



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ
ΕΡΓΑΣΤΗΡΙΟ ΣΥΣΤΗΜΑΤΩΝ ΤΕΧΝΗΤΗΣ ΝΟΗΜΟΣΤΝΗΣ ΚΑΙ ΜΑΘΗΣΗΣ

Auditory Insights into Visual Scenes: A Modular Approach Leveraging Audio Separation and Advanced Language Models

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΤΟΥ

Δημήτριου Κοκκίνη

Επιβλέπων: Αθανάσιος Βουλόδης
Επίκουρος Καθηγητής Ε.Μ.Π

Συνεπιβλέπουσα: Παρασκευή Τζούβελη
Μέλος Ε.ΔΙ.Π Ε.Μ.Π

Αθήνα, Μάρτιος 2025



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

Auditory Insights into Visual Scenes: A Modular Approach Leveraging Audio Separation and Advanced Language Models

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΤΟΥ

Δημήτριου Κοκκίνη

Επιβλέπων: Αθανάσιος Βουλόδημος
Επίκουρος Καθηγητής Ε.Μ.Π

Συνεπιβλέπουσα: Παρασκευή Τζούβελη
Μέλος Ε.ΔΙ.Π Ε.Μ.Π

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 26^η Μαρτίου 2025.

(Υπογραφή)

(Υπογραφή)

(Υπογραφή)

.....

Αθανάσιος Βουλόδημος

Επίκουρος Καθηγητής Ε.Μ.Π

.....

Γεώργιος Στάμου

Καθηγητής Ε.Μ.Π

.....

Ανδρέας-Γεώργιος Σταφυλοπάτης

Ομότιμος Καθηγητής Ε.Μ.Π

Αθήνα, Μάρτιος 2025

(Υπογραφή)

.....
ΔΗΜΗΤΡΙΟΣ ΚΟΚΚΙΝΗΣ

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Copyright © Δημήτριος Κοκκίνης, 2025.

Με την επιφύλαξη παντός δικαιώματος. All rights reserved

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.

στους ανθρώπους μου

Περίληψη

Η παραγωγική τεχνητή νοημοσύνη βρίσκεται στο επίκεντρο της προσοχής λόγω σημαντικών εξελίξεων, όπως η καθοδηγούμενη παραγωγή εικόνας, που έχουν επιτρέψει την διαδεδομένη εφαρμογή της σε πολλούς τομείς. Παρόλο που το κείμενο ως μέσο καθοδήγησης είναι αρκετά άμεσο και συνεκτικό, στερείται της φυσικής σύνδεσης που υπάρχει ανάμεσα στον οπτικό και ακουστικό κόσμο, αδυνατώντας να εκφράσει όλο το φάσμα των πληροφοριών. Για το λόγο αυτό, αρκετές έρευνες έχουν προτείνει μοντέλα παραγωγής εικόνας από ήχο, τα οποία ωστόσο, είναι περιορισμένα σε απλοϊκούς ήχους και δεν έχουν την ικανότητα να παράξουν εικόνα από μια σύνθετη ακουστική σκηνή. Η επεξεργασία της ακουστικής πληροφορίας, απαιτεί μια περίπλοκη διαδικασία που συμπεριλαμβάνει τη συλλογιστική ως μέσο για το συμπερασμό λεπτομερειών. Πρόσφατα, τα μεγάλα γλωσσικά μοντέλα έχουν παρουσιάσει ικανότητες συλλογιστικής και έτσι έχουν ενσωματωθεί σε μοντέλα με εκτεταμένες πολυτροπικές δυνατότητες, τα οποία αποτελούμενα από πολυτροπικούς κωδικοποιητές και μεγάλα γλωσσικά μοντέλα, διαθέτουν την ικανότητα να κατανοούν τον ήχο και να εξάγουν χρήσιμα συμπεράσματα από το περιεχόμενό του.

Στη διπλωματική αυτή, αρχικά ερευνάται η ικανότητα των μεγάλων γλωσσικών μοντέλων ήχου (ALLMs) να παράγουν μια λογική και οπτικά λεπτομερή περιγραφή. Στη συνέχεια, μελετάται μια νέα προσέγγιση στη παραγωγή εικόνας που αξιοποιεί ALLMs και προτείνεται ένα πλαίσιο, σχεδιασμένο για τη διαμόρφωση ευρηματικών οπτικών περιγραφών από σύνθετες ακουστικές εισόδους. Η προτεινόμενη αυτή μέθοδος, αξιοποιεί ένα μοντέλο διαχωρισμού ήχου, ένα γλωσσικό μοντέλο ήχου, ένα μεγάλο γλωσσικό μοντέλο και ένα μοντέλο παραγωγής εικόνας. Αρχικά ο δοσμένος ήχος αποσυντίθενται στις επιμέρους πηγές του και ερμηνεύεται από το γλωσσικό μοντέλο ήχου, το οποίο μεταφράζει το ακουστικό περιεχόμενο σε κειμενικές περιγραφές. Έπειτα, οι περιγραφές αυτές δίνονται ως είσοδοι στο μεγάλο γλωσσικό μοντέλο για τη λογική σύνδεσή τους σε μια οπτική σκηνή από την οποία θα μπορούσε να είχε προκύψει ο ήχος. Η τελική αυτή περιγραφή χρησιμοποιείται για τη καθοδήγηση της παραγωγής εικόνας. Η αποδοτικότητα των εκάστοτε μεθόδων, αξιολογείται από μια σειρά ποσοτικών αλλά και ποιοτικών μετρικών.

Αξιοποιώντας την εσωτερική γνώση και τη συλλογιστική που διαθέτουν τα μεγάλα γλωσσικά μοντέλα, η εργασία αυτή έχει ως σκοπό τη παράκαμψη των περιορισμών που επιβάλλουν τα σχετικά μικρά σύνολα δεδομένων ήχου, καταφέροντας να αποφανθούν λεπτομέρειες από τον δοσμένο ήχο μέσω συλλογιστικής και παράγοντας μια αληθοφανή περιγραφή της οπτικής σκηνής.

Λέξεις Κλειδιά

Πολυτροπικά Μεγάλα Γλωσσικά Μοντέλα, Γλωσσικά Μοντέλα Ήχου, Προτροπή, Διαχωρισμός Καθολικού Ήχου, Σταθερή Διάχυση, Παραγωγή Εικόνας, Όραση Υπολογιστών

Abstract

Generative AI has garnered significant interest due to its remarkable advancements in guided image generation, which have enabled its widespread application across various fields. Although text is straightforward, it lacks the inherent connection that exists between visual and auditory realms, being unable to convey the entire spectrum of information. To that end, a number of works have introduced audio-guided image generation. However, they are limited to plain sounds and struggle to picture an intricate acoustic scene. Auditory information requires a complex process involving reasoning to infer missing details. Recently, Large Language Models have exhibited reasoning abilities and have been integrated as the cognitive powerhouse of models with multi-modal extended abilities. These models, consisting of multi-modal encoders and LLM, have the capacity to understand audio and reason about its content.

This thesis first investigates the capability of Audio Large Language Models (ALLMs) to produce a coherent and visually detailed description. It further explores a novel approach to image generation by leveraging ALLMs and introduces a structured framework designed to transform complex auditory inputs into meaningful and imaginative visual representations. The proposed pipeline integrates multiple components, including an audio source separation model, an audio language model (ALM), a large language model (LLM), and an image generation model. By first decomposing mixed audio into distinct sources, the ALM interprets and translates auditory information into textual descriptions, which are subsequently refined by the LLM to enhance contextual understanding. The resulting structured textual representation is then used to guide the image generation model, producing images that align semantically with the original audio input. The efficacy of the methodologies is assessed through a combination of quantitative and qualitative measures.

By leveraging the internal knowledge and linguistic reasoning of LLMs, this research aims at alleviating the limitations imposed by constrained audio datasets, managing to infer visual details that can be deduced from the input audio, and producing a plausible description of the visual scene.

Keywords

Multi-Modal Large Language Models, Audio Language Models, Prompting, Universal Sound Separation, Stable Diffusion, Image Generation, Computer Vision

Ευχαριστίες

Η διπλωματική αυτή αποτέλεσε μια προσωπική Οδύσσεια, με την ολοκλήρωση της οποίας νιώθω ως άνθρωπος αλλά και ως μηχανικός πιο μεστός. Για τη πορεία αυτή θα ήθελα αρχικά να ευχαριστήσω τον Λευτέρη Τσώνη, ο οποίος είχε την υπομονή και μου έδωσε το χώρο να εκφράσω τις ιδέες μου έτσι ώστε να σμιλευτεί σιγά σιγά το θέμα. Είμαι επίσης ιδιαίτερα ευγνώμων για τη βοήθεια που μου παρείχε η κ.Τζούβελη, η οποία πίστεψε σε εμένα σε μια φάση που τα κύματα φαινόταντουσαν αδιόρατα και παρείχε την κατάλληλη ενθάρρυνση για την επίτευξη αυτού του στόχου. Δεν μπορώ φυσικά να παραλείψω τους γονείς μου, που έγιναν η σχεδιά πάνω στην οποία στηρίχτηκα, τα αδέρφια μου, που μου έδιναν την ώθηση για να συνεχίσω, αλλά και τους κοντινούς μου ανθρώπους και τη κοπέλα μου, οι οποίοι έκαναν το ταξίδι αυτό πιο ευχάριστο.

Κοκκίνης Δημήτριος
Μάρτιος 2025

Περιεχόμενα

Περιεχόμενα	13
Κατάλογος Σχημάτων	15
Κατάλογος Πινάκων	19
1 Εκτεταμένη περίληψη στα Ελληνικά	21
2 Introduction	53
2.1 Motivation	53
2.2 Goals and Contributions	54
3 Theoretical Background	55
3.1 Machine Learning	55
3.1.1 Supervised Learning	55
3.1.2 Unsupervised Learning	56
3.1.3 Reinforcement Learning	56
3.2 Deep Learning	56
3.2.1 The neuron	56
3.2.2 Feedforward Neural Networks	58
3.2.3 Convolutional Networks	58
3.2.4 U-Net Architecture	60
3.3 Transformers	61
3.3.1 Architecture	61
3.3.2 Self-Attention	62
3.4 Audio Source Separation	63
3.4.1 Sound	63
3.4.2 Audio Representations	63
3.4.3 Traditional Methods	65
3.4.4 Deep Learning Approaches	66
3.4.5 Evaluation Metrics for Source Separation	68
3.5 Diffusion Models	69
3.5.1 Generative Models	69
3.5.2 Diffusion Process	70
3.5.3 Improvements	72

4	Related Work	73
4.1	Language Models	73
4.1.1	Large Language Models	73
4.1.2	Prompting	75
4.2	Guided Image Generation	77
4.2.1	Text-to-Image	77
4.2.2	Audio-to-Image	78
4.3	Audio Language Models	80
4.3.1	Computer Audition	80
4.3.2	Audio Large Language Models (ALLMs)	81
4.4	Universal Sound Separation	84
5	Proposal	87
6	Experiments	91
6.1	Dataset Preparation	91
6.2	Experiment A: Prompting Techniques Comparison	93
6.2.1	Qualitative Results	95
6.2.2	Evaluation Methods	95
6.2.3	Quantitative Results	98
6.3	Experiment B: Comparison with Proposed Method	101
6.4	Experiment C: Generating Images from Movie Scenes	107
6.4.1	Dataset	108
6.4.2	Human Evaluation	108
6.4.3	CLIPscore	111
7	Conclusion	113
7.0.1	Discussion	113
7.0.2	Limitations	113
7.0.3	Future Work	113
A	Additional Details	115
A.1	Prompts	115
A.2	Example Results	119
A.3	Human Evaluation	121

Κατάλογος Σχημάτων

1.0.1 Νευρώνας (Αντίληπτρο) [23]	22
1.0.2 Η υπερβολική εφαπτομένη επεκτείνει την έξοδο στο εύρος -1 σε 1	23
1.0.3 Πράξη της συνέλιξης	25
1.0.4 Διεσταλμένη Συνέλιξη	25
1.0.5 Αρχιτεκτονική U-Net απο [59]	26
1.0.6 Αρχιτεκτονική Μετασχηματιστή απο [75]	27
1.0.7 Ένα ηχητικό κύμα αποτελείται απο συμπίεσεις και αραιώσεις	28
1.0.8 Μετατροπή Αναλογικού Σήματος σε Ψηφιακό	28
1.0.9 Παράδειγμα Δυαδικής Μάσκας απο [52]	30
1.0.10 Αρχιτεκτονική Wave-U-Net απο [69]	30
1.0.11 Διαδικασία GAN από [62]	31
1.0.12 Η εμπρόσθια και η αντίστροφη διαδικασία [35]	32
1.0.13 Τραμμικό(πάνω) και Συνημιτονικό Πρόγραμμα (κάτω) απο [54]	32
1.0.14 Βασική (zero-shot) προτροπή και Προτροπή με Chain-of-Thought απο [79]	34
1.0.15 Αρχιτεκτονική Stable Diffusion απο [58]	34
1.0.16 Συσχέτιση Ζεύγους Εικόνας-Κειμένου	35
1.0.17 Παραχθείσες Εικόνες Αναμειγνύοντας Δύο Διαφορετικές Κυματομορφές Ήχου απο [70]	36
1.0.18 Αρχιτεκτονική AudioToken απο [81]	36
1.0.19 Γενική Αρχιτεκτονική Πολυτροπικών Μεγάλων Γλωσσικών Μοντέλων απο [82]	37
1.0.20 Αρχιτεκτονική Qwen-Audio [15]	38
1.0.21 ConvTasNet Pipeline [51]	40
1.0.22 Προτεινόμενη Μέθοδος	40
1.0.23 Παραγωγή Εικόνας	41
1.0.24 Ξύνολο Δεδομένων	42
1.0.25 Κατανομή Συχνότητας Ακουστικών Λεζάντων	42
1.0.26 Κατανομή του Αριθμού των Πηγών	43
1.0.27 CLAPscore	44
1.0.28 Under Separation (<i>left</i>), Equal Separation (<i>center</i>), Over Separation (<i>right</i>)	44
1.0.29 Ακρίβεια Αναγνώρισης των Ήχων	45
1.0.30 Μετρικές Περιγραφής Εικόνας	45
1.0.31 Μετρικές Αντίληψης Αριθμού Πηγών Ήχου	47
1.0.32 Ακρίβεια Αναγνώρισης Πηγών Ήχου	47
1.0.33 Μέσο Σκορ Συνοχής για κάθε Μέθοδο	48
1.0.34 Μέσο Σκορ Οπτικής Λεπτομέρειας για κάθε Μέθοδο	48

