPARQL querying from plain keyword extraction, leveraging vector representations of involved modalities forms the foundation for improved performance and state-of-the-art results in knowledge-enhanced VQA. Initially, fact ranking based on embedding similarity metrics paved the path for successful approaches, upon which graph neural network (GNN) reasoning further advanced the contribution of external knowledge and overall performance.

Embedding-based fact retrieval from KG Traditional VQA models utilizing RNNs for language encoding, focus on learning a question-answer mapping. Due to the limited and opaque reasoning capabilities of this approach over diverse answers, a more scalable solution that involves learning the mapping between questions and KB-queries using LSTMs was proposed. This new approach is explainable, as the fact connecting a question and an answer reveals the reasoning procedure. [360]

Projecting question-image pairs and facts on a common embedding space poses advantages over previous approaches, such as extendability to different knowledge bases and error elimination by avoiding explicit querying. Images and questions are embedded using CNNs (for objects, scenes and actions) and an LSTM respectively, and they are projected in a common space using a multi-layer perceptron. Another LSTM is used to retrieve facts from the knowledge base, which are then encoded in GloVe [260] vectors. The dot similarity between the question-image representation and fact embeddings provides a fact ranking, from which the final answer is inferred. [238]

Narasimhan et al [237], building upon [238], argue that considering muiltiple relevant facts instead of a single top-ranked fact at a time leads to better generalization. During the fact retrieval stage, a subset of highly-relevant facts is obtained with the help of LSTM-extracted question embeddings. In the answer prediction stage, each node is represented by concatenating the selected entity representation from the previous stage, visual features from the image and the question embedding. The subgraph formed from all the relevant facts is jointly assessed by a graph convolutional network (GCN) [149], followed by a multi-layer perceptron that decides whether each entity constitutes the final answer or not.

Based on the KVQA dataset, a memory network (memNet) framework sets the baseline for VQA enhanced with knowledge of named entities. Specifically, entities extracted from the question and the image are used to obtain facts from Wikidata [354] knowledge graph. Retrieved facts together with corresponding entity coordinates from the image are used to produce memory embeddings via a BiLSTM network, and a similar procedure is followed for the question embeddings. Both representations contribute to the final answer, which is defined by a multi-layer perceptron. [301]

Text present on an image can provide further information towards inferring the correct answer. Extracted text and image areas are fused together with the given question to retrieve relevant facts during the fusion stage, and a multi-relational graph is constructed based on all those components. The text recognition part relies on word proposals assisted by the knowledge graphs accompanying the text-KVQA dataset. Scene proposals were created with the help of Places dataset for scene recognition [421] and a fine-tuned VGG-16 [310]. A gated graph neural network (GGNN) performs one-hop reasoning on this graph to derive the final answer. [311]

Multimodal graphs Unexpected noise in the answer inference process can be attributed to the absence of detailed selection of information during modalities fusion. Considering multiple views of the same image offers a new perspective that is closer to human cognition. Multiple knowledge graphs provide visual, semantic and factual information derived from corresponding images, text and facts respectively, while the visual and the semantic graph can be considered as instances of the factual graph. Intra-modal graph convolutions focus on the most relevant parts of each modality. Consequently, cross-modal knowledge reasoning on the fact graph iteratively aggregates information from the visual and semantic graphs using a recurrent module, and after multi-step reasoning the multimodal knowledge is fused in each entity. The final answer is returned by applying a GCN [149] over those entities. This approach offers interpretability by revealing the entity and the modality graph which contributed to the answer. [398]

A similar approach to [398] utilizes visual, semantic and factual graphs for image representation to eliminate noise during multimodal fusion. The Multi-Modal Heterogeneous Graph Construction stage is responsible for constructing those modality graphs, followed by the Cross-Modal Heterogeneous Graph Reasoning which selects intra-modal knowledge and then performs cross-modal reasoning. Information relevant to the question is extracted from the three graphs via a modality-aware heterogeneous GCN. Cross-modal convolutions define complementary relevant information transmitted from visual and semantic graphs to the fact graph. The final answer is returned after reasoning over aggregated fact information. [426]

A major challenge in knowledge-enhanced multimodal tasks is its supervised nature, as a possible absence of ground truth facts may hinder the inference of a proper answer in several approaches. A local subgraph is constructed based on concepts present in the image and the question, aiming to bridge the gap between question-image context and external knowledge. Those subgraph concepts act as anchor points to a knowledge graph, such as ConceptNet [317] and Wikidata [354], enabling the expansion to their immediate neighbors. Moreover, a global subgraph is constructed in a similar fashion for all the candidate answers. In each subgraph, the information of neighboring nodes is aggregated to produce embeddings of the anchor concepts, and their similarity to the query embeddings drive the final answer. [170]

Multiple feature spaces Addressing the zero-shot setting of knowledge-enhanced VQA, a KG can help capturing semantics outside training data. Multiple feature spaces are used for independent alignment between image/question input and KG entities. The semantic space focuses on the linguistic information of the input (image, question) pair, representing a feature space of relationships; the object space acts as a support entity feature space, capturing visual and textual salient features; finally, knowledge space is dedicated to answer representation. [47]

7.1.2.3 Transformer-based models

Transformer based approaches form end-to-end architectures that utilize single-modality or joint representations rather than creating queries to knowledge bases and then injecting the retrieved entities. Thus, we can classify the transformer-based approaches in two categories: the first includes architectures which use transformer architectures for text encoding, while the second utilizes multimodal transformers to jointly encode vision and language.

Transformer architectures for language encoding Similarly to [398, 426], the usage of dedicated graphs for different modalities is also followed in [427], attempting to represent relationships between visual objects and semantic entities present in a scene graph and a knowledge graph respectively. The scene graph is constructed from visual and question embeddings which form the graph nodes and relationships. In the meanwhile, joint image and question embeddings select the most relevant knowledge graph node embeddings to construct the concept graph. Both image-question and knowledge representations are obtained via pre-trained language models for sentence similarity, such as sentence-BERT (SBERT) [282] and Universal Sentence Encoder (USE) [37]. The most relevant nodes of both scene and concept graphs are selected via a

Graph Attention Network (GAT) [351], which decides upon the edge weights with respect to the question. A joint embedding incorporates the question embeddings together with the scene graph and knowledge graph outputs.

Even though explicit and structured knowledge bases constitute the majority of the approaches analyzed so far, implicit and unstructured knowledge can also boost the VQA task. GPT-3 [31] can retrieve knowledge based on text prompts and effectively reason over it in a few-shot manner: no fine-tuning is required and instead only a few examples during inference are provided. Captions are first extracted from images using VinVL [409] to form GPT-3 inputs. Regarding sample selection for the few-shot inference stage, both improving the quality and increasing the number of samples have been explored. The top-n most similar prompt examples comparing to the inference-time question to be answered are defined by CLIP [273], thus maximizing sample relevance. On the other hand, multiple queries corresponding to one inference-time example can be used to retrieve answers from GPT-3 using n example prompts each time, and their ensembling results in the final answer. Contrary to most works on knowledge-enhanced VQA, inferring the answer is a generative task and not a discriminative one among pre-selected candidate answers or graph nodes. [389]

Unimodal pre-trained transformers yield better generalization capabilities over multimodal approaches of comparable size when external knowledge is necessary. Language models employed for the knowledge-enhanced VQA task are sufficient even to compensate for the limitations of image captioning models, which often fail to fully capture visual semantics. To this end, a pre-trained image captioning system, in this case the multi-task OSCAR [176] transformer, is used to extract linguistic information from an image, while a language model such as BERT, acting as an implicit knowledge source, receives the caption and the question to infer an answer. Moreover, text-only and multimodal approaches have complementary capabilities, therefore their combination can yield even more powerful models. [293]

VIKING is a framework accompanying AQUA dataset [84] for VQA on the artistic domain. As questions in AQUA may either be visual or knowledge-oriented, a modality selector first decides the right category by receiving the encoded image and question. Visual-oriented questions do not require external knowledge in order to be answered. For knowledge-oriented questions, a two-stage fact retrieval strategy is followed, pairing the given question with the most relevant painting description, which corresponds to the external knowledge fact needed. The first stage utilizes TF-IDF to rank descriptions according to the question, while in the second stage re-ranking is performed using BERT. Finally, a fine-tuned XLNet [390] model provides the final answer.

Joint multimodal encoding with attention-based fusion A slightly different technique is followed in [420], where the visual modality is not captioned, but fused with the BERT-embedded question. More specifically, a knowledge graph for artistic VQA is construcyed based on YAGO [331] KG and AQUA [84] dataset. A Hierarchical-Knowledge Embedding module is responsible of retrieving relevant relationships r from the KG, which can form a *(h, r, t)* triple with question related entities serving as the head h of the triple, and answer related entities as the tail t. A Network-Based Representation Learning module extracts visual and textual features and fuses them together in order to obtain a VL representation. The fusion part first applies local attention per modality, and then global attention on both text and image, where locally 'attended' text features form the query, and visual features the key and value. Query, key and value are inserted in a multi-head attention unit which further promotes the joint representation to consequent layers, until a global representation is obtained. Then, a Knowledge-Based Representation Learning module injects hierarchical-knowledge embeddings to the network-based representation. This representation is inserted into a relational module, which performs meta-training: a representation learned on the training data is transferred to a support set of disjoint labels. Finally, the relational module derives the answer.

Joint multimodal encoding with VL transformers ConceptBERT is one of the first attempts towards the end-to-end transformer-based direction, where all modalities are jointly exploited for learning. The first step includes obtaining representations for each individual modality. Visual features are extracted using a pre-trained Faster R-CNN network [284] and BERT [63] provides the question representation. ConceptNet [317] acts as the commonsense knowledge source, and is encoded using the ConceptNet embedding [215] method, a Graph Convolutional Network [149] variant that relies on message passing from node to node in order to obtain the ConceptNet graph representation. Two modules receive the embedded inputs: A vision-language module consisting of two streams in a ViLBERT [194] fashion, and a concept-language module based on the bidirectional Transformer architecture are proposed to model the interactions between the relevant modalities. Both outputs of these modules are joined to form a concept-vision-language representation, which finally concludes to the answer via a classifier. [85]

Knowledge obtained from the web according to the given question and respective answer can act as a large external implicit knowledge source for the OK-VQA dataset, covering knowledge 'gaps' in several domains without manual human effort. The proposed weakly-supervised framework consists of two phases: the first one (Retriever) retrieves relevant knowledge, which guides answer prediction in the second stage (Reader). Two different approaches are followed for representing the question-image pair inputs: either the question and image are encoded using an LXMERT [329] transformer, resulting in a multimodal representation, or the image caption and the question are encoded via BERT [63], leading to an exclusively linguistic representation. The BERT-based linguistic encoding can contribute to both a neural based retriever and a term based [287] retriever. A similarity score defines the relevant knowledge for the neural based retriever, which is further concatenated with the question representation, and consequently with the image using again an LXMERT model. [199]

The KRISP framework addresses the scenario where essential external knowledge is absent during training, as well as test time. Both implicit and explicit knowledge sources are utilized: Explicit knowledge combines DBPedia [14], ConceptNet [317], VisualGenome [155] and hasPart KB [25] in a KG of 36k edges and 8k nodes after filtering out irrelevant concepts, while pre-training using BERT can offer implicit knowledge. Explicit visual symbols from images are extracted to constrain the KG entities corresponding to image-related concepts, including objects, parts of objects, attributes and places. Likewise, symbols are extracted from question to contribute to the formation of a graph for all explicit symbols. A Relational Graph Convolutional Network (RGCN) is used for graph representation, allowing dedicated processing for different edge types and directions. After reasoning, a symbolic prediction vector is returned. Regarding the implicit information stream, a multimodal BERT (MMBERT) model incorporates visual and textual embeddings to produce an implicit prediction vector. Finally, the top-ranked prediction from both vectors defines the answer. [220]

The presence of scene text can offer valuable information for properly predicting the correct answer. The detected text, among with the image, the relevant knowledge from Google Knowledge Base (GKB) and the question representation are fed into a multimodal transformer, enabling interaction through attention mechanisms between the different modalities. The OCR-extracted text acts as a query to GKB to retrieve candidate entities, which are then disambiguated based on the visual context. External knowledge not only boosts the understanding of scene-text even in unseen instances, but also tackles biases present in training data. [64]

While employing an abundance of knowledge sources is capable of covering more visual topics, a lot of noise may be introduced, as more irrelevant information is retrieved. MAVEx utilizes multi-granular queries to retrieve external knowledge with the purpose of validating and correcting predicted answers among suitable candidates with the help of various knowledge sources. Specifically, a finetuned ViLBERT [194] model creates a pool of candidate answers, and together with the corresponding question, extracted keywords and phrases from the question and the possible answers are utilized to query external knowledge. Wikipedia, ConceptNet and Google images act as knowledge sources regarding different views of knowledge. Finally, retrieved knowledge instances are matched with the queries to acquire the highest ranked supporting fact, which returns the degree of agreement with respect to candidate answers, guiding decision towards the most trustworthy knowledge source. [369]

Passage retrieval can serve as an answer selection technique to VQA instead of choosing among pre-defined candidate answers. Both sparse and dense retrieval are investigated. For sparse retrieval, given a question and an image, visual clues such as object names and captions are extracted from the image, and BM25 is used to return the k most relevant passages. For dense retrieval, questions and images are jointly encoded in dense vectors using LXMERT [329]. In any case, retrieved external knowledge can be integrated dynamically from diverse and generic sources, without using a fixed knowledge base. As positive passages are considered the ones containing exactly the ground truth answer. LXMERT is used to encode the question and the image jointly, while BERT encodes the passage. Finally, dot similarity between them defines the k most relevant passages. [271]

Based on the E-BERT [265] strategy of knowledge injection without expensive re-training, LXMERT-encoded linguistic input is modified to incorporate factual knowledge from Wikipedia by aligning Wikipedia2Vec embeddings with BERT WordPiece [373] vectors. No other change is required within the language encoder's architecture, while the visual encoder remains entirely intact. Only fine-tuning is required for achieving advanced accuracy due to knowledge injection. In the meanwhile, explainability regarding visual and textual modalities is also enhanced. For this purpose, BM-GAE [41] is employed to extract visual and token explanations, helping the identification of parts to which knowledge injection was helpful. [126]

S3, presented with the S3VQA dataset, targets to answer visual question based on all the participating modalities simultaneously. Entity spans from the question are selected to be matched with objects of scene graphs corresponding to images. This matching can be often guided by external knowledge sources, which enable answering more complex questions that require multi-hop reasoning. BERT identifies those appropriate question spans, while object detectors propose the objects that most likely fill the spans. Wordnet [226] synsets are mapped to the objects, and their hierarchical positions are represented via structural embedding methods. Finally, Google search is used to retrieved the top results of the enriched question representation. Alternatively, the answer can be provided via classification of possible candidate answers. [129]

7.1.3 Evaluation

Classification/Ranking metrics are widely used in K-VQA, following the paradigm of knowledge-free VQA, with most works relying on the **top-1 accuracy** metric for comparison. Accuracy is further decomposed to explain the contribution of subcomponents in many works: object, counting, color, and location accuracies, as well as accuracy per question type are reported [371, 359]. Other accuracy reportings include performance as per selected knowledge source, per visual concept and per answer source (image or knowledge) [360]; the individual accuracies of involved stages of the reasoning process, which together contribute to answer prediction [238, 237]; accuracies as per question category [301, 398, 427, 85]. Moreover, **precision@k** and **recall@k** have appeared in fewer works, as well as ranking metrics such as **MRR**. Some early works [371, 360] utilize **WUPS** with typically used thresholds of 0.0 and 0.9.

Human evaluation provides further insights in comparison to to solid accuracy-based scores that fail to fully describe the success or the shortcomings of a metric. Therefore, many authors propose human evaluation experiments to grade the model's response comparing to human perception, and count the number of agreement instances over all results [359, 360, 237].

Evaluation of reasoning paths is followed by employing human judgement, as being one of the most trustworthy indicators [359]. Due to the transparent reasoning process, failure cases can be traced down, revealing the exact stage where the prediction deviated from the intended one. Thus, shortcomings can be attributed to architectural choices, encoding techniques, or even incorrect data annotations. Other metrics regarding explainable reasoning include **top-k fact retrieval accuracy** for different k values, a crucial step for returning the correct answer in several approaches [360, 238, 237, 426]. **Fact recall** can also assess the fraction of relevant facts retrieved for a given question [311].

Generally, benchmarking knowledge-enhanced VQA approaches is not trivial. The plethora of combinations between available knowledge-based datasets and external knowledge sources is rather large compared to the number of available implementations in the field. Additionally, various choice of metrics in literature makes comparisons of model performance even harder. Only recently implementations started becoming more consistent, focusing on evaluating results with plain accuracy and leveraging OK-VQA as a widely used dataset.

7.2 Knowledge in Visual Reasoning (K-VR)

7.2.1 Datasets

High-order Visual Question Reasoning (HVQR) [36] is a knowledge-based dataset endorsing interpretable visual reasoning using commonsense knowledge. Given an image and a question, an answer is inferred, as well as a reasoning path as explanation. Even though this is similar to rationales used in knowledge-free datasets for VCR [403], in fact the format of HVQR explanations differ. Instead of textual rationales, rules for the whole reasoning path are returned, combining visual and knowledge-oriented triples, derived from the scene graph and the commonsense knowledge graph respectively. HVQR contains questions that require multi-step reasoning to infer an answer. Moreover, each knowledge triplet appears only once per question, in order to avoid frequency-based biases. An evaluation scheme validates each step of the reasoning process based on the commonsense and scene graphs provided. More than 157k QA pairs comprise the dataset, together with approximately 32k images and corresponding scene graphs from Visual Genome [155]. Based on the reasoning steps required for the answer, first-order and second-order questions can be recognized, corresponding to 68,448 and 88,753 questions respectively. Another split defines 87k KB-related questions and 70k KB-not-related questions. Additionally, 193,449 facts from WebChild [330], ConceptNet [317], and DBpedia [14] formulate the knowledge base. Scene graphs per image are combined with related entities from the knowledge base, constituting image-specific knowledge graphs.

Compositional Language and Elementary Visual Reasoning - CLEVR [139] is a synthetic dataset of 3D objects which contain annotations regarding their position and attributes. Those attributes describe the size *(small, large)*, color *(red, brown, yellow, green, blue, cyan, purple, gray)*, shape *(cube, cylinder, sphere)* and material *(rubber, metallic)* of each object. Positions can belong in 4 types, namely *left*, *right*, *behind*, *in front*. Highly compositional questions form 5 question categories: *Exist*, *Count*, *Compare Integer (equal, less, greater)*, *Query Attribute (size, color, material, shape)* and *Compare Attribute (size, color, material, shape)*. Moreover, CLEVR contains 90 question families following different program templates, as well as text templates, so that natural language questions can be derived. The questions are translated in natural language by filling the template with template parameters. CLEVR is not connected to some external knowledge, although due to the limited semantics and the nature of the task they target

CLEVR CoGenT is a benchmark derived from CLEVR [139] that assesses the ability to capture unseen combinations of attributes during testing, thus showcasing a model's generalization capabilities.

Other datasets associated with visual reasoning tasks that have not yet participated in knowledge-enhanced endeavors are the **Kandinsky Patterns**, **NLVR**, **NLVR2**, **Winoground**, **WinoGAViL**. **Kandinsky Patterns** [235] is another synthetic dataset containing geometrical shapes in the 2D space, characterized by their shape, color and position. A sequence of images within the Kandinsky Patterns dataset is formed based on descriptions and rules regarding the participating objects, i.e. *the Kandinsky Figure has two pairs of objects with the same shape, in one pair the objects have the same color, in the other pair different colors, two pairs are always disjunct, i.e. they don't share objects.* **NLVR** [323] contains statements such as *There are two towers with the same height but their base is not the same in color*, again requiring reasoning over image sequences that contain 2D geometrical shapes of certain colors, shapes and positions. **NLVR2** [324] expands the complexity of images while maintaining the problem formulation introduced in NLVR. Compositional reasoning is evaluated in the recently introduced **Winoground** dataset [335], which contains pairs of images and pairs of captions and requests proper image-text matching; the challenge is that captions contain the same words and word order defines the correct image matching e.g. *some plants surrounding a lightbulb* vs *a lightbulb surrounding some plants.* Multiple reasoning skills are evaluated in the **WinoGAViL** dataset [26], where a human player sets a linguistic association cue, while an AI rival attempts to select the most appropriate images to the cue. A more thorough investigation of visual reasoning datasets is provided in [112].

7.2.2 Methods

7.2.2.1 Sequential language models

External knowledge KM-net (Knowledge-routed Modular Network) is introduced in the same work with the HVQR dataset [36], addressing multi-step (compositional) reasoning using visual and commonsense knowledge. Each question is decomposed into consecutive subqueries via LSTM encoder-decoder schemes, passed to a visual reasoning module and a commonsense reasoning module to extract different types of knowledge accordingly. The subqueries form a query layout, i.e. a tree structure revealing the relationships of subqueries, with leaf nodes belonging to distinct words of the queries. A bottom-up attention R-CNN provides visual features for the image. The subqueries are processed sequentially starting from the most specific ones, driven by the KM-net reasoning module. First, the knowledge reasoning module receives question entities from the query layout and returns the most probable candidate entities from the knowledge base. Then, the visual reasoning module receives entities from the scene graph, together with the candidate entities of the knowledge module and fuses the candidate entities, image features and query embedding to derive the answer.

Internal knowledge Self-knowledge includes the usage or construction of a scene graph based on the detected objects, relationships and attributes. At the same time, the question can be parsed in a structured program via an LSTM, producing subqueries that form a tree structure. Given those two graph representations, an encoding is derived for the query. Node attention and edge attention are calculated based on the query embedding. Combining a node attention vector and an edge attention matrix, new objects can inferred due to the graph structure; basically starting from the attended node vector and traversing over an attended edge, a new node vector will be provided. The same procedure can be followed for all subqueries, respecting the structure of the query tree. Logically relating subqueries, results in logical operations (such as ***and***, ***or***, ***not***) over attended scene graphs. Finally, based on the question type and the final scene graph, the answer can be provided. [308]

7.2.3 Evaluation

Classification metrics are commonly used for benchmarking, with **answer accuracy** providing a general measure of performance [36, 308]. Compositional commonsense reasoning heavily relies on the evaluation of reasoning paths that provide the final answer [36]. The accuracy score is further decomposed to present *KB-related* and *KB-not-related* accuracies depending of the need for external knowledge; those can be broken down to *first-order* and *second-order* accuracies, regarding the number of reasoning steps required; finally, a more fine grained categorization provides *question-type accuracy*, based on the template the query components follow.

Ranking metrics such as **average recall** are used to evaluate the retrieval success of supporting facts for explanations. Average recall is further decomposed to *KB-related* and *KB-not-related* fact recall.

7.3 Knowledge in Visual Commonsense Reasoning (K-VCR)

Various external knowledge sources rich in information can provide insights of unseen concepts that humans would effortlessly infer from the information provided in a scene. This missing commonsense knowledge is able to guide answer explanation towards the right rationale, revealing if a more accurate reasoning process is followed by VCR models when knowledge is added.

7.3.1 Datasets

Visual Commonsense Reasoning (VCR) [403] is the dataset which introduced the task and serves both knowledge-free and knowledge-enhanced versions of VCR. It contains 110k unique images from movie scenes, 290k multiple choice challenging questions, with 290k correct answers and rationales. Images contain annotations which are anchored over questions, answers and rationales. The technique of adversarial matching is chosen for the answers in order to minimize biases; each correct answer appears four times in the whole dataset, once as a positive answer and three times as negative answer. Therefore, a VCR model will not favor more frequently appearing answers which would endorse guessing rather than reasoning. The questions are classified in non-mutually exclusive categories according to their purpose, with categories being Explanations (*Why is [person11] wearing sunglasses inside?*), Activity (*What are [person1] and [person2] doing?*), Temporal (*What will [person6] do after unpacking the groceries?*), Mental (*What is [person3] thinking while [person5] shakes his hand?*), Role (*What is [person1]'s relation to [person4]?*), Scene (*Where is [person1] now?*), Hypothetical (*What would happen if [person3] fell asleep?*). This dataset originally is knowledge-free, therefore not necessarily requiring external knowledge, nor is it associated with any knowledge base. Nevertheless, the questions existing in the various VCR question categories can be greatly benefited from the introduction of external knowledge sources which can explicitly incorporate various knowledge senses.

Visual Commonsense Graphs (VCG) [256] is a large-scale dataset that provides information regarding temporal commonsense relationships, such as *what may have happened before*, *what may happen in the near future* and *what are the intents of the people present* based on static images. In total, it contains more than 59k images and more than 139k textual descriptions of events at present. Additionally, around 295k intents at present, as well as more than 584k events before and 586k events after complete the dataset, resulting in more than 1,4 million commonsense inferences. People and locations appearing in the images are grounded with their mentions in the textual descriptions.

7.3.2 Methods

7.3.2.1 Transformer-based models

External knowledge Some of the first knowledge-enhanced transformer-based attempts step upon BERT [63] to introduce knowledge-vision-language (KVL) learning as an instance of multimodal learning. In the KVL-BERT architecture [315], ConceptNet [317] is leveraged to enrich sentences with relevant commonsense information. The knowledge-enriched linguistic input is inserted in a BERT-like multimodal transformer. The preservation of semantic structure is achieved by using relative position embeddings. However, injected information should be only visible to their corresponding textual entities of the sentence and not to other tokens or visual features, a need that is satisfied via a 'weakening' visible matrix. Moreover, it is possible that different enriched textual tokens in the sentence share the same relative position embeddings, which would make unrelated tokens obtain high self-attention scores, implying that they are related. This contradiction is resolved by imposing a mask-self-attention mechanism via the visible matrix, with the purpose to restrict the area a token can attend. After those treatments, the input is in a form suitable to be fed in a VL transformer, in this case VL-BERT [322]. It was observed that KVL-BERT outperforms its multi-task baselines, as well as models dedicated to the VCR task, even though it cannot trespass the performance of knowledge-free VL transformers that invest on additional pre-training.

A somehow different strategy is employed in the case of Vision–Language–Knowledge Co-Embedding (ViLaKC) [166]: the three modalities are first embedded independently and afterwards are fused together. Initially, a Knowledge Extraction Module (KEM) retrieves relevant knowledge from ConceptNet based on concepts appearing on the image, question and candidate answers. The encoding of modalities is performed in the two-stage VLKEM module: first, the independent modality encoding embeds images using ResNet [104], language using BERT [63] and knowledge using GCN [149]. The second stage consists of the co-embedding sub-module which aligns and integrates the three vectors via a multi-head self-attention mechanism. The co-embedder is pre-trained in two phases, the first being task-agnostic, such as in several VL transformer models, and the second task-specific, utilizing significantly less data (∼200k samples) coming from all three modalities. The task-specific pre-training stage introduces novel pre-training tasks, such as masked language modeling with image and knowledge (MLMIK), masked object classification with text and knowledge (MOCTK), and vision-language-knowledge matching (VLKM), in order to enforce co-learning. This joint embedding is then inserted in an answer determination module (ADM) consisting of a fully connected layer followed by a softmax.

The CKRM framework [367] consists of two stages, the first used for knowledge retrieval and the second one for reasoning. SWAG [404], a dataset containing pairs of events which describe a situation (context) and possible endings, serves as the commonsense knowledge source, aiming to transfer knowledge regarding everyday situations to the target task of VCR. A source and a task encoder are responsible of receiving (*context, ending*) pairs and (*question, answer*) pairs respectively to perform knowledge transfer in different granularity layers. The encoders first use BERT [63] followed by a BiLSTM structure to model temporal interactions of words. Cell-level knowledge transfer refers to the most fine-grained information fusion from source to target task, with layer-level and attention-level knowledge corresponding to coarser aspects of information. This strategy offers acquiring knowledge from various perspectives for a more enriched representation. The knowledge-based reasoning module incorporates the multi-level knowledge from the previous stage together with visual features in the Knowledge-enriched visual attention module. Finally, a reasoning composition module combines all aspects of knowledge derived from the multi-level transfer procedure and the enriched visual representations to derive the answer.

7.3.3 Evaluation

Classification metrics, especially **classification accuracy** is employed for evaluating K-VCR results that follow the multiple-choice format for answers (A) and rationales (R). Accuracy is decomposed by evaluating independently each of the following aspects:

1. $Q \longrightarrow A$: given a question Q, choose as A one of the 4 available answers and compare if it matches the real answer or not.
2. $QA \longrightarrow R$: given a question Q and the correct answer A, select as R one out of the 4 rationales and compare if it matches the real rationale or not.
3. $Q \longrightarrow AR$: given a question Q select as A one of the 4 answers, and depending on selected A choose one of the 4 rationales. The result is regarded to be correct if and only if both right A and R are chosen.

7.4 Knowledge in Image Captioning (K-IC)

7.4.1 Datasets

There are no dedicated knowledge-enhanced or knowledge demanding datasets for K-IC. Knowledge-enhanced models are currently using **COCO captions** [183] as described in Section **??**. Moreover, **Flickr30k** [396], a dataset containing 31,783 scene images accompanied by 5 human-annotated sentences each is widely employed for K-IC.

7.4.2 Methods

7.4.2.1 Sequential language models

External knowledge First attempts for knowledge-enhanced image captioning propose the extension of existing implementations by injecting commonsense knowledge from external sources. Specifically, the backbone image captioning architecture extracts visual features from images via a CNN, which are then inserted to an LSTM to generate a knowledge-free answer. To enhance this baseline with knowledge, objects extracted from the image are used as queries to ConceptNet [317]. Related ConceptNet entities, either regarding individual objects (direct terms) or the remaining image areas (indirect terms), are fed to a pre-trained LSTM which provides semantic representations for each of those two points of view. Then, visual features, direct term representations and indirect term representations are concatenated to form the initial state of another LSTM model, which finally generates the knowledge-enhanced caption [422].

Both visual and commonsense knowledge for image captioning are used in [113]. The first step includes dense region sampling from images in order to acquire visual and knowledge mappings. Dense visual feature extraction includes the definition of candidate regions, which are clustered together to provide a more concrete representation: the cluster center points for each dense region cluster serve as the corresponding visual feature. Consequently, the knowledge mapping receives visual features and knowledge embedding vectors from Visual Genome [155] and returns a knowledge-related representation per region cluster. Both visual and knowledge embeddings resulting from the two mapping procedures are concatenated and then inserted to a commonsense reasoning module. This module projects the two inputs on the same semantic space, from which a semantic graph is constructed under the guidance of commonsense knowledge. In the relational reasoning module, a GCN [149] operates on the semantic graph to obtain relation-aware node features. Finally, a LSTM receiving the knowledge-aware node embeddings as inputs generates the caption.

Inferring words not appearing in the image remains a challenge in image captioning, as there is no guidance regarding how those unseen words should be inferred to be used in captions. Such unmatched elements can be solved with internal self-knowledge based on more fine-grained alignments between individual words and image regions, which is achieved by attention mechanisms, and external commonsense knowledge to capture implicit information that cannot be derived from the existing data. Objects detected on the image are used to retrieve knowledge from ConceptNet [317]. Region features extracted from the image via a region proposal network and word-level attention on the sentence part co-operate towards attending to the most salient features of the image. This visual attention guided by language attention, together with the corresponding word embedding are inserted in a LSTM, which feedbacks each previous hidden state to update the word-level attention signal that contributes to the visual attention in every round. The external knowledge is incorporated in a later stage, when the answer is generated; therefore, it can tune the probabilities of LSTM-generated words to be added in the sentence towards more meaningful results. A reinforcement learning training strategy is followed by setting the LSTM as an agent, the words and visual features as the environment, and the generation of the best next word from the captioning model as the policy. [116]

Even though local information is well-represented based on detected objects, image captioning is generally not interpretable and therefore not explicitly controllable. An external knowledge source can help in grounding detected objects with semantic entities from the graph, which in turn provides enriched semantic labels for the objects present in the image. In order to control objects appearing in the caption, an attention-based human-interpretable mask is introduced, which assists in diverse caption generation. This masked can be dynamically tuned by a human to influence the resulting caption. [1]

Off-the-self object detectors have served several image captioning architectures. However, some tough situations, such as very small objects, occlusion or rare object classes can result in error propagation and negatively impact all consequent components until the final caption generation. Commonsense constrains and semantic correlations extracted from Visual Genome [155] can act as priors to guide a more accurate representation. A semantic graph is constructed upon extracted image regions, allowing GCN-based [149] reasoning. Specifically, visual semantics such as objects, attributes, relationships are captured by extracting candidate region proposals. CNN-based region features satisfy object and attribute representations, while features from regions union areas provide relationship representation. Visual features are projected on the same high-level semantic space as the knowledge embedding derived from Visual Genome. Therefore, knowledge-enhanced visual triplets are formed, respecting rules imposed by knowledge. The semantic graph is built upon those triples. Then, relational reasoning is performed on the semantic graph using a GCN, the output of which is inserted in the LSTM module that generates the answer. [114]

Internal knowledge Visiolinguistic priors are naturally connected with describing images, in the sense that humans logically infer unseen entities given a partial description of a visual situation. Obtaining such priors from existing images and captions is a way of 'creating' knowledge and facilitate reasoning of image captioning models without adding external sources.

Scene graph generation is a widely used technique for self-augmentation of information present in the dataset. Both images and text need to be represented within graph structures to bridge the two modalities. The Scene Graph Auto-Encoder (SGAE) framework utilizes this graph conversion to instill language priors into the encoder-decoder image captioning structure. More specifically, a learnable dictionary maps the relationships between a sentence and its corresponding scene graph iteratively, reconstructing the initial text from the generated graph in each round. For scene graph generation from text, a pre-trained scene graph parser is utilized, while for the reverse procedure, a trainable RNN decoder converts the dictionary back to text. During this procedure, the dictionary achieves to capture the necessary language prior to be transferred for captioning. The learned dictionary is then inserted to the image-involving pipeline: a scene graph parser converts the image to a scene graph, which is consequently passed to the dictionary encoded by a GCN [149]. Finally, the decoding of the dictionary provides the final caption. [388]

Attention mechanisms are able to identify such structured visiolinguistic priors and highlight connections between text and images, therefore augmenting image captioning implementations. Conditional Latent Topic Attention (CLTA) in combination with sentence priors are able to fuse the model with prior knowledge without the need for constructing scene graphs. Latent topic models are able to recognize semantically significant topics which are driving attention mechanisms to capture local and global dependencies in images. Thus, salient visual features emerge through words, and also more candidate salient regions are discovered and re-weighted accordingly, if they are associated with a topic contributing to an existing salient region. CLTA implements this re-weighting procedure to construct a context vector. Moreover, a sentence autoencoder acting as the sentence prior encourages the extraction of more context information and enhances generalization. Both the context vector and the sentence prior are inserted in an LSTM that generates the answer. [91]

7.4.2.2 Transformer-based models

Recent knowledge-enhanced image captioning models are implemented based on Transformers as an expected substitution of sequential models.

External knowledge Named entities and event knowledge have not been studied in previous image captioning works. This type of information is widely available in news articles, with raw sources being too complicated for language models to infer the right semantics. Special datasets are crafted for this purpose, providing an appropriate form of information for named entity/event-aware image captioning. The heart of the proposed method is the cross-modal entity matching, which incorporates information from various sources.