1.0.3 Μέσο Σκορ Φαντασίας για κάθε Μέθοδο	49
1.0.3 Μέσο Σκορ Καταλληλότητας Περιγραφής Εικόνας για κάθε Μέθοδο	49
1.0.3 Όμοιότητα με Ήχο (αριστερά), Οπτικά Λεπτομερές (κέντρο), Συνολικά Καλύτερη (δεξιά)	50
1.0.3 Όμοιότητα με Ήχο (αριστερά), Ρεαλιστικότητα (δεξιά)	51
1.0.3 Λεπτομέρεια	51
3.2.1 Perceptron Neuron [23]	57
3.2.2 The tanh function extends the output range from -1 to 1	57
3.2.3 Convolution operation	59
3.2.4 Architecture of a CNN from [2]	60
3.2.5 Dilated Convolution	60
3.2.6 U-Net architecture	61
3.3.1 Transformer Architecture	62
3.4.1 A sound wave is made up of compression and rarefaction	63
3.4.2 Analog to digital conversion	64
3.4.3 Spectrogram of an audio signal	65
3.4.4 Example of a binary mask operation from [52]	67
3.4.5 Wave-U-Net Architecture	68
3.5.1 Generative modeling can estimate the underlying distribution and generate new samples	69
3.5.2 GAN Pipeline from [62]	70
3.5.3 Forward and Reverse processes from [35]	71
3.5.4 Linear(<i>top</i>) and Cosine schedule(<i>bottom</i>). With the Cosine Schedule data is perturbed more gradually [54]	72
4.1.1 The word embeddings capture meaningful relations between words. As an example, a specific direction might encode the gender information or the capital of a country.	74
4.1.2 Larger models make increasingly efficient use of in-context information [9]	76
4.1.3 Standard (zero-shot) prompting and Chain-of-Thought prompting [79]	76
4.2.1 Stable Diffusion Architecture [58]	77
4.2.2 Image-Text Embeddings Alignment Illustration	78
4.2.3 Generated images by mixing two different audios in the waveform space [70]	79
4.2.4 AudioToken Architecture [81]	79
4.3.1 The general model architecture of MM-LLMs [82]	82
4.3.2 Qwen-Audio Architecture [15]	83
4.4.1 ConvTasNet Pipeline [51]	85
4.4.2 Variable source separation for a separation model with $M = 4$ outputs and input mixture with $M_a = 2$ active references from [80]	85
5.0.1 Process of Acoustic Imagery. Input audio (<i>left</i>) and generated image (<i>right</i>).	87
5.0.2 Proposed Pipeline	88
5.0.3 Image Generation Pipeline	88
6.1.1 Dataset	92

6.1.2 Label Distribution	92
6.1.3 Number of Sources Distribution	93
6.2.1 Label Distribution for the small dataset	94
6.2.2 Number of Sources Distribution for the small dataset	95
6.2.3 CLAPscore	98
6.2.4 Under Separation (<i>left</i>), Equal Separation (<i>center</i>), Over Separation (<i>right</i>) . . .	99
6.2.5 Source Perception for Each Prompting Technique	99
6.2.6 Label Accuracy	100
6.2.7 Image Description Metrics	101
6.3.1 Source Separation Metrics	103
6.3.2 Equal Separation	103
6.3.3 Label Accuracy per Number of Sources (We define as <i>Proposed_EnCLAP</i> the proposed pipeline using the EnCLAP model and <i>Proposed_Qwen</i> the proposed pipeline using the Qwen model)	104
6.3.4 Coherence Mean Score for Each Method	105
6.3.5 Visual Detail Mean Score for Each Method	106
6.3.6 Imagination Mean Score for Each Method	106
6.3.7 Image Appropriateness Mean Score for Each Method	107
6.4.1 Samples Preparation	108
6.4.2 Suitability According to Audio (<i>left</i>), Visual Detail (<i>center</i>), Overall Quality (<i>right</i>)	109
6.4.3 Similarity to Audio (<i>left</i>), Realism (<i>right</i>)	110
6.4.4 Level of Detail Mean Score for Each Method	110
6.4.5 Example Outputs for each Method	112
A.3.1Participants Gender (<i>left</i>) and Age (<i>right</i>)	121
A.3.2Participants Hearing Difficulty (<i>left</i>) and Level of English (<i>right</i>)	121
A.3.3Participants Familiarity with AI	121
A.3.4Image Results from Experiment C: ALLM Images (top row), Falcon-Zero_Shot Images (middle row), Mistral-Chain_of_Thought Images (bottom row)	123

Κατάλογος Πινάκων

1.0.1 Επιλογές Προτεινόμενου Συστήματος	46
1.0.2 CLIPscore	52
6.1.1 Dataset Statistics	93
6.2.1 Prompt Templates	94
6.2.2 Responses of each prompting technique	96
6.3.1 Proposed Pipeline Settings	104
6.4.1 CLIPscore	111
A.1.1 Prompt Templates for the Audio Perception Evaluation	115
A.1.2 Prompt Templates for Image Description Evaluation	116
A.1.3 Prompt Template for Label Accuracy of Intermediate Captions	117
A.1.4 Prompt Template for Qwen 2.5 Text Merging	117
A.1.5 Prompt Template for Falcon Text Merging	118
A.1.6 Prompt Template for Mistral Text Merging	118
A.2.1 Example Results from Experiment B for an audio clip with weak labels: [Rain, Water]	119
A.2.2 Example Results from Experiment B for an audio clip with weak labels: [Rain, Water]	120
A.3.1 Example Description Results from the Corresponding Captions	122

Κεφάλαιο 1

Εκτεταμένη περίληψη στα Ελληνικά

Θεωρητικό Υπόβαθρο

Μηχανική Μάθηση

Η μηχανική μάθηση είναι ένα πεδίο της τεχνητής νοημοσύνης, που ασχολείται με την ανάπτυξη και μελέτη στατιστικών αλγορίθμων που μπορούν να γενικεύσουν αποτελεσματικά και έτσι να εκτελούν εργασίες χωρίς ρητές οδηγίες. Ο όρος «μάθηση», δηλώνει τη διαδικασία καθορισμού ορισμένων παραμέτρων του μοντέλου από τα δεδομένα εισόδου. Οι αλγόριθμοι μηχανικής μάθησης μπορούν να ταξινομηθούν σε τρεις κύριες κατηγορίες με βάση τον τύπο της μάθησης:

- Επιβλεπόμενη Μάθηση
- Μη-Επιβλεπόμενη Μάθηση
- Ενισχυτική Μάθηση

Στην επιβλεπόμενη μάθηση, το σύνολο δεδομένων που χρησιμοποιείται για την εκπαίδευση του μοντέλου αποτελείται από παραδείγματα εισόδου-εξόδου, όπου ένα διάνυσμα χαρακτηριστικών x συνδυάζεται με μια αντίστοιχη ετικέτα y , που αντιπροσωπεύει την επιθυμητή έξοδο. Στόχος είναι η εκμάθηση μιας αντιστοίχισης $y = f(x)$ μεταξύ των δεδομένων εισόδου και της επιθυμητής εξόδου, ελαχιστοποιώντας μια συνάρτηση απώλειας. Αφού εκπαιδευτεί, το μοντέλο μπορεί να χρησιμοποιηθεί για να κάνει προβλέψεις ή να ταξινομήσει δεδομένα. Από στατιστική άποψη, ο αλγόριθμος στοχεύει να προσεγγίσει τη υποβόσκουσα συνάρτηση πυκνότητας πιθανότητας $p(y|x)$.

Στη μη επιβλεπόμενη μάθηση τα δεδομένα δεν είναι επισημασμένα. Το μοντέλο εκπαιδεύεται ώστε να ανακαλύψει τα υποκείμενα μοτίβα που διέπουν την κατανομή των δεδομένων, αναπτύσσοντας έτσι την ικανότητα να διαμορφώνει εσωτερικές αναπαραστάσεις της εισόδου. Ένας κλάδος της μη-επιβλεπόμενης μάθησης είναι η αυτοεπιβλεπόμενη μάθηση, όπου ο αλγόριθμος αυτόνομα κατασκευάζει τους δικούς του στόχους από τα δεδομένα εισόδου. Τέλος, η ημι-επιβλεπόμενη μάθηση είναι ένας αλγόριθμος μηχανικής μάθησης που βρίσκεται μεταξύ της επιβλεπόμενης και της μη επιβλεπόμενης μάθησης. Ο στόχος της ημι-επιβλεπόμενης μάθησης είναι να ταξινομήσει τα μη επισημασμένα δεδομένα από τα επισημασμένα δεδομένα.

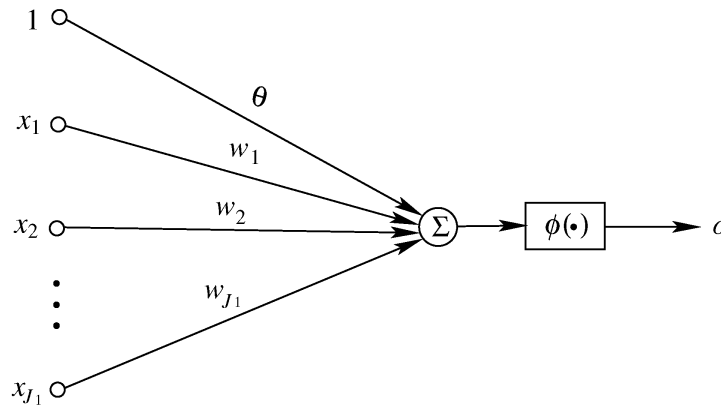
Η ενισχυτική μάθηση περιλαμβάνει μια συνεχόμενη αλληλεπίδραση με το περιβάλλον, με στόχο τη μείωση μιάς μετρικής απόδοσης. Ο αλγόριθμος μαθαίνει από ένα πρότυπο δοκιμής και σφάλματος μέσω ανταμοιβών. Εξέχοντα παραδείγματα μεθοδολογιών ενισχυτικής μάθησης περιλαμβάνουν το Q-Learning και τα Deep Q-Networks (DQN).

Βαθιά Μάθηση

Με την έλευση της βαθιάς μάθησης, ενός υποτομέα της μηχανικής μάθησης, η ιδέα ήταν να αφήσουμε το μοντέλο να μάθει τα χαρακτηριστικά από μόνο του. Ο όρος «βαθιά μάθηση» αναφέρεται στη πολυεπίπεδη δομή της διαδικασίας μάθησης, όπου κάθε επίπεδο εξάγει όλο και πιο αφηρημένα χαρακτηριστικά, οδηγώντας σε βελτιωμένη απόδοση σε διάφορους τομείς. Τα επίπεδα αυτά οργανώνονται σε δομές που ονομάζονται πολυεπίπεδα νευρωνικά δίκτυα.

Νευρώνας

Το βασικό δομικό στοιχείο των νευρωνικών δικτύων είναι ο νευρώνας, εμπνευσμένος από τον τρόπο που λειτουργούν οι νευρώνες στον ανθρώπινο εγκέφαλο. Η λειτουργία του νευρώνα είναι σαφής: δέχεται ως είσοδο ένα σύνολο εισόδων \hat{x} , πολλαπλασιασμένων με βάρη \hat{w} , και με τη προσθήκη ενός κατωφλίου θ , παράγεται μια έξοδος \hat{o} . Η διαδικασία αυτή περιγράφεται μαθηματικά ως: $\hat{o} = \phi(\theta + \hat{x}^T \hat{w})$. Τα βάρη ρυθμίζουν το μέγεθος της επιρροής κάθε εισόδου και η συνάρτηση ενεργοποίησης εισάγει μη-γραμμικότητα.



Σχήμα 1.0.1: Νευρώνας (Αντίληπτρο) [23]

Οι συναρτήσεις ενεργοποίησης εφαρμόζονται στο εσωτερικό γινόμενο του σήματος εισόδου με τα βάρη του νευρώνα. Κοινές συναρτήσεις ενεργοποίησης αποτελούν:

1. Η σιγμοειδής ή λογιστική σιγμοειδής συνάρτηση, η οποία έχει ως έξοδο αριθμούς μεταξύ 0 και 1. Ένα θέμα με τη συγκεκριμένη συνάρτηση είναι πως δεν είναι ισορροπημένη γύρω από το 0.

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

2. Η εφαπτομενική συνάρτηση ή αλλιώς υπερβολική εφαπτομένη, έχει το ίδιο σχήμα με τη σιγμοειδή, ωστόσο η έξοδος βρίσκεται μεταξύ του -1 και του 1.

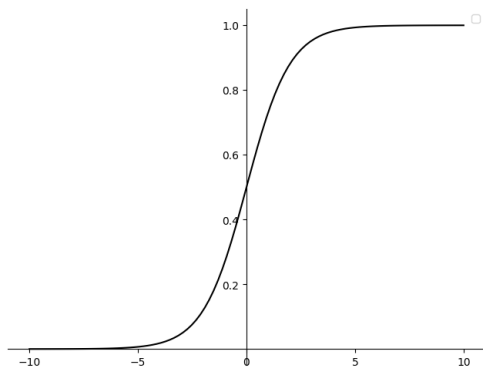
$$\tanh(x) = \frac{2}{1 + e^{-x}} - 1$$

Η έξοδος αυτή είναι κατάλληλη για την εκμάθηση με πολλαπλά επίπεδα.

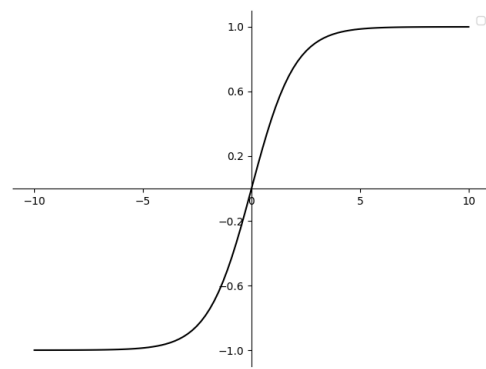
3. Η ανορθωμένη γραμμική μονάδα (ReLU) ενεργοποιεί τον νευρώνα μόνο όταν η έξοδος είναι θετική

$$A(x) = \max(0, x)$$

Μια παραλλαγή αυτής της συνάρτησης, είναι η διαρρέουσα (Leaky) ReLU, η οποία έχει μικρή κλίση για αρνητικές τιμές αντί για επίπεδη και χρησιμοποιείται στις περιπτώσεις όπου οι κλίσεις είναι αραιές.



(α') Λογιστική Σιγμοειδής



(β') Υπερβολική Εφαπτομένη

Σχήμα 1.0.2: Η υπερβολική εφαπτομένη επεκτείνει την έξοδο στο εύρος -1 σε 1

Νευρωνικά Δίκτυα Πρόσω Τροφοδότησης

Για τη διαχείριση πιο περίπλοκων κατανομών δεδομένων, η στοίβαξη πολλαπλών στρωμάτων νευρώνων μπορεί να κατασκευάσει μια βαθιά δομή γνωστή ως νευρωνικό δίκτυο πρόσω τροφοδότησης, θεμέλιο της βαθιάς μάθησης. Εκτός από το επίπεδο εισόδου και εξόδου, κάθε ενδιάμεσο στρώμα αναφέρεται ως κρυφό στρώμα. Οι νευρώνες σε ένα επίπεδο συνδέονται με όλους τους νευρώνες στο προηγούμενο επίπεδο. Αυτά τα δίκτυα ονομάζονται εμπρόσθια λόγω της ροής της πληροφορίας προς τα εμπρός, από το επίπεδο εισόδου μέσω των κρυφών επιπέδων μέχρι το επίπεδο εξόδου. Ο στόχος ενός τέτοιου δικτύου είναι να προσεγγίσει κάποια στοχευμένη συνάρτηση f^* [30], χρησιμοποιώντας τα δεδομένα εισόδου (προσαρμογή καμπύλης). Κατά την εκπαίδευση, το νευρωνικό δίκτυο μαθαίνει ένα σύνολο παραμέτρων θ που ελαχιστοποιούν μια συνάρτηση κόστους $J(\theta)$, που είναι συνάρτηση των βαρών του δικτύου. Η βελτιστοποίηση αυτή έχει ως στόχο την εύρεση του ελάχιστου της συνάρτησης κόστους χρησιμοποιώντας κάθοδο κλίσης (gradient descent).

Ο αλγόριθμος καθόδου κλίσης αναφέρεται στη διαδικασία υπολογισμού των κλίσεων της συνάρτησης κόστους και της ενημέρωσης των βαρών προς τη κατεύθυνση ελαχιστοποίησης της συναρτήσεως, σύμφωνα με τον κανόνα:

$$w(i+1) = w(i) - \eta \frac{\partial J(w^T, b)}{\partial w}$$

,όπου το b είναι ο όρος κατωφλίου και η είναι ο ρυθμός εκμάθησης, ο οποίος καθορίζει το πόσο γρήγορα συγχλίνει ο αλγόριθμος σε κάποιο ελάχιστο.

Για την ενημέρωση των βαρών σε όλα τα επίπεδα του δικτύου χρησιμοποιείται ένας αλγόριθμος, καθοριστικός στην πρόοδο της βαθιάς μάθησης, ο οποίος ονομάζεται οπισθοδιάδοση. Ο αλγόριθμος αυτός, περιλαμβάνει την επαναληπτική προσαρμογή των βαρών του δικτύου προς τη κατάλληλη κατεύθυνση.

Συνελικτικά Δίκτυα

Στο ψηφιακό κόσμο, οι εικόνες αναπαριστώνται ως δυσδιάστατοι (για ασπρόμαυρες εικόνες) ή τρισδιάστατοι (για έγχρωμες εικόνες) πίνακες αριθμητικών τιμών, που αντιπροσωπεύουν τη φωτεινότητα ενός εικονοστοιχείου (pixel). Αυτή η δομή πλέγματος είναι εξέχων χαρακτηριστικό των δεδομένων εικόνας, καθώς επιτρέπει σε πολλαπλά εικονοστοιχεία να ορίσουν ξεχωριστές περιοχές μέσα στην εικόνα. Τα παραδοσιακά πολυεπίπεδα νευρωνικά δίκτυα εμπρόσθιας τροφοδότησης, δεν έχουν τη δυνατότητα να χειριστούν τέτοιου είδους δεδομένα, διότι είναι υπολογιστικά ακριβό και δεν είναι ικανά να διατηρήσουν την εγγενή δομή της εικόνας. Τα συνελικτικά δίκτυα απεναντίας (CNN), ένας ειδικός τύπος νευρωνικών δικτύων, είναι σχεδιασμένα να επεξεργάζονται δεδομένα με τη συγκεκριμένη πλεγματοειδή δομή. Τα δίκτυα αυτά, βασίζονται στην τέλεση της διαδικασίας της συνέλιξης.

Η διακριτή δισδιάστατη συνέλιξη εμπεριέχει δύο πίνακες, τον πίνακα I που αναπαριστά το μέρος του δεκτικού πεδίου της εικόνας και τον πίνακα K , ο οποίος αναφέρεται και ως πυρήνας, που είναι το σύνολο των παραμέτρων για εκμάθηση. Η διαδικασία περιγράφεται μαθηματικά ως:

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(mn)K(i - m, j - n)$$

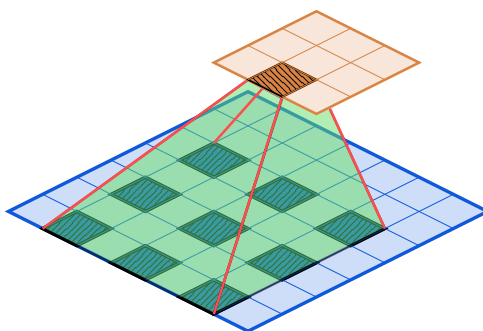
Η διαδικασία εμπρόσθιας διάδοσης περιλαμβάνει τη διάσχιση του πυρήνα κατά ύψος και κατά πλάτος της εικόνας, παράγοντας έτσι μια χωρική αναπαράσταση του δεκτικού πεδίου. Αυτό έχει ως αποτέλεσμα ένα δισδιάστατο χάρτη ενεργοποίησης που κωδικοποιεί την έξοδο του πυρήνα σε κάθε θέση της εικόνας.

Οι διαστάσεις του πυρήνα ορίζουν το δεκτικό πεδίο του συνελικτικού επιπέδου και είναι μια υπερπαραμέτρος του δικτύου. Οι πυρήνες, ή αλλιώς φίλτρα, εξάγουν χρήσιμα οπτικά χαρακτηριστικά από την εικόνα, όπως ακμές, καμπύλες και σχήματα. Καθώς προχωράμε πιο βαθιά στο δίκτυο, τα συνελικτικά στρώματα εξάγουν χαρακτηριστικά υψηλότερου επιπέδου.

Ένα τυπικό συνελικτικό δίκτυο αποτελείται από συνελικτικά στρώματα για την εξαγωγή χαρακτηριστικών, από στρώσεις συγχέντρωσης (pooling layers) και από εμπρόσθια επίπεδα για ταξινόμηση. Ο σκοπός των στρωμάτων συγχέντρωσης είναι η συστηματική μείωση των χωρικών διαστάσεων της αναπαράστασης για την ελάττωση του αριθμού των παραμέτρων και των υπολογιστικών αναγκών. Ένας ειδικός τύπος συνέλιξης είναι η διεσταλμένη συνέλιξη, η οποία εφαρμόζεται στην είσοδο με ορισμένα κενά. Μια επιπλέον παράμετρος l (ρυθμός διαστολής) υποδεικνύει την διεύρυνση του πυρήνα. Η σημαντικότητα της διεσταλμένης συνέλιξης έγκυται στην διεύρυνση του δεκτικού πεδίου, χωρίς να επιβαρύνεται η μνήμη ενώ ταυτόχρονα διατηρείται η ευκρίνεια.

Figure 1 illustrates matrix multiplication. It shows three matrices: I (8x8), K (3x3), and $I * K$ (8x8). Matrix I contains values 0, 1, and 2, with some cells highlighted in orange and blue. Matrix K contains values 1, 0, and 1, with all cells highlighted in teal. Matrix $I * K$ contains values 1, 2, 3, and 4, with some cells highlighted in blue and orange. Dashed lines indicate the element-wise multiplication and summation process.

Σχήμα 1.0.3: Πράξη της συνέλιξης



Σχήμα 1.0.4: Διεσταλμένη Συνέλιξη

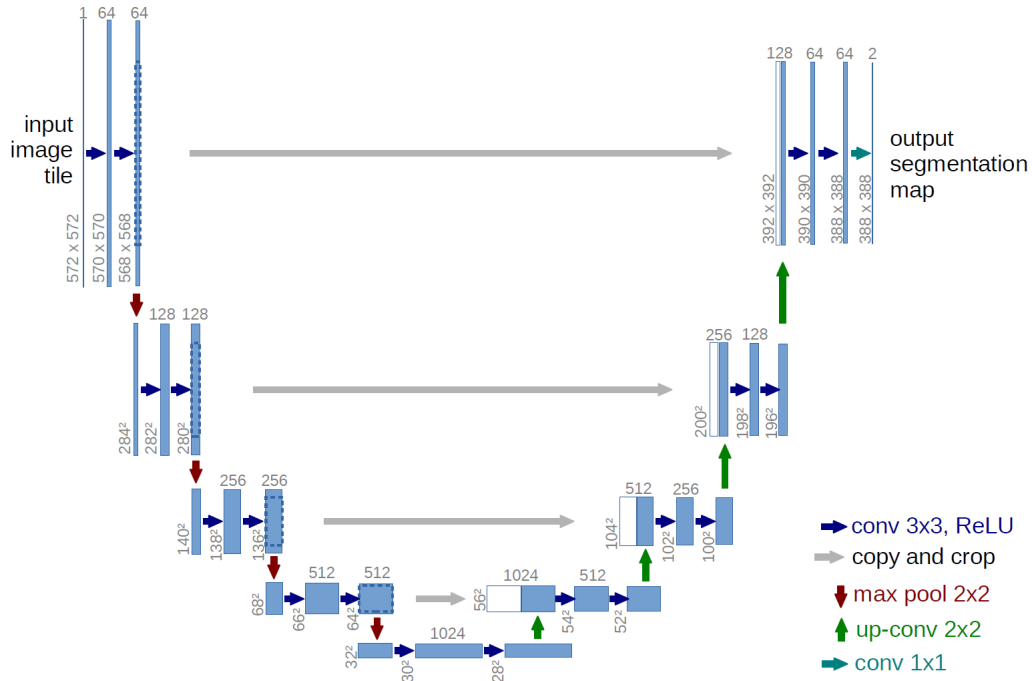
Αρχιτεκτονική U-Net

Η αρχιτεκτονική U-Net [59] αποτελείται από έναν κωδικοποιητή, ο οποίος υποδειγματοληπτεί και εξάγει χρήσιμα χαρακτηριστικά από την εικόνα εισόδου μέσα από συνεχόμενα συνελικτικά επίπεδα και επίπεδα μέγιστης συγκέντρωσης (max pooling), και έναν αποκωδικοποιητή που χρησιμοποιείται για την υπερδειγματοληψία. Μέσω του αποκωδικοποιητή, το πεδίο όρασης των συνελικτικών πυρήνων αυξάνεται. Παρόλο που αρχικά προτάθηκε για κατάτμηση εικόνων, έγινε αναπόσπαστο κομμάτι ενός μεγάλου αριθμού μοντέλων που ειδικεύονται στην όραση υπολογιστών και στις διεργασίες ήχου.

Μετασχηματιστές

Ακολουθιακά δεδομένα όπως οι χρονοσειρές, ο ήχος ή το κείμενο απαιτούν ένα ειδικευμένο δίκτυο, το οποίο θα έχει την ικανότητα να συλλαμβάνει μακροπρόθεσμες εξαρτήσεις. Αρκετές αρχιτεκτονικές δικτύων προτάθηκαν για να λύσουν το συγκεκριμένο πρόβλημα, όπως τα Αναδρομικά Νευρωνικά Δίκτυα και τα LSTMs, ωστόσο τα δίκτυα αυτά είχαν αρκετούς περιορισμούς. Με την εισαγωγή των μετασχηματιστών [75], το πεδίο της τεχνητής νοημοσύνης άλλαξε ριζικά.

Αρχιτεκτονική



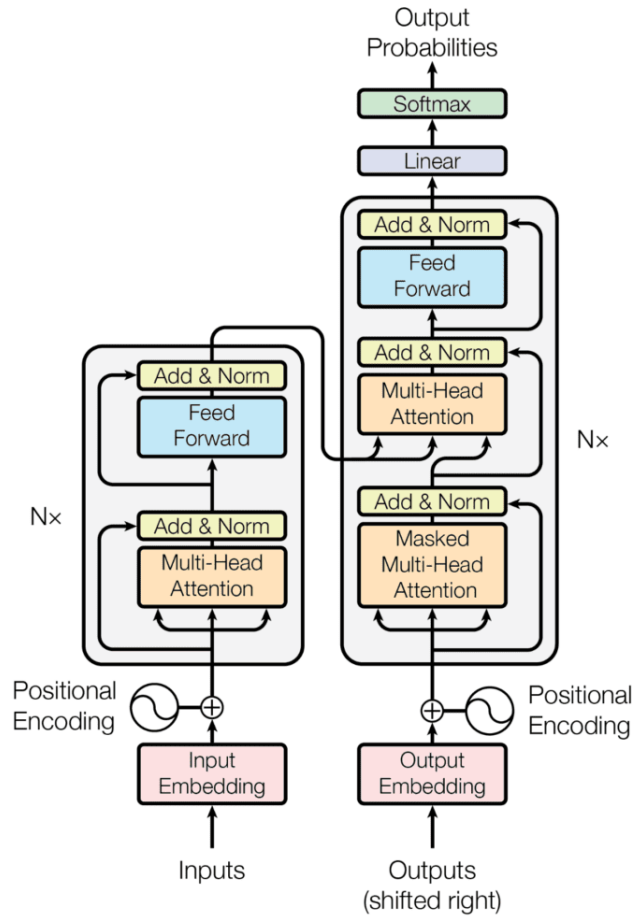
Σχήμα 1.0.5: Αρχιτεκτονική U-Net απο [59]

Η γενική αρχιτεκτονική του μετασχηματιστή φαίνεται στο παρακάτω σχήμα και αποτελείται από έναν κωδικοποιητή και έναν αποκωδικοποιητή. Ο κωδικοποιητής περιέχει 6 στοιβαγμένα επίπεδα, το καθένα από τα οποία αποτελείται από ένα πολυκεφαλικό μπλοκ αυτο-προσοχής και ένα δίκτυο προώθησης, και αντιστοιχεί την ακολουθία εισόδου σε μια ακολουθία συνεχών αναπαραστάσεων που τροφοδοτείται στον αποκωδικοποιητή. Ο αποκωδικοποιητής αποτελείται επίσης από 6 στοιβαγμένα στρώματα, με μια επιπρόσθετη μονάδα μάσκας πολυκεφαλικού μπλοκ αυτο-προσοχής στην αρχή κάθε στρώματος.

Σε αντίθεση με προηγούμενες αρχιτεκτονικές, ο μετασχηματιστής λαμβάνει όλη την είσοδο μαζί και όχι σε ξεχωριστά βήματα. Για την κωδικοποίηση της θέσης κάθε πληροφορίας, χρησιμοποιεί κωδικοποίηση με επίγνωση της θέσης.

Αυτο-προσοχή

Αρχικά, η ακολουθία εισόδου διαιρείται σε ατομικά στοιχεία που ονομάζονται tokens. Κάθε ένα αντιστοιχίζεται με ένα διάνυσμα που κωδικοποιεί το νόημα και τη θέση του στο κείμενο. Ξεκινώντας με τρεις πίνακες, query(Q), key(K), value(V), το διάνυσμα κάθε token πολλαπλασιάζεται με τους πίνακες query και key για να παραχθούν τα διανύσματα query και key. Η αυτο-προσοχή είναι η διαδικασία της προσοχής σε συγκεκριμένα σημεία της εισόδου και του υπολογισμού της ομοιότητας ανάμεσα σε κάθε στοιχείο. Μια βαθμολογία προσοχής υπολογίζεται ανάμεσα σε κάθε διάνυσμα query και key, κλιμακωμένη με έναν παράγοντα $\frac{1}{\sqrt{d_k}}$ όπου d_k είναι η διάσταση του πίνακα key. Το αποτέλεσμα περνάει μέσα από τη συνάρτηση softmax για να παραχθεί μια κατανομή πιθανότητας. Χρησιμοποιώντας τα αποτελέσματα αυτά, τα πολλαπλασιάζουμε με τα διανύσματα value για να πάρουμε το τελικό αποτέλεσμα. Η διαδικασία αυτή είναι ένα κεφάλι προσοχής.



Σχήμα 1.0.6: Αρχιτεκτονική Μετασχηματιστή από [75]

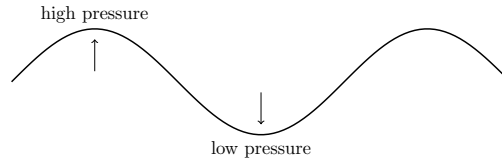
$$Attention(K, Q, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Διαχωρισμός Πηγών Ήχου

Ο διαχωρισμός ήχου αναφέρεται στη διαδικασία της ανάκτησης των επιμέρους πηγών ήχου από ένα ακουστικό μείγμα. Αυτό το πρόβλημα είναι θεμελιώδες σε αρκετές διεργασίες επεξεργασίας ακουστικών σημάτων, καθώς η ανάλυση και η επεξεργασία των απομονωμένων ηχητικών πηγών φέρνει καλύτερα αποτελέσματα. Αν και το επίκεντρο αυτού του ερευνητικού τομέα είναι η βελτίωση της ακουστικής ομιλίας, ο διαχωρισμός ομιλιών και ο διαχωρισμός μουσικών οργάνων, τα τελευταία χρόνια ο γενικευμένος τομέας του διαχωρισμού καθολικού ήχου έχει γνωρίσει αρκετή πρόοδο.

Ήχος

Ένα κύμα ήχου είναι οι αλληπάλληλες αυξομειώσεις της πίεσης του αέρα σε μια καθορισμένη απόσταση ή χρόνο [1].