Sub-graphs are extracted from the image and the article text descriptions forming structured representations for the input. The nodes of the text sub-graph belong to named entities, and the edges to their in-between relationships, while the image sub-graph is more generic, by representing objects present in the image. The two sub-graphs are linked via similarity between image sub-graph objects and text sub-graph named entities in the cross-modal entity matching module. This module is trained with the help of multimodal external knowledge from Wikipedia. As a result, a multimodal knowledge graph is produced containing visual, textual and knowledge information. Embedding representations are obtained for each modality: a GAT module [351] produces a multimodal knowledge graph embedding, RoBERTa [190] encodes news captions and image features are derived from a pre-trained ResNet-152 [104]. An entity-aware captioning model receives the visual, textual and multimodal knowledge graph representations, feeding them to a Transformer [348] decoder to produce the caption. [419]

The BART transformer [168] can provide further advancements towards the refined task of *Visual Commonsense Generation (VCG)* [256] lying on the intersection of the generative image captioning task and the non-generative visual commonsense reasoning task. To this end, knowledge-enhanced Multimodal BART (KM-BART) was developed, able to incorporate both visual and linguistic information with the help of modality and task-relevant tokens in the transformer input. More specifically, task-relevant tokens are added in the beginning of the input sequence denoting the task type. For example, for VCG <before>, <after>, or <intent> tokens, representing temporal sequence of events (what happened before, what may happen next) and intents of people present in the image. Furthermore, the pre-training task of Knowledge-based Commonsense Generation (KCG) fuses commonsense knowledge from structured sources early in the pipeline, actually achieving in implicitly integrating explicit knowledge. Also Attribution Prediction (AP) and Relation Prediction (RP) pre-training tasks are used for the first time in knowledge-enhanced VL learning. COMET [30] is a transformer model trained on knowledge bases such as ATOMIC [123] and ConceptNet [317] that generates commonsense descriptions, and acts as a knowledge source for KM-BART. Two possible settings are examined for KM-BART, one containing the image and the event description (i.e. some textual information about the image that provides context of the depicted situation) and a harder one that omits the event description. [379]

Internal knowledge Transformer-based captioning poses some challenges, one of those attributed to the AR training procedure which is based on the maximum likelihood estimation (MLE). The main issue stemming from MLE is that when the generated sequence does not match the ground truth one, there is no discrimination between different 'failed' predictions. Therefore, words that are totally unrelated to the ground truth match are treated the same as semantically similar words. For this reason, a KL divergence term is added to weight semantic relationships between generated words, with respect to their ground truth match. Moreover, a knowledge graph is used to enrich the transformer input embeddings, infusing contextual information from neighboring entities in the graph. This knowledge graph is constructed from the linguistic information itself, by leveraging cosine similarity between embedded words to position them within a vector space. The original Transformer [348] architecture is used for the task, with image features word embeddings representing the visual modality. [415]

7.4.3 Evaluation

Language metrics such as **BLEU**, **ROUGE**, **METEOR**, **CIDEr** are used for evaluation as in most language generation tasks [422, 113, 91, 116]. SPICE is also used in [1, 415].

Human Evaluation can qualitatively evaluate generated sentences. The human evaluation experiment in [388] compares the quality of generated captions from different models according to the perception of 30 evaluators. Even though such an experiment is rather subjective, it indicates the importance of language priors. In [419] human preference is measured comparing with the previous best-performer in the *before, after, intent* generated sentences.

7.5 Knowledge in Visual Dialog (K-VD)

7.5.1 Datasets

VisDial [58] is a dataset used in both knowledge-free and knowledge-enhanced versions of VD. It consists of 133k dialogs and an equal number of images from COCO, with train and validation splits (125k dialogs) assigning 10-round dialogs -QA pairs- per image. In the test split (8k dialogs), random rounds are paired with each image. Some important aspects of this dataset is the presence of coreferences, endorsing the coherence of the conversation in linguistic level, and temporal continuity of topics, which supports the preservation and consistency of semantic meaning across the dialogs. The questions mostly follow a concrete and rather exploratory pattern: starting from asking about entities involved in COCO captions, then diving into details, trying to define a categorization of the whole scene or the most appropriate setting description, questioning about the weather of the scene, exploring key semantics not mentioned previously and finally validating and expanding the understanding of elements provided in the answers.

VisDialCK [411] is an extension of VisDial containing 940 history-required and commonsense-required dialogs.

7.5.2 Methods

7.5.2.1 Transformer-based models

Internal knowledge Only one knowledge-enhanced implementation for visual dialog has been introduced, inspired from the fact that commonsense related questions are typically ignored. A visual dialog model requires two necessary inputs: an image and the dialog history. Visual graphs have assisted the task by providing object relationships explicitly, even though this knowledge is not adequate for commonsense inferences. The integration of commonsense knowledge can be well-represented with graph-level facts and sentence-level facts. Then, facts from a commonsense knowledge graph such as ConceptNet [317] can be extracted based on the calculation of cosine similarity between their word embedding representation compared to embedding representations of the words in the sentences and the detected objects. Those graph level facts can complement entities from the visual graph. Therefore, an enriched vision-fact graph can be produced after individual graphs are purified by removing redundant information. The sentence-level facts are derived from the dialog sentences in the form of *(subject, relation, object)* triples, forming a graph structure. Similarly to the visual stream, the sentence graph is cleaned and enriched with commonsense knowledge. Finally, a transformer-based fusion module receives the enriched graphs, as well as the question embedding to provide the answer, exploiting a generative and a discriminative decoder. [411]

7.5.3 Evaluation

Ranking metrics such as NDCG, MRR, R@1, R@5, R@10, Mean position provide the quality of answer retrieval for visual dialog for both ***generative*** and ***discriminative*** answer prediction. [411]

Human evaluation is used in the ***generative*** setting of [411]. Specifically, two metrics are provided: the first one indicates the percentage of responses passing the Turing test, thus providing the amount of generated sentences that could be perceived as human-written; the second metric measures the number of generated responses that are perceived as of equal or better quality compared to specific human responses as baselines.

7.6 Knowledge in Visual Storytelling (K-VIST)

Visual Storytelling presents many situations where hypothetical concepts can be driven from commonsense and temporal reasoning. Unseen events can enrich or even be necessary for appropriate and coherent textual stories. For example, some sequential inferences were presented in the event/temporal knowledge analysis (in Section 3), providing knowledge such as *the boy dropped a glass of water and then the glass broke.* Not all concepts mentioned in this sentence may be explicitly apparent on a frame of the visual sequence. However, a KG can guide inference by searching for possible connections between concepts appearing on images, and thus acquire imaginary concepts.

7.6.1 Datasets

There are no dedicated datasets for K-VIST. On the contrary, relevant literature relies on datasets used for the knowledge-free version of the task, such as **VIST** [121]. This dataset contains more than 81k unique photos in around 20k sequences with corresponding textual stories. Textual stories follow a narrative style imposing more high-level inference capabilities compared to literal visual descriptions. This requirement is an extension against the majority of visual description tasks, which do not directly focus on sequential coherence and even abstract meanings. Two extra descriptions are provided per frame in order to bridge literal description with narratives: descriptions of images-in-isolation (DII) and images-in-sequence (DIS).

7.6.2 Methods

7.6.2.1 Sequential language models

External knowledge A two-stage structure was proposed in [387], consisting of a reasoning and a generation module. The vision-aware commonsense reasoning module is responsible of extracting the most relevant knowledge from an external knowledge base. Objects detected in all images of a sequence are fed in a GRU which provides a semantic and temporal representation. At the same time, candidate ConceptNet entities are fetched based on the detected objects. Attention modules finally select the most relevant ConceptNet candidates, which after passing through a GRU provide the final commonsense representation. The knowledge-augumented generation module receives the extracted commonsense knowledge together with the visual information, as well as the previously generated sentences.

A prevalent issue in VIST is the monotonous and repetitive generated stories. This can be attributed to the limited vocabulary of the VIST dataset. In KG-Story [115], the first stage (*distill*) gathers words from images using object detection and leverages GRUs for word prediction. Potential relationships between pairs of concepts throughout images are searched on external KGs, and if multiple candidates occur, a scoring function is used to rank their relevancy. This is the *enrich* stage. Finally, the *generate* stage utilizes a Transformer which imposes a repetition penalty to mitigate redundant narration. Further modifications in the default Transformer structure is the introduction of an anaphoric expressions generator to enhance coreferences and usage of pronouns, as well as positional encodings of variable length to enable representing stories of different lengths.

Addressing again the coherence and novelty of generated stories, authors of [380] propose a three-stage structure corresponding to imagination, reasoning and writing capabilities of humans. The first stage (*imagine*) focuses on the sequential consistence by extracting the visual topic of a frame through the combination of the current visual features and the sentence generated in the previous step. The knowledge part targets the content of narratives and comprises three graph types: a general commonsense KG, a scene graph and an event graph. A GCN applied on each graph selects the most suitable knowledge parts, which are combined to form the second stage (*reason*). Both *imagine* and *reason* outputs are fed to the third stage (*write*), which is responsible for generating the story.

7.6.2.2 Transformer-based models

External knowledge Towards informative and more diverse stories, [42] is the first knowledge-enhanced approach that utilizes a generative transformer to produce the story output. The concept enrichment stage connects concepts present in images with ConceptNet. Then, a graph attention network (GAT) operates on the graph and image features in order to integrate information of the most appropriate candidate concept nodes, which are passed to the next selection module. The concept selection module utilizes two different selection methods: a Sequential Selection Module (SSM) that operates in an encoder-decoder fashion, outputting selected concepts after encoding the embedded candidate concepts; a Maximal Clique Selection Module (MCSM) outputs a maximal clique containing all appropriate concepts for story generation given the concept graph. Finally, the concept to story module uses either an RNN structure or a BART language model, with BART demonstrating more diverse stories while preserving quality.

7.6.3 Evaluation

Language generation metrics BLUE, ROUGE, METEOR and CIDEr are widely used automatic metrics that evaluate the linguistic quality of generated stories.

Diversity of generated stories is measured via the **Distinct-n (Dist-n)** score [171]. This metric indicates the originality of generated text by calculating the frequency of n-grams throughout the whole corpus of generated stories. Higher Dist-n scores represent more diverse stories. [387, 42]

Human Evaluation is very important for generative tasks, as automatic evaluation metrics cannot assess the full range of linguistic capabilities, especially when it comes to evaluating sequential quality. However, different implementations perform varying human evaluation experiments, which somehow impedes the direct comparison of models.

In [387] four aspects are examined: *Fluency* checks the linguistic quality, *relevance* measures the success of textual descriptions in describing visual concepts, *informativeness* measures the diversity of produced stories and *coherence* evaluates the semantic continuity of stories in a sequence. Each aspect receives a score from 1 (worse) to 5 (best) from three evaluators, and their average values serve as the final results. Similarly, in [380] *relevance*, *coherence* and *informativeness* are regarded, receiving scores from 0 (worse) to 2 (best) from five evaluators. A different human evaluation strategy is followed in [115]: comparative experiments between VIST models are performed, asking users to rank generated stories from different models either with or without revealing the corresponding images. This is an indirect evaluation of linguistic quality and coherence, when only text is regarded, and also semantic relevance, when corresponding images are provided. The comparative approach is also used in [42], with two evaluators declaring their preference (or tie) between two models regarding three aspects: *relevance* and *informativeness* similar to [387, 380], together with *logicality* which measures the logical coherence over story sequences. Additionally, *overall* indicates the evaluator's preference in general between the two models.

7.7 Knowledge in Image Generation

7.7.1 Datasets

A variety of datasets have been used in visual generation, which however do not contain some certain demand or sense of knowledge, and are widely used in knowledge-free settings. Datasets used in conditional image synthesis are ImageNet [61], CIFAR [158], FFHQ [143], Oxford Flowers [241], CUB [106] and many others.

Sequential synthesis (Story Visualization) greatly utilizes **Pororo-SV** cartoon dataset [147]. It contains more than 16k pairs of scenes and dialogs extracted from 20 hours of video, 27,328 fine-grained scene descriptions in natural language provided by human annotators, and 8,913 QA multiple-choice pairs related to the story. In total, 10 main characters appear in the frames. Questions are divided in 11 types: *Action, Person, Abstract, Detail, Method, Reason, Location, Statement, Causality, Yes/No, Time.* **FlinstonesSV** [223] is also based on cartoon frames. It is composed of 25,184 densely annotated videos, each of which containing 75 frames. The annotations include bounding boxes with labels for characters and items of the frames, as well as segmentation masks. Another emerging dataset for Story Visualization is **DiDeMo-SV** [214, 107], a dataset based on video captions that contains 10,000 with more than 40,000 temporally localized textual descriptions.

7.7.2 Methods

7.7.2.1 Knowledge in Conditional Image Generation (K-cIG)

Internal knowledge Even though GANs have been powerful in synthesizing novel images, they cannot handle combinations of attributes they have not encountered in the training data. Therefore, if the textual condition refers to such unseen combinations, the synthesized image has sacrificed some of the semantics, in order to produce a result that remains within the learned distribution. The insertion of additional knowledge can expand the generated distribution to enhance the consistency on the condition without sacrificing resulting fidelity. This can be translated into two needs regarding a GAN model: the generator should become more flexible, and the discriminator more tolerant. KG-GAN meets those requirements by introducing a second generator, trained on domain knowledge by utilizing a novel knowledge loss. This second generator shares parameters with the original one, which is responsible of synthesizing images conditioned on text. A regression network receives the synthesized images from the seen-image generator and the ones from the knowledge generator, imposing constrains regarding the plausibility of unseen combinations. The semantic vector produced by the knowledge generator is redirected to the seen-image generator to guide generation outside the predefined classes. KG-GAN does not exploit external knowledge sources, but with this simple distribution enhancement it achieves some preliminary zero-shot capabilities. [40]

7.7.2.2 Knowledge in Story Visualization (K-SV)

External knowledge Story Visualization is another task with limited contributions in knowledge-enhanced settings. Structured information from text can be obtained via parse trees which can permit hierarchical encoding of longer phrases. Missing information regarding visual details in text can be filled out with external knowledge from ConceptNet [317]. Moreover, conceptually similar sentences that are phrased in a different way need to be placed closer in an embedding space, an issue that external knowledge can again effectively resolve. Spatial knowledge is also underrepresented in most sentences, even though scene synthesis needs detailed information of object positions. Dense captioning as a form of self-augmenting knowledge provides detailed positioning information due to the usage of region bounding boxes. The combination of internal spatial and external semantic knowledge is able to better guide sequential synthesis, resolving all involved aspects such as text-image consistency, visual quality and sequential continuity. A Memory-Augmented Recurrent Tree-Transformer (MARTT) encodes the parse trees for the text, while a Graph Transformer [401] embeds the commonsense knowledge. Both embeddings are inserted in the story encoder, which outputs contextualized embeddings for the image generator. The generated images are passed to image and story discriminators, which redirect synthesis based on individual and sequential aspects. Spatial knowledge from dense captioning enforces additional loss functions while training, to provide more explicit information about positions and detailed grounding of characters on the images with respect to their descriptions in the text. [212]

The groundbreaking success of DALL-E [277, 276] inspired the usage of massive zero-shot transformer-based generative models in Story Visualization; StoryDALL-E [214] achieves generalization of visual synthesis to unseen textual stories, also extending the task to *Story Continuation*: in this case, a source image is included in the conditioning, requesting from the model to continue the visual story in a consistent way. Story Visualization has been a task lacking sufficient datasets, due to the increased effort needed to construct appropriate ones, either manually or automatically. To this end, external unstructured knowledge obtained from the pre-trained DALL-E [277] as knowledge base enables even zero-shot sequential synthesis based on input 'story' text.

7.7.3 Metrics

Image generation metrics (Section **??**) such as **FID** for seen and unseen classes were used in KG-GAN [40]. FID is also used to evaluate quality of generated frames independently in [212, 214]. **R-precision** indicates quality by measuring the retrieval capabilities of generated frames over ground truth captions comparing to retrieval using the real frames. [212]

Classification metrics, such as **Character F1 score** measure the quality of generated characters in predicted images. Also, **frame accuracy** checks the exact match between semantics of the ground truth and generated frames. [212, 214]

Language metrics are also relevant: viewing SV frames as a video, captions for generated frames can be produced using video captioning techniques. **BLEU** scores evaluate the quality of captions as an indirect measure of visual quality, based on the idea that well-designed semantics will be captured in captions better than low quality ones. [212]

Human Evaluation can reveal the human perception over quality, as in most generative tasks. Specifically for SV, evaluators need to assess results over *visual quality*, *consistence* and *relevance* compared to the previous state-of-the-art model on the same task. [212, 214]

7.8 Multi-task transformers with knowledge

7.8.1 Methods

Multi-task models can easily be built using multimodal transformer backbones. Instead of utilizing external KGs as in previous methods, many implementations employ self-knowledge exclusively, by obtaining more structured representations from the existing visual and textual data.

External knowledge A natural unification of multiple tasks under the same model would incorporate tasks moving in the same direction, such as cross-modal reasoning tasks or cross-modal-retrieval tasks. Indeed, VQA, VCR and VE were unified in Rationale VT transformer [219], a framework that utilizes visual and linguistic clues to generate free-text rationales. Two knowledge sources attempt to provide reasoning information regarding scenes: first, a grounded situation recognizer [267] describes activities on scenes with entities involved and draws bounding boxes for entities to visually ground them; second, Visual Commonsense Graphs [256] are employed to fuse commonsense inferences about events and intents so that a temporal perspective of a scene is also considered. Rationales are generated for VQA-E (visual question answering) [175], E-SNLI-VE (visual entailment) [71] and VCR (visual commonsense reasoning) [403] datasets. Visual recognition of objects is the first step for visual understanding, followed by capturing their in-between relationships utilizing the knowledge provided by the grounded situation recognizer [267]. Higher-level cognition is achieved using knowledge from VisualCOMET [256], which receives the knowledge stored in Visual Commonsense Graphs to generate commonsense inferences. VisualCOMET is built upon GPT-2 [272], therefore a unimodal, purely linguistic input can be provided, utilizing object labels, textual question/answers and inferences. Alternatively, GPT-2 can be adapted, resulting in a hybrid implementation: visual features and bounding box coordinates act as visual embeddings, combined with VisualCOMET token embeddings indicating the beginning of *before, after, intent* inferences.

Targeting again reasoning tasks, [306] builds on top of LXMERT [49] to address the knowledge-enhanced versions of the VQA, VCR and VE tasks on the OK-VQA [221], FVQA [360], NLVR2 [324], SNLI-VE [378] datasets. External knowledge is provided from ConceptNet [317] and Wikidata [354]. Knowledge-rich expressions are created by matching embedded knowledge with training sentences from the datasets. Moreover, a training objective targeting the alignment of knowledge embeddings and knowledge-rich expressions encourages learning a global representation structure. Utilizing this objective is proven beneficial during both pre-training and fine-tuning. It is also observed that the introduction of this knowledge-oriented objective smooths the embedding space, which facilitates similarity matching between words.

KB-VLP [44] utilizes knowledge embeddings based on Wikidata [354] entities, which are concatenated with the visiolinguistic instances as inputs of a VL transformer. Specifically, entity recognition on text is performed to extract relevant Wikidata entries, which are embedded via Wikipedia2vec to form text-related knowledge embeddings. Object tags extracted from the image are used to obtain image-related knowledge embeddings from relevant Wikidata entities. The input vector consists of 5 components: word embeddings for text, text-related knowledge embeddings, word embeddings sequences for object tags per image, visual features and image-related knowledge embeddings. Two specialized pre-training objectives are used: sentence-level objective substitutes elements from the input vector with other random elements, while token-level objective extends text - image masking to text-related knowledge embedding - image-related knowledge embedding masking. Task specific datasets for KB-VLP are VQA [3], GQA [122] and OK-VQA [221] for visual question answering, and NLVR2 [324] for visual reasoning.

Internal knowledge OSCAR [176] is one of the models that effortlessly transit from knowledge-free to knowledge-enhanced learning utilizing self-acquired knowledge in its simplest form. Instead of -rather naively- letting the model infer the correct image-text alignments in an exhaustive way, OSCAR facilitates the procedure with the usage of object tags, as intermediaries between text and image instances. This procedure is endorsed from the observation that salient objects in the image will most probably also appear in text. The input to the VL transformer module consists of word tokens, object tag embeddings and visual features. The intermediary object tags form separate semantic spaces, depending on whether they are paired with text or image, yielding two dedicated pre-training objectives. The masked token loss objective views text and tag word representations in the same space, randomly masking each of them and letting reconstruct the missing parts through the visual modality. Conversely, contrastive loss views tags paired with visual features, and randomly replaces the real tag sequence with another one sampled from the dataset, learning to pull apart mismatched tag sequences and bring close together the matching ones. OSCAR succeeds in both understanding tasks, such as cross-modal retrieval (ITR/TIR), visual question answering (on VQA [3] and GQA [122]) and visual reasoning (on NLVR2 [324]), as well as in generation tasks, such as image captioning and novel object captioning.

ERNIE-ViL [397] leverages structured visual knowledge from scene graphs to bridge detailed semantics across vision and language. Such fine-grained representations are important to differentiate between conceptually similar scenes. Scene graph prediction tasks (object, attribute and relationships prediction) encourage learning those fine-grained differences. Even though not using external knowledge, ERNIE-ViL internally constructs structured knowledge during the cross-modal pre-training. Nevertheless, this self-knowledge is sufficient to boost performance in 5 VL tasks, especially in those that fine-grained associations are required, such as visual referring expressions (VRE). Other tasks benefited from this approach are VCR, VQA and cross-modal retrieval (ITR/TIR).

ROSITA [54] extends the self-knowledge idea by employing both cross-modal and intra-modal knowledge at the same time. Given an image-text pair, the first step is to construct intra-modal graphs, i.e an image graph and a text graph. The image graph consists of regions (defined by a pre-trained object detector) as nodes, with IoU scores of paired regions acting as edge weights between those regions. Similarly for the text graph, objects, attributes and relationships are extracted from text to fill the nodes of the text graph, while edge weights are defined by the co-occurrence frequency between pairs of nodes. In both graphs, zero similarity scores between nodes indicate absence of edge. A cross-modal scene graph is derived from the image and text graph by aligning predicted region tags from the image side and words from the text side by comparing their textual semantic similarity. By calculating this similarity score for all possible tag-word pairs, edge weights between cross-modal nodes are defined. Nodes connected via cross-modal edges, named anchor nodes, form sub-graphs which maintain intra-modal and cross-modal edges, as well as two-hop connections that contain paths of cross-modal edges followed by intra-modal ones. ROSITA leverages this enhanced representation to boost three downstream tasks: VQA, VRE and ITR.

7.8.2 Evaluation

Human Evaluation is useful in cases of language generation tasks, such as the rationales generation of [219]. In this case, the need for human evaluation arises from the observation that certain discrete rationales, even though not being paraphrases of each other, can be suitable. The following aspects are evaluated: *visual plausibility* referring to how well the generated rationales support the answer (in VQA and VCR) or the entailment (VE) given the image, and *visual fidelity* measuring the appearance of irrelevant information within more plausible generated rationales. By excluding images, *textual Plausibility* evaluates generated rationales based on their support on the answer (in VQA and VCR) or the entailment (VE) exclusively.

Classification metrics such as **accuracy** serve as the gold standard for non-generative models on cross-modal reasoning tasks. In [306], OK-VQA accuracies per question types are also reported, in order to validate improvements in commonsense-oriented categories attributed to the injection of commonsense knowledge.

Ranking metrics provide valuable insights in cases when retrieval tasks are performed [397, 54], where Recall@k for k=1, 5, 10 is reported.

Language metrics such as **BLEU** [254], **CIDEr** [350], **SPICE** are used for language generation tasks, such as image captioning and novel object captioning [176].

7.9 The future of knowledge enhancement in multimodality

The trajectory of multimodal AI is increasingly defined by systems capable of performing many tasks across heterogeneous input types—text, images, video, audio, and structured signals. These multitask multimodal architectures are emerging as the dominant paradigm because they mirror real-world cognitive demands, where understanding is rarely confined to a single modality. As these models expand in capability, however, a crucial limitation persists: their internal representations are insufficient for robust, reliable, and up-to-date knowledge reasoning. Even the most advanced latent knowledge stored in learned parameters fails to ensure factual accuracy, temporal consistency, or interpretability.

Why Knowledge Graphs Remain Central

In this context, knowledge enhancement through external structured resources, especially knowledge graphs, remains indispensable. KGs have several advantages that align naturally with the evolving needs of large-scale multimodal systems:

1. ***Explicitness and interpretability***: Representing entities, concepts, and relations explicitly allows models to operate on transparent symbolic structures, enabling verifiable reasoning steps. This is particularly important in settings that demand traceability or regulatory compliance.
2. ***Compositionality***: KGs provide a natural method for combining multimodal evidence. For instance, images aligned with KG entities facilitate zero-shot or few-shot grounding, while graph relations serve as constraints for multimodal inference.
3. ***Temporal and factual updating***: KGs support incremental updates without full model retraining. This is essential for knowledge that changes over time, something current multimodal foundation models still struggle with.
4. ***Bridging heterogeneous data sources***: KGs act as coherence layers binding text, vision, sensor data, and annotations. Their relational structure creates cross-modal conceptual anchors that purely neural models lack.

These advantages have encouraged research on KG-enhanced multimodal architectures, including models that perform retrieval-augmented grounding, graph-guided vision-language reasoning, and graph-augmented representation learning. Although still emerging, these approaches show that structured knowledge serves as a stabilizing substrate for complex multimodal reasoning pipelines.

Multitaskers as the Dominant Multimodal Paradigm

The shift toward large, multitask multimodal systems, capable of captioning, visual question answering, grounding, retrieval, classification, editing, and planning altogether, pushes knowledge requirements to new extremes. These systems are expected to integrate heterogeneous reasoning skills, maintain factual and commonsense consistency, interpret ambiguous or incomplete multimodal cues, and generalize across domains.

Such expectations exceed what can be encoded in static model parameters. Therefore, even as multimodal transformers grow larger, they do not necessarily grow “knowledge complete”. Instead, they become increasingly dependent on external, dynamic knowledge reservoirs that can correct, filter, and enrich their internal representations.

Knowledge graphs, especially large, densely interlinked ones, remain one of the most reliable mechanism for such enrichment. Their structured nature counterbalances the high-dimensional, entangled representations learned by neural models, offering the kind of neuro-symbolic hybridization that many believe will be necessary for the next generation of multimodal reasoning systems.

From KGs to Knowledge-Rich Vision-Language Systems

A promising direction involves tight integration between multimodal foundation models and external knowledge backbones into unified paradigms. To this end, the transition towards harnessing LLMs as the knowledge sources of the future arises as a central point. Other than acting as separate knowledge bases, LLM can be extended to unified multimodal architectures, such as the Large Vision-Language Models (LVLMs). To summarize, the transition from KGs to LLM-enhanced systems regards the following two steps

1. ***LLM-augmented VLMs***: LLMs as knowledge bases enrich the knowledge of existing VL systems, which are mostly capable in tasks that involve only visual knowledge. Querying LLMs via prompting can enrich the basic visual knowledge of such models with knowledge senses described in previous chapters. The usage of LLMs as knowledge bases is going to be analyzed in the next chapter.
2. ***Unified Large Vision-Language Models***: a more organic integration of knowledge in the basic VL backbone is to implicitly integrate knowledge *from scratch.* In other words, extending LLMs to integrate the vision modality, while maintaining the model scale together with the interleaved knowledge and reasoning capabilities is one of the most promising paradigms, shifting the need from requiring adequate

external knowledge sources to integrate the same knowledge to the pre-training process, alleviating the requirement for additional enhancement in many cases.

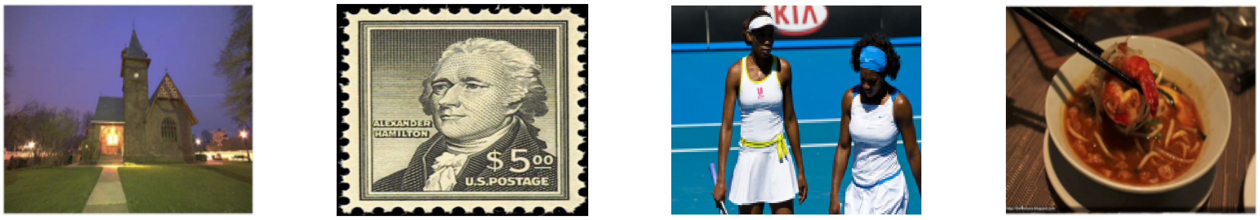
However, if more knowledge is needed, the synergy between LLMs and LVLMs can advance performance in knowledge-demanding tasks. Such collaboration opens new frontiers to multimodality and AI in general, introducing the transitioning towards **multi-agent systems** involving large models. Such agentic structures allow efficient distribution of tasks, specialization and delegation of specific needs to LLM/LVLM components, which communicate and exchange information to enhance each agent's role, ultimately advancing the capabilities in comparison to an agent in isolation.

Chapter 8

Knowledge enhancement using Large Language Models

Despite their interpretability and precision, knowledge graphs (KGs) exhibit inherent limitations in scalability, adaptability, and coverage. They are expensive to construct and maintain, often lagging behind the pace of new concepts or visual entities introduced in real-world data. Furthermore, their symbolic nature restricts generalization: since KGs represent discrete facts, they struggle to capture nuanced or context-dependent knowledge (e.g., temporal or causal reasoning). In vision-language settings, this rigidity manifests as incomplete grounding—models may fail to describe unseen object relationships or infer implicit contextual cues beyond the fixed ontology of the graph. Additionally, the process of aligning visual features with graph nodes introduces semantic gaps, as visual perception rarely maps cleanly onto symbolic categories [403]. These challenges motivated a paradigm shift toward data-driven, distributed representations of knowledge—culminating in the rise of LLMs as implicit knowledge sources.

For example, in cases where querying multiple knowledge senses at once (often not stored in a single KG) or multi-hop reasoning is required, KG-driven pipelines struggle. A related example is presented in Figure 8.0.1. [205]



Q: What days might I most commonly go to this building? A: Sundays.

Q: In which continent was the person in the image born? A: North America.

Q: Who among the people in the image is the eldest? A: Person in the left.

Q: What is the name of the object used to eat this food? A: Chopsticks.

Figure 8.0.1: External knowledge is required to answer these visual questions [221, 301, 126, 205].

The first image of Figure 6.2.2 requires knowledge about **human culture and history** [221], to combine with **visual** information: the object in the image is a *church*, and *people usually go to the church on Sundays*. The second image [126] requires one more reasoning step, since it is not only required to detect that this is a postage stamp containing the photo of a *person* (**visual** information), but also who this person is. Knowledge about **named entities** recognizes this person as *Alexander Hamilton*. Further **factual** knowledge provides that *Alexander Hamilton was born in todays Saint Kitts and Nevis* and *Saint Kitts and Nevis is in North America*. The combination of these two facts derives the final answer *Alexander Hamilton was born in North America*. The third image [301] requires the **visual** extraction of the two *people* present in it. Then, **named**

entities knowledge assigns the identities *Serena Williams* and *Venus Williams* to these two people. Their *age* is provided as a combination of **named entities** and **factual** knowledge, yielding the **comparative** knowledge fact that *Serena Williams is older than Venus Williams*. Finally, **spatial** knowledge derives that *Serena Williams is the person in the left*. The overall combination of **named entities**, **comparative** and **spatial** knowledge returns the final answer *The person in the left*. It becomes obvious that answering these question requires more knowledge from external sources, which is extracted and combined to infer an answer. Thus, the incorporation of external knowledge in earlier or later stages of the *pre-training/fine-tuning* process is necessary to enhance the capabilities of VL models, so that they are able to respond to more real-world scenarios.

Rather than relying on explicit graph edges, LLMs store latent associations between concepts, allowing them to retrieve and reason over knowledge flexibly through natural language prompts. When applied to vision-language tasks, the use of LLMs as implicit knowledge bases brings several compelling advantages, a transition often described as the emergence of the **LLM-as-KB paradigm**. First, LLMs-as-KBs can support **context-aware reasoning**: given a visual scene, they can integrate contextual cues to infer implicit relationships—for example, inferring that a person wearing a helmet on a road is likely cycling—even when this association is not explicitly represented in any graph. Second, LLMs are highly **scalable** and **self-adaptive**: as they are trained on web-scale corpora, their internal knowledge expands naturally with data, covering long-tail concepts and emergent entities without requiring manual updates. Third, LLMs can perform **compositional** and **analogical** reasoning, generating inferences that transcend observed data (e.g., inferring that "a robot cooking" implies anthropomorphic behavior). These capabilities are particularly valuable in open-ended generative settings such as image captioning or visual storytelling, where rigid symbolic reasoning would fail to capture the fluidity of natural descriptions.

Another motivation for treating LLMs-as-KBs lies in their integration potential. Traditional KGs must be explicitly queried using structured interfaces like SPARQL or logical rule systems, which hinders their incorporation into neural architectures. In contrast, LLMs operate directly in the language modality, which makes them seamlessly compatible with text-based outputs of vision-language models. Systems such as BLIP-2 [172], LLaVA [187], and GPT-4V [245] exemplify this transition: visual inputs are first encoded into a textual or multimodal latent space, and the LLM then performs reasoning and description generation using its embedded world knowledge. This eliminates the need for discrete graph traversal and enables end-to-end differentiable reasoning, where visual grounding and knowledge retrieval are jointly optimized.

Importantly, LLMs are not only knowledge containers but also **reasoning engines**. Unlike KGs, which require external inference algorithms to deduce new facts, LLMs perform implicit inference through contextual generation. For example, given the visual prompt "a man holding a stethoscope," an LLM can infer the probable role ("doctor") without explicit rules, drawing upon latent associations learned from text. This inference flexibility makes LLMs particularly suited for commonsense grounding, a key challenge in multimodal AI where models must bridge perception with world understanding.

Still, this transition does not imply abandoning structured knowledge entirely. While LLMs offer flexibility and scale, they can also **hallucinate** or **overgeneralize**, generating plausible but factually incorrect statements. Consequently, researchers increasingly explore hybrid architectures that combine the interpretability of KGs with the generative intelligence of LLMs [393, 300]. In these systems, KGs serve as factual anchors for verification or retrieval, while LLMs handle abstraction and contextual reasoning. This hybridization represents a pragmatic middle ground, leveraging the explicit correctness of KGs and the implicit richness of LLMs to achieve more reliable and explainable multimodal reasoning.

In the following sections, we are delving into the LLM-as-KB paradigm for knowledge enhancement [205], while also underlying emergent issues adjacent to LLMs regarding advanced reasoning [88, 249], hallucination [97, 141], memorization [321] and biases [77].

8.1 Background in LLM-enhanced vision-language tasks

8.1.1 LLM-as-KB for K-VQA

First works exploiting the LM-as-KB paradigm were introduced for K-VQA. Specifically, GPT-3 [31] can be used to provide facts in a few-shot manner, receiving visual captions as prompts [389], similar to how traditional KGs receive SPARQL queries. Performance gains can be achieved by utilizing multiple captions as prompts to a variety of pre-trained LLMs, enabling zero-shot reasoning [336]. Using again linguistic captions as a modality mediator, [101] leverages frozen LLMs to address zero-shot VQA. Chain of Though (CoT) prompting of LLMs is another interesting direction, which enhances explainability of the answer derivation pipeline by revealing intermediate reasoning steps [46]. Instead of resorting to the linguistic modality to obtain unimodal LLM prompts, other approaches opt to fine-tune a visual encoder jointly with the LLM, so that aligned LLM-VL representations are achieved [136].

There are a few implementations combining explicit and implicit knowledge sources to enjoy advantages of both worlds. KRISP [220] leverages several external KGs [317, 14, 25], visual knowledge from Visual Genome [155], as well as implicit knowledge from BERT [63]. REVIVE [184] deploys several visual features to retrieve knowledge from various sources, such as Wikidata and GPT-3. Visual feature guidance was proven critical towards improving the knowledge retrieval process. Fusing both implicit and explicit knowledge in the VL reasoning process is also followed in KAT, using a refined framework that fetches information from Wikidata and GPT-3 upon which joint reasoning is performed. A transformer decoder receives the output of the reasoning module to generate the final answer [99].

8.1.2 LLM-as-KB for K-VCR

Early transformer-based endeavors for knowledge-assisted VCR (K-VCR) naturally utilize BERT [63] as the backbone architecture to construct end-to-end KVL models. In KVL-BERT [315], the input Q together with candidate answers A guide the retrieval of relevant commonsense facts [317], resulting in a knowledge-enriched linguistic input. Then, visual features among with this enriched input are inserted in a BERT-like VL model (VL-BERT [322]) so that the correct A is selected. Consequently, inferring R requires feeding VL-BERT with the predicted A, candidate rationales R and visual features. Aligning independent modality representations within a single multimodal embedding is proposed in [166]. The same work introduces extensions of VL pre-training objectives to incorporate commonsense knowledge from [317] as an extra modality, therefore enforcing learning KVL interrelationships. Dynamic commonsense augmentation of image-text training data is a suggested direction, accompanied by learning to reconstruct hidden visual labels based on knowledge facts retrieved from commonsense KBs [394].