Σχήμα 1.0.7: Ένα ηχητικό κύμα αποτελείται από συμπίεσεις και αραιώσεις

Αναπαραστάσεις Ήχου

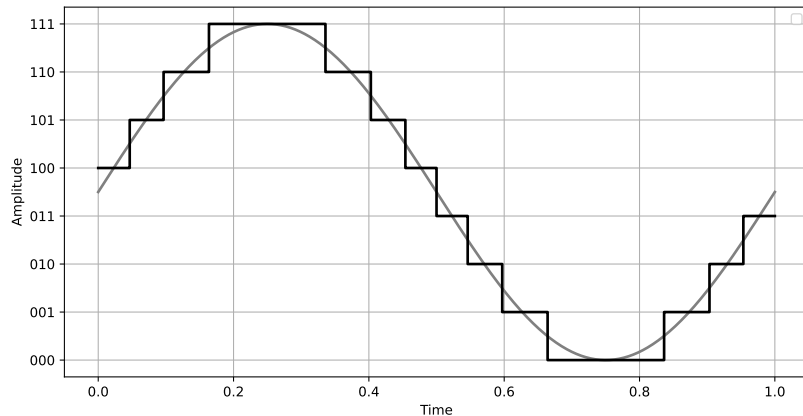
Στην πιο ακατέργαστη μορφή του, ο ήχος αποθηκεύεται ως μια κυματομορφή στον υπολογιστή, σε ψηφιακή μορφή. Η διαδικασία μετατροπής ενός αναλογικού σήματος στη ψηφιακή μορφή του ονομάζεται μετατροπή Αναλογική-σε-Ψηφιακή [33]. Το αναλογικό σήμα $x_a(t)$ αρχικά δειγματοληπτείται σε τακτά χρονικά διαστήματα, πολλαπλάσια της περιόδου δειγματοληψίας T_n , παράγοντας έτσι ένα σήμα συνεχούς χρόνου $x(n)$.

$$x(n) = x_a(nT_s)$$

Μετάπειτα, ο κβαντοποιητής αντιστοιχεί τα συνεχή πλάτη σε ένα διακριτό σύνολο τιμών και το σήμα μετατρέπεται σε ένα ψηφιακό πίνακα στον υπολογιστή.

$$\hat{x}(n) = Q[x(n)]$$

Ο ρυθμός δειγματοληψίας εκφράζει τον αριθμό των δειγμάτων που λαμβάνονται ανά δευτερόλεπτο και μετριέται σε Hz. Για ένα σήμα με ρυθμό δειγματοληψίας sr , η μέγιστη συχνότητα που μπορεί να ανακατασκευαστεί αξιόπιστα είναι $f_N = \frac{sr}{2}$, και ονομάζεται συχνότητα Nyquist.



Σχήμα 1.0.8: Μετατροπή Αναλογικού Σήματος σε Ψηφιακό

Μια αναπαράσταση στο χώρο χρόνου-συχνότητας είναι ένας δισδιάστατος πίνακας που αναπαριστά το περιεχόμενο των συχνοτήτων του σήματος κατά το χρόνο. Οι πιο σημαντικές τέτοιες αναπαραστάσεις είναι ο μετασχηματισμός Φουριέρ βραχύ-χρόνου (STFT), το λογαριθμικό φασματογράφημα (Log Spectrogram) και το φασματογράφημα σε κλίμακα Mel (Mel Spectrogram).

Παραδοσιακές Μέθοδοι

Πριν την άφιξη της βαθιάς μάθησης, για το διαχωρισμό ήχου αξιοποιούνταν κλασικοί αλγόριθμοι μηχανικής μάθησης και επεξεργασίας ψηφιακών σημάτων, κυρίως για το πρόβλημα του τυφλού διαχωρισμού. Στο πρόβλημα του τυφλού διαχωρισμού, όπως προτάθηκε από [34], ένα σύνολο πηγών $s_1(t), \dots, s_N(t)$ αναμειγνύονται γραμμικά από έναν πίνακα A . Ο όρος 'τυφλός' προέρχεται από το γεγονός ότι δεν γνωρίζουμε τίποτα για τις πηγές ή τη διαδικασία ανάμειξης, το μόνο που γνωρίζουμε είναι οι N υπερθέσεις αυτών $x_1(t), \dots, x_N(t)$.

Ένας τρόπος να κατηγοριοποιήσουμε τις μεθόδους αυτές είναι με βάση τον αριθμό των αισθητήρων που χρησιμοποιούνται για την σύλληψη των πηγών ήχου. Έτσι, έχουμε την υπερ-καθορισμένη περίπτωση όπου ο αριθμός των αισθητήρων είναι μεγαλύτερος ή ίσος με τον αριθμό των πηγών, και την υπο-καθορισμένη περίπτωση όπου ο αριθμός των αισθητήρων είναι μικρότερος από τον αριθμό των πηγών.

Στη πρώτη περίπτωση, η πιο διαδεδομένη μέθοδος για το πρόβλημα διαχωρισμού ήχου αποτελεί η ανάλυση ανεξαρτήτων συνιστωσών (ICA), [5], όπου δοσμένου σήματος $\mathbf{x} = \mathbf{A}\mathbf{s}$ με $\mathbf{x} = [x_1 x_2 \dots x_n]^T$ ένα n -διάστατο διάνυσμα, $\mathbf{s} = [s_1 s_2 \dots s_p]^T$ ένα τυχαίο διάνυσμα με p ανεξάρτητα σήματα πηγών και $\mathbf{A} = [a_1 a_2 \dots a_p]^T$ ένας $n \times p$ άγνωστος αντιστρέψιμος πίνακας με γραμμικά ανεξάρτητες στήλες, η μέθοδος ICA στοχεύει στην εύρεση ενός πίνακα \mathbf{W} τέτοιου ώστε $\mathbf{u} = \mathbf{W}\mathbf{x}$.

Για την υπο-καθορισμένη περίπτωση, οι μέθοδοι που χρησιμοποιούνταν κυρίως ήταν η ανάλυση ανεξαρτήτου υποχώρου ISA [10], η παραγοντοποίηση μη-μηδενικού πίνακα NMF [48], όπου αποσυνθέτει έναν γνωστό μη-μηδενικό πίνακα \mathbf{V} στο γινόμενο δύο μη-μηδενικών πινάκων \mathbf{W} και \mathbf{H} έτσι ώστε

$$\mathbf{V} = \mathbf{W}\mathbf{H}$$

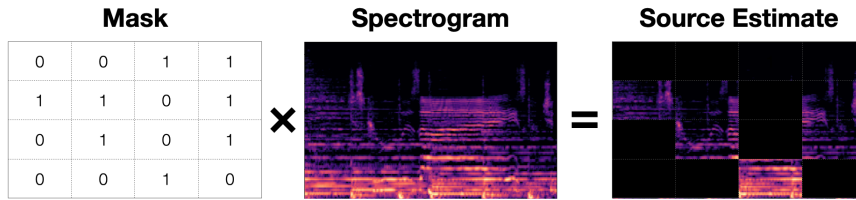
,και τα συστήματα CASA, που χρησιμοποιεί υπολογιστικά μοντέλα με αρχές ακουστικών ομαδοποιήσεων.

Μέθοδοι Βαθιάς Μάθησης

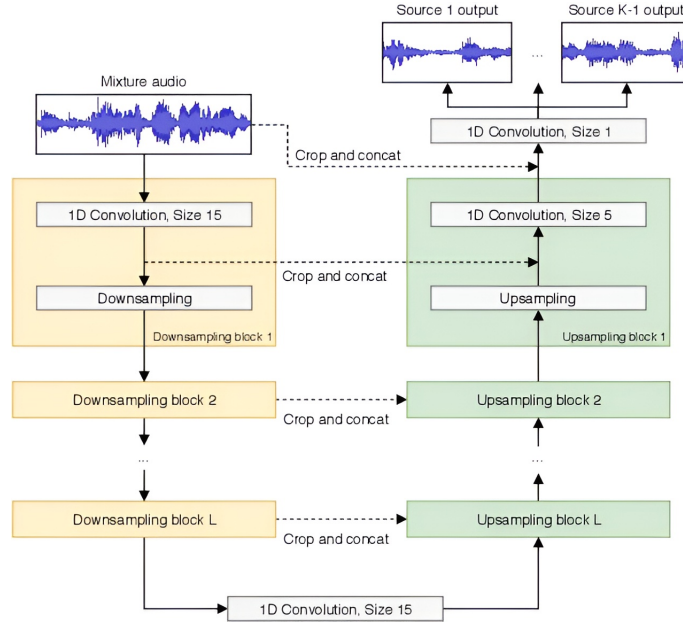
Οι μέθοδοι βαθιάς μάθησης χωρίζονται κυρίως στις μεθόδους που λειτουργούν στο χώρο του χρόνου-συχνότητας και στις μεθόδους που λειτουργούν απευθείας στη χρονική κυματομορφή και δεν χρειάζονται κάποια ενδιάμεση αναπαράσταση.

Οι αλγόριθμοι που λειτουργούν στο χώρο χρόνου-συχνότητας χρησιμοποιούν μάσκες. Μια μάσκα, $\hat{M}_i \in [0.0, 1.0]^{T \times F}$, που αναπαριστά την i^{th} πηγή S_i , είναι ένας πίνακας που περιέχει τιμές και πολλαπλασιάζεται κατά στοιχείο με το φασματογράφημα του μείγματος $|Y|$ για την απόκτηση μιας εκτίμησης $S_i = \hat{M}_i \odot |Y|$. Κάθε τιμή, καθορίζει το ποσοστό της ενέργειας του αρχικού μείγματος που συμβάλλει μια πηγή. Επομένως, σκοπός είναι η εκτίμηση μιας μάσκας που θα προσεγγίζει καλύτερα τη πηγή-στόχο.

Στο χώρο του χρόνου, το πλεονέκτημα είναι πως διατηρείται η πληροφορία για τη φάση του ήχου, που συχνά χάνεται στις προηγούμενες μεθόδους. Διάφοροι τρόποι έχουν προταθεί για το πρόβλημα του διαχωρισμού ήχου οι οποίοι λειτουργούν στο χώρο του χρόνου, ένας από αυτούς είναι οι μέθοδοι που βασίζονται στην αρχιτεκτονική WaveNet [73], όπως η δουλειά των [69].



Σχήμα 1.0.9: Παράδειγμα Δυαδικής Μάσκας απο [52]



Σχήμα 1.0.10: Αρχιτεκτονική Wave-U-Net απο [69]

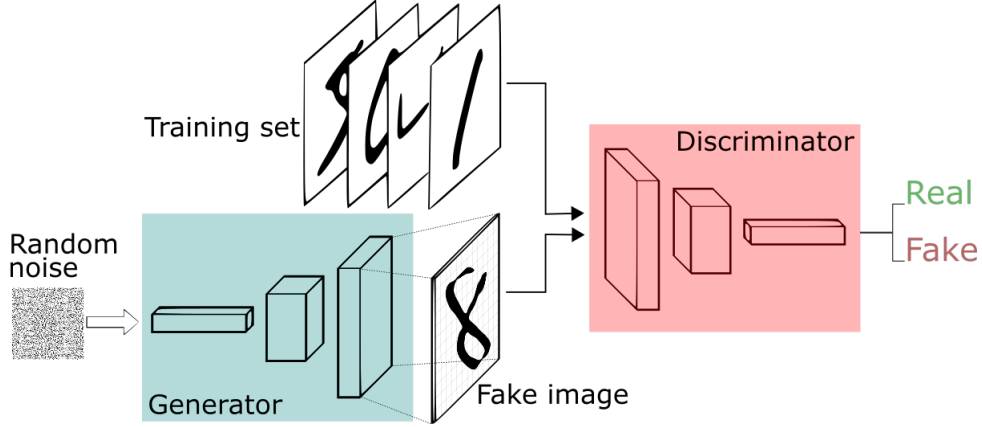
Ο κωδικοποιητής υπο-δειγματοληπτει τη κυματομορφή ενώ λαμβάνει τα εξέχοντα ιεραρχικά χαρακτηριστικά και ο αποκωδικοποιητής αυξάνει την ανάλυση για να ανακατασκευάσει τις διαχωρισμένες πηγές. Άλλες σημαντικές μέθοδοι είναι το TasNet [50], αλλά και η βελτίωσή του, Conv-TasNet [51] που αξιοποιεί ένα χρονικό συνελικτικό δίκτυο TCN [46]. Στο τομέα του διαχωρισμού μουσικής, μια εξέχουσα δουλειά είναι το Demucs [18], που ενσωματώνει LSTMs στο λαιμό (bottleneck) του U-Net, καταφέροντας έτσι να συλλάβει μακροπρόθεσμες χρονικές εξαρτήσεις στο μουσικό σήμα.

Μοντέλα Διάχυσης

Παραγωγικά Μοντέλα

Σκοπός της παραγωγικής μοντελοποίησης είναι η ανάπτυξη στατιστικών μοντέλων που όχι μόνο αναγνωρίζουν μοτίβα στα δεδομένα, αλλά έχουν τη δυνατότητα να μάθουν την υποβόσκουσα κατανομή δεδομένων και να παράξουν νέα δείγματα με βάση τη κατανομή αυτή. Κάποια απο τα βασικά παραγωγικά μοντέλα είναι ο μεταβλητός αυτο-κωδικοποιητής (variational autoencoder) [44], που προβάλλει τα δεδομένα εισόδου σε ένα χώρο χαμηλής διαστατικότητας και μαθαίνει να ανακατασκευάζει τη λανθάνουσα αναπαράσταση ελαχιστοποιώντας μια συνάρτηση σφάλματος, και τα παραγω-

γικά ανταγωνιστικά δίκτυα GANs [31]. Τα μοντέλα αυτά αποτελούνται από ένα δίκτυο γεννήτρια (generator), που μαθαίνει να παράγει αληθοφανή δείγματα από καθαρό θόρυβο, και από ένα δίκτυο διαχωρισμού, που εκπαιδεύεται ώστε να διαχωρίζει αληθινά από κάλπικα δείγματα. Τα δύο αυτά δίκτυα εκπαιδεύονται με έναν ανταγωνιστικό τρόπο, προσπαθώντας το ένα να ξεγελάσει το άλλο.



Σχήμα 1.0.11: Διαδικασία GAN από [62]

Διαδικασία Διάχυσης

Τα μοντέλα διάχυσης πρωτοεμφανίστηκαν το 2015 όταν ο Sohl-Dickstein [63], εμπνευσμένος από τη θερμοδυναμική, καινοτόμησε τη διαδικασία διάχυσης για τη παραγωγή νέων δειγμάτων. Τα μοντέλα διάχυσης έχουν τρεις κύριες μορφές: πιθανοτικά μοντέλα διάχυσης αποθρομβοποίησης (DDPMs), παραγωγικά μοντέλα βασισμένα στο σκορ (SGMs) και τις στοχαστικές διαφορικές εξισώσεις (Score SDEs).

Η κύρια ιδέα της μεθόδου, είναι η προοδευτική διαταραχή της κατανομής των δεδομένων μέσω μιας επαναληπτικής διαδικασίας και η εκμάθηση της αντιστροφής της διαδικασίας, αποκτώντας την αρχική κατανομή. Αυτό έχει ως αποτέλεσμα ένα παραγωγικό μοντέλο που είναι ικανό να δημιουργεί νέα δείγματα, παρόμοια με τα δεδομένα εκπαίδευσης. Το πλαίσιο αποτελείται από δύο διαδικασίες που μοντελοποιούνται με δύο Μαρκοβιανές αλυσίδες. Στην εμπρόσθια διαδικασία, ξεκινώντας από την αρχική κατανομή των δεδομένων $q(\mathbf{x}^{(0)})$, εισάγεται θόρυβος σε σταδιακά βήματα με βάση ένα προκαθορισμένο πρόγραμμα, έως ότου η δομή έχει πλήρως εξαλειφθεί.