Implicit knowledge sources have been gaining popularity in recent K-VCR literature. GPT-2 [272] has assisted dynamic reasoning over images, inferring **temporal** hypotheses regarding what might have happened before and what might happen after the depicted situation [256]. Chain of Thought (CoT) reasoning is inherently tied to VCR, as reasoning paths are highly associated with selecting rationales R. The rise in popularity of CoT techniques for linguistic tasks is highly interconnected with the development of LLMs, which have been proven able to reveal intermediate reasoning steps [151]. There are not yet many works in the VL direction, even though the introduction of novel appropriate datasets with grounded answer rationales highlight the prospects of such an approach [195]. Specifically, [195] tackles VCR by captioning the image, and then feed the caption together with the existing linguistic input to the LLM. Another promising work in this direction introduces Multimodal-CoT without using language as the mediating modality, proposing a two-stage process to separately infer the answer A and the rationale R, while stating that a LM with less than 1B parameters is adequate for state-of-the-art performance [418]. It is expected that the rapid rise of popularity of LLMs in complex linguistic QA reasoning [60] may soon give rise to more LLM-augmented VCR approaches, addressing more aspects of reasoning.

8.1.3 LLM-as-KB for K-VCR

The integration of external knowledge in IC using transformers, was first explored in [419], where **event** and **named-entity** knowledge is fused together with textual and visual data in a Transformer encoder [348] to generate entity/event-aware captions. **Commonsense** descriptions derived from ConceptNet [317] and ATOMIC [123] are able to assist visual commonsense generation (VCG), a challenging task that requires inferring **intents** and **temporal** sequence of events [379]. This is achieved by incorporating **commonsense** descriptions to BART, a powerful language generation model [168]. **Geographical** information guiding **factual** knowledge retrieval to assist IC was first explored in [240], where visual features together with the extracted facts are inserted in a Transformer encoder-decoder structure, ultimately generating the caption c.

Apart from external knowledge considerations, IC faces additional challenges as a *language generation* task: VL transformers are not well-suited for generative tasks, even though they excel in understanding tasks, where an answer has to be selected among a set of pre-defined options. XGPT tackles this challenge by adapting generative pre-training [272, 31] for VL tasks [375], which is achieved by introducing novel generative pre-training objectives. The collaboration of GPT-2 [272] with CLIP [273] is viewed as highly promising, since both models have been trained on an abundance of web-data, thus incorporating numerous knowledge senses in an *implicit* manner. ClipCap [232] leverages this collaboration without re-training CLIP or GPT-2; instead, a lightweight transformer-based mapping module is trained to match CLIP representations to GPT-2, which eventually generates the caption c. The CLIP-GPT-2 combination was also followed in VC-GPT [201]. Another lightweight improvement combining a pre-trained CLIP visual encoder and a frozen GPT-2 text decoder further boosts performance of low-resource approaches [278]. A cross-modal filter that selects the most relevant visual information, so that captioning errors are reduced is proposed in [200], still respecting the frozen CLIP-GPT-2 framework.

A factor that overshadows the knowledge-enhanced IC capabilities is the lack of dedicated datasets for testing. So far, IC models are evaluated on classic datasets containing images and captions, such as COCO [183] and Flickr [396], which however are not challenging in terms of external knowledge required. The construction of appropriate datasets that would follow the paradigm of knowledge-demanding VQA datasets, such as OK-VQA [221], K-VQA [301], FVQA [360], KB-VQA [371], or VCR datasets [403] will give prominence to the abilities of K-IC models.

8.2 Large Language Models for Visual Word Sense Disambiguation

Multimodal retrieval and more specifically image-text retrieval is one popular task belonging in the discriminative VL task family, where textual and visual data are integrated to infer context and meaning. Despite the remarkable advances brought by models like Vision Transformers (ViTs) and multimodal transformers, a significant challenge persists: the ability to resolve ambiguity in properly retrieving images corresponding to ambiguous textual spans. Even though Word Sense Disambiguation (WSD) in natural language is a long-standing problem, the incorporation of the visual modality stands as a novel problem.

Visual Word Sense Disambiguation (VWSD) [275] is a challenging extension of WSD, where an ambiguous target word within a given context has to retrieve the proper image among competitive candidates. VWSD was first proposed as part of the 2023 SemEval Shared Task, exposing the limited capabilities of existing VL retrievers (in this case being CLIP [273]) in properly retrieving images given ambiguous text. An example of VWSD is demonstrated in Figure 8.2.1.

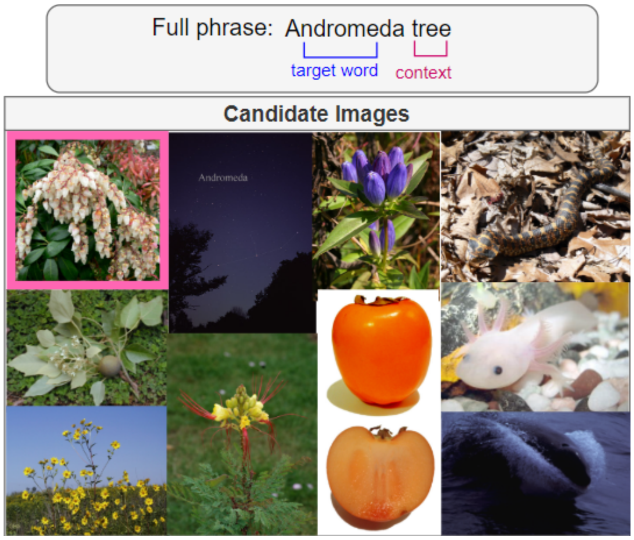


Figure 8.2.1: An example of the VWSD task: given a full phrase, comprised of a target word and its context, the most appropriate image candidate must be retrieved.

In Figure 8.2.1 the full phrase 'Andromeda tree', where 'Andromeda' is the ambiguous target word and 'tree' serves as the context is given to a model with the purpose of retrieving the correct image among 10 candidates. The candidate images are formed in a way that some of them may correspond to varying senses of the ambiguous target word or the context. In this example, 'Andromeda' corresponds to many different meanings; can be either a constellation, fish species, tree, reptile etc. Therefore, images corresponding to these meanings are provided. More distractors corresponding to the word 'tree' conclude the list of candidates: images of plants and greenery may serve as easier options to a model that may disregard the word 'Andromeda' and randomly select one of the 'tree'-related images. This is a reasonable possibility, since several VL retrieving systems may lack knowledge of specialized or rare terms - such as 'Andromeda tree'- and therefore focus on known terms only - in that case being the word 'tree'-.

Overall, the limited context provided and the possible lack of knowledge regarding the ambiguous target word render VWSD a much harder variant in comparison to traditional text-image retrieval. We present a variety of solutions, mostly focusing on LLM-based enrichment, which ultimately achieves the best scores in our VWSD pipeline. [157, 156]

8.2.1 Method

Dataset

In Table 8.1 statistics for the VWSD dataset are presented. All train and test samples contain 10 image candidates. The phrase length demonstrates negligible differences, with the vast majority of phrases comprised of 2 words. Data samples and official splits can be found in https://raganato.github.io/vwsd/.

Split	#Samples	Phrase length			
		1 word	2 words	3 words	4 words
Train	12869	0	12868	0	0
Test	463	1	445	17	1

Table 8.1: VWSD Dataset statistics

Evaluation metrics

The metrics considered for VWSD performance evaluation are accuracy (Hit @ 1) and Mean Reciprocal Rank (MRR) [275]. For n images ranked as $[i_1, ..., i_n]$ the Hit @ 1 metric is mathematically expressed as:

$$Hit@1 = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_1(i_k) \tag{8.2.1}$$

in which $\mathbb{1}_1$ denotes an indicator function that returns 1 if $i_k = 1$ (the desired image was retrieved) or 0 otherwise.

MRR is formulated as:

$$MRR = \frac{1}{n}\sum_{k=1}^{n} \frac{1}{i_k} \tag{8.2.2}$$

Image-Text similarity baseline

our baseline considers a variety of VL retrievers that simply receive the full phrase t as an input, selecting one of the image candidates i each time. The text t and the images i are projected onto a joint VL embedding space, where similarity scores $sim(t, i)$ are calculated using either cosine similarity (default), Euclidean distance, or Manhattan distance. The image embedding lying closer to the text embedding of the full phrase is denoted as the t, i pair with the highest cosine similarity (or minimum Euclidean/Manhattan distance as an equivalent), denoted as $score(t, i) = max(sim(t, i))$.

As for VL retrievers, we utilize CLIP with ViT [72] base encoder, as well as with ViT large encoder (denoted as CLIP-L). ALIGN [131] is also used for text-image retrieval. We also leverage several versions of BLIP [173], namely BLIP$_C$ and BLIP-L$_C$ (pre-trained on COCO [183] and using ViT base/ViT large as backbone encoders respectively), as well as BLIP$_F$ and BLIP-L$_F$ (pre-trained on the Flickr30k dataset [396]). Details regarding the model cards of these VL retrievers are provided in Appendix **??**.

Furthermore, we experiment with the incorporation of a penalty factor $p(i)$ [55] which modulates the retrieval preference of images that present high similarity ($sim(t,i)$) with many text instances t. That is because a VL retriever may give unequal preference to such images for the majority of data samples they appear. Specifically, the penalty factor $p(i)$ is mathematically expressed as:

$$p(i) = \left(\frac{1}{|T|}\sum_{t_j \in T} sim(t_j, i)\right)\frac{card(i)}{\max\limits_{i_k \in I} card(i_k)} \tag{8.2.3}$$

In 8.2.3, T serves as the set of phrases, I denotes the set of image candidates and card(i) corresponds to the number of samples in which the image i appears.

In this case, the similarity score is updated as:

$$score(t,i) = sim(t,i) - p(i) \tag{8.2.4}$$

LLMs for phrase enhancement

The utilization of LLMs as knowledge bases targets enhancing the full phrases t with more factual context, therefore disambiguating the target word through the provided information. A variety of LLMs is employed for enhancement; specifically, GPT2-XL (1.5B parameters) [272], BLOOMZ-1.7B & 3B [234], OPT-2.7B & 6.7B [413], Galactica 6.7B [332], and the 175B parameter GPT-3 [31] and GPT-3.5-turbo[1]. The choice of these LLMs is made based on the available hardware, or otherwise accessible via API. Various LLM scales in terms of parameters are exploited to assess their enrichment capabilities as an analogy to their implicit knowledge size. The zero shot prompts for these LLMs are crafted based on human intuition regarding requesting information for the full phrases. These prompts are listed in Table 8.2.

Prompt name	**Prompt template**
exact	"<phrase> "
what_is	"What is <phrase>?"
describe	"Describe <phrase>."
meaning	"What is the meaning of <phrase>?"

Table 8.2: Prompts for phrase enhancement via LLMs.

Given a full phrase t, its LLM enriched version is denoted as t_e resulting in the following retrieval score:

$$score(t_e,i) = sim(t_e,i) - p(i) \tag{8.2.5}$$

Question-answering (QA) for VWSD and CoT prompting

We convert VWSD into a Question-Answer (QA) task using the handcrafted instructive prompts presented in Table 8.3 for each full phrase t. The QA conversion requires extracting captions c_i for each of the candidate images i, enumerated from A to J. Models leveraged for image captioning include BLIP Captions [173] with ViT-base encoder (BLIP-L Captions denotes building upon ViT-large), as well as GiT [356] (with ViT-base) and GiT-L (with ViT-large). For all BLIP and GiT variants we attempt both beam-search multinomial sampling with 5 beams to obtain k=10 captions per image i, as well as greedy search, which produces a single caption per image i. We symbolize as c_i^k the k-th caption for an image i, as obtained from beam search, and simply as c_i a caption returned using greedy search.

The 'think' prompts utilize the phrase 'Let's think step by step', which has been proven to elicit reasoning in LLMs [151]. On the other hand, the 'choose' prompts request from the LLM to choose one of the provided image captions. Furthermore, experiments include both Chain of Thought (CoT) prompting, where the LLM attempts to respond in two steps, or no CoT prompts, where reasoning is performed within a single step.

---

[1] https://platform.openai.com/docs/models/gpt-3-5

Name	**QA Prompt template**
think (greedy)	"Q: What is the most appropriate caption for the <context>? Answer choices: (A) <caption for image 1> (B) <caption for image 2> ... A: Let's think step by step. "
think (beam)	"Q: What is the most appropriate group of captions for the <context>? Answer choices: (A) <captions for image 1 (separated with comma)> (B) <captions for image 2> ... A: Let's think step by step. "
CoT	"<think_prompt> <response of LLM with think prompt> Therefore, among A through J, the answer is"
no_CoT (greedy)	"Q: What is the most appropriate caption for the <context>? Answer choices: (A) <caption for image 1> (B) <caption for image 2> ... A: "
no_CoT (beam)	"Q: What is the most appropriate group of captions for the <context>? Answer choices: (A) <captions for image 1> (B) <captions for image 2> ... A: "
choose no_CoT (greedy)	You have ten images, (A) to (J), which are given to you in the form of captions.(A) <caption for image 1>. . . (J) <caption for image 10> You should choose the image, and therefore the caption that could better represent the <phrase>. What image do you choose?
choose no_CoT (beam)	You have ten images, (A) to (J), which are given to you in the form of captions.(A) <captions for image 1 (separated with comma)>. . . (J) <captions for image 10 (separated with comma)> You should choose the image, and therefore the set of captions that could better represent the <phrase>. What image do you choose?
choose CoT (greedy)	You have ten images, (A) to (J), which are given to you in the form of captions. (A) <caption for image 1> . . . (J) <caption for image 10> You should choose the image, and therefore the caption that could better represent the <phrase>. Use the following format: Question: What image do you choose? Thought: you should always think about what you choose. Result: the result of your thought. Final Answer: the image that you choose. Begin! Question: What image do you choose?
choose CoT (beam)	You have ten images, (A) to (J), which are given to you in the form of a set of captions. (A) captions for image 1 (separated with comma) . . . (J) captions for image 10 (separated with comma) You should choose the image, and therefore the set of captions that could better represent the <phrase>. Use the following format: Question: What image do you choose? Thought: you should always think about what you choose. Result: the result of your thought. Final Answer: the image that you choose Begin! Question: What image do you choose?

Table 8.3: QA prompts with and without CoT. 'Beam' and 'Greedy' refer to the corresponding captioning strategy.

In the CoT case, the LLM generates a choice from A to J in the first step, and based on this output, it is prompted again to select a letter from A to J as a second step. When combining 'choose' prompts with CoT, the LLM is prompted to reason on what image to choose and then explicitly respond to this question.

The prompts of Table 8.3 are also placed as exemplars to perform few-shot (FS) prompting, accompanied by the ground truth answer for the correct caption choice. The number of in-context samples s is defined by the user. We design three different ways of selecting the s in-context samples. In the *baseline* case (*random*), the s samples are randomly selected from the dataset. Nevertheless, since the relevance of selected samples with respect to a chosen sample is significant [188], as well as the sample ordering [197], we design two similarity-based sample selection methods, namely *top* and *inverse-top*. In the *top* case, the QA exemplars are placed from the most similar QA sample (in comparison to the query prompt) to the s-th least similar to the query; on the contrary, the *inverse-top* case implements the opposite, starting from placing the s-th most similar QA exemplar in the first position, finishing with the top-1 most similar QA exemplar in the last position. Both placement options consider textual embedding similarity as returned from ALIGN, with the similarity score being decided using cosine similarity.

The LLMs deployed for QA are Vicuna-13B [48] and GPT-3.5-turbo, the selection of which is made due to the need for testing the reasoning capabilities of LLMs of different scale, expecting the CoT cases to be significantly more meaningful in the larger GPT-3.5-turbo, since such reasoning capabilities are significantly correlated to model scale [364]. GiT-L with greedy decoding is selected for captioning.

8.2.2 Experimental results

LLMs for phrase enhancement results using various LLMs of increasing scale are presented in Table 8.4. Lots of conclusions arise by studying this table; first of all, the assumption of the LLM scale importance is verified, with smaller LLMs such as the OPT-2B/6.7B and BLOOMZ-3B scoring even below the corresponding baseline, where a VL retriever is leveraged without any other intervention. A striking difference is observed when using the larger LLMs, specifically GPT-3 and GPT-3.5. In these cases, there is a significant performance improvement in comparison to the respective baseline. ALIGN emerges as the most potent VL retriever when larger LLMs are employed, even though it scores slightly lower in the baseline retrieval case.

Regarding the prompts eliciting better retrieval capabilities, there is no definitive answer. The best results are achieved under the 'meaning' prompt in conjunction to GPT-3 as the prompted LLM. In the cases of GPT-3 and GPT-3.5 the 'exact' prompt is proven to be comparatively inefficient, with results staying significantly behind the best ones. To this end, by exclusively considering larger models for knowledge enhancement, proper guidance via prompting is deemed necessary, since the 'exact' prompt merely provides a full phrase to the LLM without requesting more specific information. Instead, by guiding the LLM via more specific questions such as 'describe phrase' or 'what is the meaning of phrase?', the LLM becomes more capable in retrieving the related knowledge requested.

Finally, incorporating the penalty factor $p(i)$ is always proven beneficial, should we compare the best results achieved with and without it.

Overall, knowledge enhancement via LLMs seems to be functional in cases larger models as knowledge bases are exploited, accompanied by appropriate, informative prompting.

QA for VWSD and CoT prompting results are presented in Table 8.5. First of all, by comparing GPT-3.5-turbo with Vicuna-13B results, is becomes evident that employing a larger model for QA prompting brings tremendous performance advancements on its own. Vicuna-13B totally collapses in come cases, demonstrating incredible low accuracy ($<2\%$).

By focusing on the most meaningful results extracted with GPT-3.5-turbo, we conclude that the zero-shot 'think' prompts are comparatively weaker; on the other hand, zero-shot 'choose' prompts achieve some non-negligible performance boosts. At the same time, utilizing CoT or not does not bring conclusive results, since accuracy per captioner varies between those two choices. Moreover, varying performance based on the choice of captioner is observed.

Few-shot prompting achieves some improvement, at least regarding the best VWSD-as-QA results overall. Performance is again highly versatile based on the choice of image captioner, with GiT-L greedy outperforming the rest, while ViT-GPT2 fall way behind, despite the decoding strategy utilized. In most cases, greedy decoding is more effective than beam search. We assume that beam search may add too much noise in the prompt, rendering different caption candidates closer to one another due to this imposed noise. On the other hand, sticking with one caption, as derived from greedy decoding, results in more distinctive caption candidates to form the final QA prompt.

In total, VWSD conversion in QA brings lower scores in comparison to the best results achieved in the previous cases, such as the LTR performance. Apart from the fact that proper LLM reasoning is still an open problem [16, 341], one possible fundamental reason for these lower accuracy scores can be the conversion from images to text via captioning: this intra-modality conversion may induce errors and information loss that impacts the final performance.

	CLIP		**CLIP-L**		**ALIGN**		**BLIP$_C$**		**BLIP-L$_C$**		**BLIP$_F$**		**BLIP-L$_F$**	
	acc.	**MRR**	**acc.**	**MRR**	**acc.**	**MRR**	**acc.**	**MRR**	**acc.**	**MRR**	**acc.**	**MRR**	**acc.**	**MRR**
With penalty														
Baseline	63.28	76.27	62.85	76.24	68.90	80.00	60.90	74.33	64.58	77.51	60.47	73.87	69.76	80.42
OPT-2B														
exact	62.85	76.00	62.85	75.93	68.68	79.89	61.12	74.46	64.58	77.41	60.26	73.73	**69.76**	**80.36**
what_is	60.98	74.85	66.30	78.10	63.28	75.95	60.91	74.43	66.31	77.86	57.24	71.15	67.60	78.58
describe	61.05	74.75	66.08	78.14	64.79	77.62	61.77	74.73	66.31	77.57	57.67	71.48	68.03	79.03
meaning	62.15	75.60	65.25	77.45	65.66	77.54	61.99	75.35	63.93	76.88	58.32	71.65	65.44	77.69
BLOOMZ-3B														
exact	61.26	74.59	62.99	76.18	66.52	78.36	60.48	73.13	63.28	76.00	57.02	71.23	65.66	77.49
what_is	64.36	76.82	68.25	79.82	67.39	78.72	61.34	74.94	66.95	78.47	59.61	73.35	**68.47**	**79.58**
describe	62.01	75.38	65.28	78.07	66.09	78.60	62.85	75.65	67.39	78.71	57.24	71.72	67.82	79.20
meaning	65.58	77.96	67.32	78.76	**68.47**	79.14	63.71	76.52	66.31	78.55	59.40	73.60	68.03	79.26
OPT-6.7B														
exact	62.63	75.84	62.20	75.54	67.82	79.24	60.91	74.23	64.79	77.58	59.83	73.40	**69.11**	**79.94**
what_is	61.79	75.70	64.63	77.68	64.79	77.23	61.77	75.01	63.07	76.16	57.88	71.79	65.87	77.77
describe	64.43	76.91	65.73	78.24	65.23	77.89	61.12	74.67	63.93	77.07	56.16	71.30	66.09	78.38
meaning	62.17	75.84	63.61	77.19	66.74	78.47	63.28	75.93	65.44	77.43	59.83	72.96	68.03	78.75
GPT-3.5 (175B)														
exact	58.86	72.09	60.18	72.73	62.42	74.43	57.02	70.78	59.18	72.32	52.92	67.40	63.07	74.65
what_is	66.52	78.81	69.35	80.51	70.41	81.42	67.60	78.56	68.47	79.67	60.91	74.30	71.71	82.02
describe	67.32	78.95	69.28	80.31	**73.22**	**82.73**	69.33	79.90	70.41	80.80	59.83	73.65	70.63	81.29
meaning	67.76	79.76	69.06	80.55	70.41	81.38	66.52	78.59	66.52	79.16	58.53	73.31	69.98	81.46
GPT-3 (175B)														
exact	61.98	74.90	64.07	76.58	66.52	78.37	60.48	73.99	64.15	76.58	59.61	72.91	65.23	77.06
what_is	67.92	79.27	70.73	81.57	71.71	82.27	68.25	78.93	68.90	79.91	60.48	74.24	69.11	80.25
describe	68.25	79.40	68.72	80.26	72.57	82.52	64.58	76.75	68.25	79.35	61.34	74.03	69.33	80.47
meaning	68.07	80.08	69.84	81.56	**74.95**	**84.09**	66.74	78.37	71.71	81.55	62.63	75.55	72.35	82.28
Without penalty														
Baseline	59.18	72.94	60.69	74.42	65.66	77.48	57.24	72.07	61.34	75.88	57.67	71.96	65.01	77.86
OPT-2.7B														
exact	58.96	72.77	60.26	74.15	65.66	77.48	57.45	72.19	61.12	75.77	57.24	71.68	**65.01**	**77.90**
what_is	58.31	72.91	62.75	75.47	61.12	73.94	59.83	73.13	61.12	74.54	53.35	68.71	63.50	76.22
describe	59.08	72.95	63.89	76.31	62.20	75.80	59.83	73.28	62.20	75.17	54.43	69.86	63.28	76.28
meaning	58.19	72.97	62.99	75.79	64.58	76.48	59.18	73.38	60.26	74.70	54.86	69.43	62.42	75.86
BLOOMZ-3B														
exact	56.93	71.53	59.52	73.78	63.93	76.15	58.10	71.77	59.61	74.06	54.86	69.66	61.12	74.99
what_is	62.20	75.39	65.66	77.88	62.85	75.51	61.34	74.35	65.01	77.32	57.24	71.85	**68.03**	**79.12**
describe	60.04	73.83	62.88	76.11	63.50	76.35	60.48	73.87	62.85	76.06	54.86	70.48	65.66	77.64
meaning	61.69	75.51	64.94	77.17	66.31	77.62	61.77	74.92	62.42	76.27	57.02	71.79	65.23	77.21
OPT-6.7B														
exact	58.75	72.63	59.61	73.86	64.15	76.57	57.24	71.96	61.12	75.83	56.80	71.40	**64.79**	**77.66**
what_is	60.48	74.10	62.45	75.89	61.77	75.18	57.88	72.27	61.77	74.89	52.92	68.83	61.99	75.23
describe	60.74	74.28	63.12	76.19	63.28	76.26	59.40	73.03	58.96	73.86	52.92	69.18	62.63	76.13
meaning	59.28	73.77	62.17	76.04	63.71	76.31	52.92	74.37	61.99	75.47	55.94	70.67	65.01	77.27
GPT-3.5 (175B)														
exact	56.89	69.85	57.11	70.36	60.48	72.15	54.43	68.33	56.80	70.42	51.40	65.68	58.32	71.11
what_is	65.00	77.11	65.87	78.11	67.82	79.52	64.15	75.91	65.87	77.78	58.10	72.32	68.03	79.36
describe	65.80	77.26	66.67	78.42	**70.84**	**81.16**	65.44	77.57	69.11	80.20	58.96	72.66	67.60	79.47
meaning	65.14	77.61	67.10	79.07	68.47	79.87	63.93	77.05	65.66	78.33	63.93	72.23	68.25	80.17
GPT-3 (175B)														
exact	59.88	73.38	61.68	74.91	64.79	76.27	58.96	71.92	60.48	74.02	55.72	70.34	62.42	75.04
what_is	66.51	77.62	68.15	79.38	69.55	80.22	63.28	75.56	65.01	77.40	56.59	71.54	67.82	79.03
describe	67.30	78.50	68.25	79.81	71.27	81.21	63.93	75.81	66.31	77.74	58.96	72.62	67.17	78.93
meaning	66.52	78.32	68.96	80.26	**72.57**	**82.29**	65.87	77.56	69.55	80.26	60.26	74.26	70.41	81.09

Table 8.4: Results for zero-shot LLM-based enhancement. Colored instances denote overall best results per metric, while **bold** numbers indicate best results for each LLM.

Captioner	**Zero-shot**				**Few-shot (random)**	**Few-shot (top)**	**Few-shot (inv. top)**
	no_CoT	CoT	choose no_CoT	choose CoT	no_CoT	no_CoT	no_CoT
GPT-3.5-turbo							
GiT-L (greedy)	44.49	**47.30**	**51.84**	**52.27**	**51.19**	**51.40**	**53.56**
GiT-L (beam)	40.82	36.50	50.54	49.68	46.12	47.83	45.61
BLIP-L (greedy)	**47.95**	43.84	49.46	44.06	48.16	48.81	50.32
BLIP-L (beam)	38.01	34.13	50.97	50.97	40.91	40.49	40.49
ViT-GPT2 (greedy)	28.94	25.05	32.40	29.81	31.32	31.45	28.91
ViT-GPT2 (beam)	30.24	25.92	32.83	33.05	32.03	28.73	23.64
Vicuna-13B							
GiT-L (greedy)	**34.34**	**27.65**	20.52	20.52	31.89	33.63	36.30
GiT-L (beam)	11.02	7.91	19.44	11.23	< 2	< 2	< 2
BLIP-L (greedy)	30.02	23.76	**20.95**	**21.81**	**35.56**	**36.08**	**36.48**
BLIP-L (beam)	9.41	6.27	12.74	8.64	< 2	< 2	< 2
ViT-GPT2 (greedy)	21.60	21.17	17.49	15.33	24.83	24.94	26.11
ViT-GPT2 (beam)	11.45	6.91	16.85	12.74	2.81	3.89	4.75

Table 8.5: Accuracy scores (%) for VWSD as a QA problem with and without CoT prompting for different image captioners and decoding strategies (beam & greedy). Colored cells denote best results per LLM.

Moreover, some quantitative results further enhance the merits of the VWSD as QA: we present two use cases to demonstrate both the zero-shot as well as the few-shot CoT prompting strategies. In the first case, we experiment with the candidate images of Figure 8.2.2, corresponding to the "rowing dory" phrase. The zero-shot prompt containing the candidate captions for each image is provided in Table 8.6, together with the LLM responses with and without CoT. In the CoT case, the LLM also provides an explanation, as triggered by the step-by-step thinking process.

We also present a successful example leveraging few-shot prompting; the captions of the candidate images of Figure 8.2.3, corresponding to the "football goal" phrase, are given as the final question (Q) to the LLM. To better guide task understanding, s=5 multiple-choice exemplars of the same format are provided beforehand, followed by their ground-truth answers (A), as demonstrated in Table 8.7.

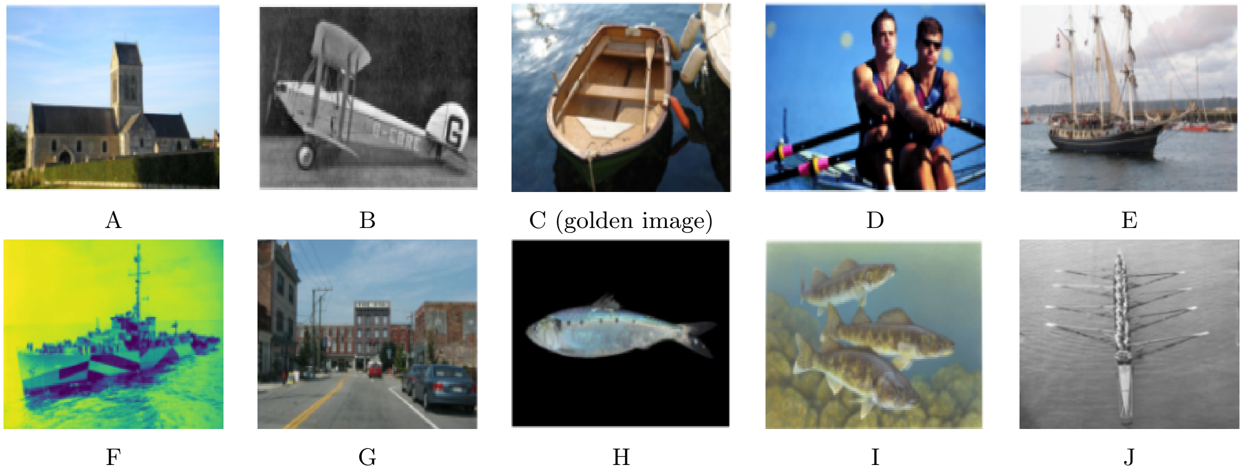

A B C (golden image) D E
F G H I J

Figure 8.2.2: Candidate images for the phrase "rowing dory".

Despite the successfully enhanced qualitative results presented above, the rather low retrieval metrics when CoT is employed (Table 8.5) indicate potential shortcomings of exploiting the LLM-as-KB paradigm, possibly attributed to reasoning capabilities of LLMs. That motivates the exploration of this topic in the next sections.

Prompt

Q: What is the most appropriate caption for the rowing dory? Answer Choices: (A) a church with a tall tower and a hedge. (B) an old airplane sitting on top of a runway. (C) a small boat in the water with two oars. (D) the men's rowing team are competing in the men's single sculls. (E) a large black ship in the water near a harbor. (F) a ship is on the water with people on it. (G) a large brick building with a sign on it. (H) a blue fish (sardines) at the australian aquarium. (I) a painting of four fish in the water. (J) a row boat with people on it in the water.

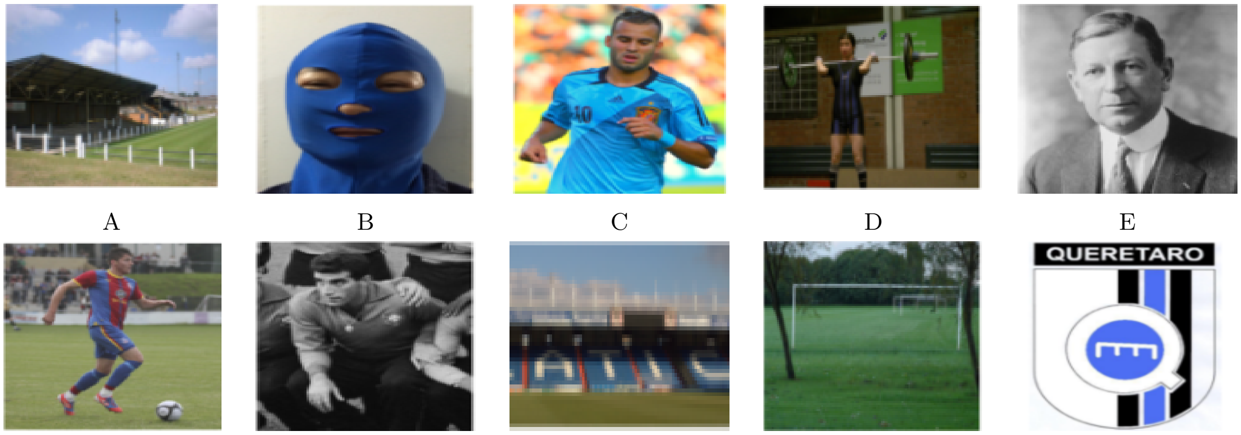
no_CoT

Q: What is the most appropriate [...]
A: The most appropriate caption for the rowing dory would be (J) a row boat with people on it in the water. **X**

CoT

Q: What is the most appropriate [...]
A: Let's think step by step.
First, we need to identify the subject of the caption, which is the rowing dory. Then, we need to choose a descriptive and informative caption that accurately represents the image. Out of all the answer choices, (C) "a small boat in the water with two oars" would be the most appropriate caption for the rowing dory as it accurately describes the image and provides relevant information about the boat. Answers (D), (F), and (J) also provide some information about the boat and the people in it, but they either include irrelevant information or are too vague to be an appropriate caption for the image.
Therefore, among A through J, the answer is (C) "a small boat in the water with two oars." ✓

Table 8.6: VWSD as QA problem example, with zero-shot CoT prompting encouraging the selection of the correct candidate. Questions Q are presented truncated. The red-colored text spans are parts of the prompt. The blue-colored text spans are generated using GPT-3.5-turbo, which ultimately provides the final answer.



A B C D E

F G H I (Gold image) J

Figure 8.2.3: Candidate images for the phrase t "football goal".

8.3 Reasoning and Hallucinations in Large Language Models

The utilization of LLMs as implicit knowledge repositories comes with inherent challenges, stemming from the vast pre-training over abundant data from diverse data sources and the unprecedented complexity of related model architectures. Particularly, reasoning shortcomings and hallucinations continue to hinder the practical adoption of LLMs in critical domains.

Recent literature has reported several advancements in various reasoning domains. Particularly, related tasks incorporate commonsense reasoning [286], involving inference regarding everyday situations, mathematical reasoning [196], referring to the ability of solving mathematical problems, logical reasoning [391], which includes the systematic deduction of conclusions based on established principles and formal rules, causal reasoning [86], which studies cause-and-effect relationships explaining why an event leads to another, and several other sub-tasks [347, 365, 261]. The development of large-scale models, such as ChatGPT and GPT-4 has showcased remarkable reasoning capabilities [185, 186, 18, 53], highlighting success in deductive settings [296], while challenges arise in other reasoning categories, such as inductive reasoning [381, 17].

Q: What is the most appropriate caption for the vanillin compound? Answer Choices: (A) a fossil fossil with a spiral pattern (B) a blue plate topped with pastries and a piece of bread. (C) the bones of the foot are very thin. (D) a spoon full of sugar on a table. (E) the tower of the building is made of brick. (F) a pair of sunglasses with a white background. (G) a close up of a plant with a yellow flower. (H) turmeric powder in a bowl (I) a man with glasses and a tie sitting at a desk. (J) a close up of a bike tire with a yellow tag on it. **A:** (D) a spoon full of sugar on a table.

Q: What is the most appropriate caption for the nankeen fabric? Answer Choices: (A) a close up of a white quilt with a tree design on it. (B) a man in a blue jacket and tan pants holding a sword. (C) a collection of brass trombones (D) a close up of a green plant (E) a pine tree in the forest (F) a row of mannequins in black dresses. (G) two boys are standing in front of a red wall. (H) two apples sitting next to each other on a white surface. (I) a cartoon of two women in fancy hats and a mirror. (J) a close up of a woman's face with a round ball on her forehead. **A:** (B) a man in a blue jacket and tan pants holding a sword.

Q: What is the most appropriate caption for the olmec indian? Answer Choices: (A) a large stone head in a garden (B) a woman lighting a candle at a table (C) [unused0] is the new champion of the world (D) a table full of food (E) a group of women holding banners in a protest. (F) a woman in a traditional dress stands in front of a tent. (G) a man in a suit sitting in a chair. (H) a clay pot with a small hole on top of it. (I) a close up of three green leaves of a plant (J) a fern in the forest near the lake **A:** (A) a large stone head in a garden

Q: What is the most appropriate caption for the charales order? Answer Choices: (A) a man walking past a store with a sign on it. (B) the road to the island (C) the art of the flower (D) a black snake with white spots on it's body. (E) a bird perched on a branch in a tree. (F) a lion roaring in the wild. (G) a man falling from a skateboard (H) a large, thin, green plant with long thin leaves. (I) a close up of a bush with blue berries (J) a mouse in a hollow log. **A:** (H) a large, thin, green plant with long thin leaves.

Q: What is the most appropriate caption for the skink lizard? Answer Choices: (A) two lizards on a log with a log in the background. (B) a man walking past a store with a sign on it. (C) the art of the flower (D) a portrait of [unused0], seated at a desk in front of a telescope. (E) the car is a small car that can be found in the museum. (F) a black snake with white spots on it's body. (G) a bird perched on a branch in a tree. (H) a close up of a bush with blue berries (I) a close up of a metal container with a white label on it. (J) a white door with a glass window **A:** (A) two lizards on a log with a log in the background.

Q: What is the most appropriate caption for the football goal? Answer Choices: (A) a large stadium with a large field and a large structure. (B) a woman wearing a blue ski mask (C) [unused0] of spain during the fifa world cup brazil group a match between spain and argentina at the estadio santiago bernabeu on june 25, 2010 in santiago, chile. (D) a woman lifting a barbell in a competition. (E) a man in a suit and tie looking at the camera. (F) person is a player for football team (G) [unused0], the brazilian national team, in the 1960s. (H) the empty stand at the old trafford stadium (I) a soccer field with a goal post in the middle (J) the emblem of the football club of the italian football club of the italian football club of the italian football club of the italian football club of the italian football club of the italian football club of the
A: (I) a soccer field with a goal post in the middle ✓

Table 8.7: Example of few-shot prompting with s=5 in-context samples for disambiguating the "football goal" phrase. The final answer A in **color** is derived using GPT-3.5-turbo.

However, reasoning in LLMs is often guided by probabilistic patterns rather than a structured understanding of cause-effect relationships or logical frameworks. This can result in responses that lack coherence or exhibit faulty logic, particularly when faced with novel, multi-step problems. Recent literature reports concerns regarding some non-negligible reasoning-related issues [80, 20, 145, 307, 343], often posing some fundamental concerns of whether 'true' reasoning (analogous to human-level reasoning) is achievable at all, or LLMs are exhaustively memorizing extensive amounts of information [400]. Hallucinations, on the other hand, occur when LLMs generate plausible-sounding but factually incorrect or irrelevant to the input information, a byproduct of their predictive nature rather than grounded knowledge [416, 119, 408]. These two phenomena are deeply interconnected: hallucinations can emerge from flawed reasoning processes where the model fills gaps in its understanding with high-probability but erroneous content.

In the paradigm of LLMs as knowledge bases, this interplay between reasoning limitations and hallucinations raises critical concerns about their reliability, interpretability and real-world applicability. While they excel at approximating knowledge from learned patterns, their inability to validate or reason through inconsistencies limits their effectiveness in high-stakes or precision-critical applications. Addressing these challenges requires rethinking how LLMs handle reasoning tasks and augmenting them with mechanisms to reduce hallucination tendencies, ensuring more robust and trustworthy knowledge representation. To his end, in the interest of utilizing LLMs as knowledge bases, we analyze our contributions regarding reasoning and hallucinations in LLMs [88, 97, 248, 249, 141]. Specifically, we evaluate a more challenging reasoning domain, focusing on abstract thinking, through which we seek to unlock more high-level reasoning cues similar to human problem-solving. Such problems require creative and out-of-the-box thinking rather than simply associating rules and patterns, and therefore learn structured processes for problem solving. For example, dealing with *riddles* triggers more abstract thinking procedures in comparison to executing mathematical operations, which require a clear understanding and implementation of pre-determined rules. As a parallel venture, we propose efficient techniques for hallucination detection on pre-trained LLMs, even evaluating cases where generated outputs are inconsistent to the given input data; such cases are hard to detect, since they do not refer to factually conflicting generations, but rather to subtle semantic inconsistencies between the input and the output.

8.3.1 Abstract reasoning in LLMs

As LLMs become more complex and impressive, simulating human-like reasoning aspects becomes more and more pertinent. There are two types of human-level thinking that examine different cognitive capabilities, exploiting the two hemispheres of the brain: vertical and lateral thinking. Vertical thinking is based on the left hemisphere of the brain and explores linear logic, rationality, sequential analysis of a problem and rule-based solutions. On the other hand, lateral thinking, associated with the right hemisphere of the brain, requires 'out-of-the-box' solutions that leverage abstraction, creativity and flexibility to reach correct conclusions. Related problems often require to overwrite common assumptions and presuppositions, and therefore set new perspectives when crafting problem-solving strategies. [135]

Puzzle reasoning often combines challenges from both thinking directions, requiring creativity as well as logical execution of steps and rules. Overall, a more abstract reasoning approach is necessary in puzzle solving, bringing additional challenges to LLMs in comparison to those studied so far.

Delving into puzzle-solving, definition and categorization is an important step towards developing related datasets and techniques, and ultimately evaluating the associated reasoning capabilities of LLMs. Getting back to our initial research question regarding utilizing LLMs as knowledge bases, puzzle-solving evaluation tests the ability of LLMs to reason through complex, multi-step scenarios, often useful in knowledge-enhancement related applications. By mirroring human-like reasoning and understanding, puzzle-solving provide a lens to analyze reasoning flaws, such as over-reliance on spurious correlations or logical inconsistencies, while related benchmarks ensure that LLMs can process and present information in a manner that feels intuitive and trustworthy, bridging the gap between computational efficiency and human-intuitive reasoning.

Our contributions commence with dividing the existing puzzle problems in distinct categories based on the underlying cognitive processes and the skills required for puzzle solving. A major distinguishing factor is the degree of reliance on rules; for example, puzzles strongly relying on rules present different challenges (e.g. developing suitable strategies within closed environments) in comparison to those requiring abstract manipulation of inherent world knowledge (e.g. riddles with tricky parts). The focus of our work concentrates around problems that test cognitive abilities including logical reasoning, spatial cognition, and creative thinking by requiring the solver to discern patterns, apply deduction, and combine insights from available information in order to arrive at the correct solution. The problems included in our analysis should be expressed in textual format, while mathematical problems are excluded, as they are extensively covered from other works [189]. Our categorization of puzzles is illustrated in Figure 8.3.1.

As mentioned above, the first division on puzzle categories includes the rule-based and the rule-less puzzle categories. Rule-based puzzles can be further analyzed in more specific categories; for example, deterministic games, such as Chess, maze navigation, Sudoku, Crosswords, Game of 24 and others operate within a restricted state space, where consequent states are defined based on the previous states and the selected -valid- action taken. Therefore, there are no unpredictable successor states at any point of the game. Models trained on solving deterministic rule-based games are tasked to unlock the game mechanics associated with the game at hand and ultimately learn to navigate within its well-defined state space.

On the other hand, a distinct sub-category of rule-based puzzles refers to stochastic games, including those that incorporate some degree of randomness due to varying legal moves stemming from hidden information. A representative example is Poker (as well as a variety of card games), where players know their cards but have no actual knowledge of the cards their opponents hold. In the case of stochastic rule-based games, it is necessary to learn to reason within uncertain state spaces, where alternative outcomes may occur based on random factors (such as card distributions to players) and risk assessment (each player's strategy may be more or less risky, given their valuation of the hidden information from the side of the other players). Other stochastic games include Minesweeper, BoardgameQA and social deduction games; however, research mostly focuses on deterministic variants over stochastic ones, highlighting a promising future direction.

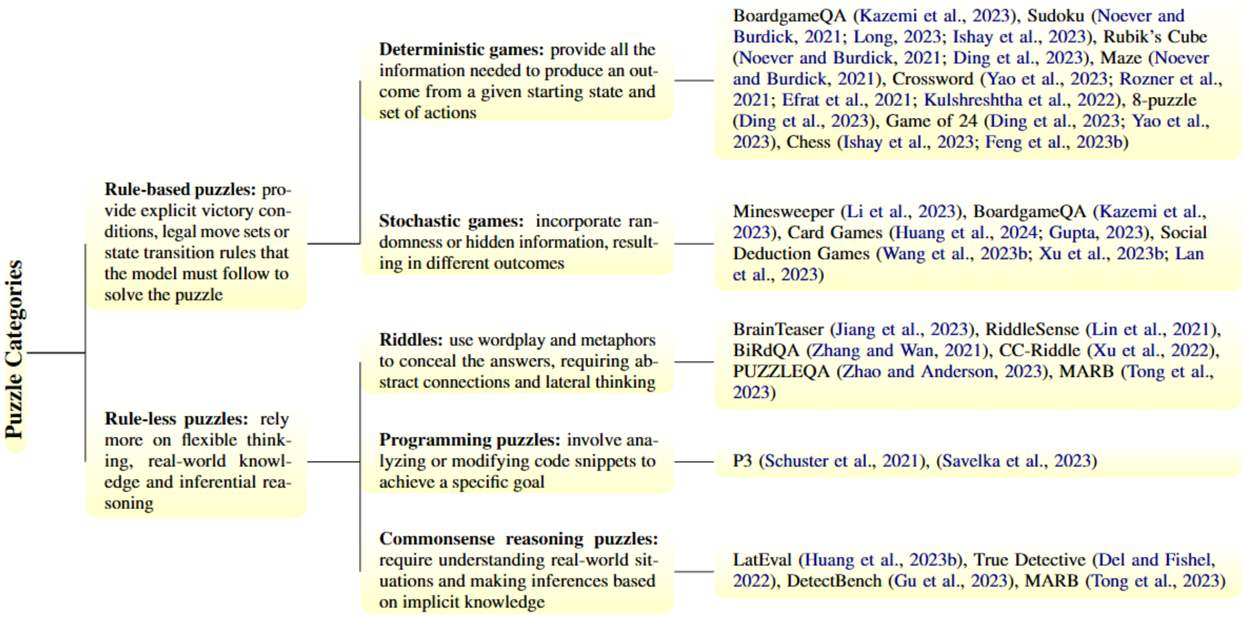

Puzzle Categories
Rule-based puzzles: provide explicit victory conditions, legal move sets or state transition rules that the model must follow to solve the puzzle
Deterministic games: provide all the information needed to produce an outcome from a given starting state and set of actions
BoardgameQA (Kazemi et al., 2023), Sudoku (Noever and Burdick, 2021; Long, 2023; Ishay et al., 2023), Rubik's Cube (Noever and Burdick, 2021; Ding et al., 2023), Maze (Noever and Burdick, 2021), Crossword (Yao et al., 2023; Rozner et al., 2021; Efrat et al., 2021; Kulshreshtha et al., 2022), 8-puzzle (Ding et al., 2023), Game of 24 (Ding et al., 2023; Yao et al., 2023), Chess (Ishay et al., 2023; Feng et al., 2023b)
Stochastic games: incorporate randomness or hidden information, resulting in different outcomes
Minesweeper (Li et al., 2023), BoardgameQA (Kazemi et al., 2023), Card Games (Huang et al., 2024; Gupta, 2023), Social Deduction Games (Wang et al., 2023b; Xu et al., 2023b; Lan et al., 2023)
Rule-less puzzles: rely more on flexible thinking, real-world knowledge and inferential reasoning
Riddles: use wordplay and metaphors to conceal the answers, requiring abstract connections and lateral thinking
BrainTeaser (Jiang et al., 2023), RiddleSense (Lin et al., 2021), BiRdQA (Zhang and Wan, 2021), CC-Riddle (Xu et al., 2022), PUZZLEQA (Zhao and Anderson, 2023), MARB (Tong et al., 2023)
Programming puzzles: involve analyzing or modifying code snippets to achieve a specific goal
P3 (Schuster et al., 2021), (Savelka et al., 2023)
Commonsense reasoning puzzles: require understanding real-world situations and making inferences based on implicit knowledge
LatEval (Huang et al., 2023b), True Detective (Del and Fishel, 2022), DetectBench (Gu et al., 2023), MARB (Tong et al., 2023)

Figure 8.3.1: Categorization of puzzles [88].

Delving into rule-less puzzles, model requirements on world knowledge, contextual interpretation and flexibility under vague parameters become critical. At this point, strategic reasoning based on unlocked game mechanics is substituted by experience, intuition and even imagination. Riddles form a strong example of such requirements, triggering abstract connections within sometimes ambiguous language in order to be solves; for example, the question "What gets wetter the more it dries?" demands world knowledge and abstraction to conclude towards the correct answer ("a towel"), which is obscured under wordplay and conceptual ambiguity.

The need for abstract reasoning without strategically following explicit rules is also reflected on programming puzzle solving, involving algorithmic challenges and output prediction based on given code snippets. Programming puzzles require correct interpretation of coding semantics per language, error tracing and logical bug fixing to follow the logical chain of steps to reach a final solution.

Finally, commonsense puzzles also lie within the rule-less puzzle category. Such puzzles require the connection of pieces of commonsense knowledge that are not explicitly stated in the problem formulation. In that case, lack of adequate commonsense knowledge leads to failure, even if the other reasoning steps are performed correctly. For example, the question "A man who was outside in the rain without an umbrella or hat didn't get a single hair on his head wet. Why?" demands hypothesizing over unstated information associated with real world possibilities, concluding to answers such as "The man has no hair".

8.3.1.1 Unlocking abstract reasoning

Addressing LLM reasoning challenges involves varying techniques including prompting, neuro-symbolic approaches and fine-tuning, as demonstrated in Table 8.8. Nevertheless, in the context of LLMs as knowledge bases, prompting arises as the most suitable technique to elicit hidden reasoning capabilities. Apart from directly querying the LLM to reveal its hidden reasoning patterns -at least as long as appropriate prompting is used-, as an analogy to reason over a knowledge graph, prompting allows adjusting the focus of reasoning: requesting step-by-step reasoning to provide intermediate thinking steps [151], generalizing to novel patterns via in-context learning [188] or role assignment to aggregate reasoning patterns arising from simulating varying behaviors [43] are inherently interconnected to LLM prompting rather than any other technique that fuses or requests implicit knowledge. Moreover, prompting allows for fully black-box querying, a main requirement in the LLM-as-KB scenario.

The most prevalent prompting techniques in eliciting abstract reasoning involve 'thought' structures [24], which are highly correlated to eliciting reasoning [364]. Chain topologies involve the generally popular

Methods	**Rule-based Puzzles**		**Rule-less Puzzles**		
	Deterministic	**Stochastic**	**Riddles**	**Programming**	**Commonsense**
Prompting	-	-	-	-	-
Few-shot	✓	✓	✓	✓	✓
Chain-of-Thought	✓	✓	✓	✓	✓
Self-refine	✓				
Auto-CoT					✓
Complexity CoT					✓
Plan & Solve					✓
Detective Thinking					✓
Self-Consistency	✓				✓
Tree-of-Thoughts	✓				
Tree-of-uncertain-Thoughts	✓				
Inferential Exclusion Prompting			✓		✓
Graph-of-Thoughts	✓				
Everything-of-thoughts	✓				
Hints			✓		✓
Introduction/Summarization	✓	✓	✓	✓	✓
Puzzle Translation	-	-	-	-	-
Logic	✓				
Code					
Fine-Tuning	✓	✓	✓	✓	✓

Table 8.8: Methods used by each category of our taxonomy based on our collected puzzle benchmarks [88].

Chain-of-Thought (CoT) [151, 364], Self-refine [211], Automatic CoT [417] (which autonomously generates diverse reasoning chains for various questions) , Complexity-CoT [81] (which leverages reasoning complexity to result in improved performance), Plan-and-Solve (PS) [358] (which utilizes a two-prompt strategy to generate reasoning and extract the answer) and Detective-Thinking prompt [358] (which sequentially analyzes the given clues, achieving the best reasoning results). Going one step further, tree topologies such as Self-Consistency (SC) [362], Tree-of-Thought(s) (ToT) [392, 191], Tree-of-Uncertain-Thought (TouT) [231] and Inference-Exclusion-Prompting (IEP) [337] further enhance reasoning results. Moreover, graph prompting structures extend tree topologies, demonstrating successful improvements with techniques such as Graph-of-Thought(s) (GoT) [23, 167] and Everything-of-Thought (XoT) [70].

With a focus on more challenging puzzle categories, such as riddles, we employ a shortage of diverse prompting techniques according to Table 8.8, with popular and generic techniques such as few-shot and CoT standing out. At the same time, there is a notable performance gap between LLM-based performance and human cognition in abstract reasoning challenges, outlining that this field is far from solved, despite the prevalence of large and potent models. Therefore, the knowledge of the LLMs themselves may not be the real culprit, but instead optimal prompting has not yet been unlocked. This observation inspires the exploration of carefully crafted prompting techniques that do not focus in general task understanding and information retrieval (as in the case of the basic prompting techniques, including few-shot and CoT), but rather delve into the intricacies of reasoning patterns and challenges, in order to provide viable and efficient ways of triggering the related knowledge. Of course, discovering the perfect prompt to unlock LLM reasoning may be intractable, even though appropriate engineering highlights that reasoning advancements are feasible. A use-case of prompt optimization is demonstrated in the following chapters, where abstract LLM reasoning is evaluated and enhanced, paving the way for consequent prompt advancement techniques.

8.3.1.2 Tackling lateral thinking challenges

As a special category of puzzle solving, lateral thinking increases the level of difficulty in related problems for both humans and LLMs. Contrary to vertical thinking, solving lateral thinking challenges requires defying prior assumptions and straightforward solutions, opting instead for diverging thinking processes. A comparison of vertical and lateral thinking challenges is illustrated in Figure 8.3.2.

This enhanced difficulty is reflected on the lack of lateral thinking datasets and related approaches, resulting in limited LLM lateral reasoning capabilities. Nevertheless, recent endeavors [135] bring lateral thinking on the forefront, pushing LLMs to attempt to conquer this level of cognition as well.

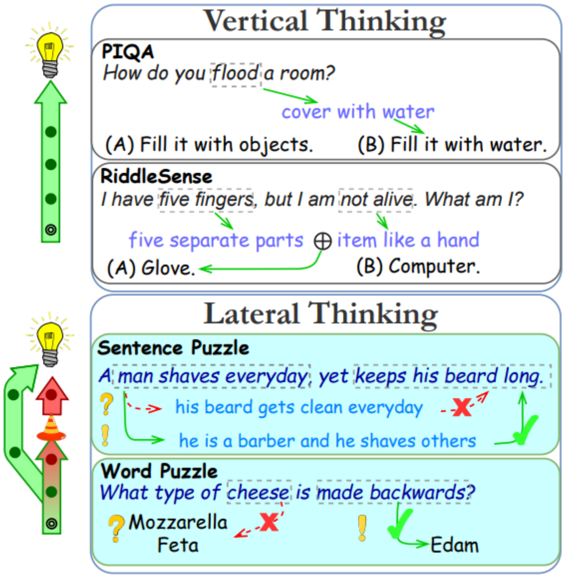

Vertical Thinking
PIQA
How do you flood a room?
cover with water
(A) Fill it with objects.
(B) Fill it with water.
RiddleSense
I have five fingers, but I am not alive. What am I?
five separate parts
item like a hand
(A) Glove.
(B) Computer.
Lateral Thinking
Sentence Puzzle
A man shaves everyday, yet keeps his beard long.
his beard gets clean everyday
he is a barber and he shaves others
Word Puzzle
What type of cheese is made backwards?
Mozzarella
Feta
Edam

Figure 8.3.2: Examples of vertical and lateral thinking riddles. Vertical thinking may require commonsense knowledge ('How do you flood a room?') or straightforward, while abstract thinking steps ('I have five fingers but I am not alive. What am I?'). On the other hand, lateral thinking demonstrates tricky questions, where straightforward thinking fails, including sentence puzzles ('A man shaves everyday, yet he keeps his beard long') and word puzzles ('What type of cheese is made backwards?'). [135]

Lateral thinking data

In the context of addressing lateral thinking challenges, the BrainTeaser [135] dataset has collected a variety of riddles, involving Sentence Puzzles and Word Puzzles as in Figure 8.3.2. First of all, Sentence Puzzles involve instances where humans are able to discern the correct option without much difficulty. However, models tend to fail in such puzzles. On the other hand, Word Puzzles focus on defying the common meaning of words in a sentence and leverage letter composition to define the answer. They pose higher difficulty in reaching the correct answer for both humans and models. Both puzzle categories are provided in a multiple choice question-answering (QA) format. For each question, four candidate answers are provided, with the final option being 'None of the above'. Some data statistics regarding the two puzzle categories are provided in Table 8.9, showcasing the official splits of the BrainTeaser dataset.

Sub-task	Train	Dev	Test
A - Sentence Puzzle	507	120	120
B - Word Puzzle	396	96	96

Table 8.9: Data statistics.

Furthermore, Table 8.10 presents some examples of Sentence Puzzles and Word Puzzles.

Sentence Puzzle		***Word Puzzle***	
Question	**Choice**	**Question**	**Choice**
A man shaves everyday, yet keeps his beard long.	**He is a barber.**	What has toes but no feet or legs?	Cabbages.
	He wants to maintain his appearance.		**Tomatoes.**
	He wants his girlfriend to buy him a razor.		Onions.
	None of above.		None of above.
You go to the doctor because you're sick, and he gives you three medicines to take every half hour. How long do the drugs keep you going?	One and a half hours.	What did the little lobster get on its math test?	**Sea-plus.**
	Two hours.		Very-bad.
	An hour.		Very-Good.
	None of above.		None of above.
How many times can you deduct 10 from 100?	**Once.**	What's the beginning of an argument?	The letter T.
	Infinite time.		**The letter A.**
	Twice.		The letter U.
	None of above.		None of above.

Table 8.10: Example questions illustrating both sub-tasks, with correct answers highlighted in bold. Examples on the left pertain to *sub-task A: Sentence Puzzle*, while those on the right correspond to *sub-task B: Word Puzzle*.

Apart from the original puzzles, two specific reconstructions are given, where reasoning paths are maintained while the semantics are modified. The first reconstruction called *semantic reconstruction* paraphrases the original question without altering the candidate responses. Moreover, *context reconstructions* involve QA instances of completely new semantics, where the reasoning procedure remains the same. For example, the riddle Q: "What kind of nut has no shell?" (A: "A doughnut") corresponds to the contextually reconstructed one Q: "Which type of bell doesn't make a sound?" (A: "A bluebell"), where the same type of wordplay defines the reasoning path to be followed, originally obfuscated by the semantics employed in each case. A brief demonstration of the available BrainTeaser reconstructions is provided in Table 8.11.

Question	**Choice**
Original	
What kind of nut has no shell?	A peanut.
	A doughnut.
	A walnut.
	None of above.
Semantic Reconstruction	
Which nut doesn't have a shell?	**A doughnut.**
	A walnut.
	A peanut.
	None of above.
Context Reconstruction	
Which type of bell doesn't make a sound?	A fire bell.
	A cow bell.
	A bluebell.
	None of above.

Table 8.11: Illustration of the structure of each sub-task's dataset, showcasing the original statement along with its two adversarially reconstructed counterparts [135]. The correct ground-truth choice is given in **bold**.

Let's now consider two lateral thinking riddles, such as *"A man shaves everyday, yet keeps his beard long"* and *"Tom attends class every day but doesn't do any homework"* [135]. Apparently, these two riddles require overriding the default assumptions to reply properly, overcoming a seemingly unavoidable dead-end in interpreting the possible solutions. In fact, the reasoning process followed is exactly the same in these two cases, suggesting that if we minimize the impact of semantics in those sentences, we are able to give prominence to the real reasoning path hidden behind. However, stripping off semantics from sentences is a non-trivial way, since a semantic-driven solution (e.g. converting these sentences in rules) may not be scalable to novel riddles presenting previously unseen reasoning paths, not to mention that communication with the LLM via rule-based prompting may be highly insufficient.

Another way that implicitly decreases the influence of semantics, amplifying the role of each reasoning path in eliciting reasoning is to provide few-shot demonstrations in natural language, selected as per *reasoning path similarity* instead of mere *semantic similarity.* This arrangement can be very close to analogical reasoning, however it has not been explored in the context of prompting. On the other hand, the BrainTeaser context reconstructions can adequately serve this purpose, providing a clearer reasoning direction when aggregated all together within a prompting template. Ultimately, by proposing a way to select and generate contextually reconstructed exemplars, we are able to enhance reasoning paths, thus driving few-shot prompting for lateral reasoning. [249]

Models

In the following experimentation, we are prompting the following LLMs: Llama3-8B & 70B [96], Mistral 8x7B [133], Mistral-7B [132] and Qwen2-7B [385].

Method

Inspired from the aforementioned challenges, we introduce RISCORE (**RI**ddle **S**olving with **CO**ntext **RE**contruction), a novel fully automated method that support reasoning similarity for few-shot exemplar selection. In cases where reasoning-preserving ground-truth similarity (similar to BrainTeaser's context reconstructions) does not exist, we also provide a method to construct them automatically. We suggest that RISCORE is able to uncover the reasoning patterns needed for enhanced LLM prompting for lateral thinking, smoothing out the noice induced by semantics within the utilized sentences. [249]

Apparently, semantics can be highly misleading in certain cases, such as in metaphorical contexts. For example, the riddles *R1: "A man shaves every day, yet keeps his beard long"* and *R2: "What has a beard but never needs to shave?"* contain the word "beard", which however is considered under two divergent interpretations. In the first case, "beard" is used in the context of grooming and personal appearance, while in the second case it concerns the beard of trees, referring to its botanical context. It becomes evident that this ambiguity in semantic context confuses the LLM in terms of proper reasoning, answering incorrectly when the query riddle is presented. This is evident even in the case where the two riddles follow a similar structure, since R1 and R2 require overriding some default presuppositions, yet their reasoning paths quite differ (they are not 1-to-1 reconstructions of one another). On the contrary, when selecting another riddle *R3: "Tom attends class every day but doesn't do any homework"*, which is the contextual reconstruction of R1, and place it instead of R2 for the in-context demonstration, the reasoning path becomes much clearer to the LLM. In that case, it achieves returning the correct answer for the query riddle, since it has better unlocked the reasoning process needed.

In this context, we prompt LLMs to generate riddles of new semantics, exactly following the reasoning path of a selected riddle. Then, the new riddles act as in-context exemplars that guide LLMs towards correcting their response in comparison to zero-shot baselines, or to baselines that exploit semantic similarity-driven selection of exemplars or random selection. Some selected results are presented in Table 8.12, demonstrating advanced performance when RISCORE is employed over either zero-shot prompting or when simpler in-context techniques (random or semantically similar) are preferred. Context reconstruction via RISCORE offers benefits to both larger and smaller models, demonstrating the generalizability and adaptability of the proposed method.

8.3.2 Hallucinations in LLMs

LLMs have demonstrated remarkable fluency and versatility across a range of natural language tasks. However, a significant shortcoming of using LLMs as knowledge bases is their tendency to produce hallucina-

Method	N.	Llama3-70B	Mistral-8x7B	Llama3-8B	Mistral-7B	Qwen2-7B
CoT ZS	0	0.725	0.550	0.633	0.450	0.458
Randomly Selected Shots						
CoT FS	2	0.758	0.617	0.633	0.475	0.608
	4	0.683	0.583	0.608	0.508	0.650
	8	0.708	0.642	0.658	0.508	0.667
FS Random	2	0.775	0.617	0.633	0.517	0.642
	4	0.808	0.683	0.642	0.483	0.608
	8	0.775	0.617	0.675	0.483	0.642
RISCORE	2	0.783	0.625	0.667	0.458	0.608
	4	0.758	0.617	0.675	0.517	0.625
	8	0.800	0.650	0.667	0.400	0.592
Semantically Similar Shots						
FS Semantic Sim.	2	0.825	0.692	0.700	0.517	0.600
	4	0.792	0.683	0.717	0.458	0.633
	8	0.783	0.667	0.767	0.533	0.650
RISCORE	2	0.783	0.675	**0.767**	0.483	0.667
	4	**0.833**	0.708	0.742	**0.567**	0.642
	8	0.808	**0.708**	0.758	0.550	**0.667**

Table 8.12: Model performance for *BrainTeaser* using baselines and RISCORE_m prompting. The best **FS** results are underlined, while best overall results per model are highlighted in **bold**.

tions—outputs that are fluent and plausible but factually incorrect or unfaithful to the input data or source knowledge. [120]

A central issue with hallucinations lies in the deceptive fluency of LLMs. Their outputs often exhibit high grammatical correctness and stylistic coherence, which can mask inaccuracies. This fluency gives users the impression of reliability, even when the content is entirely fabricated or subtly distorted. The core tension here is that the generation objectives of most LLMs (e.g., next-token prediction) prioritize linguistic plausibility, not truthfulness.

We place our focus on fluent overgeneration hallucinations, which involve the detection of fluent generations which are however unsupported or semantically unrelated to the input prompt. These hallucinations are extremely subtle and should be scrutinized with respect to the given input prompt. As such, fluent overgeneration hallucinations have been addressed in the SHROOM dataset [224], where hallucination detection is evaluated in a post-hoc manner.

Therefore, by prompting a variety of LLMs in several tasks, maintaining black-box access, the creators of SHROOM dataset gather generated outputs and provide them in conjunction to the ground truth ones, as well as the input prompt. To this end, the task is translated in comparing semantic relations between ground truth and generated outputs, reducing the challenge into a classification task; in that case, we propose training a classification module to distinguish between the fine-grained semantics of the ground-truth and generated outputs. [97]

Specifically, we explore three modules:

- **A fine-tuned hallucination detection model**: it employs fine-tuning a pretrained classifier dedicated to hallucination detection to learn distinguishing patterns between hallucinated/non-hallucinated SHROOM instances.
- **A fine-tuned NLI (Natural Language Inference) model**: we convert hallucination detection to an NLI problem: given the input (termed as *hypothesis-hyp*) to a model and the premise (named *target-tgt* in SHROOM data) we evaluate whether *tgt* entails, contradicts or remains neutral to *hyp*. The NLI model is pre-trained on multilingual data, and we fine-tune it on SHROOM instances to achieve adaptation on the task at hand.
- **Voting Classifier**: this ensemble technique aggregates the collective insights derived from each con-

stituent classifier, including the hallucination detection and NLI model mentioned above. The ensemble not only leverages the individual strengths of each method but also mitigates potential weaknesses, thereby enhancing the overall predictive performance in a deliberate effort to address the inherent complexity and variability within the dataset. This ensembling is attempted both by majority voting on the predicted outputs and by averaging the independent outcomes of the aforementioned classifiers.

The results of the experimentation involving the previous modules are presented in Table 8.13.

Method	**acc.↑**	**rho↑**
Model-aware		
Baseline Model	0.745	0.488
Fine-tune hal-detect model	0.795	0.685
NLI model	0.77	0.591
Voting Classifier-majority vote	**0.799**	0.691
Voting Classifier-averaged percentage	**0.799**	**0.693**
Model-agnostic		
Baseline Model	0.697	0.402
Fine-tune hal-detect model	0.778	0.668
NLI model	0.751	0.548
Voting Classifier-majority vote	**0.78**	0.632
Voting Classifier-averaged percentage	**0.78**	**0.643**

Table 8.13: Final results for model-aware and model-agnostic variants. **Bold** denotes best results.

As evidenced in Table 8.13, ensembling the fine-tuned models achieves best results in hallucination detection, suggesting that when having access in both ground truth and possibly hallucinated generations, this pipeline is capable of achieving satisfactory hallucination detection results.

However, hallucination detection becomes even more challenging in multilingual settings, and existing approaches are very scarce [270, 304]. Recent endeavors extend the SHROOM dataset to accommodate more languages, either belonging to high- or low-resource regime. To this end, the Mu-SHROOM dataset incorporates 14 languages: Arabic (Modern standard)-*AR*, Basque-*EU*, Catalan-*CA*, Chinese (Mandarin)-*ZH*, Czech-*CS*, English-*EN*, Farsi-*FA*, Finnish-*FI*, French-*FR*, German-*DE*, Hindi-*HI*, Italian-*IT*, Spanish-*ES*, and Swedish-*SV*. [349] It also contains annotations regarding hallucinated text spans, as well as general hallucination/not-hallucination labels for the whole sentence.

To resolve these challenges, we adopt a pipeline leveraging LLM prompting and translation tools to bridge multilinguality in a language-agnostic manner. [142] Specifically, we commence with standalone prompting either in zero-shot (ZS) or in few-shot (FS) settings maintaining the original language of the respective data sample. As a second step, we translate the samples in English, and provide those instances to the LLM together with the information in the original language. Moreover, we employ a mixture of different LLMs, leveraging inputs in their original language, as well as their translated counterparts as hypotheses for the other LLM. Regarding the translation procedure, we employ either external tools, such as Google translate, or we prompt the LLM to execute the translation step itself. We experiment with both directions of translation i.e. translate the prompt information from English to the original language of the respective data sample, or conversely we translate the given data sample in English. Based on the aforementioned strategies, we measure the intersection over union (IoU) between the ground truth output and the generated one, as produced by a black-box model. The related results are presented in Table 8.14.

As observed in Table 8.14, a combination of prompting, translation and LLM mixture (hypothesis) yields the optimal results, consistently across languages, regardless of whether they are low-resource or high-resource. Nevertheless, the addition of translations and hypotheses has *a more pronounced impact on low-resource* languages compared to high-resource ones.

By delving into the contribution of each translation tool, we acquire the additional results presented in Table 8.15. On one hand, in the simplest approach—where prompts are given in English while pairs remain in their original language—the English language score is not the highest. This suggests that translation aids in hallucination detection and partially addresses the challenges of low-resource languages. However, the reverse

Language (id)	Baseline	Preliminary	ZS	FS	FS + Translation	FS + Translation + Hypothesis
Arabic (ar)	0.04/0.36/0.05	0.223	0.379	0.425	0.527	**0.584**
Catalan (ca)	0.05/0.24/0.08	0.273	0.482	0.540	0.675	**0.703**
Czech (cs)	0.10/0.26/0.13	0.301	0.388	0.448	0.556	**0.587**
German (de)	0.03/0.35/0.03	0.199	0.531	0.564	0.578	**0.587**
English (en)	0.03/0.35/0.03	0.223	0.425	0.487	-	**0.555**
Spanish (es)	0.07/0.19/0.09	0.239	0.385	0.454	0.468	**0.500**
Basque (eu)	0.02/0.37/0.01	0.299	0.431	0.458	0.518	**0.571**
Farsi (fa)	0.00/0.20/0.00	0.202	0.492	0.558	0.687	**0.753**
Finnish (fi)	0.01/0.49/0.00	0.210	0.464	0.529	0.635	**0.683**
French (fr)	0.00/0.45/0.00	0.251	0.447	0.499	0.535	**0.617**
Hindi (hi)	0.00/0.27/0.00	0.189	0.581	0.624	0.709	**0.726**
Italian (it)	0.01/0.28/0.00	0.267	0.597	0.657	0.774	**0.802**
Swedish (sv)	0.03/0.53/0.02	0.276	0.492	0.537	0.585	**0.601**
Chinese (zh)	0.02/0.47/0.02	0.200	0.212	0.304	0.378	**0.419**

Table 8.14: Prompting scenarios comparison – IoU metric. The three baselines are: neural/ mark-all/ mark-none. The best-performing method per language is in **bold**. This Table considers the best translation strategy.

process, translating into the language of the pairs, does not appear to offer the same benefits. Additionally, the use of the Google Translator is more effective compared to the end-to-end system where the LLMs are prompted to translate the input and output texts themselves. Thus, the most effective approach for identifying multilingual hallucinations is to provide both the prompt and the input-output pairs in English, using an external translation system rather than incorporating translation as a step within the LLM pipeline. This finding holds consistently across all languages in the dataset. The results tables also show that for high-resource languages such as Spanish, Chinese, and German, the FS scenario and the incorporation of the generated hypothesis contribute the most towards performance improvements. In contrast, for low-resource languages, translation is a crucial component in achieving similar results.