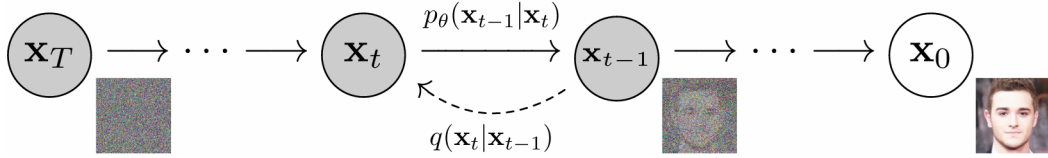
$$q(\mathbf{x}^{(0...T)}) = q(\mathbf{x}^{(0)}) \prod_{t=1}^T q(\mathbf{x}^{(t)} | \mathbf{x}^{(t-1)})$$

Η αντίστροφη διαδικασία, περιλαμβάνει την αντιστροφή της εμπρόσθιας διαδικασίας μαθαίνοντας πυρήνες μετάβασης παραμετροποιημένους από ένα νευρωνικό δίκτυο.

$$p(\mathbf{x}^{(T)}) = \pi(\mathbf{x}^{(T)})$$

$$p(\mathbf{x}^{(0...T)}) = p(\mathbf{x}^{(0)}) \prod_{t=1}^T p(\mathbf{x}^{(t)} | \mathbf{x}^{(t-1)})$$

Κατά την εκπαίδευση, η μέση τιμή και η συνδιακύμανση του Γκαουσιανού πυρήνα διάχυσης χρειάζονται να εκτιμηθούν.



Σχήμα 1.0.12: Η εμπρόσθια και η αντίστροφη διαδικασία [35]

Στα χρόνια που ακολούθησαν, πολλές ερευνητικές δουλειές προσέφεραν βελτιώσεις στα μοντέλα διάχυσης. Απο τις βελτιώσεις αυτές ξεχωρίζουν η απλοποίηση του μεταβλητού ορίου [35], η επιτάχυνση της διαδικασίας δειγματοληψίας με την εισαγωγή του μοντέλου Denoising Diffusion Implicit Model [64] και η εισαγωγή του συνημιτονικού προγράμματος θορύβου που έχει ως αποτέλεσμα μια πιο βαθμιαία καταστροφή της κατανομής των δεδομένων [54].



Σχήμα 1.0.13: Γραμμικό(πάνω) και Συνημιτονικό Πρόγραμμα (κάτω) απο [54]

Σχετική Βιβλιογραφία

Γλωσσικά Μοντέλα

Ένα γλωσσικό μοντέλο είναι ένα υπολογιστικό μοντέλο, ικανό να συλλαμβάνει τη δομή και τις στατιστικές ιδιότητες του ανθρώπινου κειμένου, επιτρέποντάς του με αυτό το τρόπο να παράγει πρωτότυπα δείγματα κειμένου. Συγκεκριμένα, ένα γλωσσικό μοντέλο εκπαιδεύεται σε ένα μεγάλο αριθμό κειμενικών παραδειγμάτων, έτσι ώστε να εκτελεί μια κατανομή πιθανότητας σε ένα σύνολο προκαθορισμένων λέξεων που ονομάζεται λεξιλόγιο (vocabulary). Επομένως, η βασική λειτουργία ενός γλωσσικού μοντέλου είναι η εκτίμηση της πιθανότητας ενός token ή ακολουθία απο tokens να εμφανιστούν σε μια ακολουθία απο tokens.

$$P(w_1, w_2, \dots, w_n) = P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_1, w_2) \dots P(w_n|w_1, w_2, \dots, w_{n-1})$$

Μεγάλα Γλωσσικά Μοντέλα

Η εγκαθίδρυση της αρχιτεκτονικής των μετασχηματιστών ως βασική μονάδα, μεταμόρφωσε το πεδίο των γλωσσικών μοντέλων. Η παραλληλοποίησή τους επέτρεψε την ανάλυση μεγαλύτερης εισόδου και το εκτενές δεκτικό τους πεδίο διευκόλυνε την παραγωγή κειμένου πιο κοντά στα ανθρώπινα δεδομένα. Πρόσφατα, τα Μεγάλα Γλωσσικά Μοντέλα (LLM), έχουν έρθει στην επιφάνεια ως η νέα τάση στο τομέα της επεξεργασίας φυσικής γλώσσας. Τα μοντέλα αυτά, τα οποία βασίζονται στην

αρχιτεκτονική των μετασχηματιστών, εκπαιδεύονται σε ένα ογκώδες σύνολο απο δεδομένα κειμένου και έχουν την ικανότητα να κατανοούν τις σημασιολογικές σχέσεις της ανθρώπινης γλώσσας.

Γενικά, τα γλωσσικά μοντέλα μπορούν να ομαδοποιηθούν σε τρεις κατηγορίες:

- **Αυτο-Παλινδρομικά** (Auto-regressive) μοντέλα όπως η σειρά των GPT μοντέλων, όπου εκπαιδεύονται μέσω μιας διαδικασίας πρόβλεψης του επόμενου token.
- **Αυτο-Κωδικοποιητές** (Auto-encoding) όπως το μοντέλο Bert, που εκπαιδεύονται με μοντελοποίηση μάσκας.
- **Ακολουθία-σε-Ακολουθία** (sequence-to-sequence) μοντέλα, όπως το T5, που σχεδιάζονται για διεργασίες όπως μετάφραση, περίληψη και παραγωγή κειμένου.

Η επιτυχία των μεγάλων γλωσσικών μοντέλων τα τελευταία χρόνια βασίζεται εν μέρη στη κλιμάκωση του μεγέθους τους, καθώς έχει βρεθεί πως όσο ο αριθμός των παραμέτρων αυξάνεται, τόσο αυξάνεται και η απόδοσή τους. Αν και η τάση αυτή αρχίζει και αμφισβητείται τον τελευταίο καιρό, με τα μοντέλα να αποκτούν μια αδράνεια ως προς τη περαιτέρω βελτίωσή τους, είναι σίγουρο πως η αλματώδης άνοδος του αριθμού των παραμέτρων υπήρξε καταλυτική για τη επιτυχία τους σε διάφορους τομείς. Μια απο αυτές τις επιτυχίες είναι η ικανότητα των μοντέλων αυτών να εμφανίζουν συμπεριφορές για τις οποίες δεν εκπαιδεύτηκαν ρητά. Τέτοιες συμπεριφορές, περιλαμβάνουν την ικανότητα των μοντέλων να μαθαίνουν απο τα συμφραζόμενα (in-context learning) αλλά και να παρουσιάζουν σημάδια εγγενούς συλλογισμού, κάτι που φυσικά αποτελεί έναν απο τους πρωταρχικούς στόχους της τεχνητής νοημοσύνης.

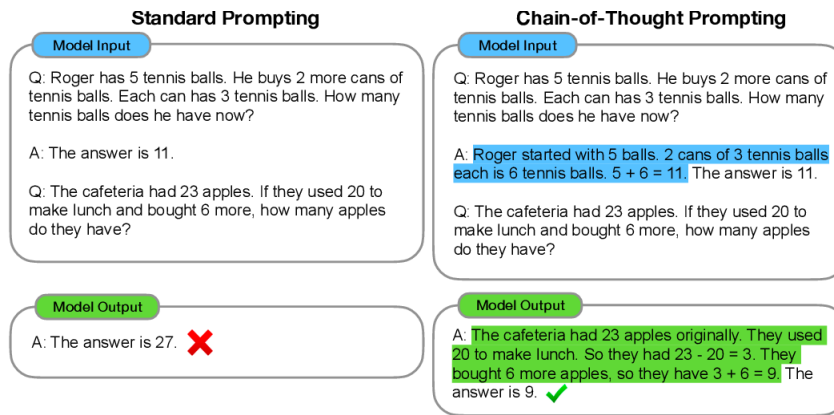
Μέθοδοι Προτροπής

Προτροπή ονομάζουμε την κειμενική οδηγία που δίνουμε ως είσοδο στο γλωσσικό μοντέλο, ώστε να το καθοδηγήσουμε προς μια επιθυμητή έξοδο. Η μηχανική της προτροπής (prompt engineering), αναφέρεται στην σύνθεση οδηγιών φυσικής γλώσσας, με σκοπό την εκμαίευση γνώσης απο τα μοντέλα με έναν δομημένο τρόπο [76]. Στον πυρήνα, η μέθοδος αυτή βασίζεται στην υπόθεση ότι η απόδοση του μεγάλου γλωσσικού μοντέλου, μπορεί να επηρεαστεί απο τον τρόπο με τον οποίο έχει δομηθεί η είσοδος.

Παραδείγματα τέτοιων τεχνικών αποτελούν η μέθοδος μηδενικής-βολής (zero-shot), όπου η πληροφορία δίνεται άμεσα στον μοντέλο, μερικών-βολών (few-shot) [9], όπου στο μοντέλο παρέχονται και μερικά παραδείγματα αναφορικά με την επιθυμητή έξοδο και αλυσίδα-σκέψης (chain-of-thought) [79], όπου η οδηγία προς το μοντέλο δομείται σε μια σειρά ενδιάμεσων βημάτων προτού καταλήξει στο επιθυμητό αποτέλεσμα, έτσι ώστε το μοντέλο να μπορέσει να εμβαθύνει καλύτερα στο δοσμένο πρόβλημα.

Καθοδηγούμενη Παραγωγή Εικόνας

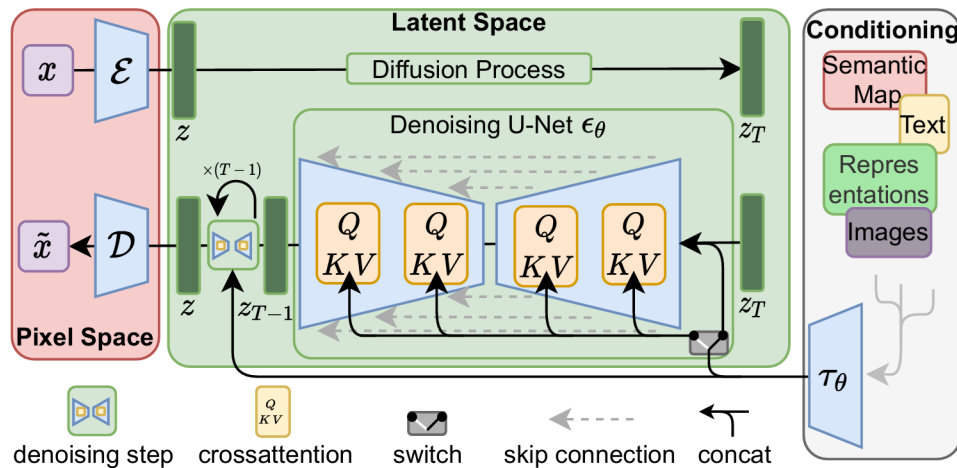
Ένας απο τους νεοφυής και απο τους πιο ενδιαφέρων τομείς της παραγωγικής τεχνητής νοημοσύνης, είναι η σύνθεση εικόνας, και συγκεκριμένα η υπο-κατεύθυνση παραγωγή μιας εικόνας. Στο πέρας των χρόνων, η παραγωγή εικόνων έχει εξελιχθεί απο την απεικόνιση απλών προσώπων στη δημιουργία περίπλοκων και οπτικά εντυπωσιακών γραφικών, που σε ορισμένες περιπτώσεις συναγωνίζονται τον χαρακτήρα της ανθρώπινης δημιουργίας και αυτοτελούς έμπνευσης.



Σχήμα 1.0.14: Βασική (zero-shot) προτροπή και Προτροπή με Chain-of-Thought απο [79]

Κείμενο-προς-Εικόνα

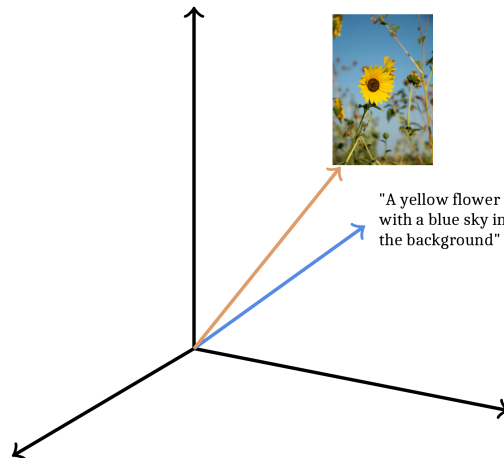
Η αμεσότητα και η απλότητα της κειμενικής πληροφορίας αλλά και η πληθώρα των συνόλων δεδομένων που πραγματεύονται κειμενικές περιγραφές, κατέστη σημαντική για τη χρήση του μέσου αυτού ως τρόπου κατεύθυνσης της παραγωγής εικόνων. Με την εισαγωγή των Λανθανόντων Μοντέλων Διάχυσης (Latent Diffusion Models) [58], τα οποία σε αντίθεση με τα προηγούμενα μοντέλα που λειτουργούσαν στο χώρο των εικονοστοιχείων λειτουργούν σε έναν υποβόσκιον χώρο ενός αυτοκωδικοποιητή, αλλά και με την εισαγωγή των μοντέλων διάχυσης καθοδηγούμενα απο κείμενο, η διάδοση των μοντέλων αυτών σε πρακτικές εφαρμογές παραγωγής εικόνας υπήρξε ραγδαία.



Σχήμα 1.0.15: Αρχιτεκτονική Stable Diffusion απο [58]

Για την καθοδήγηση της παραγωγής εικόνας μέσω κειμενικών προτροπών, η είσοδος αρχικά κωδικοποιείται σε μια σημασιολογικά αντίστοιχη αναπαράσταση που είναι στην ίδια κατεύθυνση με την αναπαράσταση με την αντίστοιχη εικόνα. Αυτό επιτυγχάνεται με τη χρήση του μοντέλου CLIP [56], το οποίο εκπαιδεύεται με αντιθετική μάθηση ώστε να κωδικοποιεί παρόμοια ζεύγη εικόνας-κειμένου σε κοντική απόσταση σε έναν λανθανών χώρο. Τα διανύσματα αυτά, επονομαζόμενα ως embeddings, ενσωματώνονται στο δίκτυο του μοντέλου μέσω επιπέδων διασταυρούμενης προσοχής

(cross-attention).



Σχήμα 1.0.16: Συσχέτιση Ζεύγους Εικόνας-Κειμένου

Ήχος-προς-Εικόνα

Παρόλο που η αμεσότητα του κειμένου βοηθάει αρκετά στην απευθείας κατεύθυνση της εικόνας προς το επιθυμητό αποτέλεσμα, η κειμενική γλώσσα δεν είναι φυσικά συζευγμένη με τον οπτικό μας κόσμο, έτσι έπεται μια απώλεια πληροφοριών κατά τη προσπάθεια σύνδεσης των δύο αυτών μέσων. Από την άλλη πλευρά, ο ήχος είναι άμεσα συνδεδεμένος με την εικόνα, επιτυγχάνοντας μια πιο βέλτιστη απεικόνιση των λεπτομερειών της οπτικής σκηνής που είναι δύσκολο να εκφραστούν μέσω κειμένου. Αρκετοί έχουν προσπαθήσει να γεφυρώσουν τους δύο κόσμους του ήχου και της εικόνας, βασιζόμενοι στην εκμάθηση μιας κοινής αναπαράστασης.