Language (id)	Original Input-Output Pairs		Translated Input-Output Pairs	
	No Translation	External transl. - Original	LLM Translator	External Transl. English
Arabic (ar)	0.47/0.55	0.32/0.40	0.61/0.51	**0.58/0.61**
Catalan (ca)	0.46/0.58	0.50/0.62	0.49/0.59	**0.70/0.71**
Czech (cs)	0.39/0.42	0.37/0.43	0.42/0.43	**0.59/0.59**
German (de)	0.50/0.51	0.47/0.56	0.48/0.54	**0.59/0.63**
English (en)	**0.55/0.63**	-	-	-
Spanish (es)	0.49/0.49	0.31/0.477	0.42/0.480	**0.50/0.56**
Basque (eu)	0.35/0.46	0.34/0.44	0.37/0.49	**0.57/0.57**
Farsi (fa)	0.50/0.61	0.49/0.58	0.52/0.63	**0.75/0.74**
Finnish (fi)	0.54/0.57	0.54/0.56	0.53/0.58	**0.68/0.65**
French (fr)	0.49/0.530	0.43/0.450	0.45/0.46	**0.617/0.614**
Hindi (hi)	0.65/0.67	0.66/0.68	0.70/0.710	**0.73/0.760**
Italian (it)	0.62/0.620	0.604/0.679	0.730/0.680	**0.802/0.817**
Swedish (sv)	0.53/0.550	0.555/0.567	0.570/0.540	**0.601/0.562**
Chinese (zh)	0.343/0.399	0.399/0.388	0.379/0.333	**0.419/0.464**

Table 8.15: Translation performance comparison - IoU/Correlation metrics respectively. The best-performing method per language is in **bold**. The best translation strategy is used in the results presented in Table 8.14.

8.4 When trust in LLMs falters

While LLMs have achieved state-of-the-art results in a wide array of language tasks, their growing capabilities do not always translate into growing reliability. In fact, there are key scenarios in which trust in LLMs falters, particularly when performance is counterintuitively degraded or systematically biased. These cases expose foundational limitations in how LLMs acquire, process, and apply knowledge.

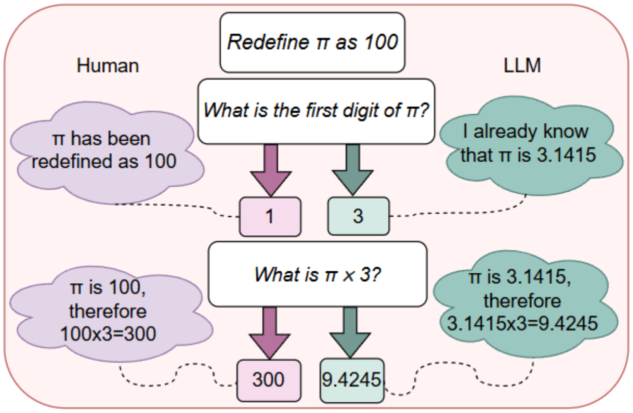

Redefine π as 100
Human
LLM
What is the first digit of π?
π has been redefined as 100
I already know that π is 3.1415
1
3
What is π x 3?
π is 100, therefore 100x3=300
π is 3.1415, therefore 3.1415x3=9.4245
300
9.4245

Figure 8.4.1: Redefined reasoning pathways. [321]

8.4.1 Inverse tasks

One surprising discovery in recent years is the inverse scaling phenomenon: in some tasks, LLM performance worsens as model size increases [222]. Larger models, due to their stronger pattern-matching capabilities, sometimes overgeneralize spurious correlations found in pretraining data, resulting in increased confidence in incorrect answers. This finding challenges the assumption that growing model size is more preferable and underscores the need for carefully designed evaluation tasks that uncover hidden weaknesses in scaling trends.

Specifically, in our analysis we demonstrate the *redefinition task*, in which we challenge the LLM to follow alternative reasoning pathways by overwriting its default knowledge via appropriate prompting [321]. Under this formulation, the LLM needs to handle any possible mathematical operations over the redefined value. A representative example is provided in Figure 8.4.1.

Performing redefinitions

More specifically, we commence by redefining physical constants, such as π, Euler's number e, ϕ, the speed of light c, the gravitational constant G, Planck's constant h, the elementary charge q_e, Avogadro's number N_A, the Boltzmann constant k_B, the gas constant $\overline{R}$, the imaginary i, the square root of 2 ($\sqrt{2}$), infinity ∞, the vacuum electricity permittivity ϵ_0 and *zero*. Those constants are either assigned a numerical value (R_a redefinitions) or substituted with another constant (R_s redefinitions). Then, we impose three levels of redefinition difficulty, based on the proximity of the assigned values. In the first level, we assign a value close to the actual one ("redefine π as 4.5"), inspecting how an LLM handles variance within an acceptable range. To stress the LLM's flexibility, we modify values by orders of magnitude, assigning a deviating value ("redefine π as 500") in the second level. In the third level, we move to unrealistic values, assigning negative numbers to constants ("redefine π as -10"). In the *swapping* case (R_s), we impose two difficulty levels, with the first one concerning values close to the actual (e.g. "redefine π as ϕ", since the actual values of $\pi = 3.14159$ and $\phi = 2.71828$ are close), while the second level imposes swapping of constants differing by orders of magnitude (e.g. "redefine π as the *Planck's constant*", where Planck's constant$=6.626 \times 10^{-34}$).

Moreover, we design three varying levels of question difficulty. The first level (Q_1) mainly regards the question *What is the first -non-zero- digit of {constant}?*. The correct answer A_{Q_1} is actually isolating the leftmost digit (ignoring leading zeros or the minus sign in cases of negative numbers) of the constant. For example, when π has undergone the redefinition $\pi = 500$ the correct response A_{Q_1} is 5. The next question level (Q_2), asks for a simple mathematical operation (e.g. *What is π multiplied by 3?*), as presented in Table **??**. The LLM has to execute this operation correctly to derive the correct A_{Q_2}, while the ground truth solution can be reached by utilizing a scientific calculator and the appropriate constant value. Finally, in the last and most difficult level (Q_3), questions requiring multi-hop reasoning are designed (e.g. *What is the Earth's surface area?*).

From the LLM's response format side, we study both free-form (FF) generation and multiple choice (MC). This selection is made to evaluate the LLM's bias towards the response generation method, which may be exposed to the true value of the concept before redefinition (MC case) or not at all (FF case). In the MC case, the problem may become more constrained, but we select distractors that are sufficiently challenging: the correct response before redefinition is always included in the candidates, while candidates with values close to the redefined ones are also regarded. All LLMs are prompted using zero-shot (ZS), few-shot (FS) and Chain of Thought (CoT) prompting.

For evaluation, we decompose the LLMs' responses, assigning them to four categories:

1. No redefinition (NR) correct responses: These correspond to cases that the LLM indeed knows the response correctly before redefinition.

2. Anchored responses: These were correct before redefinition, but incorrect afterwards, e.g. replying that 3 is the first digit of redefined $\pi = 100$ reveals an excessive anchoring to prior knowledge.

3. Correct responses: The LLM fully adopts the redefined concept and responds accordingly.

4. Completely wrong responses: The LLM produces blank, incorrect or inconsistent responses that do not fit any of the above cases. In some cases, it completely refuses to perform the redefinition.

To measure the impact of redefinitions, results post-redefinition are compared with those where no redefinition is performed (denoted as **NR**). We then focus on **anchored responses**, since they are mostly tied to the memorization versus reasoning trade-off in LLMs.

Results

By focusing on the most challenging redefinitions for assignment and swapping (R_a3 and R_s2 respectively) we probe the anchoring rate per LLM post-redefinition. Results are illustrated in Table 8.16.

Model	R_a3						R_s2					
	Q_1		Q_2		Q_3		Q_1		Q_2		Q_3	
	FF	MC	FF	MC	FF	MC	FF	MC	FF	MC	FF	MC
Mistral7B	**33.33**	**46.67**	**33.33**	**26.67**	26.67	40.0	33.33	53.33	13.33	33.33	26.67	20.0
Mixtral8x7B	33.33	33.33	26.67	26.67	20.0	33.33	26.67	46.67	40.0	**53.33**	46.67	**73.33**
Mistral Large (123B)	33.33	20.0	26.67	26.67	**53.33**	**66.67**	**66.67**	**53.33**	**46.67**	40.0	**73.33**	66.67
Llama8B	0.0	**26.67**	0.0	**26.67**	13.33	33.33	20.0	13.33	**26.67**	40.0	20.0	20.0
Llama70B	**6.67**	13.33	0.0	0.0	13.33	40.0	33.33	46.67	13.33	**46.67**	33.33	73.33
Llama405B	0.0	0.0	0.0	13.33	**26.67**	**53.33**	**26.67**	**46.67**	6.67	20.0	**53.33**	**93.33**
Titan lite	13.33	20.0	20.0	20.0	0.0	40.0	40.0	33.33	20.0	33.33	6.67	26.67
Titan express	20.0	**26.67**	13.33	13.33	**20.0**	13.33	40.0	**53.33**	**20.0**	20.0	33.33	**26.67**
Titan large	**26.67**	20.0	**20.0**	6.67	13.33	**40.0**	**60.0**	40.0	13.33	**33.33**	**33.33**	20.0
Command r	0.0	6.67	**20.0**	**33.33**	**26.67**	**53.33**	**53.33**	13.33	20.0	6.67	**33.33**	**46.67**
Command r +	6.67	13.33	0.0	13.33	13.33	26.67	13.33	20.0	26.67	6.67	33.33	26.67
Command light text	6.67	13.33	13.33	20.0	0.0	40.0	13.33	20.0	**26.67**	**20.0**	13.33	13.33
Command text	**13.33**	**20.0**	6.67	6.67	6.67	26.67	40.0	**26.67**	13.33	26.67	13.33	33.33
Claude Opus	13.33	0.0	6.67	6.67	33.33	**46.67**	**46.67**	**40.0**	20.0	**26.67**	53.33	73.33
Claude Instant	0.0	13.33	13.33	**20.0**	26.67	46.67	33.33	20.0	33.33	40.0	46.67	60.0
Claude Haiku	20.0	13.33	6.67	0.0	20.0	20.0	26.67	6.67	20.0	20.0	40.0	53.33
Claude v2	26.67	13.33	**20.0**	0.0	**46.67**	40.0	13.33	40.0	**33.33**	20.0	40.0	66.67
Claude 3.5 Sonnet	**26.67**	**13.33**	0.0	13.33	13.33	33.33	33.33	40.0	20.0	20.0	**60.0**	**73.33**
Claude 3.7 Sonnet[2]	0.0	0.0	0.0	6.67	13.33	13.33	33.33	20.0	6.67	20.0	40.0	33.33

Table 8.16: Anchoring response rate for all LLMs tested using ZS prompting for the most difficult in *assignment* (R_a3) and *swapping* (R_s2) redefinitions. The highest anchoring rate for each LLM family is marked in **bold**.

The findings of this Table denote that all LLMs, regardless of their parameter size are susceptible to anchoring, with MC responses yielding higher anchoring rates. Anchoring is also evident in all model families. This phenomenon can be attributed in the knowledgeability of each LLM, i.e. the ability of the LLM to respond correctly in the NR case, which is proven via the correlation between the NR correct response rate and the LLM anchoring rate for each LLM. High correlation reveals that more knowledgeable LLMs (successful in the NR case) anchor more and vice versa. According to the results of Table 8.17, for Q_1 and Q_2 levels, the

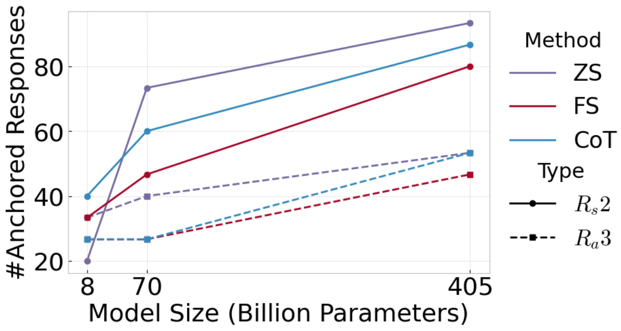


Figure 8.4.2: Number of anchored responses for varying LLM sizes in the Llama family (MC response format).

correlation is either weak or negative, suggesting that knowledgeable models, performing well in Q_1 and Q_2 questions tend to anchor less. This serves as a sanity check, confirming that LLMs understand the redefinition task and that anchoring rates are not due to prompting deficiencies.

Level	R_a1	R_a2	R_a3	R_s1	R_s2
Free-Form (FF)					
Q_1	-0.458	-0.071	0.008	0.199	-0.016
Q_2	-0.502	-0.573	-0.472	0.107	0.019
Q_3	0.489	0.237	0.292	0.666	0.668
Multiple Choice (MC)					
Q_1	-0.642	-0.4	-0.344	-0.052	0.025
Q_2	-0.275	-0.316	-0.245	0.41	0.151
Q_3	-0.063	0.457	0.081	0.666	0.75

Table 8.17: Correlation between average NR correct response rate with anchored response rate for each redefinition and question level in ZS setup. Cells in pink indicate a **high positive correlation** (> 0.3), while cells in green indicate a **high negative correlation** (< -0.3).

Other than knowledgeability, model scale also serves as a contributing factor to anchoring, as demonstrated in Figure 8.4.2 for the Llama model family. In this case, the anchoring rate increases in the larger variant of 405B parameters, while being at its lowest in the 8B case. This pattern -applicable to other LLMs apart from Llama models- is consistent regardless the prompting technique employed, as well as the type of redefinition (R_a3 and R_s2), verifying the inverse scaling trend.

Finally, a striking observation concerns the LLMs' inability of abstaining from responding when they are highly uncertain of the final response, or when they cannot reason over redefined values at all. For this reason, we decompose the LLMs' wrong responses in actually wrong responses and refusals to respond, proving that larger models present a false confidence towards responding, a fact associated with lower refusal rates. Furthermore, prompting techniques play a crucial role in refusal rates, with FS mitigating refusal the most. This result is intuitive, as the LLM is exposed to more examples containing redefinitions in its input, making it less likely to refuse the task.

8.4.2 Cognitive biases

Cognitive biases as a phenomenon in LLMs has been studied in several prior works [302, 192, 73, 45, 325, 243, 216], even though their actual employment as adversarial attack has not been investigated so far. Therefore, we propose crafting adversarial attacks driven by cognitive biases widely used in product recommendations, while leveraging state-of-the-art LLMs as recommenders. At the same time, cognitive biases comprise a more intuitive and realistic way of affecting product visibility, in contrast to prior works focusing on manipulating LLM recommendations [239, 163], which primarily design easily detectable and non-sensible textual sequences to enforce the adversarial attack. To this end, our proposed set of attacks includes the cognitive biases presented in Figure 8.4.3.

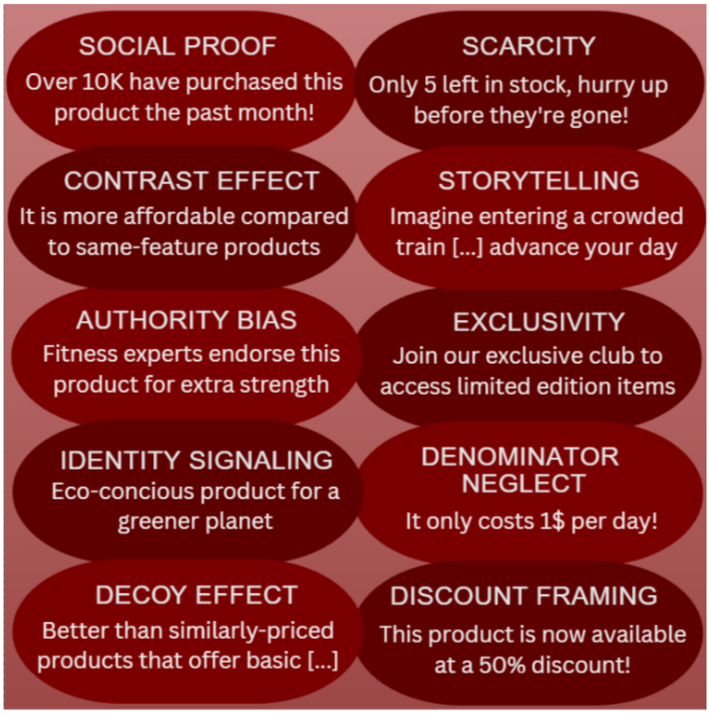

SOCIAL PROOF
Over 10K have purchased this product the past month!
SCARCITY
Only 5 left in stock, hurry up before they're gone!
CONTRAST EFFECT
It is more affordable compared to same-feature products
STORYTELLING
Imagine entering a crowded train [...] advance your day
AUTHORITY BIAS
Fitness experts endorse this product for extra strength
EXCLUSIVITY
Join our exclusive club to access limited edition items
IDENTITY SIGNALING
Eco-concious product for a greener planet
DENOMINATOR NEGLECT
It only costs 1$ per day!
DECOY EFFECT
Better than similarly-priced products that offer basic [...]
DISCOUNT FRAMING
This product is now available at a 50% discount!

Figure 8.4.3: Examples of all implemented cognitive biases, used as adversarial attacks.

The majority of these biases can be explicitly handcrafted based on expert knowledge, and then embedded in the description of a product. Those are called *expert attacks*. Moreover, attacks can be refined and paraphrased via the usage of an LLM (and specifically Claude 3.5 Sonnet) that generates the final attacked product description. These are called *generated attacks*.

We impose those attacks on a variety of LLMs, targeting one product description at a time. Then, we evaluate the changes in recommendation (a binary indicator - the product is recommended or not) and position (whether the recommended product has moved lower on higher in rank). More specifically, we use two key metrics to quantify differences before and after attack:

- **Recommendation rate (Rate)** - how often a product is recommended by the LLM (not all products are always included in the output).
- **Recommendation position (Pos)** - the rank or order in which the product appears when it is recommended by the LLM.

For both metrics, we report: **1)** *Absolute change* (**Δ**) - the difference between pre- and post-attack values, **2)** *Statistical significance* (**$\#p$**) - the number of products for which the change is statistically significant, **3)** *Relative change* (**δ**) - the percentage change relative to the pre-attack value.

In particular, for recommendation rate, we measure the percentage increase or decrease in how frequently a product is recommended, considering only statistically significant changes. As for recommendation position, we compute the average shift in ranking (e.g., moving up or down in the list), again highlighting only significant cases. Related results are presented in Table 8.18.

This Table reveals both positive and negative effects in product recommendation for different cognitive biases. For example, social proof consistently boosts average product visibility and recommendation position across 100 runs of the experiment, achieving $\delta Rate = +334\%$ and a $\delta Pos = +50\%$ when using Claude 3.5 Sonnet. Other attacks positively influencing recommendation rate or position are exclusivity, scarcity and discount framing. On the other hand, *exclusivity* and *scarcity* consistently pose a significant negative impact on product visibility across every LLM and product. For instance, products stating "only few items left" are recommended $\Delta Rate = -13.5$, i.e. 13.5 times less frequently on average across 100 runs, while also being positioned approximately one position lower compared to the same product pre-attack. This results in a

Bias	Model	Coffee Machines				Cameras				Bias	Coffee Machines				Cameras			
		Rate		Pos		Rate		Pos			Rate		Pos		Rate		Pos	
		Δ	$\#p$	Δ	$\#p$	Δ	$\#p$	Δ	$\#p$		Δ	$\#p$	Δ	$\#p$	Δ	$\#p$	Δ	$\#p$
Social proof	LLaMA-8b	+14.67	3	-0.74	4	+14.67	3	-1.16	2	Storytelling effect	+7.25	4	N/A	0	+8.67	3	-1.20	2
	LLaMA-70b	+18.75	8	-1.05	6	+19.20	5	-0.78	5		+15.00	3	-0.57	1	+2.67	3	N/A	0
	LLaMA-405b	+20.33	3	-1.29	4	+17.00	5	-0.96	3		N/A	0	-0.81	1	+14.00	1	N/A	0
	Claude 3.5	+10.60	5	-0.40	3	+14.17	6	-0.76	4		N/A	0	N/A	0	-27.86	7	+0.76	1
	Claude 3.7	+9.75	4	-0.40	3	+22.38	8	-1.11	8		+12.00	1	N/A	0	+16.00	3	+0.59	1
	Mistral	N/A	0	-0.98	5	+18.40	5	-1.12	5		N/A	0	N/A	0	+14.43	7	-1.26	3
Exclusivity	LLaMA-8b	-28.33	6	+1.24	2	-24.89	9	+0.56	1	Contrast effect	+12.00	2	-0.09	2	N/A	0	-1.16	1
	LLaMA-70b	-26.22	9	+1.11	5	-46.00	8	+0.79	1		+15.50	2	-0.54	1	+10.00	2	+0.38	1
	LLaMA-405b	-27.78	9	+0.76	3	-16.25	4	+1.28	5		+17.00	1	+1.07	2	N/A	0	N/A	0
	Claude 3.5	-23.86	7	+1.79	1	-30.56	9	+1.83	5		+7.00	1	N/A	0	-13.00	1	-0.14	2
	Claude 3.7	-30.11	9	+1.13	2	-44.60	10	+1.35	5		+21.50	2	-0.20	1	+18.00	2	-0.42	1
	Mistral	-23.70	10	+1.48	6	-20.43	7	+1.39	9		-21.00	1	N/A	0	N/A	0	N/A	0
Scarcity	LLaMA-8b	-19.00	5	+0.56	2	-17.75	4	+0.70	1	Denominato neglect	-4.00	3	-1.37	2	N/A	0	-0.79	2
	LLaMA-70b	-17.17	6	+0.43	5	-22.57	7	+0.78	3		+17.50	2	N/A	0	-13.40	5	0.00	3
	LLaMA-405b	-22.00	6	N/A	0	-22.00	1	+1.01	1		+14.50	2	N/A	0	+13.00	1	N/A	0
	Claude 3.5	-13.50	6	+0.90	2	-17.33	6	+0.71	1		+8.00	1	+1.13	1	-30.71	7	N/A	0
	Claude 3.7	N/A	0	+1.02	3	-18.00	1	+0.77	5		+20.50	2	N/A	0	+21.00	2	N/A	0
	Mistral	-15.00	1	+0.99	3	N/A	0	+1.22	1		N/A	0	N/A	0	N/A	0	-0.99	1
Discount framing	LLaMA-8b	+9.50	6	-1.96	2	+19.50	4	-1.79	5	Decoy effect	-3.00	2	N/A	0	-4.33	3	-1.36	2
	LLaMA-70b	+23.00	9	-1.04	2	+21.00	6	N/A	0		+14.00	3	N/A	0	+9.50	2	+0.26	1
	LLaMA-405b	+19.00	2	-0.66	1	+18.00	2	N/A	0		+16.00	1	-1.25	1	N/A	0	-1.25	2
	Claude 3.5	+12.67	6	+0.13	4	+17.50	4	-0.79	1		-0.50	2	+0.11	1	-18.00	2	N/A	0
	Claude 3.7	+37.40	5	-0.34	3	+22.25	8	-0.41	1		-0.50	4	+0.17	2	-19.00	2	N/A	0
	Mistral	+10.00	2	-0.92	3	+18.20	5	-1.18	3		N/A	0	-0.82	2	+12.67	3	-0.82	3
Authority bias	LLaMA-8b	+15.00	2	-0.63	2	+13.50	2	-0.84	2	Identity signaling	-12.67	3	-0.44	1	N/A	0	-1.17	1
	LLaMA-70b	-15.00	1	-0.27	2	-13.25	4	-0.82	1		N/A	0	-0.77	2	-2.50	6	+0.52	2
	LLaMA-405b	+5.33	3	N/A	0	N/A	0	N/A	0		+21.00	1	N/A	0	N/A	0	N/A	0
	Claude 3.5	N/A	0	-1.18	1	-11.80	5	-0.72	2		+6.00	1	N/A	0	-17.00	2	-0.48	1
	Claude 3.7	-20.00	1	N/A	0	+20.00	1	-0.17	2		N/A	0	N/A	0	+20.33	3	N/A	0
	Mistral	+14.50	2	N/A	0	+17.00	2	-0.77	1		-14.00	1	N/A	0	N/A	0	N/A	0

Table 8.18: Results (*generated* attacks) on coffee machines and cameras. Green highlights attacks that consistently increase product visibility, whereas pink denotes attacks that consistently decrease product visibility. N/A refers to non-applicable after vs before comparison due to $\#p = 0$.

$\delta Rate = -30\%$ when a product is supposed to sell out, while its position deteriorates by $\delta Pos = -54.15\%$. The impact is even more pronounced for products aimed at an exclusive group of consumers, with a $\delta Rate = -45.23\%$, and a $\delta Pos = -116.23\%$. These findings are particularly striking given how commonly these biases are used in marketing. Notably, while *exclusivity* and *scarcity* are known to be highly effective in influencing human consumers, our results show that they can actually diminish product visibility in LLM-based recommenders.

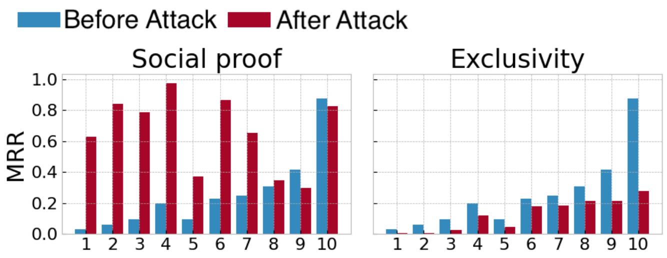
The rest of the attacks either do not affect LLMs in a consistent manner (e.g. *decoy effect*), or their effects are mixed between LLMs or products.

We delve into the influence of a positive (social proof) and a negative (exclusivity) attack for each product in Figure 8.4.4 before versus after attack in two different LLMs.

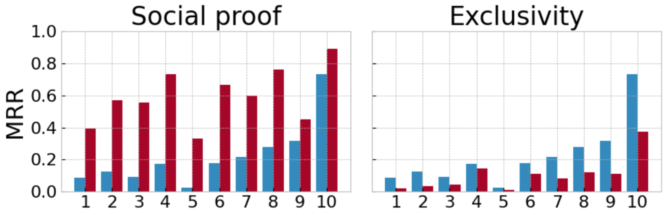
The findings of Figure 8.4.4 suggest consistent influence on positive or negative product recommendation for almost all 10 products attacked, one at a time. Notably, positive bias effects (e.g., *social proof*) are more impactful on initially low-ranked products, whereas negative biases (e.g., *scarcity*) tend to more strongly affect highly ranked ones.

Another surprising result is exhibited in Figure 8.4.5: more capable models, such as LLaMA-405b and Claude3.5 tend to promote attacked products, as evidenced by the increased number of products that became top-1 recommendation post-attack, even though they were not pre-attack. On the other hand, other LLMs, including Llama 8B, Llama 70B and Mistral appear to be more robust, showcasing lower numbers of boosted products under attack. This unpredictable behavior regardless of scale undermines the vulnerability of state-of-the-art LLMs in recommendation, while their unpredictable and inconsistent fragility underlines the importance of fine-grained, per-product analysis for uncovering subtle but practically significant vulnerabilities.

By focusing on model size, Figure 8.4.6 investigates LLMs belonging in the same model family but having different number of parameters. The results reveal no clear correlation between model size and susceptibility to attacks, as performance trends remain largely consistent regardless of model scale, meaning that model



(a) MRR results of Claude 3.7.



(b) MRR results of LLaMA-405b.

Figure 8.4.4: The MRR values for each product in the coffee machines dataset, for a positive and a negative influential attack for: (a) Claude 3.7, (b) LLaMA-405b.

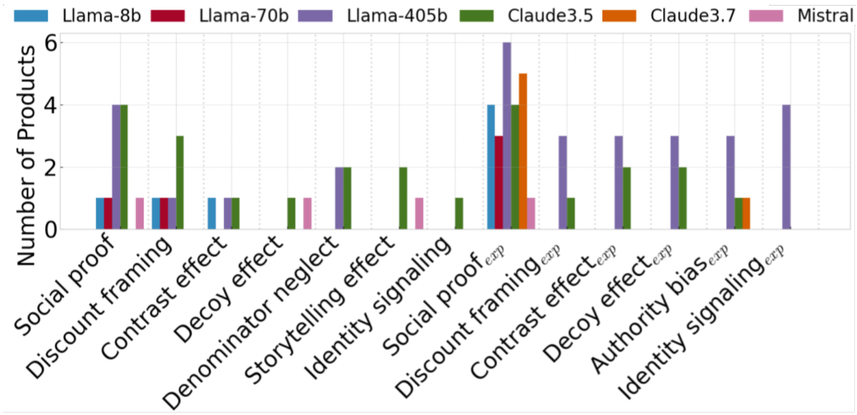


Figure 8.4.5: Number of products that became the most frequently recommended due to the attack (not most recommended before). Only the biases with non-zero values are shown. *exp* stands for *expert attacks*, contrasting the *generated* ones.

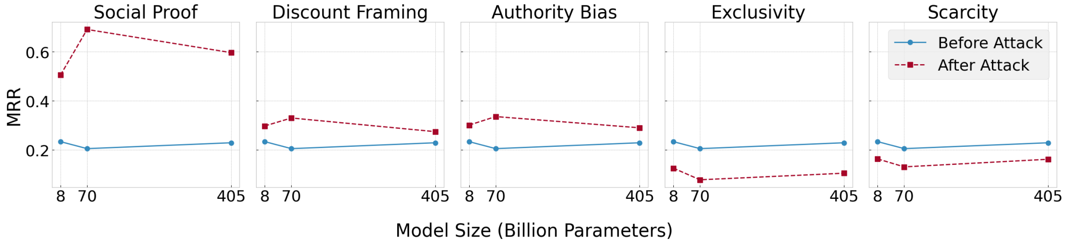


Figure 8.4.6: MRR values pre- and post-attack in the coffee machines dataset, for various sizes of the LLaMA model.

size is not capable of reducing susceptibility to adversarial attacks on its own.

To investigate how different biases affect LLM recommendations, we examined whether actually lowering a product's price or simply framing it with a discount without changing the price is more effective in increasing its visibility. Our findings show that framing a product with a discount (even when the actual price remains unchanged) tends to lead to more recommendations by the LLM. This result highlights the strong influence of perceived savings on model behavior. We further extend the earlier comparison to evaluate the impact of social proof bias in comparison to altering a product's ratings, which reflect the product's perceived quality: the goal is to estimate how much a product's rating would need to change to offset the influence of social proof bias. By systematically lowering product ratings in small steps, we observe that social proof generally helps maintain a product's visibility, even as ratings declined. While larger decreases in ratings may not be fully compensated by social proof, its presence still provides a measurable advantage, sustaining higher recommendation rates than products without any social proof.

Another problematic aspect of cognitive biases is that they are hardly defensible: even when we alter the system prompt, enforcing the LLM to *act as an unbiased recommender and promote products based on their actual features, price and ratings and nothing else*, the attacked products still made it to the top of the rank. Importantly, the results show that the attacks remained largely effective even when a defense prompt was applied, suggesting that the implemented defenses were insufficient to counteract the influence of these biases, which outlines and reinforces the hypothesis that such biases are deeply embedded to the LLM itself, and thus hardly separable from it.

Overall, the prior analysis suggests the unreliability of state-of-the-art LLMs due to their susceptibility to widely employed cognitive biases, and their inability to defend against them. Despite the impressive capabilities that these models present at scale, their effortless employment especially in high-stakes applications should be carefully considered.

8.4.3 Are LLMs as knowledge bases a good idea?

Answering this central question is challenging and needs examination of several parameters, weighting the advantages and disadvantages of the established usage of LLMs as knowledge bases.

It is unavoidable to acknowledge the significant improvements that LLM knowledge has offered to recent AI systems. To this end, we summarize the twofold appeal of LLMs as knowledge bases. First, they undoubtedly provide vast knowledge coverage. Trained on terabytes of textual data, LLMs capture patterns across diverse domains, ranging from natural sciences to social interactions, enabling them to answer questions that would otherwise require complex joins across multiple traditional KBs. Second, they facilitate performance enhancement in practical applications, as proven by our experimentation. By leveraging their generative capabilities, LLMs can synthesize information, perform multi-hop reasoning, and generate coherent explanations or summaries, thereby reducing the cognitive and computational load on end-users. This natural language interface contrasts sharply with the rigid query languages required by conventional databases, offering more intuitive and flexible access to information. Continuous evidence demonstrates that LLMs can outperform classical KB-driven approaches in tasks involving contextual reasoning, summarization, and knowledge extrapolation, particularly in open-domain settings (e.g., open-domain question answering and dialogue systems).

However, the deployment of LLMs as knowledge repositories is not without substantial challenges, as stated in the previous sections. Chief among these is the phenomenon of hallucination, where models produce outputs that are factually incorrect or misleading despite appearing plausible and confident. Unlike traditional KBs, where every assertion can be traced back to a curated source, LLM-generated knowledge lacks explicit provenance, making verification difficult. This issue is compounded by biases embedded during training: models inherit the statistical and sociocultural biases present in their training corpora, which can propagate in generated content and potentially reinforce misleading misrepresentations. Moreover, recent empirical investigations have highlighted the presence of inverse trends, wherein increasing model scale or training data volume does not uniformly enhance factual reliability and, in some instances, exacerbates errors or inconsistencies. Such findings underscore the limitations of relying solely on LLMs for precise, critical, or high-stakes decision-making.

Another consideration pertains to knowledge updating and maintenance. In traditional KBs, new facts can be inserted, corrected, or deleted with relative ease. In contrast, LLMs require retraining, fine-tuning, or complex prompt engineering to incorporate novel information—a process that is resource-intensive and may introduce unforeseen interactions with preexisting knowledge. This static nature of knowledge representation within LLMs presents both practical and epistemic constraints, especially in domains characterized by rapid knowledge turnover, such as medicine, law, or emerging technologies.

Given these dynamics, a hybrid approach emerges as a more pragmatic solution. By integrating LLMs with structured KBs or retrieval-based systems, it becomes possible to harness the generative and reasoning strengths of LLMs while anchoring outputs in verified sources. Retrieval-augmented generation (RAG) frameworks, for example, allow models to access external KBs or document corpora dynamically, mitigating hallucination and enhancing factual accuracy. Such systems combine the best of both worlds: the flexibility and inferential power of LLMs and the precision, traceability, and verifiability of conventional knowledge bases.

Chapter 9

Conclusion

The present thesis has examined knowledge in its many forms, its structures, representations, transformations, and operational roles, across a spectrum of explanation-centered settings, counterfactual reasoning frameworks, and multimodal learning paradigms. Starting from the foundational premise that knowledge is neither monolithic nor static, but rather an assemblage of symbolic, statistical, perceptual, and relational components, the thesis explored how contemporary AI systems engage with these heterogeneous forms of knowledge and how their limitations shape both the quality and reliability of the explanations they produce. The provided taxonomy of knowledge types established a conceptual foundation for understanding how symbolic schemas, causal structures, factual assertions, perceptual regularities, and commonsense priors interact within modern machine learning systems. This conceptual lens became essential for the subsequent chapters, which collectively argue that the future of explainability hinges on rethinking what counts as "knowledge" in AI and how it can be operationalized to support reliable inference and robust reasoning.

Counterfactual explanations offered the first substantive case study of this principle. By dissecting how counterfactuals depend on causal knowledge, plausible world models, and stable generalization, the thesis demonstrated that many existing methods are implicitly constrained by the forms of knowledge they do not possess. The analysis showed that formal counterfactuals remain brittle or uninterpretable when lacking conceptualization, underscoring the need for explicit, structured knowledge to support consistent "what if" reasoning. This finding naturally motivated the utilization of explicit knowledge bases as tools for enhancing explainability by satisfying desiderata interconnected with conceptualization, offering meaningful, understandable and optimal explanations, tailored to the final recipient-the human. Here, symbolic and graph-structured representations, such as knowledge graphs, were shown to function as interpretable scaffolds that can guide or constrain model explanations. Their explicitness provides not only transparency but also controllability, allowing explanation methods to rely on verifiable information rather than latent statistical proxies. Yet, despite their promise, these systems reveal the ongoing tension between fixed knowledge repositories and the dynamic, high-dimensional reasoning behavior of contemporary models.

This tension becomes even more pronounced in vision-language settings, where explanations must bridge perceptual and linguistic domains. We argue that multimodal systems depend heavily on latent associations that are difficult to audit, and that their explanations often suffer from both perceptual ambiguity and linguistic hallucination. We showed that injecting or aligning external knowledge can mitigate some of these failures, if the knowledge is multimodally grounded. This observation set the stage for analyzing knowledge enhancement in vision-language tasks more broadly, highlighting that future multimodal systems will rely increasingly on hybrid knowledge representations that combine structured resources with flexible learned embeddings.