Στο **Sound2Scene** [70], χρησιμοποιώντας οπτικοακουστικά ζεύγη από το VGGSound [12], $\mathcal{D} = \{V_i, A_i\}_{i=1}^N$, όπου V_i είναι ένα στιγμιότυπο βίντεο και A_i είναι ήχος, στόχος είναι να μάθει ο κωδικοποιητής ήχου να εξάγει χαρακτηριστικά ήχου $\mathbf{z}^A = f_A(A)$ που είναι ευθυγραμμισμένα με τα οπτικά χαρακτηριστικά $\mathbf{z}^V = f_V(V)$, προσπαθώντας να ελαχιστοποιηθεί το InfoNCE loss [74]. Μετά την εκπαίδευση, τα διανύσματα ήχου μπορούν να αξιοποιηθούν για τη καθοδήγηση της παραγωγής εικόνας.

Μια άλλη σημαντική δουλειά είναι το **AudioToken** [81], το οποίο αξιοποιεί ένα Stable Diffusion για τη παραγωγή εικόνας. Ένα δίκτυο κωδικοποίησης ήχου αρχικά παράγει ένα διάνυσμα ήχου, το οποίο προβάλλεται στο χώρο του κειμένου και έτσι μπορεί να χρησιμοποιηθεί σε ένα text-guided μοντέλο διάχυσης.

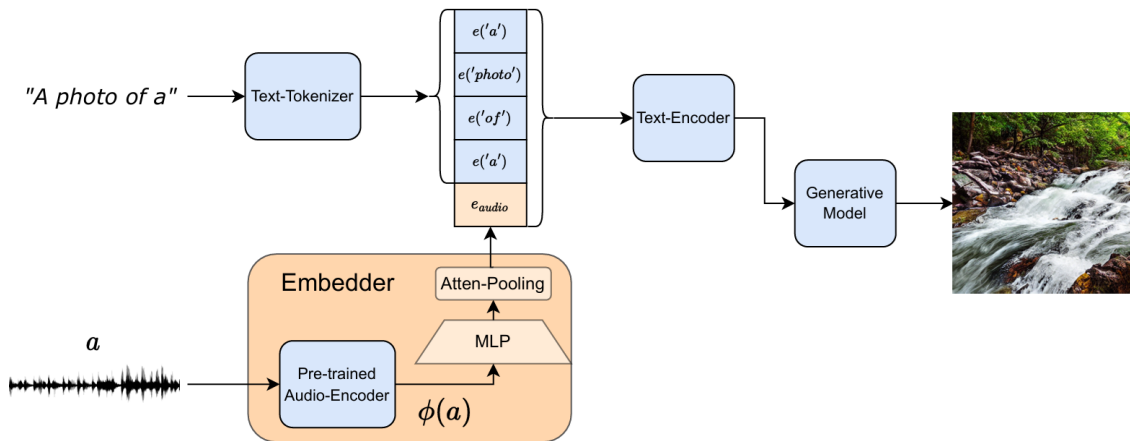
Τέλος, το Sonic Diffusion [7], εκτός από την κοινή αναπαράσταση εικόνας-ήχου, αξιοποιούν και στρώματα διασταυρούμενης προσοχής μέσω των οποίων τα tokens ήχου, μπορούν να ενσωματωθούν στη καθοδήγηση της παραγωγής εικόνας.

Οι προηγούμενες μέθοδοι ωστόσο έχουν κάποιους εγγενείς περιορισμούς, όπως:

- Η παραγωγή εικόνων βασίζεται σε απλούς ήχους που προέρχονται από συγκεκριμένες πηγές και όχι από περίπλοκες σκηνές



Σχήμα 1.0.17: Παραχθείσες Εικόνες Αναμειγνύοντας Δύο Διαφορετικές Κυματομορφές Ήχου απο [70]



Σχήμα 1.0.18: Αρχιτεκτονική AudioToken απο [81]

- Δεν διαθέτουν τη συλλογιστική που χρειάζεται η επεξεργασία μιας σύνθετης ακουστικής σκηνής, με διάφορους, πιθανώς επικαλυπτόμενους ήχους

Γλωσσικά Μοντέλα Ήχου

Η ανάδειξη των μεγάλων γλωσσικών μοντέλων ως ένα ισχυρό εργαλείο στο τομέα της επεξεργασίας φυσικής γλώσσας έφερε στην επιφάνεια ένα εύλογο ερώτημα, αναφορικά με την επέκταση των δυνατοτήτων των μοντέλων αυτών σε επιπλέον μέσα, όπως είναι ο ήχος. Στο παρελθόν, η συνύπαρξη γλώσσας και ήχου εμφανιζόταν σε προβλήματα κλασικής μηχανικής μάθησης, όπως η ταξινόμηση ακουστικής σκηνής, ο εντοπισμός ακουστικών γεγονότων, η αυτόματη περιγραφή ήχου κ.λ.π. Μέχρι πρόσφατα, τα προβλήματα αυτά προσεγγίζοντας με συγκεκριμένες αρχιτεκτονικές νευρωνικών δικτύων. Ωστόσο, καθώς οι εργασίες αυτές συνυπήρχαν στη διασταύρωση κειμένου-ήχου, αρκετοί στράφηκαν στα Μεγάλα Γλωσσικά Μοντέλα και εκμεταλλεύτηκαν τις γλωσσικές τους γνώσεις για να δημιουργηθεί μια νέα κατηγορία μοντέλων, τα Μεγάλα Γλωσσικά Μοντέλα Ήχου, τα οποία ανήκουν στην ευρύτερη κατηγορία των πολυτροπικών μεγάλων γλωσσικών μοντέλων. Τα μεγάλα γλωσσικά μοντέλα ήχου εκπαιδεύονται είτε μέσω αντιθετικής μάθησης, είτε μέσω πρόβλεψης του επόμενου token.

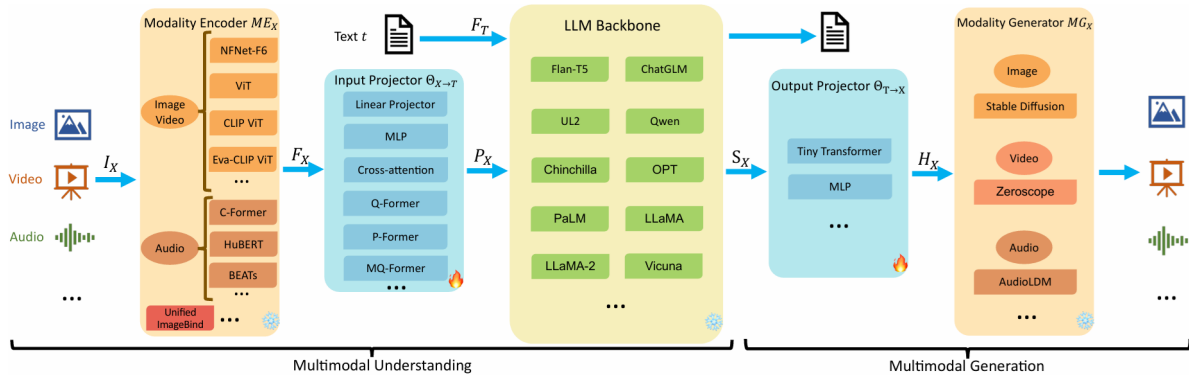
Τα δίκτυα που εκπαιδεύονται με αντιθετική μάθηση, αξιοποιούν έναν κωδικοποιητή ήχου και κειμένου σε μια δομή πύργου, ώστε να μάθουν έναν κοινό πολυτροπικό λανθάνων χώρο. Μετά την εκπαίδευσή τους, μπορούν να χρησιμοποιηθούν για ταξινόμηση ή ανάκτηση. Παράδειγμα τέτοιου μοντέλου είναι το CLAP [24].

Τα μοντέλα που εμπίπτουν στη κατηγορία της μάθησης μέσω πρόβλεψης του επόμενου token, πλαισιώνουν κάθε πρόβλημα ήχου ως πρόβλημα ερωτοαπαντήσεων και παράγουν απαντήσεις ανοιχτού τύπου. Δοσμένου ενός ζεύγους δεδομένων (\mathbf{a}, \mathbf{x}) , όπου \mathbf{a} υποδεικνύει την ακολουθία ήχου και \mathbf{x} την ακολουθία κειμένου, ο στόχος της εκπαίδευσης είναι η μεγιστοποίηση της πιθανότητας:

$$\mathcal{P}_{\theta}(x_t | \mathbf{x}_{<t}, \text{Encoder}_{\phi}(\mathbf{a})),$$

,όπου τα θ και ϕ δηλώνουν τις εκπαιδευσιμες παραμέτρους του LLM και του κωδικοποιητή ήχου [15].

Παρακάτω φαίνεται η γενική αρχιτεκτονική ενός πολυτροπικού μεγάλου γλωσσικού μοντέλου.



Σχήμα 1.0.19: Γενική Αρχιτεκτονική Πολυτροπικών Μεγάλων Γλωσσικών Μοντέλων απο [82]

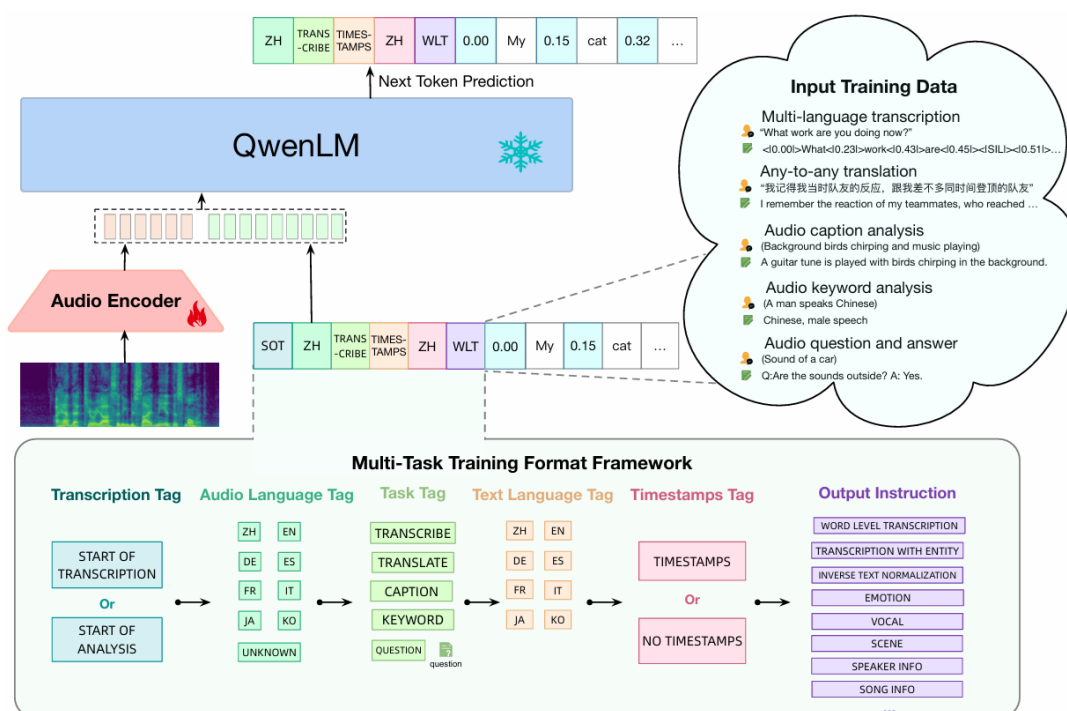
Αποτελούνται απο:

- Κωδικοποιητή Μέσου
- Προβολέα Εισόδου
- Μεγάλο Γλωσσικό Μοντέλο
- Προβολέα Εξόδου
- Γεννήτρια Μέσου

Τα κύρια στοιχεία είναι τα τρία πρώτα. Ο κωδικοποιητής αναλαμβάνει τη κωδικοποίηση των πολυτροπικών δεδομένων σε μια κατάλληλη αναπαράσταση, ο προβολέας εισόδου (Input Projector), προβάλλει τις κωδικοποιημένες αναπαραστάσεις των δεδομένων εισόδου στο χώρο του LLM ώστε να μπορούν να χρησιμοποιηθούν ως είσοδος στο μεγάλο γλωσσικό μοντέλο. Το LLM αποτελεί το νοητικό στάδιο της διαδικασίας, όπου με βάση τις γνώσεις που έχει αποκτήσει κατά την εκπαίδευσή του, έχει την ικανότητα να απαντήσει κατάλληλα στις δοσμένες ερωτήσεις. Σε ορισμένες περιπτώσεις, το LLM υποβάλλεται σε μια περαιτέρω εκπαίδευση των παραμέτρων του fine-tuning,

με διάφορες τεχνικές, ώστε να βελτιωθεί η απόδοσή του στις επεκταμένες απαιτήσεις των ερωτήσεων ήχου. Το LLM μπορεί στο τέλος να παράγει είτε κείμενο, είτε συγκεκριμένα tokens τα οποία θα αξιοποιηθούν για την περαιτέρω παραγωγή πολυτροπικών εξόδων.

Μια απο τις πρώτες προσπάθειες, η οποία εκμεταλλεύτηκε τη μάθηση μεταφοράς (transfer learning) προσεγγίζοντας όλες τις διεργασίες ήχου ως παραγωγή κειμένου, είναι το μοντέλο Pengi [19]. Το μοντέλο αυτό είναι ικανό να δεχθεί ένα ηχητικό κλιπ στην είσοδο, μαζί με μια κειμενική προτροπή και να παράγει κείμενο. Μια μεγάλη επιτυχία στο κλάδο της συλλογιστικής με ήχο, σημείωσε το μοντέλο LTU [29], το οποίο επικεντρώνεται στην βελτίωση της ικανότητας του δικτύου να αντιλαμβάνεται και να κατανοεί τα ηχητικά σήματα σε ένα βαθύτερο επίπεδο. Το Qwen-Audio [15] είναι επίσης απο τα πιο μεγάλα και ισχυρά μοντέλα στον τομέα αυτό. Συμπεριλαμβάνει πάνω απο 30 διεργασίες, καθιστώντας εφικτές τις ικανότητες κατανόησης καθολικών ήχων. Αξιοποιεί το Whisper-large-v2 [57] ως κωδικοποιητή ήχου και το Qwen-7B [3] ως το LLM.



Σχήμα 1.0.20: Αρχιτεκτονική Qwen-Audio [15]

Παρόλο που τα πολυτροπικά αυτά μοντέλα έχουν επιδείξει αρκετές προοπτικές σε διάφορους τομείς που αφορούν τον ήχο, τίθεται ακόμα προς συζήτηση η ικανότητά τους να συλλογίζονται με βάση τον ήχο κάτι που κωλυσιεργεί τη γενική κατανόηση ήχου.

Διαχωρισμός Καθολικού Ήχου

Ο διαχωρισμός καθολικού ήχου αναφέρεται στη διαδικασία κατασκευής μοντέλων, τα οποία δεν περιορίζονται σε μια συγκεκριμένη κατηγορία ήχων, και στοχεύει στο διαχωρισμό ενός αγνώστου αριθμού πηγών, δοσμένου ενός αυθαίρετου μείγματος. Σε αντίθεση με άλλες μορφές διαχωρισμού, όπου το σύνολο δεδομένων είναι περιορισμένο στο διαχωρισμό ομιλίας ή μουσικών οργάνων για παράδειγμα, ο καθολικός ήχος περιλαμβάνει μια πληθώρα ακουστικών πηγών.