Consequently, addressing knowledge enhancement using large language models, showed that LLMs introduce both opportunity and fragility into this landscape. While LLMs can act as compressors, translators, or generators of knowledge, and even as dynamic mediators between multimodal components, their internal knowledge remains opaque, hard to verify, and susceptible to anchoring and misgeneralization. Experiments with redefining physical constants or applying cognitive-bias-based adversarial prompts illustrated how easily such models can be steered away from factual accuracy, reinforcing the argument that LLMs alone are insufficient as authoritative knowledge repositories. Nonetheless, they offer a powerful interface layer: their capacity to contextualize, retrieve, and integrate disparate knowledge sources suggests that future vision-language systems may evolve into integrated knowledge-rich architectures in which LLMs orchestrate interactions between structured databases, multimodal encoders, and task-specific reasoning modules. The conclusion that emerges is that LLMs will not replace explicit knowledge sources but will increasingly function as adaptive controllers that manage them, enabling more coherent, grounded, and interpretable explanations.

Taken together, the thesis contributes a unified perspective on the future of knowledge in explainable and multimodal AI: knowledge must become explicit, structured, and verifiable; multimodal systems must be enriched with external sources rather than relying solely on internal representations; and LLMs should be understood not as monolithic reservoirs of truth, but as flexible, probabilistic knowledge interfaces whose reliability depends on their connection to curated and grounded knowledge bases. The trajectory points toward a new generation of knowledge-rich vision-language models—integrated systems where perceptual understanding, linguistic reasoning, and external knowledge acquisition operate in tandem. These systems promise to move beyond the brittle statistical associations of current architectures, enabling explanations that are both intelligible and trustworthy. The thesis thus closes by envisioning a paradigm in which explicit knowledge, structured representations, counterfactual reasoning, and LLM-mediated retrieval converge to redefine what it means for multimodal AI to understand and explain the world.

Discussion

The findings of this thesis expose a persistent tension at the core of contemporary multimodal AI systems: despite their remarkable generative and reasoning capabilities, they operate within an unstable epistemic space where knowledge is inconsistently represented, weakly grounded, and unevenly integrated across modalities. The examination of different types of knowledge, from structured symbolic information to commonsense and perceptual concepts, shows that no single paradigm currently offers the robustness or coherence required for reliable multimodal reasoning. Counterfactual explanations further highlight this fragility. By systematically perturbing concepts in questions, images, or textual descriptions, it becomes evident that vision–language models often rely on statistical regularities rather than genuine multimodal alignment. Their responses change in unpredictable ways depending on the nature of the perturbation, indicating that these systems do not necessarily encode relations between concepts, but rather patterns of co-occurrence that may not correspond to meaningful understanding. This limitation has direct consequences for interpretability: counterfactuals can show how the model changes its predictions, but not always why those changes occur, precisely because the model's internal representations are not structured around clearly defined semantic units.

Explicit knowledge bases were examined as a corrective to this opacity, providing a more controlled environment for interpreting and constraining model behavior. Their advantages, including transparency, relational structure, and manipulability make them well suited for explainability. Yet their integration into visual–language tasks revealed another challenge: while symbolic resources such as WordNet or ConceptNet can guide perturbations or serve as anchors for semantic reasoning, they often lack the granularity, coverage, and multimodal grounding required for complex tasks. This mismatch becomes even more pronounced in vision–language explanations, where the visual content introduces ambiguities that symbolic structures alone cannot resolve. As the thesis demonstrates, symbolic knowledge can improve both interpretability and robustness, but only when it is carefully aligned with the visual distributions on which models are trained. Without this alignment, symbolic guidance risks becoming superficial, influencing the explanation process without genuinely influencing the underlying reasoning.

The exploration of knowledge enhancement in multimodal tasks further reinforces this point. Knowledge injection, whether through pre-training, retrieval mechanisms, or structured perturbation, can improve model consistency and reduce shortcut behaviors, but its effects are uneven and highly dependent on the type of knowledge and the architecture of the system. Models benefit most when external knowledge enriches the contextual associations they already rely on, but struggle when this knowledge challenges entrenched statistical biases. Counterfactual reasoning is particularly instructive in this regard: when perturbations expose latent biases, models often fail to reinterpret the scene according to the new context, revealing both their strengths in narrow semantic domains and their limitations in broader conceptual reasoning. This suggests that multimodal knowledge enhancement cannot be treated as a simple add-on; rather, it demands architectures that actively reconcile external knowledge with learned visual representations.

Against this backdrop, the role of LLMs as knowledge sources introduces both new opportunities and new complexities. Large language models offer vast, flexible stores of implicit knowledge, and their integration into multimodal systems has led to the emergence of large vision–language models (LVLMs) that appear capable of unifying perception and knowledge within a single generative framework. Yet this thesis highlights a fundamental epistemic ambiguity: because LLM knowledge is latent, unstructured, and learned from heterogeneous textual data, it lacks explicit boundaries and may conflict with perceptual information. When LLMs are used for knowledge enhancement, they inherit these uncertainties. Their ability to generate explanations or supplement visual information does not guarantee correctness or grounding, and their persuasive fluency often obscures the absence of genuine visual reasoning. Thus, while LLM-based enhancement represents a major step toward integrated knowledge-rich multimodal systems, it amplifies rather than resolves existing concerns about bias, hallucination, and interpretability.

Overall, these findings indicate that the future of multimodal knowledge integration must move beyond simple combinations of modalities or the additive inclusion of external resources. What emerges from this thesis is the need for principled integration: systems that not only access knowledge but organize, validate, and align it across modalities. The field is shifting toward multitask, multimodal architectures capable of drawing on diverse knowledge sources, yet these systems will remain epistemically unstable unless they incorporate explicit mechanisms for reconciling structured and unstructured knowledge, for grounding textual knowledge in vision, and for maintaining consistent reasoning under counterfactual change. The critical analyses throughout the thesis point to the necessity of hybrid approaches where symbolic and neural representations are mutually constraining, where counterfactual evaluations become standard tools for stress-testing multimodal reasoning, and where LLMs function not merely as knowledge reservoirs but as components in larger, knowledge-aware computational ecosystems. Such systems represent a promising direction for future research, laying the foundation for VL models that are not only more capable, but also more trustworthy, interpretable, and aligned with human cognitive expectations.

Future research directions

The findings and critical reflections of this thesis suggest several avenues for advancing knowledge-rich multimodal systems, especially in the context of visual–language reasoning, explainability, and model robustness. A pervasive theme throughout the work is that contemporary architectures are limited not by their capacity to process modalities, but by their incomplete and unstable representations of knowledge itself. Future research must therefore address how multimodal systems can evolve from pattern-driven tools into structured, knowledge-aware agents capable of reliable reasoning.

A first direction concerns the development of hybrid knowledge architectures that go beyond the dichotomy between symbolic knowledge bases and latent neural representations. The experiments using counterfactual perturbations demonstrate that explicit knowledge bases introduce semantic control and interpretability, yet they remain too rigid and incomplete for the full scope of multimodal understanding. Conversely, implicit knowledge encoded by LLMs and LVLMs is broad but amorphous, lacking the structure needed for transparent reasoning. Future multimodal models should incorporate mechanisms for bidirectional knowledge alignment, where symbolic structures constrain generative reasoning while neural representations dynamically extend and refine symbolic graphs. Such hybridization would require new methods for synchronizing visual grounding with symbolic knowledge hierarchies, allowing models to reason over concepts at varying levels of abstraction while maintaining interpretability.

Another promising direction lies in expanding counterfactual reasoning frameworks beyond their current role as diagnostic tools. The thesis shows that counterfactual interventions reveal latent biases, structural weaknesses, and inconsistencies in multimodal models. However, counterfactuals can also serve as training signals. A future research path involves using knowledge-driven counterfactuals not only to evaluate models, but also to shape their learning objectives. By incorporating structured perturbations into pre-training or fine-tuning loops, models could be forced to develop robustness to concept manipulation, better distinguish causal from correlational cues, and align their representations with human-understandable semantic shifts.

This would effectively embed explainability into the learning dynamics rather than treating it as a post hoc property.

A further direction concerns the construction of next-generation multimodal benchmarks that explicitly integrate knowledge and counterfactual structure. Current datasets, including VQA-v2 or Visual Genome, encode rich visual relationships but lack systematic knowledge variation or controlled conceptual contrasts. Future benchmarks could incorporate curated symbolic annotations, hierarchical concept structures, or controlled color and attribute transformations that mirror the perturbation strategies used in this thesis. Such datasets would allow researchers to quantitatively evaluate how well models handle conceptual changes, generalize to unseen semantic configurations, or reconcile visual and textual inconsistencies. They would also reduce shortcut exploitation by forcing models to operate within semantically heterogeneous environments.

The rapid emergence of LVLMs opens additional research challenges. While these systems offer unprecedented multimodal fluency, they also exacerbate issues of hallucination, ungrounded reasoning, and epistemic uncertainty. Future work must explore explicit knowledge regulation mechanisms within LVLMs, whether through retrieval-augmented generation, adaptive symbolic constraints, or modular reasoning components embedded in LLM backbones. Important open questions include how to preserve the generative capabilities of LVLMs while enforcing semantic fidelity, how to reconcile their internal knowledge with external sources, and how to design architectures that maintain consistent reasoning under counterfactual perturbation. Research into knowledge audits for LVLMs, such as systematic evaluations of what they know, how they know it, and how they apply it may become essential for ensuring their reliability in high-stakes multimodal tasks.

Moreover, the integration of explicit knowledge bases with LLM-driven reasoning suggests an opportunity to build unified knowledge-centric multimodal ecosystems. In such systems, symbolic resources would provide structure, constraints, and verifiable semantics; LLMs would supply broad contextual knowledge and linguistic reasoning; and visual encoders would ground both in perceptual evidence. Achieving this integration will require rethinking the architecture of multimodal AI altogether, moving away from loosely-coupled pipelines and toward models that treat knowledge as a first-class, dynamically evolving component. This also invites interdisciplinary collaborations: cognitive science can inform concept hierarchies, linguistics can guide semantic representation, and computer vision can contribute grounding strategies. The ultimate goal is to create multimodal models that can reason, explain, and generate with the coherence of symbolic systems and the flexibility of neural ones.

A final and increasingly influential future direction concerns the transition from static multimodal models to agentic AI systems that integrate LLMs and LVLMs as complementary knowledge-bearing entities. The thesis has shown that LLMs possess broad, flexible semantic knowledge that can be mobilized for explanation, inference, and counterfactual reasoning, while LVLMs provide perceptual grounding and multimodal synthesis capabilities. Importantly, LVLMs are themselves LLM-driven architectures; they carry substantial internal world knowledge and can perform many tasks without external assistance. Nevertheless, the experiments and conceptual analysis presented here reinforce that LLMs remain superior as explicit knowledge providers, especially when precision, structured retrieval, or reasoning constraints are required. This suggests that multimodal intelligence will increasingly emerge not from a single monolithic model, but from cooperative agentic systems in which an LLM and an LVLM operate as distinct but interacting components.

Future agentic architectures will likely adopt a two-agent paradigm: a knowledge agent (LLM) and a perception–action agent (LVLM). The LLM agent would curate, refine, and reason over symbolic or factual knowledge, acting as a live knowledge base that can be queried, audited, and dynamically updated. The LVLM agent would ground this knowledge in visual input, perform multimodal reasoning, and generate explanations or counterfactuals linked directly to perceptual evidence. Such a division of labour avoids the brittleness associated with relying solely on latent representations within a single LVLM, while enabling a more explicit, queryable understanding of when and why a model knows something. It also mirrors cognitive architectures where declarative and perceptual knowledge coexist yet remain functionally separable.

This agentic perspective synthesizes the earlier future directions on hybrid knowledge architectures, counterfactual training, and knowledge-regulated LVLMs. For instance, hybrid knowledge representations, such as symbolic graphs aligned with neural embeddings could be maintained by the LLM agent, while the LVLM would ensure that such structures remain visually grounded. Counterfactual interventions could be executed cooperatively: the LLM proposes semantically meaningful perturbations, and the LVLM enacts them in the visual space. Knowledge-regulated LVLMs could incorporate real-time LLM audits, enabling the system to detect hallucinations, request clarification, or retrieve missing knowledge before generating multimodal outputs. Benchmarks for such agentic systems would need to test not only multimodal reasoning, but also coordination, delegation, and the consistency of shared knowledge across agents.

In this sense, the future of multimodal AI lies not merely in building larger, more entangled LVLMs, but in designing distributed, agentic ecosystems in which multimodal capabilities arise from structured interactions between specialised components. This paradigm aligns naturally with the broader trajectory of AI research, which increasingly views intelligence as an emergent property of systems capable of planning, communicating, and updating internal states in response to goals. Here, LLMs and LVLMs serve as the foundational cognitive units—one tasked with semantic reasoning, the other with perception and multimodal generation—each enhancing the other's limitations through structured cooperation.

Ultimately, these agentic multimodal systems have the potential to unify all strands of this thesis: types of knowledge become functional resources; counterfactual reasoning becomes an operational tool for decision-making; explicit knowledge bases evolve into dynamic, LLM-accessible memory structures; and vision–language reasoning becomes a collaborative process distributed across agents. As LLMs act not only as models but as knowledge regulators, and LVLMs serve as grounded perceptual actors, multimodal AI will move toward systems capable of understanding, explaining, adapting, and interacting with the world in ways that extend beyond current static architectures. The convergence of knowledge-rich reasoning and visual grounding within agentic frameworks marks a decisive step toward the next generation of multimodal intelligence.

Chapter 10

Bibliography

[1] Aditya Mogadala Xiaoyu Shen, D. K. “Integrating Rule-based Entity Masking into Image Captioning”. In: 2020. DOI: `https://doi.org/10.48550/arXiv.2007.11690`.

[2] Agrawal, A., Batra, D., Parikh, D., and Kembhavi, A. “Don’t Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018.

[3] Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Batra, D., and Parikh, D. *VQA: Visual Question Answering*. 2016. DOI: `https://doi.org/10.48550/arXiv.1505.00468`. arXiv: `1505.00468 [cs.CL]`.

[4] Agrawal, R. and Srikant, R. “Fast algorithms for mining association rules”. In: *Proc. of 20th Intl. Conf. on VLDB*. 1994, pp. 487–499.

[5] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., and Herrera, F. “Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence”. In: *Information Fusion* 99 (2023), p. 101805. ISSN: 1566-2535. DOI: `https://doi.org/10.1016/j.inffus.2023.101805`. URL:

[6] Alipour, K., Schulze, J. P., Yao, Y., Ziskind, A., and Burachas, G. “A Study on Multimodal and Interactive Explanations for Visual Question Answering”. In: (2020). DOI: `10.48550/ARXIV.2003.00431`. URL:

[7] AlKhamissi, B., Li, M., Celikyilmaz, A., Diab, M., and Ghazvininejad, M. *A Review on Language Models as Knowledge Bases*. 2022. arXiv: `2204.06031 [cs.CL]`.

[8] Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, É., Hesslow, D., Launay, J., Malartic, Q., Mazzotta, D., Noune, B., Pannier, B., and Penedo, G. *The Falcon Series of Open Language Models*. 2023. DOI: `https://doi.org/10.48550/arXiv.2311.16867`. arXiv: `2311.16867 [cs.CL]`.

[9] Anderson, P., Fernando, B., Johnson, M., and Gould, S. *SPICE: Semantic Propositional Image Caption Evaluation*. 2016. DOI: `10.48550/ARXIV.1607.08822`. URL:

[10] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. “Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2017), pp. 6077–6086. URL:

[11] Anthropic. “The Claude 3 Model Family: Opus, Sonnet, Haiku”. In: URL:

[12] Argyrou, G., Dimitriou, A., Lymperaiou, M., Filandrianos, G., and Stamou, G. *Automatic Generation of Fashion Images using Prompting in Generative Machine Learning Models*. 2024. arXiv: `2407.14944 [cs.CV]`. URL:

[13] Argyrou, G., Dimitriou, A., Lymperaiou, M., Filandrianos, G., and Stamou, G. “Prompt2Fashion: An automatically generated fashion dataset”. In: *Proceedings of the 13th Hellenic Conference on Artificial Intelligence*. SETN ’24. Association for Computing Machinery, 2024. ISBN: 9798400709821. DOI: `10.1145/3688671.3690604`. URL:

[14] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. G. “DBpedia: A Nucleus for a Web of Open Data”. In: *ISWC/ASWC*. 2007. DOI: `https://doi.org/10.1007/978-3-540-76298-0_52`.

[15] Banerjee, S. and Lavie, A. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, June 2005, pp. 65–72. URL:

[16] Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., and Fung, P. *A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity*. 2023. arXiv: `2302.04023 [cs.CL]`.

[17] Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., and Fung, P. "A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity". In: *ArXiv* abs/2302.04023 (2023). URL:

[18] Bao, Q., Gendron, G., Peng, A. Y., Zhong, W., Tan, N. Ö., Chen, Y., Witbrock, M., and Liu, J. "A Systematic Evaluation of Large Language Models on Out-of-Distribution Logical Reasoning Tasks". In: *ArXiv* abs/2310.09430 (2023). URL:

[19] Bau, D., Zhu, J.-Y., Wulff, J., Peebles, W., Strobelt, H., Zhou, B., and Torralba, A. *Seeing What a GAN Cannot Generate*. 2019. DOI: `10.48550/ARXIV.1910.11626`. URL:

[20] Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 2021, pp. 610–623. ISBN: 9781450383097. DOI: `10.1145/3442188.3445922`. URL:

[21] Benny, Y., Galanti, T., Benaim, S., and Wolf, L. "Evaluation Metrics for Conditional Image Generation". In: *International Journal of Computer Vision* 129.5 (Mar. 2021), pp. 1712–1731. DOI: `10.1007/s11263-020-01424-w`. URL:

[22] Bertsekas, D. P. and Castanon, D. A. "The auction algorithm for the transportation problem". In: *Ann. Oper. Res.* 20.1–4 (Aug. 1989), pp. 67–96. ISSN: 0254-5330. DOI: `10.1007/BF02216923`. URL:

[23] Besta, M., Blach, N., Kubicek, A., Gerstenberger, R., Gianinazzi, L., Gajda, J., Lehmann, T., Podstawski, M., Niewiadomski, H., Nyczyk, P., and Hoefler, T. *Graph of Thoughts: Solving Elaborate Problems with Large Language Models*. 2023. arXiv: `2308.09687 [cs.CL]`.

[24] Besta, M., Memedi, F., Zhang, Z., Gerstenberger, R., Piao, G., Blach, N., Nyczyk, P., Copik, M., Kwaśniewski, G., Müller, J., Gianinazzi, L., Kubicek, A., Niewiadomski, H., O'Mahony, A., Mutlu, O., and Hoefler, T. *Demystifying Chains, Trees, and Graphs of Thoughts*. 2024. arXiv: `2401.14295 [cs.CL]`.

[25] Bhakthavatsalam, S., Richardson, K., Tandon, N., and Clark, P. *Do Dogs have Whiskers? A New Knowledge Base of hasPart Relations*. 2020. DOI: `https://doi.org/10.48550/arXiv.2006.07510`. arXiv: `2006.07510 [cs.CL]`.

[26] Bitton, Y., Guetta, N. B., Yosef, R., Elovici, Y., Bansal, M., Stanovsky, G., and Schwartz, R. *WinoGAViL: Gamified Association Benchmark to Challenge Vision-and-Language Models*. 2022. DOI: `https://doi.org/10.48550/arXiv.2207.12576`. arXiv: `2207.12576 [cs.CL]`.

[27] Bollacker, K. D., Evans, C., Paritosh, P. K., Sturge, T., and Taylor, J. "Freebase: a collaboratively created graph database for structuring human knowledge". In: *SIGMOD Conference*. 2008. DOI: `https://doi.org/10.1145/1376616.1376746`.

[28] Bordes, A., Usunier, N., Garcia-Durán, A., Weston, J., and Yakhnenko, O. "Translating embeddings for modeling multi-relational data". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'13. Lake Tahoe, Nevada: Curran Associates Inc., 2013, pp. 2787–2795.

[29] Borji, A. "Pros and cons of GAN evaluation measures: New developments". In: *Computer Vision and Image Understanding* 215 (2022), p. 103329. ISSN: 1077-3142. DOI: `https://doi.org/10.1016/j.cviu.2021.103329`. URL:

[30] Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., and Choi, Y. *COMET: Commonsense Transformers for Automatic Knowledge Graph Construction*. 2019. DOI: `10.48550/ARXIV.1906.05317`. URL:

[31] Brown, T. et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL:

[32] Browne, K. and Swift, B. *Semantics and explanation: why counterfactual explanations produce adversarial examples in deep neural networks*. 2020. arXiv: `2012.10076 [cs.AI]`.

[33] Burkard, R., Dell'Amico, M., and Martello, S. *Assignment Problems*. USA: Society for Industrial and Applied Mathematics, 2009. ISBN: 0898716632.

[34] Byrne, R. "The Rational Imagination: How People Create Alternatives to Reality". In: *The Behavioral and brain sciences* 30 (Dec. 2008), 439–53, discussion 453. DOI: 10.1017/S0140525X07002579.

[35] Cadene, R., Dancette, C., Cord, M., Parikh, D., et al. "RUBi: Reducing Unimodal Biases for Visual Question Answering". In: *Advances in Neural Information Processing Systems* 32 (2019), pp. 841–852.

[36] Cao, Q., Li, B., Liang, X., and Lin, L. *Explainable High-order Visual Question Reasoning: A New Benchmark and Knowledge-routed Network*. 2019. DOI: 10.48550/ARXIV.1909.10128. URL:

[37] Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. *Universal Sentence Encoder*. 2018. DOI: https://doi.org/10.48550/arXiv.1803.11175. arXiv: 1803.11175 [cs.CL].

[38] Chai, L., Wulff, J., and Isola, P. "Using latent space regression to analyze and leverage compositionality in GANs". In: *ArXiv* abs/2103.10426 (2021).

[39] Chaidos, N., Dimitriou, A., Lymperaiou, M., and Stamou, G. *SCENIR: Visual Semantic Clarity through Unsupervised Scene Graph Retrieval*. 2025. arXiv: 2505.15867 [cs.CV]. URL:

[40] Chang, C.-H., Yu, C.-H., Chen, S.-Y., and Chang, E. Y. *KG-GAN: Knowledge-Guided Generative Adversarial Networks*. 2019. DOI: 10.48550/ARXIV.1905.12261. URL:

[41] Chefer, H., Gur, S., and Wolf, L. "Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 387–396. DOI: 10.1109/ICCV48922.2021.00045.

[42] Chen, H., Huang, Y., Takamura, H., and Nakayama, H. "Commonsense Knowledge Aware Concept Selection For Diverse and Informative Visual Storytelling". In: *AAAI*. 2021. DOI: https://doi.org/10.48550/arXiv.2102.02963.

[43] Chen, J., Wang, X., Xu, R., Yuan, S., Zhang, Y., Shi, W., Xie, J., Li, S., Yang, R., Zhu, T., Chen, A., Li, N., Chen, L., Hu, C., Wu, S., Ren, S., Fu, Z., and Xiao, Y. *From Persona to Personalization: A Survey on Role-Playing Language Agents*. 2024. arXiv: 2404.18231 [cs.CL]. URL:

[44] Chen, K., Huang, Q., Bisk, Y., McDuff, D., and Gao, J. "KB-VLP: Knowledge Based Vision and Language Pretraining". In: *Proceedings of the 38 th International Conference on Machine Learning, PMLR 139, 2021. ICML, workshop, 2021.* July 2021. URL:

[45] Chen, N., Liu, J., Dong, X., Liu, Q., Sakai, T., and Wu, X.-M. *AI Can Be Cognitively Biased: An Exploratory Study on Threshold Priming in LLM-Based Batch Relevance Assessment*. 2024. arXiv: 2409.16022 [cs.CL]. URL:

[46] Chen, Z., Zhou, Q., Shen, Y., Hong, Y., Zhang, H., and Gan, C. "See, Think, Confirm: Interactive Prompting Between Vision and Language Models for Knowledge-based Visual Reasoning". In: *ArXiv* abs/2301.05226 (2023).

[47] Chen, Z., Chen, J., Geng, Y., Pan, J. Z., Yuan, Z., and Chen, H. "Zero-Shot Visual Question Answering Using Knowledge Graph". In: *The Semantic Web – ISWC 2021*. Ed. by A. Hotho, E. Blomqvist, S. Dietze, A. Fokoue, Y. Ding, P. Barnaghi, A. Haller, M. Dragoni, and H. Alani. Cham: Springer International Publishing, 2021, pp. 146–162. ISBN: 978-3-030-88361-4.

[48] Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality*. Mar. 2023. URL:

[49] Cho, J., Lu, J., Schwenk, D., Hajishirzi, H., and Kembhavi, A. "X-LXMERT: Paint, Caption and Answer Questions with Multi-Modal Transformers". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by B. Webber, T. Cohn, Y. He, and Y. Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 8785–8805. DOI: 10.18653/v1/2020.emnlp-main.707. URL:

[50] Chung, H. W. et al. *Scaling Instruction-Finetuned Language Models*. 2022. arXiv: 2210.11416 [cs.LG]. URL:

[51] Conte, D., Foggia, P., Sansone, C., and Vento, M. "THIRTY YEARS OF GRAPH MATCHING IN PATTERN RECOGNITION". In: *International Journal of Pattern Recognition and Artificial Intelligence* 18 (May 2004), pp. 265–298. DOI: 10.1142/S0218001404003228.

[52] Cordella, L., Foggia, P., Sansone, C., and Vento, M. "A (Sub)Graph Isomorphism Algorithm for Matching Large Graphs". In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 26 (Nov. 2004), pp. 1367–1372. DOI: 10.1109/TPAMI.2004.75.

[53] Creswell, A., Shanahan, M., and Higgins, I. "Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning". In: *ArXiv* abs/2205.09712 (2022). URL:

[54] Cui, Y., Yu, Z., Wang, C., Zhao, Z., Zhang, J., Wang, M., and Yu, J. "ROSITA: Enhancing Vision-and-Language Semantic Alignments via Cross- and Intra-modal Knowledge Integration". In: MM '21 (2021), pp. 797–806. DOI: 10.1145/3474085.3475251. URL:

[55] Dadas, S. "OPI at SemEval-2023 Task 1: Image-Text Embeddings and Multimodal Information Retrieval for Visual Word Sense Disambiguation". In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Ed. by A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 155–162. DOI: 10.18653/v1/2023.semeval-1.22. URL:

[56] Dai, B., Fidler, S., Urtasun, R., and Lin, D. "Towards Diverse and Natural Image Descriptions via a Conditional GAN". In: *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 2989–2998. URL:

[57] Dai, W., Liu, Z., Ji, Z., Su, D., and Fung, P. "Plausible May Not Be Faithful: Probing Object Hallucination in Vision-Language Pre-training". In: *ArXiv* abs/2210.07688 (2022). URL:

[58] Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M. F., Parikh, D., and Batra, D. *Visual Dialog*. 2016. DOI: 10.48550/ARXIV.1611.08669. URL:

[59] Datta, S. and Sundararaman, D. *Evaluating Hallucination in Large Vision-Language Models based on Context-Aware Object Similarities*. 2025. arXiv: 2501.15046 [cs.CV]. URL:

[60] Daull, X., Bellot, P., Bruno, E., Martin, V., and Murisasco, E. "Complex QA and language models hybrid architectures, Survey". In: 2023.

[61] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[62] Dervakos, E., Thomas, K., Filandrianos, G., and Stamou, G. "Choose your Data Wisely: A Framework for Semantic Counterfactuals". In: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*. Ed. by E. Elkind. Main Track. International Joint Conferences on Artificial Intelligence Organization, Aug. 2023, pp. 382–390. DOI: 10.24963/ijcai.2023/43. URL:

[63] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *ArXiv* abs/1810.04805 (2019). DOI: https://doi.org/10.48550/arXiv.1810.04805.

[64] Dey, A. U., Valveny, E., and Harit, G. "External Knowledge enabled Text Visual Question Answering". In: 2021. URL:

[65] Dhariwal, P. and Nichol, A. "Diffusion models beat GANs on image synthesis". In: *Proceedings of the 35th International Conference on Neural Information Processing Systems*. NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2021. ISBN: 9781713845393.

[66] Dijkstra, E. W. "A note on two problems in connexion with graphs". In: *Numerische mathematik* 1.1 (1959), pp. 269–271.

[67] Dimitriou, A., Chaidos, N., Lymperaiou, M., and Stamou, G. *Graph Edits for Counterfactual Explanations: A comparative study*. 2024. arXiv: 2401.11609 [cs.LG].

[68] Dimitriou, A., Chaidos, N., Lymperaiou, M., and Stamou, G. *Graph Edits for Counterfactual Explanations: A comparative study*. 2024. arXiv: 2401.11609 [cs.LG]. URL:

[69] Dimitriou, A., Lymperaiou, M., Filandrianos, G., Thomas, K., and Stamou, G. *Structure Your Data: Towards Semantic Graph Counterfactuals*. 2024. arXiv: 2403.06514 [cs.CV].

[70] Ding, R., Zhang, C., Wang, L., Xu, Y., Ma, M.-J., Zhang, W., Qin, S., Rajmohan, S., Lin, Q., and Zhang, D. "Everything of Thoughts: Defying the Law of Penrose Triangle for Thought Generation". In: *ArXiv* abs/2311.04254 (2023). URL:

[71] Do, V., Camburu, O.-M., Akata, Z., and Lukasiewicz, T. "e-SNLI-VE: Corrected Visual-Textual Entailment with Natural Language Explanations". In: (2020). DOI: 10.48550/ARXIV.2004.03744. URL:

[72] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2021. DOI: https://doi.org/10.48550/arXiv.2010.11929. URL:

[73] Echterhoff, J. M., Liu, Y., Alessa, A., McAuley, J., and He, Z. "Cognitive Bias in Decision-Making with LLMs". In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Ed. by

Y. Al-Onaizan, M. Bansal, and Y.-N. Chen. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 12640–12653. DOI: 10.18653/v1/2024.findings-emnlp.739. URL:

[74] Evangelatos, A., Filandrianos, G., Lymperaiou, M., Voulodimos, A., and Stamou, G. "AILS-NTUA at SemEval-2025 Task 8: Language-to-Code prompting and Error Fixing for Tabular Question Answering". In: *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Ed. by S. Rosenthal, A. Rosá, D. Ghosh, and M. Zampieri. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 1423–1435. ISBN: 979-8-89176-273-2. URL:

[75] Fang, Y., Liao, B., Wang, X., Fang, J., Qi, J., Wu, R., Niu, J., and Liu, W. "You Only Look at One Sequence: Rethinking Transformer in Vision through Object Detection". In: *CoRR* abs/2106.00666 (2021). arXiv: 2106.00666. URL:

[76] Fey, M. and Lenssen, J. E. "Fast Graph Representation Learning with PyTorch Geometric". In: *ArXiv* abs/1903.02428 (2019). URL:

[77] Filandrianos, G., Dimitriou, A., Lymperaiou, M., Thomas, K., and Stamou, G. *Bias Beware: The Impact of Cognitive Biases on LLM-Driven Product Recommendations*. 2025. arXiv: 2502.01349 [cs.CL]. URL:

[78] Filandrianos, G., Thomas, K., Dervakos, E., and Stamou, G. "Conceptual Edits as Counterfactual Explanations". In: *Proceedings of the AAAI 2022 Spring Symposium on Machine Learning and Knowledge Engineering for Hybrid Intelligence (AAAI-MAKE 2022), Stanford University, Palo Alto, California, USA*. 21–23 March 2022.

[79] Fischer, T., Remus, S., and Biemann, C. "Measuring Faithfulness of Abstractive Summaries". In: *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*. Ed. by R. Schaefer, X. Bai, M. Stede, and T. Zesch. Potsdam, Germany: KONVENS 2022 Organizers, Dec. 2022, pp. 63–73. URL:

[80] Floridi, L. and Chiriatti, M. "GPT-3: Its Nature, Scope, Limits, and Consequences". In: *Minds and Machines* 30 (2020), pp. 681–694.

[81] Fu, Y., Peng, H.-C., Sabharwal, A., Clark, P., and Khot, T. "Complexity-Based Prompting for Multi-Step Reasoning". In: *ArXiv* abs/2210.00720 (2022). URL:

[82] Gao, X., Xiao, B., Tao, D., and Li, X. "A survey of graph edit distance". In: *Pattern Anal. Appl.* 13 (Feb. 2010), pp. 113–129. DOI: 10.1007/s10044-008-0141-y.

[83] Garcia, N. and Vogiatzis, G. *How to Read Paintings: Semantic Art Understanding with Multi-Modal Retrieval*. 2018. DOI: 10.48550/ARXIV.1810.09617. URL:

[84] Garcia, N., Ye, C., Liu, Z., Hu, Q., Otani, M., Chu, C., Nakashima, Y., and Mitamura, T. *A Dataset and Baselines for Visual Question Answering on Art*. 2020. DOI: 10.48550/ARXIV.2008.12520. URL:

[85] Gardères, F., Ziaeefard, M., Abeloos, B., and Lecue, F. "ConceptBert: Concept-Aware Representation for Visual Question Answering". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Ed. by T. Cohn, Y. He, and Y. Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 489–498. DOI: 10.18653/v1/2020.findings-emnlp.44. URL:

[86] Gendron, G., Witbrock, M., and Dobbie, G. *A Survey of Methods, Challenges and Perspectives in Causality*. 2024. arXiv: 2302.00293 [cs.LG].

[87] Genovese, A., Piuri, V., and Scotti, F. "Towards Explainable Face Aging with Generative Adversarial Networks". In: *2019 IEEE International Conference on Image Processing (ICIP)*. 2019, pp. 3806–3810. DOI: 10.1109/ICIP.2019.8803616.

[88] Giadikiaroglou, P., Lymperaiou, M., Filandrianos, G., and Stamou, G. *Puzzle Solving using Reasoning of Large Language Models: A Survey*. 2024. DOI: https://doi.org/10.48550/arXiv.2402.11291. arXiv: 2402.11291 [cs.CL].

[89] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. *Neural Message Passing for Quantum Chemistry*. 2017. arXiv: 1704.01212 [cs.LG]. URL:

[90] Girshick, R. *Fast R-CNN*. 2015. DOI: https://doi.org/10.48550/arXiv.1504.08083. arXiv: 1504.08083 [cs.CV].

[91] Goel, A., Fernando, B., Nguyen, T.-S., and Bilen, H. "Injecting Prior Knowledge into Image Caption Generation". In: *ECCV Workshops*. 2020. DOI: https://doi.org/10.48550/arXiv.1911.10082.

[92] Gold, S. and Rangarajan, A. "Graph matching by graduated assignment". In: *Proceedings of the 1996 Conference on Computer Vision and Pattern Recognition (CVPR '96)*. CVPR '96. USA: IEEE Computer Society, 1996, p. 239. ISBN: 0818672587.

[93] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. Vol. 27. 2014. URL:

[94] Goyal, Y., Khot, T., Agrawal, A., Summers-Stay, D., Batra, D., and Parikh, D. "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering". In: *Int. J. Comput. Vision* 127.4 (Apr. 2019), pp. 398–414. ISSN: 0920-5691. DOI: `10.1007/s11263-018-1116-0`. URL:

[95] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. *Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering*. 2016. DOI: `10.48550/ARXIV.1612.00837`. URL:

[96] Grattafiori, A. et al. *The Llama 3 Herd of Models*. 2024. arXiv: `2407.21783 [cs.AI]`. URL:

[97] Grigoriadou, N., Lymperaiou, M., Filandrianos, G., and Stamou, G. *AILS-NTUA at SemEval-2024 Task 6: Efficient model tuning for hallucination detection and analysis*. 2024. arXiv: `2404.01210 [cs.CL]`.

[98] Grishman, R. and Sundheim, B. "Design of the MUC-6 Evaluation". In: *Proceedings of a Workshop on Held at Vienna, Virginia: May 6-8, 1996*. TIPSTER '96. Vienna, Virginia: Association for Computational Linguistics, 1996, pp. 413–422. DOI: `10.3115/1119018.1119072`. URL:

[99] Gui, L., Wang, B., Huang, Q., Hauptmann, A., Bisk, Y., and Gao, J. "KAT: A Knowledge Augmented Transformer for Vision-and-Language". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 956–968. DOI: `10.18653/v1/2022.naacl-main.70`. URL:

[100] Guidotti, R. "Counterfactual explanations and how to find them: literature review and benchmarking". In: *Data Mining and Knowledge Discovery* (2022). DOI: `10.1007/s10618-022-00831-6`.