Λόγω ανεπάρκειας συνόλων δεδομένων αλλά και λόγω ενός εξαιρετικά απαιτητικού στόχου συνολικά, ο διαχωρισμός καθολικού ήχου δεν έχει γνωρίσει, μέχρι στιγμής, την ανάπτυξη που έχουν δει οι άλλες μορφές διαχωρισμού. Είναι ωστόσο λογικό, αν συλλογιστεί κανείς τον υποκειμενικό χαρακτήρα που έχει το ζήτημα του διαχωρισμού καθολικών ήχων και τη μεταβλητότητά του ανάλογα την εκάστοτε κατάσταση. Ακόμα και ο τομέας της ανάλυσης ακουστικής σκηνής απο τον άνθρωπο εξελίσσεται, προσπαθώντας να καταλάβουμε πως οι άνθρωποι έχουν την ικανότητα να ξεχωρίζουν μια συγκεκριμένη πηγή ήχου δεχόμενοι ένα μεγάλο αριθμό ακουστικών συχνοτήτων.

Απο τις πρώτες προσπάθειες είναι η δουλειά των Kavalero et al. [41], οι οποίοι επιμελήθηκαν ένα νέο σύνολο δεδομένων και το χρησιμοποίησαν για να πειραματιστούν με διάφορες αρχιτεκτονικές διαχωρισμού βασισμένες στη μάσκα. Μια απο τις αρχιτεκτονικές που εξέτασαν ήταν μια βελτιωμένη έκδοση του ConvTasNet (TDCN) [51], αναφερόμενο ως TDCN++. Το μοντέλο ConvTasNet, λειτουργεί στο χώρο του χρόνου, χρησιμοποιώντας ένα πλαίσιο κωδικοποιητή-αποκωδικοποιητή με βαθιά συνελικτικά στρώματα. Διαθέτει τρία στάδια: κωδικοποίηση, διαχωρισμός, αποκωδικοποίηση. Ο κωδικοποιητής αρχικά μετατρέπει την κυματομορφή εισόδου σε μια αναπαράσταση, που χρησιμοποιείται για την εκτίμηση μασκών για κάθε πηγή μέσω ενός χρονικού συνελικτικού δικτύου, και οι κυματομορφές των πηγών ανακατασκευάζονται μέσω του κωδικοποιητή.

Το FUSS baseline model [80], χρησιμοποιεί μια βελτιωμένη έκδοση του μοντέλου αυτού, με μακρύτερες skip-residual συνδέσεις, επίπεδο κανονικοποίησης χαρακτηριστικών αντί για καθολική κανονικοποίηση και μια παράμετρο κλιμάκωσης μετά απο κάθε πυκνό στρώμα, αρχικοποιημένη σε ένα εκθετικά φθίνων βαθμωτό ίσο με 0.9^ℓ , όπου ℓ είναι ο δείκτης του επιπέδου. Το μοντέλο αυτό, είναι ικανό να διαχωρίζει μείγμα ήχων με ένα μεταβλητό αριθμό πηγών. Αρχικά, ο δοσμένος ήχος μετατρέπεται σε ένα φασματογράφημα με παράθυρο $32ms$ και $8ms$ αναπήδηση, και το πλάτος της αναπαράστασης αυτής παρέχεται στο δίκτυο TDCN++, για την εύρεση M μασκών. Οι μάσκες αυτές, πολλαπλασιάζονται με το φασματογράφημα εισόδου και για να παραχθούν οι κυματομορφές των διαχωρισμένων πηγών εφαρμόζεται ένας αντίστροφος STFT. Τέλος, εφαρμόζεται ένα επίπεδο συνοχής μείγματος στις διαχωρισμένες κυματομορφές:

$$\hat{s}_m = s_m + \frac{1}{M}(x - \sum_{m'} s_{m'})$$

,το οποίο προβάλλει τις διαχωρισμένες αυτές πηγές έτσι ώστε να αθροίζονται συνολικά στο αρχικό μείγμα εισόδου. Τα δείγματα του FUSS συνόλου δεδομένων έχουν απο 1 έως 4 πηγές ήχου, οπότε επιλέχθηκε το $M = 4$.

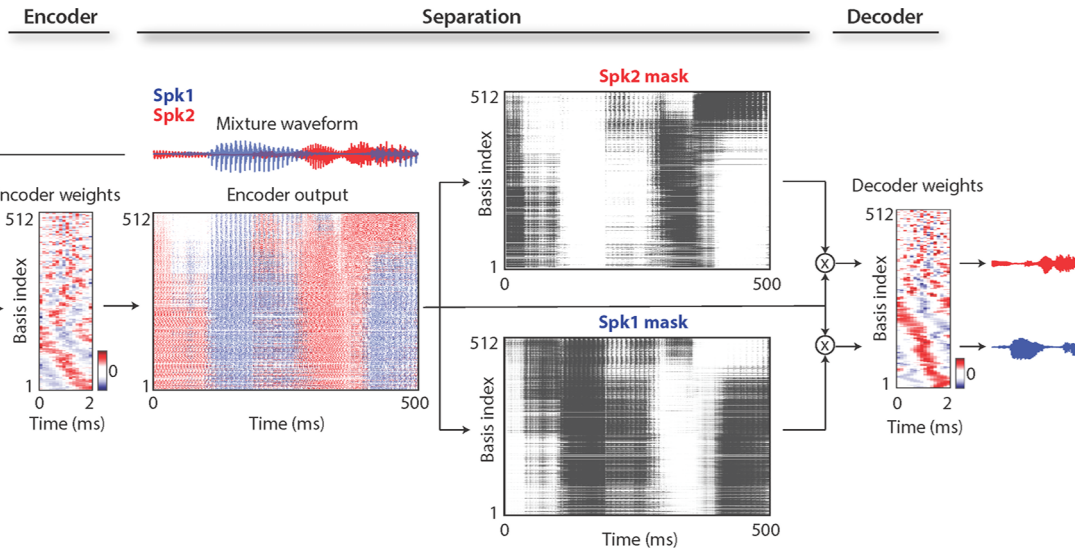
Η συνάρτηση απώλειας που εφαρμόζουν είναι:

$$\mathcal{L}(s, \hat{s}) = \min_{\pi \in \Pi} [\sum_{m_a=1}^{M_a} \mathcal{L}_{SNR}(s_{m_a}, \hat{s}_{\pi(m_a)}) + \sum_{m_0=M_a+1}^M \mathcal{L}_0(x, \hat{s}_{\pi(m_0)})],$$

όπου $\mathcal{L}_{SNR}(y, \hat{y}) = 10 \log_{10}(\|y - \hat{y}\|^2 + \tau \|y\|^2)$ είναι η απώλεια ανα πηγή και $\mathcal{L}_{SNR}(x, \hat{y}) = 10 \log_{10}(\|\hat{y}\|^2 + \tau \|x\|^2)$ είναι η απώλεια για μη-ενεργή διαχωρισμένη πηγή.

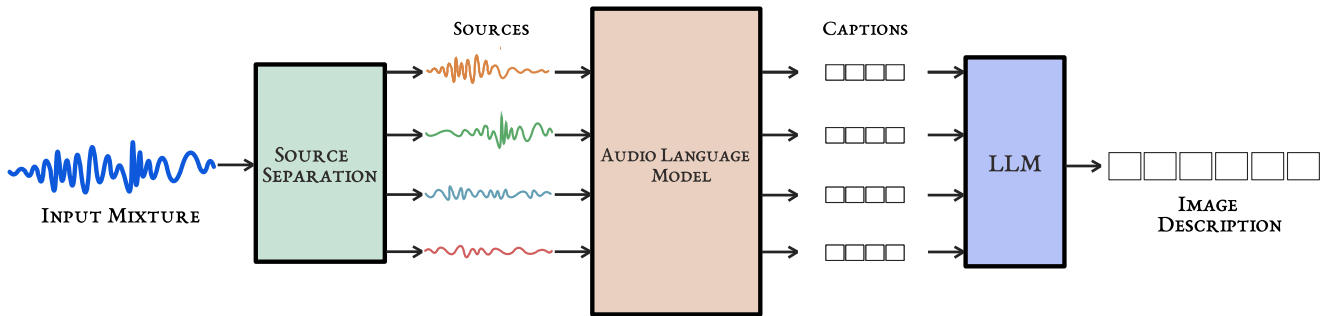
Πρόταση

Η διαδικασία παραγωγής εικόνας βασισμένης αποκλειστικά στην ακουστική πληροφορία δεν έχει εξερευνηθεί εκτενώς και τα μοντέλα που εξετάσαμε προηγουμένως για παραγωγή εικόνας, δεν έχουν



Σχήμα 1.0.21: ConvTasNet Pipeline [51]

την δυνατότητα συλλογισμού, κάτι που πιστεύουμε πως είναι αναγκαίο στην επεξεργασία και κατανόηση μιας σύνθετης ακουστικής σκηνής. Η οπτικοποίηση μιας εικόνας με βάση τον ήχο γενικά, καταλήγει σε τρία βασικά στάδια. Αρχικά, ο ήχος συλλαμβάνεται και επεξεργάζεται ώστε να τεκμηριωθούν οι επιμέρους πηγές του ήχου. Στη συνέχεια, υπάρχει ένα στάδιο συλλογισμού, όπου οι ξεχωριστοί ήχοι και οι σχέσεις μεταξύ τους αντιστοιχίζονται σε οπτικές οντότητες, περιγράφοντας έτσι μια λεπτομερή εικόνα που στη συνέχεια οπτικοποιείται. Κάποιες από τις οπτικές λεπτομέρειες της περιγραφής δεν υπάρχουν στον ήχο, αλλά μπορούν να αποφανθούν από πρότερη γνώση του φυσικού κόσμου. Για παράδειγμα, αν ακούσουμε τον ήχο μιας αγελάδας, είναι εύλογο να σκεφτούμε πως υπάρχει μια φάρμα κοντά ακόμα και αν αυτή η οπτική πληροφορία δεν υπάρχει στον ήχο.



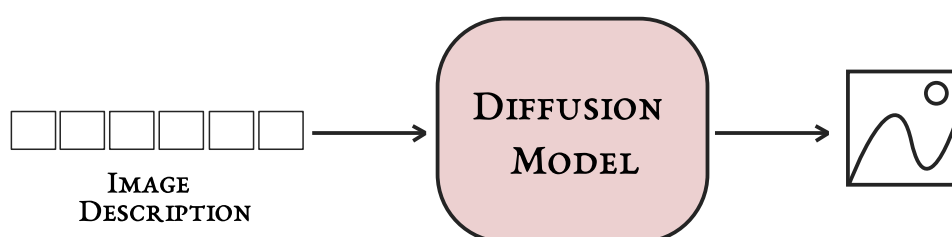
Σχήμα 1.0.22: Προτεινόμενη Μέθοδος

Τα τρέχον μεγάλα γλωσσικά μοντέλα ήχου, δυσκολεύονται να ανταποκριθούν σε περίπλοκες ακουστικές εισόδους με πιθανώς επικαλυπτόμενους ήχους, και δεν εκμεταλλεύονται πλήρως τις γλωσσικές συλλογιστικές δυνατότητες του LLM, ώστε να συμπεράνουν κατάλληλα οπτικές λεπτομέρειες από

τον ήχο.

Για αυτό το σκοπό, αρχικά εξετάζουμε την ικανότητα ενός μεγάλου γλωσσικού μοντέλου ήχου να παράγει μια ουσιαστική και οπτικά αποτελεσματική περιγραφή εικόνας. Στη συνέχεια, προτείνουμε μια μέθοδο, με στόχο τη βελτίωση της περιγραφής μιας οπτικής σκηνής με βάση τον ήχο, με απώτερο σκοπό τη βελτίωση των τελικών εικόνων.

Αρχικά, το δοσμένο ηχητικό κλιπ περνάει από ένα μοντέλο διαχωρισμού ήχου, ώστε να εξαχθούν οι επιμέρους πηγές. Αυτό το στάδιο, απλοποιεί την αναγνώριση των επιμέρους ήχων από το γλωσσικό μοντέλο ήχου, καθιστώντας δυνατή την επεξεργασία πιο σύνθετων εισόδων. Στη συνέχεια, οι κυματομορφές των πηγών δίνονται ως εισόδοι στο γλωσσικό μοντέλο ήχου ώστε να αναγνωριστούν οι ήχοι και να παραχθούν οι λεζάντες που αντιστοιχούν στον εκάστοτε ήχο. Τέλος, αξιοποιείται ένα μεγάλο γλωσσικό μοντέλο, ώστε να συμπτύξει τις επιμέρους λεζάντες και να συλλογιστεί μια εύλογη περιγραφή οπτικής σκηνής όπου συνδυάζονται. Η περιγραφή αυτή, μπορεί ύστερα να δοθεί ως είσοδος σε ένα μοντέλο παραγωγής εικόνας, για να παραχθεί η τελική εικόνα.



Σχήμα 1.0.23: Παραγωγή Εικόνας

Πειραματικά Αποτελέσματα

Τα πειράματα έχουν σκοπό την αξιολόγηση ενός μεγάλου γλωσσικού μοντέλου ήχου να παράγει πιστές περιγραφές εικόνες με βάση τον δοσμένο ήχο, και η σύγκρισή του με τη δική μας μεθοδολογία. Διαιρούμε τα πειράματα σε τρία στάδια:

- Αρχικά, εξετάζουμε διαφορετικές τεχνικές προτροπής χρησιμοποιώντας ως πολυτροπικό μοντέλο το Qwen-Audio, ώστε να βρούμε αν μπορούν να βελτιωθούν τα αποτελέσματα χωρίς περαιτέρω εκπαίδευση και να επιλεγεί μια αντιπροσωπευτική εκδοχή του
- Στη συνέχεια, παράγουμε κειμενικές περιγραφές χρησιμοποιώντας το επιλεγθέν μοντέλο ήχου και τη δική μας μεθοδολογία και συγκρίνουμε τα αποτελέσματα
- Τέλος, διαλέγουμε τα καλύτερα δείγματα, παράγουμε εικόνες και διεξάγουμε ανθρώπινη αξιολόγηση

Προετοιμασία Συνόλου Δεδομένων

Για το σκοπό της εργασίας, χρειαζόμασταν δείγματα σύνθετων ήχων, με πολλαπλές (και πιθανώς επικαλυπτώμενες) πηγές. Για τον λόγο αυτό, στραφήκαμε σε ένα σύνολο δεδομένων που χρησιμοποιείται για διαχωρισμό ήχου, το FUSS [80]. Το Free Universal Sound Separation (FUSS) σύνολο

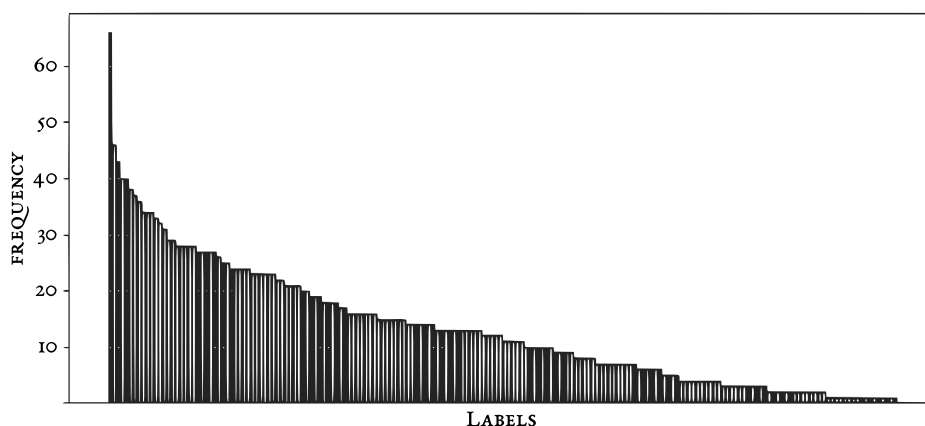
δεδομένων, περιέχει μείγματα ήχων, συγκεντρωμένων απο το Freesound Dataset (FSD50K) [26], το οποίο είναι ένα σύνολο δεδομένων που περιέχει ακουστικά γεγονότα επισημειωμένα με λεζάντες απο την οντολογία του AudioSet [27]. Κάθε δείγμα, περιέχει ένα ακουστικό μείγμα δέκα δευτερολέπτων απο έναν έως τέσσερις προστιθέμενους ήχους, οι οποίοι συνελίσσονται με προσομοιωμένες χρουστικές αποκρίσεις δωματίου. Κάθε χρουστική απόκριση, αντιστοιχεί σε ένα ορθογώνιο δωμάτιο τυχαίου μεγέθους και τυχαία υλικά τοίχου, όπου ένα μικρόφωνο και έως τέσσερις ήχοι, τοποθετούνται σε τυχαίες θέσεις. Με αυτό το τρόπο, τα τελικά δείγματα έχουν αντήχηση.