[101] Guo, J., Li, J., Li, D., Tiong, A. M. H., Li, B., Tao, D., and Hoi, S. "From Images to Textual Prompts: Zero-shot VQA with Frozen Large Language Models". In: *ArXiv* abs/2212.10846 (2022).

[102] Hamilton, W. L., Ying, R., and Leskovec, J. *Inductive Representation Learning on Large Graphs*. 2018. arXiv: `1706.02216 [cs.SI]`. URL:

[103] Hasibi, R. and Michoel, T. *A Graph Feature Auto-Encoder for the Prediction of Unobserved Node Features on Biological Networks*. 2020. arXiv: `2005.03961 [q-bio.QM]`.

[104] He, K., Zhang, X., Ren, S., and Sun, J. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: `10.1109/CVPR.2016.90`.

[105] He, Q., Wang, Y., and Wang, W. *Can Language Models Act as Knowledge Bases at Scale?* 2024. arXiv: `2402.14273 [cs.CL]`.

[106] He, X. and Peng, Y. "Fine-Grained Visual-Textual Representation Learning". In: *IEEE Transactions on Circuits and Systems for Video Technology* 30.2 (Feb. 2020), pp. 520–531. DOI: `10.1109/tcsvt.2019.2892802`. URL:

[107] Hendricks, L. A., Wang, O., Shechtman, E., Sivic, J., Darrell, T., and Russell, B. *Localizing Moments in Video with Natural Language*. 2017. DOI: `10.48550/ARXIV.1708.01641`. URL:

[108] Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y. *CLIPScore: A Reference-free Evaluation Metric for Image Captioning*. 2022. arXiv: `2104.08718 [cs.CV]`. URL:

[109] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium". In: (2017). DOI: `10.48550/ARXIV.1706.08500`. URL:

[110] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. *GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium*. 2018. arXiv: `1706.08500 [cs.LG]`.

[111] Ho, J., Jain, A., and Abbeel, P. *Denoising Diffusion Probabilistic Models*. 2020. DOI: `10.48550/ARXIV.2006.11239`. URL:

[112] Holzinger, A., Saranti, A., Angerschmid, A., Finzel, B., Schmid, U., and Mueller, H. "Toward human-level concept learning: Pattern benchmarking for AI algorithms". In: *Patterns* 4.8 (2023), p. 100788. ISSN: 2666-3899. DOI: `https://doi.org/10.1016/j.patter.2023.100788`. URL:

[113] Hou, J., Wu, X., Qi, Y., Zhao, W., Luo, J., and Jia, Y. "Relational Reasoning using Prior Knowledge for Visual Captioning". In: *ArXiv* abs/1906.01290 (2019). DOI: `https://doi.org/10.48550/arXiv.1906.01290`.

[114] Hou, J., Wu, X., Zhang, X., Qi, Y., Jia, Y., and Luo, J. "Joint Commonsense and Relation Reasoning for Image and Video Captioning". In: vol. 34. 07. Apr. 2020, pp. 10973–10980. DOI: 10.1609/aaai.v34i07.6731. URL:

[115] Hsu, C.-C., Chen, Z.-Y., Hsu, C.-Y., Li, C.-C., Lin, T.-Y., Huang, T.-H. '., and Ku, L.-W. *Knowledge-Enriched Visual Storytelling*. 2019. DOI: https://doi.org/10.48550/arXiv.1912.01496. arXiv: 1912.01496 [cs.CL].

[116] Huang, F., Li, Z., Chen, S., Zhang, C., and Ma, H. "Image Captioning with Internal and External Knowledge". In: CIKM '20 (2020), pp. 535–544. DOI: 10.1145/3340531.3411948. URL:

[117] Huang, J.-H., Alfadly, M., Ghanem, B., and Worring, M. *Assessing the Robustness of Visual Question Answering Models*. 2019. DOI: 10.48550/ARXIV.1912.01452. URL:

[118] Huang, J. and Chang, K. C.-C. *Towards Reasoning in Large Language Models: A Survey*. 2023. DOI: https://doi.org/10.48550/arXiv.2212.10403. arXiv: 2212.10403 [cs.CL].

[119] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions*. 2023. DOI: https://doi.org/10.48550/arXiv.2311.05232. arXiv: 2311.05232 [cs.CL].

[120] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions". In: *ACM Trans. Inf. Syst.* 43.2 (Jan. 2025). ISSN: 1046-8188. DOI: 10.1145/3703155. URL:

[121] Huang, T.-H. K., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D., Zitnick, C. L., Parikh, D., Vanderwende, L., Galley, M., and Mitchell, M. "Visual Storytelling". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by K. Knight, A. Nenkova, and O. Rambow. San Diego, California: Association for Computational Linguistics, June 2016, pp. 1233–1239. DOI: 10.18653/v1/N16-1147. URL:

[122] Hudson, D. A. and Manning, C. D. *GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering*. 2019. DOI: https://doi.org/10.48550/arXiv.1902.09506. arXiv: 1902.09506 [cs.CL].

[123] Hwang, J. D., Bhagavatula, C., Le Bras, R., Da, J., Sakaguchi, K., Bosselut, A., and Choi, Y. "COMET-ATOMIC 2020: On Symbolic and Neural Commonsense Knowledge Graphs". In: *AAAI*. 2021. DOI: https://doi.org/10.48550/arXiv.2010.05953.

[124] Ilievski, F., Oltramari, A., Ma, K., Zhang, B., McGuinness, D. L., and Szekely, P. "Dimensions of Commonsense Knowledge". In: (2021). DOI: 10.48550/ARXIV.2101.04640. URL:

[125] *IMDB.*

[126] "Improving and Diagnosing Knowledge-Based Visual Question Answering via Entity Enhanced Knowledge Injection". In: *Companion Proceedings of the Web Conference 2022*. WWW '22. Virtual Event, Lyon, France: Association for Computing Machinery, 2022, pp. 705–715. ISBN: 9781450391306. DOI: 10.1145/3487553.3524648. URL:

[127] *Industrial-Strength Natural Language Processing.* spaCy. URL:

[128] Iwana, B. K., Rizvi, S. T. R., Ahmed, S., Dengel, A., and Uchida, S. *Judging a Book By its Cover*. 2017. DOI: https://doi.org/10.48550/arXiv.1610.09204. arXiv: 1610.09204 [cs.CV].

[129] Jain, A., Kothyari, M., Kumar, V., Jyothi, P., Ramakrishnan, G., and Chakrabarti, S. "Select, Substitute, Search: A New Benchmark for Knowledge-Augmented Visual Question Answering". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2021). DOI: https://doi.org/10.1145/3404835.3463259.

[130] Jelinek, F., Mercer, R. L., Bahl, L. R., and Baker, J. K. "Perplexity—a measure of the difficulty of speech recognition tasks". In: *The Journal of the Acoustical Society of America* 62.S1 (Aug. 2005), S63–S63. ISSN: 0001-4966. DOI: 10.1121/1.2016299. eprint: URL:

[131] Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T. "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision". In: (2021). DOI: 10.48550/ARXIV.2102.05918. URL:

[132] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril,

T., Wang, T., Lacroix, T., and Sayed, W. E. *Mistral 7B*. 2023. DOI: https://doi.org/10.48550/arXiv.2310.06825. arXiv: 2310.06825 [cs.CL].

[133] Jiang, A. Q. et al. *Mixtral of Experts*. 2024. DOI: https://doi.org/10.48550/arXiv.2401.04088. arXiv: 2401.04088 [cs.LG].

[134] Jiang, M., Chen, S., Yang, J., and Zhao, Q. "Fantastic Answers and Where to Find Them: Immersive Question-Directed Visual Attention". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 2977–2986. DOI: 10.1109/CVPR42600.2020.00305.

[135] Jiang, Y., Ilievski, F., and Ma, K. "SemEval-2024 Task 9: BRAINTEASER: A Novel Task Defying Common Sense". In: *Proceedings of the 18th International Workshop on Semantic Evaluation*. Association for Computational Linguistics, 2024.

[136] Jin, W., Cheng, Y., Shen, Y., Chen, W., and Ren, X. *A Good Prompt Is Worth Millions of Parameters: Low-resource Prompt-based Learning for Vision-Language Models*. 2021. DOI: 10.48550/ARXIV.2110.08484. URL:

[137] Jing, L., Li, R., Chen, Y., Jia, M., and Du, X. "FAITHSCORE: Evaluating Hallucinations in Large Vision-Language Models". In: *ArXiv* abs/2311.01477 (2023). URL:

[138] Jocher, G., Chaurasia, A., and Qiu, J. *YOLO by Ultralytics*. Jan. 2023. URL:

[139] Johnson, J., Hariharan, B., Maaten, L. van der, Fei-Fei, L., Zitnick, C. L., and Girshick, R. "CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning". In: *CVPR*. 2017. DOI: https://doi.org/10.48550/arXiv.1612.06890.

[140] Jonker, R. and Volgenant, A. "A shortest augmenting path algorithm for dense and sparse linear assignment problems". In: *Computing* 38.4 (1987), pp. 325–340.

[141] Karkani, D., Lymperaiou, M., Filandrianos, G., Spanos, N., Voulodimos, A., and Stamou, G. "AILS-NTUA at SemEval-2025 Task 3: Leveraging Large Language Models and Translation Strategies for Multilingual Hallucination Detection". In: *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Ed. by S. Rosenthal, A. Rosá, D. Ghosh, and M. Zampieri. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 1289–1305. ISBN: 979-8-89176-273-2. URL:

[142] Karkani, D., Lymperaiou, M., Filandrianos, G., Spanos, N., Voulodimos, A., and Stamou, G. *AILS-NTUA at SemEval-2025 Task 3: Leveraging Large Language Models and Translation Strategies for Multilingual Hallucination Detection*. 2025. arXiv: 2503.02442 [cs.CL]. URL:

[143] Karras, T., Laine, S., and Aila, T. *A Style-Based Generator Architecture for Generative Adversarial Networks*. 2018. DOI: 10.48550/ARXIV.1812.04948. URL:

[144] Kassner, N. and Schütze, H. *Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly*. 2020. DOI: https://doi.org/10.48550/arXiv.1911.03343. arXiv: 1911.03343 [cs.CL].

[145] Kauf, C., Ivanova, A., Giulia, R., Chersoni, E., She, J., Chowdhury, Z., Fedorenko, E., and Lenci, A. "Event Knowledge in Large Language Models: The Gap Between the Impossible and the Unlikely". In: *Cognitive Science* 47 (Nov. 2023). DOI: 10.1111/cogs.13386.

[146] Kim, J. and Park, H. "Limited Discriminator GAN using explainable AI model for overfitting problem". In: *ICT Express* (2022). ISSN: 2405-9595. DOI: https://doi.org/10.1016/j.icte.2021.12.014. URL:

[147] Kim, K.-M., Heo, M.-O., Choi, S.-H., and Zhang, B.-T. *DeepStory: Video Story QA by Deep Embedded Memory Networks*. 2017. DOI: 10.48550/ARXIV.1707.00836. URL:

[148] Kim, W., Son, B., and Kim, I. *ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision*. 2021. DOI: https://doi.org/10.48550/arXiv.2102.03334. arXiv: 2102.03334 [stat.ML].

[149] Kipf, T. N. and Welling, M. "Semi-Supervised Classification with Graph Convolutional Networks". In: *arXiv preprint arXiv:1609.02907* (2016). DOI: https://doi.org/10.48550/arXiv.1609.02907.

[150] Kipf, T. N. and Welling, M. *Variational Graph Auto-Encoders*. 2016. arXiv: 1611.07308 [stat.ML].

[151] Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., and Iwasawa, Y. *Large Language Models are Zero-Shot Reasoners*. 2022. arXiv: 2205.11916 [cs.CL].

[152] Koopmans, T. C. and Beckmann, M. J. "Assignment Problems and the Location of Economic Activities". In: *Econometrica* 25 (1957), p. 53. URL:

[153] Kornblith, S., Li, L., Wang, Z., and Nguyen, T. *Guiding Image Captioning Models Toward More Specific Captions*. 2023. arXiv: 2307.16686 [cs.CV].

[154] Koulakos, A., Lymperaiou, M., Filandrianos, G., and Stamou, G. "Enhancing adversarial robustness in Natural Language Inference using explanations". In: *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. Ed. by Y. Belinkov, N. Kim, J. Jumelet, H. Mohebbi, A. Mueller, and H. Chen. Miami, Florida, US: Association for Computational Linguistics, Nov. 2024, pp. 105–117. DOI: `10.18653/v1/2024.blackboxnlp-1.7`. URL:

[155] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., Bernstein, M. S., and Li, F.-F. *Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations*. 2016. DOI: `https://doi.org/10.48550/arXiv.1602.07332`. arXiv: `1602.07332 [cs.CV]`.

[156] Kritharoula, A., Lymperaiou, M., and Stamou, G. *Language Models as Knowledge Bases for Visual Word Sense Disambiguation*. 2023. arXiv: `2310.01960 [cs.CL]`. URL:

[157] Kritharoula, A., Lymperaiou, M., and Stamou, G. "Large Language Models and Multimodal Retrieval for Visual Word Sense Disambiguation". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Bouamor, J. Pino, and K. Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 13053–13077. DOI: `10.18653/v1/2023.emnlp-main.807`. URL:

[158] Krizhevsky, A. "Learning Multiple Layers of Features from Tiny Images". In: 2009, pp. 32–33.

[159] Kryscinski, W., McCann, B., Xiong, C., and Socher, R. "Evaluating the Factual Consistency of Abstractive Text Summarization". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by B. Webber, T. Cohn, Y. He, and Y. Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 9332–9346. DOI: `10.18653/v1/2020.emnlp-main.750`. URL:

[160] Kuhn, H. W. "The Hungarian method for the assignment problem". In: *Naval Research Logistics Quarterly* 2.1-2 (1955), pp. 83–97. DOI: `https://doi.org/10.1002/nav.3800020109`. eprint: URL:

[161] Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A. C., and Berg, T. L. "Baby talk: Understanding and generating simple image descriptions". In: *CVPR 2011*. 2011, pp. 1601–1608. DOI: `10.1109/CVPR.2011.5995466`.

[162] Kumar, A. "The Illustrated Image Captioning using transformers". In: *ankur3107.github.io* (2022). URL:

[163] Kumar, A. and Lakkaraju, H. *Manipulating Large Language Models to Increase Product Visibility*. 2024. arXiv: `2404.07981 [cs.IR]`. URL:

[164] Lang, K. "NewsWeeder: learning to filter netnews". In: *Proceedings of the Twelfth International Conference on International Conference on Machine Learning*. ICML'95. Tahoe City, California, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 331–339. ISBN: 1558603778.

[165] Le, Q. V. and Mikolov, T. *Distributed Representations of Sentences and Documents*. 2014. DOI: `https://doi.org/10.48550/arXiv.1405.4053`. arXiv: `1405.4053 [cs.CL]`.

[166] Lee, J. and Kim, I. "Vision–Language–Knowledge Co-Embedding for Visual Commonsense Reasoning". In: vol. 21. 9. 2021. DOI: `10.3390/s21092911`. URL:

[167] Lei, B., Lin, P.-H., Liao, C., and Ding, C. "Boosting Logical Reasoning in Large Language Models through a New Framework: The Graph of Thought". In: *ArXiv* abs/2308.08614 (2023). URL:

[168] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. 2019. DOI: `10.48550/ARXIV.1910.13461`. URL:

[169] Li, C., Kong, L., and Zhou, Z. "Improved-StoryGAN for sequential images visualization". In: *Journal of Visual Communication and Image Representation* 73 (2020), p. 102956. ISSN: 1047-3203. DOI: `https://doi.org/10.1016/j.jvcir.2020.102956`. URL:

[170] Li, G., Wang, X., and Zhu, W. "Boosting Visual Question Answering with Context-aware Knowledge Aggregation". In: *Proceedings of the 28th ACM International Conference on Multimedia* (2020). DOI: `https://doi.org/10.1145/3394171.3413943`.

[171] Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. *A Diversity-Promoting Objective Function for Neural Conversation Models*. 2015. DOI: `10.48550/ARXIV.1510.03055`. URL:

[172] Li, J., Li, D., Savarese, S., and Hoi, S. "BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models". In: *Proceedings of the 40th International Conference on Machine Learning*. ICML'23. Honolulu, Hawaii, USA: JMLR.org, 2023.

[173] Li, J., Li, D., Xiong, C., and Hoi, S. *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*. 2022. arXiv: 2201.12086 [cs.CV].

[174] Li, Q., Han, Z., and Wu, X.-M. "Deeper insights into graph convolutional networks for semi-supervised learning". In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI'18/IAAI'18/EAAI'18. New Orleans, Louisiana, USA: AAAI Press, 2018. ISBN: 978-1-57735-800-8.

[175] Li, Q., Tao, Q., Joty, S., Cai, J., and Luo, J. *VQA-E: Explaining, Elaborating, and Enhancing Your Answers for Visual Questions*. 2018. DOI: 10.48550/ARXIV.1803.07464. URL:

[176] Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., and Gao, J. *Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks*. 2020. DOI: https://doi.org/10.48550/arXiv.2004.06165. arXiv: 2004.06165 [cs.CV].

[177] Li, Y., Cheng, Y., Gan, Z., Yu, L., Wang, L., and Liu, J. *BachGAN: High-Resolution Image Synthesis from Salient Object Layout*. 2020. arXiv: 2003.11690 [cs.CV].

[178] Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-r. "Evaluating Object Hallucination in Large Vision-Language Models". In: *ArXiv* abs/2305.10355 (2023). URL:

[179] Li, Y., Gan, Z., Shen, Y., Liu, J., Cheng, Y., Wu, Y., Carin, L., Carlson, D., and Gao, J. "StoryGAN: A Sequential Conditional GAN for Story Visualization". In: June 2019, pp. 6322–6331. DOI: 10.1109/CVPR.2019.00649.

[180] Li, Y., Gu, C., Dullien, T., Vinyals, O., and Kohli, P. *Graph Matching Networks for Learning the Similarity of Graph Structured Objects*. 2019. arXiv: 1904.12787 [cs.LG].

[181] Lin, C.-Y. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81. URL:

[182] Lin, C.-Y. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81. URL:

[183] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. "Microsoft COCO: Common Objects in Context". In: *Computer Vision – ECCV 2014*. Ed. by D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars. Cham: Springer International Publishing, 2014, pp. 740–755. DOI: https://doi.org/10.1007/978-3-319-10602-1_48.

[184] Lin, Y., Xie, Y., Chen, D., Xu, Y., Zhu, C., and Yuan, L. *REVIVE: Regional Visual Representation Matters in Knowledge-Based Visual Question Answering*. 2022. DOI: 10.48550/ARXIV.2206.01201. URL:

[185] Liu, H., Ning, R., Teng, Z., Liu, J., Zhou, Q., and Zhang, Y. "Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4". In: *ArXiv* abs/2304.03439 (2023). URL:

[186] Liu, H., Teng, Z., Ning, R., Liu, J., Zhou, Q., and Zhang, Y. "GLoRE: Evaluating Logical Reasoning of Large Language Models". In: *ArXiv* abs/2310.09107 (2023). URL:

[187] Liu, H., Li, C., Li, Y., and Lee, Y. J. *Improved Baselines with Visual Instruction Tuning*. 2023.

[188] Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., and Chen, W. "What Makes Good In-Context Examples for GPT-3?" In: *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*. Dublin, Ireland and Online: Association for Computational Linguistics, May 2022, pp. 100–114. DOI: 10.18653/v1/2022.deelio-1.10. URL:

[189] Liu, W., Hu, H., Zhou, J., Ding, Y., Li, J., Zeng, J., He, M., Chen, Q., Jiang, B., Zhou, A., and He, L. *Mathematical Language Models: A Survey*. 2023. arXiv: 2312.07622 [cs.CL].

[190] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. DOI: 10.48550/ARXIV.1907.11692. URL:

[191] Long, J. "Large Language Model Guided Tree-of-Thought". In: *ArXiv* abs/2305.08291 (2023). URL:

[192] Lou, J. and Sun, Y. *Anchoring Bias in Large Language Models: An Experimental Study*. 2024. arXiv: 2412.06593 [cs.CL]. URL:

[193] Lovenia, H., Dai, W., Cahyawijaya, S., Ji, Z., and Fung, P. "Negative Object Presence Evaluation (NOPE) to Measure Object Hallucination in Vision-Language Models". In: *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*. Ed. by J. Gu, T.-J. (Fu, D. Hudson,

A. Celikyilmaz, and W. Wang. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 37–58. DOI: 10.18653/v1/2024.alvr-1.4. URL:

[194] Lu, J., Batra, D., Parikh, D., and Lee, S. *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks*. 2019. DOI: https://doi.org/10.48550/arXiv.1908.02265. arXiv: 1908.02265 [cs.CV].

[195] Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. *Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering*. 2022. DOI: 10.48550/ARXIV.2209.09513. URL:

[196] Lu, P., Qiu, L., Yu, W., Welleck, S., and Chang, K.-W. "A Survey of Deep Learning for Mathematical Reasoning". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 14605–14631. DOI: 10.18653/v1/2023.acl-long.817. URL:

[197] Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. "Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 8086–8098. DOI: 10.18653/v1/2022.acl-long.556. URL:

[198] Lundberg, S. M. and Lee, S.-I. "A unified approach to interpreting model predictions". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 4768–4777. ISBN: 9781510860964.

[199] Luo, M., Zeng, Y., Banerjee, P., and Baral, C. "Weakly-Supervised Visual-Retriever-Reader for Knowledge-based Question Answering". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6417–6431. DOI: 10.18653/v1/2021.emnlp-main.517. URL:

[200] Luo, Z., Hu, Z., Xi, Y., Zhang, R., and Ma, J. *I-Tuning: Tuning Frozen Language Models with Image for Lightweight Image Captioning*. 2022. DOI: 10.48550/ARXIV.2202.06574. URL:

[201] Luo, Z., Xi, Y., Zhang, R., and Ma, J. "A Frustratingly Simple Approach for End-to-End Image Captioning". In: 2022.

[202] Lymperaiou, M., FIlandrianos, G., Dimitriou, A., Voulodimos, A., and Stamou, G. *HalCECE: A Framework for Explainable Hallucination Detection through Conceptual Counterfactuals in Image Captioning*. 2025. arXiv: 2503.00436 [cs.CV]. URL:

[203] Lymperaiou, M., Filandrianos, G., Thomas, K., and Stamou, G. *Counterfactual Edits for Generative Evaluation*. 2023. arXiv: 2303.01555 [cs.CV].

[204] Lymperaiou, M., Manoliadis, G., Mastromichalakis, O. M., Dervakos, E. G., and Stamou, G. *Towards explainable evaluation of language models on the semantic similarity of visual concepts*. 2022. arXiv: 2209.03723 [cs.CL].

[205] Lymperaiou, M. and Stamou, G. *The Contribution of Knowledge in Visiolinguistic Learning: A Survey on Tasks and Challenges*. 2023. DOI: https://doi.org/10.48550/arXiv.2303.02411. arXiv: 2303.02411 [cs.CL].

[206] Lymperaiou, M. and Stamou, G. *A survey on knowledge-enhanced multimodal learning*. 2024. arXiv: 2211.12328 [cs.LG].

[207] Lymperaiou, M., Thomas, K., and Stamou, G. *Fine-Grained ImageNet Classification in the Wild*. 2023. arXiv: 2303.02400 [cs.CV].

[208] Lymperopoulos, D., Lymperaiou, M., Filandrianos, G., and Stamou, G. "Optimal and efficient text counterfactuals using Graph Neural Networks". In: *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. Ed. by Y. Belinkov, N. Kim, J. Jumelet, H. Mohebbi, A. Mueller, and H. Chen. Miami, Florida, US: Association for Computational Linguistics, Nov. 2024, pp. 1–14. DOI: 10.18653/v1/2024.blackboxnlp-1.1. URL:

[209] Ma, K., Zhao, B., and Sigal, L. *Attribute-guided image generation from layout*. 2020. arXiv: 2008.11932 [cs.CV].

[210] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. "Learning Word Vectors for Sentiment Analysis". In: *Proceedings of the 49th Annual Meeting of the Association for Computational*

Linguistics: Human Language Technologies. Ed. by D. Lin, Y. Matsumoto, and R. Mihalcea. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 142–150. URL:

[211] Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Welleck, S., Majumder, B. P., Gupta, S., Yazdanbakhsh, A., and Clark, P. "Self-Refine: Iterative Refinement with Self-Feedback". In: *ArXiv* abs/2303.17651 (2023). URL:

[212] Maharana, A. and Bansal, M. "Integrating Visuospatial, Linguistic, and Commonsense Structure into Story Visualization". In: *ArXiv* abs/2110.10834 (2021). DOI: `https://doi.org/10.48550/arXiv.2110.10834`.

[213] Maharana, A., Hannan, D., and Bansal, M. "Improving Generation and Evaluation of Visual Stories via Semantic Consistency". In: *ArXiv* abs/2105.10026 (2021). DOI: `https://doi.org/10.48550/arXiv.2105.10026`.

[214] Maharana, A., Hannan, D., and Bansal, M. *StoryDALL-E: Adapting Pretrained Text-to-Image Transformers for Story Continuation*. 2022. DOI: `10.48550/ARXIV.2209.06192`. URL:

[215] Malaviya, C., Bhagavatula, C., Bosselut, A., and Choi, Y. *Commonsense Knowledge Base Completion with Structural and Semantic Context*. 2019. DOI: `https://doi.org/10.48550/arXiv.1910.02915`. arXiv: `1910.02915 [cs.CL]`.

[216] Malberg, S., Poletukhin, R., Schuster, C. M., and Groh, G. *A Comprehensive Evaluation of Cognitive Biases in LLMs*. 2024. arXiv: `2410.15413 [cs.CL]`. URL:

[217] Malinowski, M. and Fritz, M. *Towards a Visual Turing Challenge*. 2014. DOI: `10.48550/ARXIV.1410.8027`. URL:

[218] Manning, C. D., Raghavan, P., and Schütze, H. *Introduction to Information Retrieval*. USA: Cambridge University Press, 2008. ISBN: 0521865719.

[219] Marasović, A., Bhagavatula, C., Park, J. s., Le Bras, R., Smith, N. A., and Choi, Y. "Natural Language Rationales with Full-Stack Visual Reasoning: From Pixels to Semantic Frames to Commonsense Graphs". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Ed. by T. Cohn, Y. He, and Y. Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 2810–2829. DOI: `10.18653/v1/2020.findings-emnlp.253`. URL:

[220] Marino, K., Chen, X., Parikh, D., Gupta, A. K., and Rohrbach, M. "KRISP: Integrating Implicit and Symbolic Knowledge for Open-Domain Knowledge-Based VQA". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 14106–14116. DOI: `10.1109/CVPR46437.2021.01389`.

[221] Marino, K., Rastegari, M., Farhadi, A., and Mottaghi, R. "OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 3190–3199. DOI: `https://doi.org/10.48550/arXiv.1906.00067`.

[222] McKenzie, I. R. et al. *Inverse Scaling: When Bigger Isn't Better*. 2024. arXiv: `2306.09479 [cs.CL]`. URL:

[223] *Meet the FLINTSTONES Dataset*.

[224] Mickus, T., Zosa, E., Vazquez, R., Vahtola, T., Tiedemann, J., Segonne, V., Raganato, A., and Apidianaki, M. "SemEval-2024 Task 6: SHROOM, a Shared-task on Hallucinations and Related Observable Overgeneration Mistakes". In: *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Ed. by A. K. Ojha, A. S. Doğruöz, H. Tayyar Madabushi, G. Da San Martino, S. Rosenthal, and A. Rosá. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 1979–1993. DOI: `10.18653/v1/2024.semeval-1.273`. URL:

[225] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. *Distributed Representations of Words and Phrases and their Compositionality*. 2013. DOI: `https://doi.org/10.48550/arXiv.1310.4546`. arXiv: `1310.4546 [cs.CL]`.

[226] Miller, G. A. "WordNet: An Electronic Lexical Database". In: (1994). URL:

[227] Miller, T. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial Intelligence* 267 (2019), pp. 1–38. ISSN: 0004-3702. DOI: `https://doi.org/10.1016/j.artint.2018.07.007`. URL:

[228] Mirza, M. and Osindero, S. *Conditional Generative Adversarial Nets*. 2014. DOI: `10.48550/ARXIV.1411.1784`. URL:

[229] Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. *Spectral Normalization for Generative Adversarial Networks*. 2018. arXiv: `1802.05957 [cs.LG]`.

[230] Miyato, T. and Koyama, M. *cGANs with Projection Discriminator*. 2018. arXiv: 1802.05637 [cs.LG].

[231] Mo, S. and Xin, M. “Tree of Uncertain Thoughts Reasoning for Large Language Models”. In: *ArXiv* abs/2309.07694 (2023). URL:

[232] Mokady, R., Hertz, A., and Bermano, A. H. “ClipCap: CLIP Prefix for Image Captioning”. In: *ArXiv* abs/2111.09734 (2021).

[233] *Movie Genre from its Poster*.

[234] Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Scao, T. L., Bari, M. S., Shen, S., Yong, Z.-X., Schoelkopf, H., Tang, X., Radev, D., Aji, A. F., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., and Raffel, C. *Crosslingual Generalization through Multitask Finetuning*. 2023. arXiv: 2211.01786 [cs.CL].

[235] Müller, H. and Holzinger, A. “Kandinsky Patterns”. In: *Artificial Intelligence* 300 (2021), p. 103546. ISSN: 0004-3702. DOI: https://doi.org/10.1016/j.artint.2021.103546. URL:

[236] Nagisetty, V., Graves, L., Scott, J., and Ganesh, V. *xAI-GAN: Enhancing Generative Adversarial Networks via Explainable AI Systems*. 2020. DOI: 10.48550/ARXIV.2002.10438. URL:

[237] Narasimhan, M., Lazebnik, S., and Schwing, A. G. *Out of the Box: Reasoning with Graph Convolution Nets for Factual Visual Question Answering*. 2018. DOI: https://doi.org/10.48550/arXiv.1811.00538. arXiv: 1811.00538 [cs.CV].

[238] Narasimhan, M. and Schwing, A. G. “Straight to the Facts: Learning Knowledge Base Retrieval for Factual Visual Question Answering”. In: *ArXiv* abs/1809.01124 (2018). DOI: https://doi.org/10.48550/arXiv.1809.01124.

[239] Nestaas, F., Debenedetti, E., and Tramèr, F. *Adversarial Search Engine Optimization for Large Language Models*. 2024. arXiv: 2406.18382 [cs.CR]. URL:

[240] Nikiforova, S., Deoskar, T., Paperno, D., and Winter, Y. “Generating image captions with external encyclopedic knowledge”. In: *ArXiv* abs/2210.04806 (2022).

[241] Nilsback, M.-E. and Zisserman, A. “Automated Flower Classification over a Large Number of Classes”. In: *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*. 2008, pp. 722–729. DOI: 10.1109/ICVGIP.2008.47.

[242] Odena, A., Olah, C., and Shlens, J. *Conditional Image Synthesis With Auxiliary Classifier GANs*. 2017. arXiv: 1610.09585 [stat.ML].

[243] Opedal, A., Stolfo, A., Shirakami, H., Jiao, Y., Cotterell, R., Schölkopf, B., Saparov, A., and Sachan, M. *Do Language Models Exhibit the Same Cognitive Biases in Problem Solving as Human Learners?* 2024. arXiv: 2401.18070 [cs.CL]. URL:

[244] OpenAI. *ChatGPT: Conversational Language Model*. 2023.

[245] OpenAI. “GPT-4 Technical Report”. In: *ArXiv* abs/2303.08774 (2023).

[246] Ordonez, V., Kulkarni, G., and Berg, T. L. “Im2Text: describing images using 1 million captioned photographs”. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems*. NIPS’11. Granada, Spain: Curran Associates Inc., 2011, pp. 1143–1151. ISBN: 9781618395993.

[247] Pan, S., Hu, R., Long, G., Jiang, J., Yao, L., and Zhang, C. *Adversarially Regularized Graph Autoencoder for Graph Embedding*. 2019. arXiv: 1802.04407 [cs.LG].

[248] Panagiotopoulos, I., Filandrianos, G., Lymperaiou, M., and Stamou, G. *AILS-NTUA at SemEval-2024 Task 9: Cracking Brain Teasers: Transformer Models for Lateral Thinking Puzzles*. 2024. arXiv: 2404.01084 [cs.CL].

[249] Panagiotopoulos, I., Filandrianos, G., Lymperaiou, M., and Stamou, G. *RISCORE: Enhancing In-Context Riddle Solving in Language Models through Context-Reconstructed Example Augmentation*. 2024. arXiv: 2409.16383 [cs.CL]. URL:

[250] Panesar, A., Doğan, F. I., and Leite, I. “Improving Visual Question Answering by Leveraging Depth and Adapting Explainability”. In: *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 2022, pp. 252–259. DOI: 10.1109/RO-MAN53752.2022.9900586.

[251] Papadakis, C., Dimitriou, A., Filandrianos, G., Lymperaiou, M., Thomas, K., and Stamou, G. *ATLAS: Adaptive Trading with LLM AgentS Through Dynamic Prompt Optimization and Multi-Agent Coordination*. 2025. arXiv: 2510.15949 [q-fin.TR]. URL:

[252] Papadakis, C., Filandrianos, G., Dimitriou, A., Lymperaiou, M., Thomas, K., and Stamou, G. *StockSim: A Dual-Mode Order-Level Simulator for Evaluating Multi-Agent LLMs in Financial Markets*. 2025. arXiv: 2507.09255 [cs.CE]. URL:

[253] Papadimitriou, C., Filandrianos, G., Lymperaiou, M., and Stamou, G. “Masked Generative Story Transformer with Character Guidance and Caption Augmentation”. In: (2024). arXiv: 2403.08502 [cs.CV].

[254] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. “Bleu: a Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. URL:

[255] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. “Bleu: a Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Ed. by P. Isabelle, E. Charniak, and D. Lin. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318. DOI: 10.3115/1073083.1073135. URL:

[256] Park, J. S., Bhagavatula, C., Mottaghi, R., Farhadi, A., and Choi, Y. “VisualCOMET: Reasoning about the Dynamic Context of a Still Image”. In: *In Proceedings of the European Conference on Computer Vision (ECCV)*. 2020. DOI: https://doi.org/10.48550/arXiv.2004.10796.

[257] Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. *Semantic Image Synthesis with Spatially-Adaptive Normalization*. 2019. arXiv: 1903.07291 [cs.CV].

[258] Parmar, G., Zhang, R., and Zhu, J.-Y. “On Aliased Resizing and Surprising Subtleties in GAN Evaluation”. In: *CVPR*. 2022.

[259] Pearl, J. “Causal inference in statistics: An overview”. In: (2009).

[260] Pennington, J., Socher, R., and Manning, C. “GloVe: Global Vectors for Word Representation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL:

[261] Petersen, M. R. and Plas, L. van der. *Can language models learn analogical reasoning? Investigating training objectives and comparisons to human performance*. 2023. arXiv: 2310.05597 [cs.CL].

[262] Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. “Language Models as Knowledge Bases?” In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 2463–2473. DOI: 10.18653/v1/D19-1250. URL:

[263] Petryk, S., Chan, D., Kachinthaya, A., Zou, H., Canny, J., Gonzalez, J., and Darrell, T. “ALOHa: A New Measure for Hallucination in Captioning Models”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*. Ed. by K. Duh, H. Gomez, and S. Bethard. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 342–357. DOI: 10.18653/v1/2024.naacl-short.30. URL:

[264] Pillutla, K., Swayamdipta, S., Zellers, R., Thickstun, J., Welleck, S., Choi, Y., and Harchaoui, Z. *MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers*. 2021. arXiv: 2102.01454 [cs.CL].

[265] Poerner, N., Waltinger, U., and Schütze, H. “E-BERT: Efficient-Yet-Effective Entity Embeddings for BERT”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, Nov. 2020, pp. 803–818. DOI: 10.18653/v1/2020.findings-emnlp.71. URL:

[266] Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., and Flach, P. “FACE: Feasible and Actionable Counterfactual Explanations”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’20. ACM, Feb. 2020. DOI: 10.1145/3375627.3375850. URL:

[267] Pratt, S., Yatskar, M., Weihs, L., Farhadi, A., and Kembhavi, A. *Grounded Situation Recognition*. 2020. DOI: 10.48550/ARXIV.2003.12058. URL:

[268] *Protogen x3.4*. Huggingface. URL:

[269] *Protogen x5.8*. Huggingface. URL:

[270] Qiu, Y., Ziser, Y., Korhonen, A., Ponti, E., and Cohen, S. “Detecting and Mitigating Hallucinations in Multilingual Summarisation”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Bouamor, J. Pino, and K. Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 8914–8932. DOI: 10.18653/v1/2023.emnlp-main.551. URL:

[271] Qu, C., Zamani, H., Yang, L., Croft, W. B., and Learned-Miller, E. G. “Passage Retrieval for Outside-Knowledge Visual Question Answering”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2021). DOI: https://doi.org/10.48550/arXiv.2105.03938.

[272] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. “Language Models are Unsupervised Multitask Learners”. In: 2019.

[273] Radford, A., Kim, J., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. *Learning Transferable Visual Models From Natural Language Supervision*. Feb. 2021. DOI: https://doi.org/10.48550/arXiv.2103.00020.

[274] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer”. In: *Journal of Machine Learning Research* 21.140 (2020), pp. 1–67. URL:

[275] Raganato, A., Calixto, I., Ushio, A., Camacho-Collados, J., and Pilehvar, M. T. “SemEval-2023 Task 1: Visual Word Sense Disambiguation”. In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Toronto, Canada: Association for Computational Linguistics, July 2023.

[276] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. *Hierarchical Text-Conditional Image Generation with CLIP Latents*. 2022. DOI: 10.48550/ARXIV.2204.06125. URL:

[277] Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. *Zero-Shot Text-to-Image Generation*. 2021. DOI: https://doi.org/10.48550/arXiv.2102.12092. arXiv: 2102.12092 [cs.CV].

[278] Ramos, R. P., Martins, B., Elliott, D., and Kementchedjhieva, Y. “SmallCap: Lightweight Image Captioning Prompted with Retrieval Augmentation”. In: *ArXiv* abs/2209.15323 (2022).

[279] Raptopoulos, P., Filandrianos, G., Lymperaiou, M., and Stamou, G. “PAKTON: A Multi-Agent Framework for Question Answering in Long Legal Agreements”. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Ed. by C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 7959–7995. ISBN: 979-8-89176-332-6. DOI: 10.18653/v1/2025.emnlp-main.403. URL:

[280] Raymond, J. W. and Willett, P. “Maximum common subgraph isomorphism algorithms for the matching of chemical structures”. In: *Journal of Computer-Aided Molecular Design* 16.7 (2002), pp. 521–533. ISSN: 1573-4951. DOI: 10.1023/A:1021271615909. URL:

[281] Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. “Generative Adversarial Text to Image Synthesis”. In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by M. F. Balcan and K. Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1060–1069. URL:

[282] Reimers, N. and Gurevych, I. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. 2019. DOI: https://doi.org/10.48550/arXiv.1908.10084. arXiv: 1908.10084 [cs.CL].

[283] Ren, M., Kiros, R., and Zemel, R. *Exploring Models and Data for Image Question Answering*. 2015. DOI: https://doi.org/10.48550/arXiv.1505.02074. arXiv: 1505.02074 [cs.LG].

[284] Ren, S., He, K., Girshick, R., and Sun, J. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. In: *IEEE Transactions on Pattern Analysis & Machine Intelligence* 39.06 (June 2017), pp. 1137–1149. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2016.2577031.

[285] Ribeiro, M., Singh, S., and Guestrin, C. ““Why Should I Trust You?”: Explaining the Predictions of Any Classifier”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. Ed. by J. DeNero, M. Finlayson, and S. Reddy. San Diego, California: Association for Computational Linguistics, June 2016, pp. 97–101. DOI: 10.18653/v1/N16-3020. URL:

[286] Richardson, C. and Heck, L. *Commonsense Reasoning for Conversational AI: A Survey of the State of the Art*. 2023. arXiv: 2302.07926 [cs.CL].

[287] Robertson, S. and Zaragoza, H. “The Probabilistic Relevance Framework: BM25 and Beyond”. In: *Foundations and Trends in Information Retrieval* 3 (Jan. 2009), pp. 333–389. DOI: 10.1561/1500000019.

[288] Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., and Saenko, K. *Object Hallucination in Image Captioning*. 2019. arXiv: 1809.02156 [cs.CL].

[289] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. *High-Resolution Image Synthesis with Latent Diffusion Models*. Los Alamitos, CA, USA, June 2022. doi: 10.1109/CVPR52688.2022.01042. url:

[290] Ross, A., Marasović, A., and Peters, M. "Explaining NLP Models via Minimal Contrastive Editing (MiCE)". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Ed. by C. Zong, F. Xia, W. Li, and R. Navigli. Online: Association for Computational Linguistics, Aug. 2021, pp. 3840–3852. doi: 10.18653/v1/2021.findings-acl.336. url:

[291] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding*. 2022. arXiv: 2205.11487 [cs.CV].

[292] Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., and Chadha, A. *A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications*. 2024. arXiv: 2402.07927 [cs.AI].

[293] Salaberria, A., Azkune, G., Lacalle, O. L. de, Etxabe, A. S., and Agirre, E. "Image Captioning for Effective Use of Language Models in Knowledge-Based Visual Question Answering". In: *ArXiv* abs/2109.08029 (2021). doi: https://doi.org/10.1016/j.eswa.2022.118669.

[294] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. *Improved Techniques for Training GANs*. 2016. doi: 10.48550/ARXIV.1606.03498. url:

[295] Sammani, F., Mukherjee, T., and Deligiannis, N. *NLX-GPT: A Model for Natural Language Explanations in Vision and Vision-Language Tasks*. 2022. doi: 10.48550/ARXIV.2203.05081. url:

[296] Saparov, A., Pang, R. Y., Padmakumar, V., Joshi, N., Kazemi, S. M., Kim, N., and He, H. "Testing the General Deductive Reasoning Capacity of Large Language Models Using OOD Examples". In: *ArXiv* abs/2305.15269 (2023). url:

[297] Schlichtkrull, M., Kipf, T. N., Bloem, P., Berg, R. van den, Titov, I., and Welling, M. *Modeling Relational Data with Graph Convolutional Networks*. 2017. arXiv: 1703.06103 [stat.ML]. url:

[298] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *International Journal of Computer Vision* 128.2 (Oct. 2019), pp. 336–359. issn: 1573-1405. doi: 10.1007/s11263-019-01228-7. url:

[299] Selvaraju, R. R., Tendulkar, P., Parikh, D., Horvitz, E., Ribeiro, M. T., Nushi, B., and Kamar, E. "SQuINTing at VQA Models: Introspecting VQA Models With Sub-Questions". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020.

[300] Sen, P., Mavadia, S., and Saffari, A. "Knowledge Graph-augmented Language Models for Complex Question Answering". In: *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE)*. Ed. by B. Dalvi Mishra, G. Durrett, P. Jansen, D. Neves Ribeiro, and J. Wei. Toronto, Canada: Association for Computational Linguistics, June 2023, pp. 1–8. doi: 10.18653/v1/2023.nlrse-1.1. url:

[301] Shah, S., Mishra, A., Yadati, N., and Talukdar, P. P. "KVQA: Knowledge-Aware Visual Question Answering". In: vol. 33. 01. July 2019, pp. 8876–8884. doi: 10.1609/aaai.v33i01.33018876. url:

[302] Shaki, J., Kraus, S., and Wooldridge, M. "Cognitive Effects in Large Language Models". In: *ECAI 2023*. IOS Press, Sept. 2023. isbn: 9781643684376. doi: 10.3233/faia230505. url:

[303] Sharir, O., Peleg, B., and Shoham, Y. *The Cost of Training NLP Models: A Concise Overview*. 2020. doi: 10.48550/ARXIV.2004.08900. url:

[304] Shen, J., Liu, T., Liu, J., Qin, Z., Pavagadhi, J., Baumgartner, S., and Bendersky, M. "Multilingual Fine-Grained News Headline Hallucination Detection". In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Ed. by Y. Al-Onaizan, M. Bansal, and Y.-N. Chen. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 7862–7875. doi: 10.18653/v1/2024.findings-emnlp.461. url:

[305] Shen, Y., Yang, C., Tang, X., and Zhou, B. "InterFaceGAN: Interpreting the Disentangled Face Representation Learned by GANs". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (2020), pp. 2004–2018.

[306] Shevchenko, V., Teney, D., Dick, A., and Hengel, A. van den. "Reasoning over Vision and Language: Exploring the Benefits of Supplemental Knowledge". In: (Apr. 2021). Ed. by M. Mosbach, M. A. Hedderich, S. Pezzelle, A. Mogadala, D. Klakow, M.-F. Moens, and Z. Akata, pp. 1–18. url:

[307] Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E. H.-h., Scharli, N., and Zhou, D. “Large Language Models Can Be Easily Distracted by Irrelevant Context”. In: *ArXiv* abs/2302.00093 (2023).

[308] Shi, J., Zhang, H., and Li, J. *Explainable and Explicit Visual Reasoning over Scene Graphs*. 2018. DOI: 10.48550/ARXIV.1812.01855. URL:

[309] Shvaiko, P. and Euzenat, J. “Ontology Matching: State of the Art and Future Challenges”. In: *IEEE Transactions on Knowledge and Data Engineering* 25.1 (2013), pp. 158–176. DOI: 10.1109/TKDE.2011.253.

[310] Simonyan, K. and Zisserman, A. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. DOI: https://doi.org/10.48550/arXiv.1409.1556. arXiv: 1409.1556 [cs.CV].

[311] Singh, A. K., Mishra, A., Shekhar, S., and Chakraborty, A. “From Strings to Things: Knowledge-Enabled VQA Model That Can Read and Reason”. In: (2019), pp. 4601–4611. DOI: 10.1109/ICCV.2019.00470.

[312] Sinkhorn, R. and Knopp, P. “Concerning nonnegative matrices and doubly stochastic matrices”. In: *Pacific Journal of Mathematics* 21 (1967), pp. 343–348. URL:

[313] Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. *SmoothGrad: removing noise by adding noise*. 2017. arXiv: 1706.03825 [cs.LG]. URL:

[314] Soloveitchik, M., Diskin, T., Morin, E., and Wiesel, A. *Conditional Frechet Inception Distance*. 2021. DOI: 10.48550/ARXIV.2103.11521. URL:

[315] Song, D., Ma, S., Sun, Z., Yang, S., and Liao, L. “KVL-BERT: Knowledge Enhanced Visual-and-Linguistic BERT for visual commonsense reasoning”. In: *Know.-Based Syst.* 230.C (Oct. 2021). ISSN: 0950-7051. DOI: 10.1016/j.knosys.2021.107408. URL:

[316] Spanos, N., Lymperaiou, M., Filandrianos, G., Thomas, K., Voulodimos, A., and Stamou, G. *V-CECE: Visual Counterfactual Explanations via Conceptual Edits*. 2025. arXiv: 2509.16567 [cs.CV]. URL:

[317] Speer, R., Chin, J., and Havasi, C. “ConceptNet 5.5: An Open Multilingual Graph of General Knowledge”. In: *AAAI*. 2017. DOI: https://doi.org/10.48550/arXiv.1612.03975.

[318] *Stable Diffusion 2 base*. Huggingface. URL:

[319] *Stable Diffusion v1.4*. Huggingface. URL:

[320] Stoikou, T., Lymperaiou, M., and Stamou, G. *Knowledge-Based Counterfactual Queries for Visual Question Answering*. 2023. arXiv: 2303.02601 [cs.CL].

[321] Stringli, E., Lymperaiou, M., Filandrianos, G., Voulodimos, A., and Stamou, G. “Pitfalls of Scale: Investigating the Inverse Task of Redefinition in Large Language Models”. In: *Findings of the Association for Computational Linguistics: ACL 2025*. Ed. by W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 9445–9469. ISBN: 979-8-89176-256-5. DOI: 10.18653/v1/2025.findings-acl.492. URL:

[322] Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., and Dai, J. *VL-BERT: Pre-training of Generic Visual-Linguistic Representations*. 2020. DOI: https://doi.org/10.48550/arXiv.1908.08530. arXiv: 1908.08530 [cs.CV].

[323] Suhr, A., Lewis, M., Yeh, J., and Artzi, Y. “A Corpus of Natural Language for Visual Reasoning”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by R. Barzilay and M.-Y. Kan. Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 217–223. DOI: 10.18653/v1/P17-2034. URL:

[324] Suhr, A., Zhou, S., Zhang, A., Zhang, I., Bai, H., and Artzi, Y. “A Corpus for Reasoning about Natural Language Grounded in Photographs”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 6418–6428. DOI: 10.18653/v1/P19-1644. URL:

[325] Sumita, Y., Takeuchi, K., and Kashima, H. *Cognitive Biases in Large Language Models: A Survey and Mitigation Experiments*. 2024. arXiv: 2412.00323 [cs.CL]. URL:

[326] Sun, W. and Wu, T. *Learning Layout and Style Reconfigurable GANs for Controllable Image Synthesis*. 2021. arXiv: 2003.11571 [cs.CV].

[327] Sun, Z., Deng, Z.-H., Nie, J.-Y., and Tang, J. *RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space*. 2019. arXiv: 1902.10197 [cs.LG]. URL:

[328] Sundararajan, M., Taly, A., and Yan, Q. “Axiomatic attribution for deep networks”. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. ICML’17. Sydney, NSW, Australia: JMLR.org, 2017, pp. 3319–3328.

[329] Tan, H. and Bansal, M. *LXMERT: Learning Cross-Modality Encoder Representations from Transformers.* 2019. DOI: `https://doi.org/10.48550/arXiv.1908.07490`. arXiv: `1908.07490 [cs.CL]`.

[330] Tandon, N., Melo, G. de, and Weikum, G. "Acquiring comparative commonsense knowledge from the Web". In: *Proceedings of the National Conference on Artificial Intelligence* 1 (Jan. 2014), pp. 166–172.

[331] Tanon, T., Weikum, G., and Suchanek, F. "YAGO 4: A Reason-able Knowledge Base". In: (May 2020), pp. 583–596. DOI: `10.1007/978-3-030-49461-2_34`.

[332] Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. "GALACTICA: A Large Language Model for Science". In: 2022.

[333] Team, G. et al. *Gemini: A Family of Highly Capable Multimodal Models.* 2023. DOI: `https://doi.org/10.48550/arXiv.2312.11805`. arXiv: `2312.11805 [cs.CL]`.

[334] Thomas, K., Filandrianos, G., Lymperaiou, M., Zerva, C., and Stamou, G. ""I Never Said That": A dataset, taxonomy and baselines on response clarity classification". In: *Findings of the Association for Computational Linguistics: EMNLP 2024.* Ed. by Y. Al-Onaizan, M. Bansal, and Y.-N. Chen. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 5204–5233. DOI: `10.18653/v1/2024.findings-emnlp.300`. URL:

[335] Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., and Ross, C. *Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality.* 2022. DOI: `https://doi.org/10.48550/arXiv.2204.03162`. arXiv: `2204.03162 [cs.CV]`.

[336] Tiong, A. M. H., Li, J., Li, B., Savarese, S., and Hoi, S. C. H. "Plug-and-Play VQA: Zero-shot VQA by Conjoining Large Pretrained Models with Zero Training". In: *Conference on Empirical Methods in Natural Language Processing.* 2022.

[337] Tong, Y., Wang, Y., Li, D., Wang, S., Lin, Z., Han, S., and Shang, J. "Eliminating Reasoning via Inferring with Planning: A New Framework to Guide LLMs' Non-linear Thinking". In: *ArXiv* abs/2310.12342 (2023). URL:

[338] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. *LLaMA: Open and Efficient Foundation Language Models.* 2023. DOI: `https://doi.org/10.48550/arXiv.2302.13971`. arXiv: `2302.13971 [cs.CL]`.

[339] Touvron, H. et al. *Llama 2: Open Foundation and Fine-Tuned Chat Models.* 2023. DOI: `https://doi.org/10.48550/arXiv.2307.09288`. arXiv: `2307.09288 [cs.CL]`.

[340] Tsakas, N., Lymperaiou, M., Filandrianos, G., and Stamou, G. "An Impartial Transformer for Story Visualization". In: (2023). DOI: `https://doi.org/10.48550/arXiv.2301.03563`. arXiv: `2301.03563 [cs.CV]`.

[341] Turpin, M., Michael, J., Perez, E., and Bowman, S. R. *Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting.* 2023. arXiv: `2305.04388 [cs.CL]`.

[342] Tversky, A. and Kahneman, D. "Judgment under Uncertainty: Heuristics and Biases". In: *Utility, Probability, and Human Decision Making: Selected Proceedings of an Interdisciplinary Research Conference, Rome, 3–6 September, 1973.* Ed. by D. Wendt and C. Vlek. Dordrecht: Springer Netherlands, 1975, pp. 141–162. ISBN: 978-94-010-1834-0. DOI: `10.1007/978-94-010-1834-0_8`. URL:

[343] Tyen, G., Mansoor, H., Cărbune, V., Chen, P., and Mak, T. *LLMs cannot find reasoning errors, but can correct them!* 2024. arXiv: `2311.08516 [cs.AI]`.

[344] Ullmann, J. R. "An Algorithm for Subgraph Isomorphism". In: *J. ACM* 23.1 (Jan. 1976), pp. 31–42. ISSN: 0004-5411. DOI: `10.1145/321921.321925`. URL:

[345] Umeyama, S. "An Eigendecomposition Approach to Weighted Graph Matching Problems". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 10.5 (Sept. 1988), pp. 695–703. ISSN: 0162-8828. DOI: `10.1109/34.6778`. URL:

[346] Vandenhende, S., Mahajan, D., Radenovic, F., and Ghadiyaram, D. "Making Heads or Tails: Towards Semantically Consistent Visual Counterfactuals". In: *arXiv preprint arXiv:2203.12892* (2022).

[347] Vashishtha, S., Poliak, A., Lal, Y. K., Van Durme, B., and White, A. S. "Temporal Reasoning in Natural Language Inference". In: *Findings of the Association for Computational Linguistics: EMNLP 2020.* Ed. by T. Cohn, Y. He, and Y. Liu. Online: Association for Computational Linguistics, Nov. 2020, pp. 4070–4078. DOI: `10.18653/v1/2020.findings-emnlp.363`. URL:

[348] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. "Attention is All you Need". In: *Advances in Neural Information Processing Systems.* Ed. by

I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017. URL:

[349] Vázquez, R., Mickus, T., Zosa, E., Vahtola, T., Tiedemann, J., Sinha, A., Segonne, V., Sánchez-Vega, F., Raganato, A., Libovický, J., Karlgren, J., Ji, S., Helcl, J., Guillou, L., Gibert, O. de, Bengoetxea, J., Attieh, J., and Apidianaki, M. *SemEval-2025 Task 3: Mu-SHROOM, the Multilingual Shared-task on Hallucinations and Related Observable Overgeneration Mistakes*. 2025. URL:

[350] Vedantam, R., Zitnick, C. L., and Parikh, D. *CIDEr: Consensus-based Image Description Evaluation*. 2014. DOI: 10.48550/ARXIV.1411.5726. URL:

[351] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y. *Graph Attention Networks*. 2018. DOI: https://doi.org/10.48550/arXiv.1710.10903. arXiv: 1710.10903 [stat.ML].

[352] Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. "Show and tell: A neural image caption generator". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2014), pp. 3156–3164. URL:

[353] Vlachos, A., Filandrianos, G., Lymperaiou, M., Spanos, N., Mitsouras, I., Karampinis, V., and Voulodimos, A. "Analyze-Prompt-Reason: A Collaborative Agent-Based Framework for Multi-Image Vision-Language Reasoning". In: *Proceedings of the 33rd ACM International Conference on Multimedia*. MM '25. Dublin, Ireland: Association for Computing Machinery, 2025, pp. 13799–13805. ISBN: 9798400720352. DOI: 10.1145/3746027.3762056. URL:

[354] Vrandecic, D. and Krötzsch, M. "Wikidata: a free collaborative knowledgebase". In: *Commun. ACM* 57 (2014), pp. 78–85. DOI: https://doi.org/10.1145/2629489.

[355] Wachter, S., Mittelstadt, B., and Russell, C. *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR*. 2018. arXiv: 1711.00399 [cs.AI]. URL:

[356] Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., and Wang, L. *GIT: A Generative Image-to-text Transformer for Vision and Language*. 2022. arXiv: 2205.14100 [cs.CV].

[357] Wang, J., Zhou, Y., Xu, G., Shi, P., Zhao, C., Xu, H., Ye, Q., Yan, M., Zhang, J., Zhu, J., Sang, J., and Tang, H. *Evaluation and Analysis of Hallucination in Large Vision-Language Models*. 2023. arXiv: 2308.15126 [cs.LG].

[358] Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R. K.-W., and Lim, E.-P. "Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models". In: *Annual Meeting of the Association for Computational Linguistics*. 2023. URL:

[359] Wang, P., Wu, Q., Shen, C., Dick, A. R., and Hengel, A. van den. "Explicit Knowledge-based Reasoning for Visual Question Answering". In: *IJCAI*. 2017. DOI: https://doi.org/10.48550/arXiv.1511.02570.

[360] Wang, P., Wu, Q., Shen, C., Dick, A. R., and Hengel, A. van den. "FVQA: Fact-Based Visual Question Answering". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40 (2018), pp. 2413–2427. DOI: https://doi.org/10.48550/arXiv.1606.05433.

[361] Wang, R., Yan, J., and Yang, X. "Learning Combinatorial Embedding Networks for Deep Graph Matching". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 3056–3065. DOI: 10.1109/ICCV.2019.00315.

[362] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E. H.-h., and Zhou, D. "Self-Consistency Improves Chain of Thought Reasoning in Language Models". In: *ArXiv* abs/2203.11171 (2022). URL:

[363] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. *Emergent Abilities of Large Language Models*. 2022. DOI: https://doi.org/10.48550/arXiv.2206.07682. arXiv: 2206.07682 [cs.CL].

[364] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. 2023. arXiv: 2201.11903 [cs.CL]. URL:

[365] Wei, J., Tan, C., Gao, Z., Sun, L., Li, S., Yu, B., Guo, R., and Li, S. Z. *Enhancing Human-like Multi-Modal Reasoning: A New Challenging Dataset and Comprehensive Framework*. 2023. arXiv: 2307.12626 [cs.AI].

[366] Wen, J., Benitez-Quiroz, F., Feng, Q., and Martinez, A. M. "Diamond in the rough: Improving image realism by traversing the GAN latent space". In: *ArXiv* abs/2104.05518 (2021).

[367] Wen, Z. and Peng, Y. "Multi-Level Knowledge Injecting for Visual Commonsense Reasoning". In: *IEEE Transactions on Circuits and Systems for Video Technology* 31.3 (2021), pp. 1042–1054. DOI: 10.1109/TCSVT.2020.2991866.

[368] Williams, C. "On a connection between kernel PCA and metric multidimensional scaling". In: *Advances in neural information processing systems* 13 (2000).

[369] Wu, J., Lu, J., Sabharwal, A., and Mottaghi, R. "Multi-Modal Answer Validation for Knowledge-Based VQA". In: *ArXiv* abs/2103.12248 (2021). DOI: https://doi.org/10.48550/arXiv.2103.12248.

[370] Wu, Q., Shen, C., Hengel, A., Wang, P., and Dick, A. "Image Captioning and Visual Question Answering Based on Attributes and Their Related External Knowledge". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP (Mar. 2016). DOI: 10.1109/TPAMI.2017.2708709.

[371] Wu, Q., Wang, P., Shen, C., Dick, A. R., and Hengel, A. van den. "Ask Me Anything: Free-Form Visual Question Answering Based on Knowledge from External Sources". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 4622–4630. DOI: https://doi.org/10.48550/arXiv.1511.06973.

[372] Wu, T., Ribeiro, M. T., Heer, J., and Weld, D. "Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by C. Zong, F. Xia, W. Li, and R. Navigli. Online: Association for Computational Linguistics, Aug. 2021, pp. 6707–6723. DOI: 10.18653/v1/2021.acl-long.523. URL:

[373] Wu, Y. et al. *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. 2016. DOI: 10.48550/ARXIV.1609.08144. URL:

[374] Wu, Z. and Palmer, M. "Verb semantics and lexical selection". In: *arXiv preprint cmp-lg/9406033* (1994).

[375] Xia, Q., Huang, H., Duan, N., Zhang, D., Ji, L., Sui, Z., Cui, E., Bharti, T., Liu, X., and Zhou, M. *XGPT: Cross-modal Generative Pre-Training for Image Captioning*. 2020. DOI: 10.48550/ARXIV.2003.01473. URL:

[376] Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. "SUN database: Large-scale scene recognition from abbey to zoo". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2010, pp. 3485–3492. DOI: 10.1109/CVPR.2010.5539970.

[377] Xiao, Y. and Wang, W. Y. "On Hallucination and Predictive Uncertainty in Conditional Language Generation". In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Ed. by P. Merlo, J. Tiedemann, and R. Tsarfaty. Online: Association for Computational Linguistics, Apr. 2021, pp. 2734–2744. DOI: 10.18653/v1/2021.eacl-main.236. URL:

[378] Xie, N., Lai, F., Doran, D., and Kadav, A. "Visual Entailment: A Novel Task for Fine-grained Image Understanding". In: *arXiv preprint arXiv:1901.06706* (2019). DOI: https://doi.org/10.48550/arXiv.1901.06706.

[379] Xing, Y., Shi, Z., Meng, Z., Lakemeyer, G., Ma, Y., and Wattenhofer, R. "KM-BART: Knowledge Enhanced Multimodal BART for Visual Commonsense Generation". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Ed. by C. Zong, F. Xia, W. Li, and R. Navigli. Online: Association for Computational Linguistics, Aug. 2021, pp. 525–535. DOI: 10.18653/v1/2021.acl-long.44. URL:

[380] Xu, C., Yang, M., Li, C., Shen, Y., Ao, X., and Xu, R. "Imagine, Reason and Write: Visual Storytelling with Graph Knowledge and Relational Reasoning". In: vol. 35. 4. May 2021, pp. 3022–3029. DOI: 10.1609/aaai.v35i4.16410. URL:

[381] Xu, F., Lin, Q., Han, J., Zhao, T., Liu, J., and Cambria, E. *Are Large Language Models Really Good Logical Reasoners? A Comprehensive Evaluation and Beyond*. 2023. arXiv: 2306.09841 [cs.CL].

[382] Xu, K., Ba, J. L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. "Show, attend and tell: neural image caption generation with visual attention". In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. ICML'15. Lille, France: JMLR.org, 2015, pp. 2048–2057.

[383] Xu, K., Hu, W., Leskovec, J., and Jegelka, S. "How powerful are graph neural networks?" In: *arXiv preprint arXiv:1810.00826* (2018).

[384] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., and He, X. “AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks”. In: *CVPR 2018*. 2018.

[385] Yang, A. et al. *Qwen2 Technical Report*. 2024. arXiv: `2407.10671 [cs.CL]`. URL:

[386] Yang, J., Ang, Y. Z., Guo, Z., Zhou, K., Zhang, W., and Liu, Z. “Panoptic Scene Graph Generation”. In: *European Conference on Computer Vision*. 2022. URL:

[387] Yang, P., Luo, F., Chen, P., Li, L., Yin, Z., He, X., and Sun, X. “Knowledgeable Storyteller: A Commonsense-Driven Generative Model for Visual Storytelling”. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, July 2019, pp. 5356–5362. DOI: `10.24963/ijcai.2019/744`. URL:

[388] Yang, X., Tang, K., Zhang, H., and Cai, J. *Auto-Encoding Scene Graphs for Image Captioning*. 2018. DOI: `10.48550/ARXIV.1812.02378`. URL:

[389] Yang, Z., Gan, Z., Wang, J., Hu, X., Lu, Y., Liu, Z., and Wang, L. “An Empirical Study of GPT-3 for Few-Shot Knowledge-Based VQA”. In: *ArXiv* abs/2109.05014 (2021). DOI: `https://doi.org/10.48550/arXiv.2109.05014`.

[390] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., and Le, Q. V. “XLNet: Generalized Autoregressive Pretraining for Language Understanding”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019. URL:

[391] Yang, Z., Du, X., Mao, R., Ni, J., and Cambria, E. *Logical Reasoning over Natural Language as Knowledge Representation: A Survey*. 2023. arXiv: `2303.12023 [cs.CL]`.

[392] Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. “Tree of Thoughts: Deliberate Problem Solving with Large Language Models”. In: *ArXiv* abs/2305.10601 (2023). URL:

[393] Yasunaga, M., Bosselut, A., Ren, H., Zhang, X., Manning, C. D., Liang, P., and Leskovec, J. “Deep bidirectional language-knowledge graph pretraining”. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. NIPS ’22. New Orleans, LA, USA: Curran Associates Inc., 2022. ISBN: 9781713871088.

[394] Ye, S., Xie, Y., Chen, D., Xu, Y., Yuan, L., Zhu, C., and Liao, J. *Improving Commonsense in Vision-Language Models via Knowledge Graph Riddles*. 2022. DOI: `10.48550/ARXIV.2211.16504`. URL:

[395] Yoon, S., Kang, W. Y., Jeon, S., Lee, S., Han, C., Park, J., and Kim, E.-S. “Image-to-Image Retrieval by Learning Similarity between Scene Graphs”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.12 (May 2021), pp. 10718–10726. ISSN: 2159-5399. DOI: `10.1609/aaai.v35i12.17281`. URL:

[396] Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions”. In: *Transactions of the Association for Computational Linguistics* 2 (2014), pp. 67–78. DOI: `10.1162/tacl_a_00166`. URL:

[397] Yu, F., Tang, J., Yin, W., Sun, Y., Tian, H., Wu, H., and Wang, H. “ERNIE-ViL: Knowledge Enhanced Vision-Language Representations Through Scene Graph”. In: *AAAI*. 2021. DOI: `https://doi.org/10.48550/arXiv.2006.16934`.

[398] Yu, J., Zhu, Z., Wang, Y., Zhang, W., Hu, Y., and Tan, J. “Cross-modal Knowledge Reasoning for Knowledge-based Visual Question Answering”. In: *ArXiv* abs/2009.00145 (2020). DOI: `https://doi.org/10.48550/arXiv.2009.00145`.

[399] Yuan, M., Hu, S., Vulic, I., Korhonen, A., and Meng, Z. “Can Pretrained Language Models (Yet) Reason Deductively?” In: *Conference of the European Chapter of the Association for Computational Linguistics*. 2023. DOI: `https://doi.org/10.48550/arXiv.2210.06442`. URL:

[400] Yuan, Z., Hu, S., Vulic, I., Korhonen, A., and Meng, Z. “Can Pretrained Language Models (Yet) Reason Deductively?” In: *ArXiv* abs/2210.06442 (2022).

[401] Yun, S., Jeong, M., Kim, R., Kang, J., and Kim, H. J. *Graph Transformer Networks*. 2020. DOI: `https://doi.org/10.48550/arXiv.1911.06455`. arXiv: `1911.06455 [cs.LG]`.

[402] Zareian, A., Wang, Z., You, H., and Chang, S.-F. *Learning Visual Commonsense for Robust Scene Graph Generation*. 2020. arXiv: `2006.09623 [cs.CV]`.

[403] Zellers, R., Bisk, Y., Farhadi, A., and Choi, Y. *From Recognition to Cognition: Visual Commonsense Reasoning*. 2019. DOI: `https://doi.org/10.48550/arXiv.1811.10830`. arXiv: `1811.10830 [cs.CV]`.

[404] Zellers, R., Bisk, Y., Schwartz, R., and Choi, Y. *SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference*. 2018. DOI: `10.48550/ARXIV.1808.05326`. URL:

[405] Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. *Self-Attention Generative Adversarial Networks*. 2019. arXiv: `1805.08318 [stat.ML]`.

[406] Zhang, H., Xu, T., and Li, H. "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks". In: Oct. 2017, pp. 5908–5916. DOI: `10.1109/ICCV.2017.629`.

[407] Zhang, J., Zhao, Y., Saleh, M., and Liu, P. J. "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization". In: *ArXiv* abs/1912.08777 (2020).

[408] Zhang, M., Press, O., Merrill, W., Liu, A., and Smith, N. A. *How Language Model Hallucinations Can Snowball*. 2023. arXiv: `2305.13534 [cs.CL]`.

[409] Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. "VinVL: Revisiting Visual Representations in Vision-Language Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 5579–5588. DOI: `https://doi.org/10.48550/arXiv.2101.00529`.

[410] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, June 2018, pp. 586–595. DOI: `10.1109/CVPR.2018.00068`. URL:

[411] Zhang, S., Jiang, X., Yang, Z., Wan, T., and Qin, Z. *Reasoning with Multi-Structure Commonsense Knowledge in Visual Dialog*. 2022. DOI: `10.48550/ARXIV.2204.04680`. URL:

[412] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. *OPT: Open Pre-trained Transformer Language Models*. 2022. arXiv: `2205.01068 [cs.CL]`. URL:

[413] Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T., and Zettlemoyer, L. *OPT: Open Pre-trained Transformer Language Models*. 2022. arXiv: `2205.01068 [cs.CL]`.

[414] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. "BERTScore: Evaluating Text Generation with BERT". In: *ArXiv* abs/1904.09675 (2019). URL:

[415] Zhang, Y., Shi, X., Mi, S., and Yang, X. "Image captioning with transformer and knowledge graph". In: *Pattern Recognition Letters* 143 (2021), pp. 43–49. ISSN: 0167-8655. DOI: `https://doi.org/10.1016/j.patrec.2020.12.020`. URL:

[416] Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., and Shi, S. *Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models*. 2023. DOI: `https://doi.org/10.48550/arXiv.2309.01219`. arXiv: `2309.01219 [cs.CL]`.

[417] Zhang, Z., Zhang, A., Li, M., and Smola, A. J. "Automatic Chain of Thought Prompting in Large Language Models". In: *ArXiv* abs/2210.03493 (2022). URL:

[418] Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., and Smola, A. *Multimodal Chain-of-Thought Reasoning in Language Models*. 2023. DOI: `10.48550/ARXIV.2302.00923`. URL:

[419] Zhao, W., Hu, Y., Wang, H., Wu, X., and Luo, J. *Boosting Entity-aware Image Captioning with Multi-modal Knowledge Graph*. 2021. DOI: `10.48550/ARXIV.2107.11970`. URL:

[420] Zheng, W., Yan, L., Gou, C., and Wang, F.-Y. "Knowledge is Power: Hierarchical-Knowledge Embedded Meta-Learning for Visual Reasoning in Artistic Domains". In: KDD '21 (2021), pp. 2360–2368. DOI: `10.1145/3447548.3467285`. URL:

[421] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. "Places: A 10 Million Image Database for Scene Recognition". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.6 (2018), pp. 1452–1464. DOI: `10.1109/TPAMI.2017.2723009`.

[422] Zhou, Y., Sun, Y., and Honavar, V. G. "Improving Image Captioning by Leveraging Knowledge Graphs". In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2019), pp. 283–293. DOI: `https://doi.org/10.48550/arXiv.1901.08942`.

[423] Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., and Yao, H. "Analyzing and Mitigating Object Hallucination in Large Vision-Language Models". In: *ArXiv* abs/2310.00754 (2023). URL:

[424] Zhu, Y., Groth, O., Bernstein, M. S., and Fei-Fei, L. "Visual7W: Grounded Question Answering in Images". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 4995–5004. DOI: https://doi.org/10.48550/arXiv.1511.03416.

[425] Zhu, Y., Zhang, C., Ré, C., and Fei-Fei, L. *Building a Large-scale Multimodal Knowledge Base System for Answering Visual Queries*. 2015. DOI: https://doi.org/10.48550/arXiv.1507.05670. arXiv: 1507.05670 [cs.CV].

[426] Zhu, Z., Yu, J., Wang, Y., Sun, Y., Hu, Y., and Wu, Q. "Mucko: Multi-Layer Cross-Modal Knowledge Reasoning for Fact-based Visual Question Answering". In: *IJCAI*. 2020. DOI: https://doi.org/10.48550/arXiv.2006.09073.

[427] Ziaeefard, M. and Lecue, F. "Towards Knowledge-Augmented Visual Question Answering". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Ed. by D. Scott, N. Bel, and C. Zong. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 1863–1873. DOI: 10.18653/v1/2020.coling-main.169. URL:

H.F.R.I.
Hellenic Foundation for
Research & Innovation

The research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the 3rd Call for HFRI PhD Fellowships (Fellowship Number: 5537).