Χρησιμοποιούμε 1000 δείγματα απο το σύνολο αξιολόγησης του συνόλου δεδομένων, μαζί με τις λεζάντες κάθε επιμέρους πηγής του μείγματος. Κάθε γραμμή περιέχει το όνομα του αρχείου, τον πίνακα αναπαράστασης του ήχου, τις λεζάντες και τον αριθμό των πηγών.

	filename	mixed_audio_array	source_labels	number_of_sources
0	example0000.wav	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	[Rain,Water]	1
1	example0001.wav	[0.0, 0.0, 0.0, -3.0517578e-05, -3.0517578e-05...	[Wind, Glockenspiel,Mallet_percussion,Percussi...	2
2	example0002.wav	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, -3.05...	[Harmonica,Musical_instrument,Music]	1
3	example0003.wav	[0.0, 0.0, -3.0517578e-05, 0.0, -3.0517578e-05...	[Mechanical_fan,Mechanisms, Alarm, Female_spee...	4
4	example0004.wav	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	[Hiss, Wind_chime,Chime,Bell, Tick-tock,Clock,...	4

Σχήμα 1.0.24: Σύνολο Δεδομένων

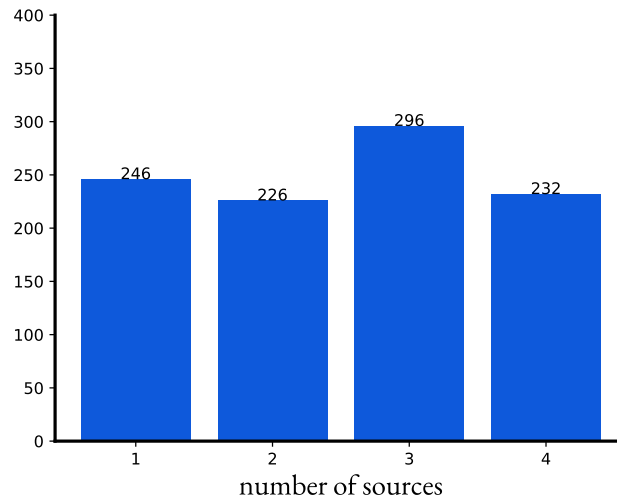
Παρακάτω φαίνεται η κατανομή της συχνότητας κάθε λεζάντας, αλλά και η κατανομή του αριθμού των πηγών.



Σχήμα 1.0.25: Κατανομή Συχνότητας Ακουστικών Λεζάντων

Πρώτο Πείραμα: Σύγκριση Τεχνικών Προτροπής

Χρησιμοποιώντας το Qwen-Audio-Chat, πειραματιστήκαμε με τρεις διαφορετικές τεχνικές προτροπής, συγκεκριμένα zero-shot, chain-of-thought και role. Για το συγκεκριμένο πείραμα, επιλέξαμε



Σχήμα 1.0.26: Κατανομή του Αριθμού των Πηγών

ένα υποσύνολο 100 δειγμάτων του αρχικού συνόλου δεδομένων.

Μέθοδοι Αξιολόγησης

Μια ποσοτική μετρική που αξιοποιείται στο συγκεκριμένο πείραμα είναι το CLAPscore. Χρησιμοποιώντας το μοντέλο CLAP [24], μετράμε τη συνημιτονική ομοιότητα των διανυσμάτων ήχου-κειμένου, η οποία εκφράζει το πόσο όμοια είναι ο ήχος με το παραχθέν κείμενο.

$$\text{cosine similarity} = S_C(A, B) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

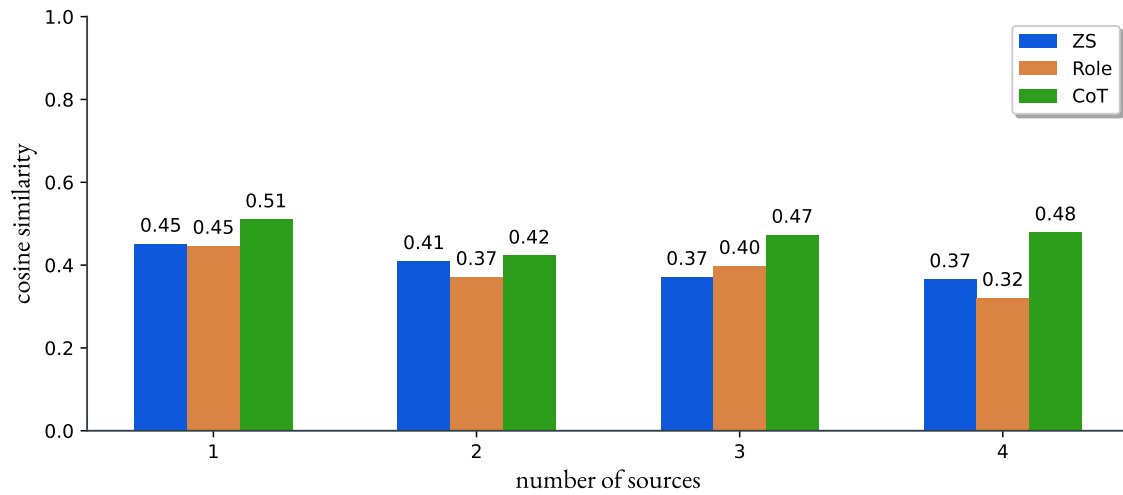
Ωστόσο, η δημιουργία μιας εικόνας, βασισμένης σε μια σύνθετη ακουστική σκηνή, είναι μια περίπλοκη και σε ορισμένες περιπτώσεις υποκειμενική διαδικασία, αποτελούμενη από επιμέρους στάδια. Αρχικά ο ήχος θα πρέπει να ληφθεί και να αναγνωριστεί επιτυχώς, και στη συνέχεια θα πρέπει να υπάρξει κατάλληλη περιγραφή της οπτικής σκηνής από την οποία θα μπορούσε να είχε προκύψει ο εν λόγω ήχος. Για αυτό το λόγο, η συνολική διαδικασία είναι δύσκολο να αξιολογηθεί με ποσοτικές μετρικές, καθιστώντας αναγκαίο ένα ποιοτικό τρόπο αξιολόγησης. Συνεπώς, ακολουθούμε τη τεχνική LLM-as-a-Judge, αξιοποιώντας ένα μεγάλο γλωσσικό μοντέλο για την απόδοση βαθμολογίας στις περιγραφές. Συγκεκριμένα, χρησιμοποιούμε το Mistral 7B-Instruct [38]. Οι μετρικές που χρησιμοποιούνται για την αξιολόγηση της συνολικής διαδικασίας είναι:

- Αντίληψη ήχου, όπου μετράμε τον αριθμό των αντιληφθέντων πηγών ήχου συγκριτικά με τον αληθινό αριθμό πηγών και βρίσκουμε το ποσοστό των δειγμάτων που το μοντέλο βρήκε λιγότερους ήχους από ότι είναι πραγματικά (US), ίσους ήχους (ES) και περισσότερους ήχους (OS).
- Αναγνώριση ήχου, όπου μετράμε την ακρίβεια της αναγνώρισης όλων των πηγών ήχου από το μείγμα εισόδου
- Συνοχή, όπου βαθμολογούμε τη περιγραφή κειμένου με βάση τη συνοχή της με σκορ από 1 έως 10

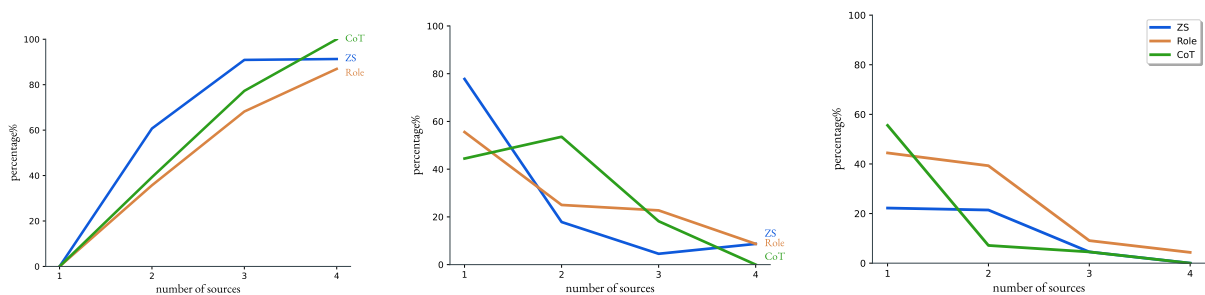
- Οπτική λεπτομέρεια, όπου βαθμολογείται απο 1 έως 10, πόσο περιγραφικό είναι το κείμενο και κατα πόσο αποδίδει οπτικά μια εικόνα
- Φαντασία, όπου αξιολογείται απο το 1 έως το 10, η ικανότητα του μοντέλου να φαντάζεται οπτικές λεπτομέρειες που δεν βρίσκονται στις ακουστικές πληροφορίες αλλά μπορούν να συναχθούν λογικά με βάση τον ήχο.
- Καταλληλότητα, που βαθμολογείται απο το 1 έως το 10 η καταλληλότητα της κειμενικής περιγραφής, ως περιγραφή εικόνας συνολικά.

Αποτελέσματα

Παρακάτω, φαίνονται τα αποτελέσματα για τις μετρικές αξιολόγησης, όπου έχουν παρουσιαστεί συναρτήσει του αριθμού των πηγών του ήχου, ο οποίος ορίζει έναν βαθμό δυσκολίας.



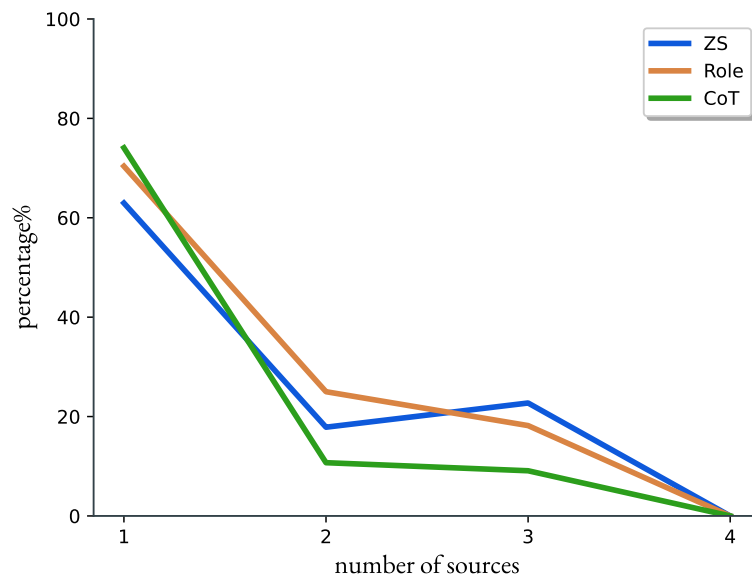
Σχήμα 1.0.27: CLAPscore



Σχήμα 1.0.28: Under Separation (*left*), Equal Separation (*center*), Over Separation (*right*)

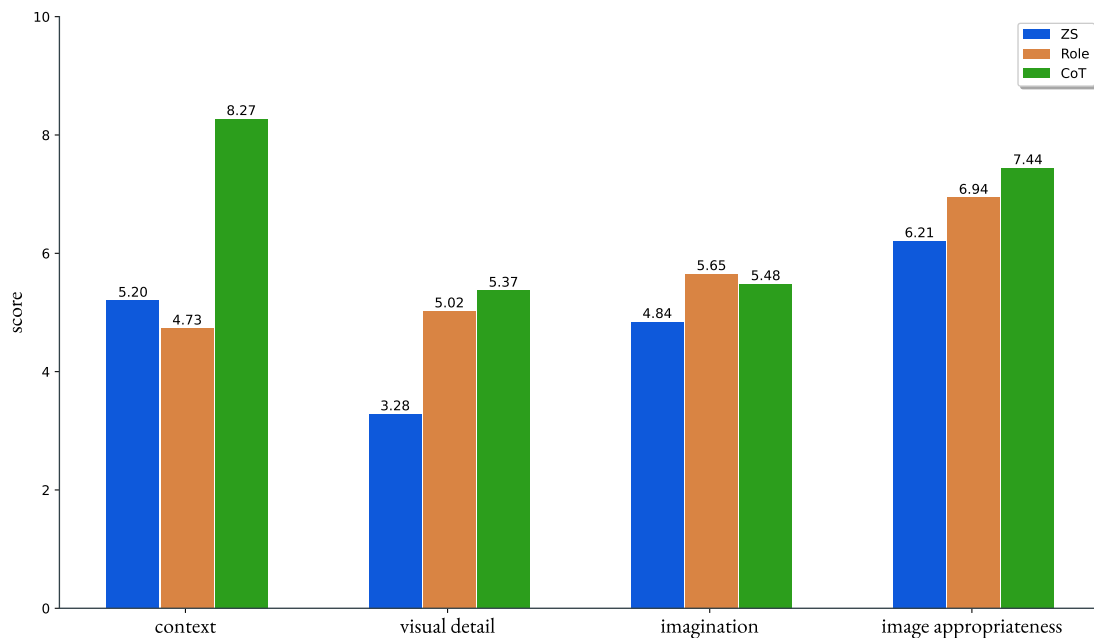
Όπως φαίνεται και απο τα γραφήματα, το μοντέλο δυσκολεύεται να αναγνωρίσει τον ακριβή αριθμό των πηγών ήχου καθώς οι πηγές ήχου αυξάνονται. Το ποσοστό OS υποδεικνύει το πόσο το μοντέλο έχει παραισθήσεις (hallucinations), διότι με βάση τις περιγραφές, βρήκε παραπάνω πηγές

ήχου απο ότι είναι στη πραγματικότητα. Σε γενικές γραμμές, όλες οι τεχνικές προτροπής αποδίδουν το ίδιο, κάτι αναμενόμενο εφόσον η προτροπή κειμένου δεν μπορεί να επιφέρει σημαντικές αλλαγές στο τρόπο αντίληψης του ήχου.



Σχήμα 1.0.29: Ακρίβεια Αναγνώρισης των Ήχων

Όπως και προηγουμένως, το μοντέλο δυσκολεύεται να αναγνωρίσει τις πηγές των ήχων, ειδικά όταν ο αριθμός των πηγών είναι υψηλός.



Σχήμα 1.0.30: Μετρικές Περιγραφής Εικόνας

Συνολικά, αξιολογώντας τις κειμενικές περιγραφές του μοντέλου, βλέπουμε πως η τεχνική Chain-of-Thought έχει το υψηλότερο σκορ ανάμεσα στις υπόλοιπες. Μόνο στη φαντασία, η τεχνική Role παρέχει καλύτερα αποτελέσματα, κάτι ίσως αναμενόμενο εφόσον αναθέτουμε στο μοντέλο το ρόλο του σκηνοθέτη, που πιθανόν να ωθεί το μοντέλο να σκεφτεί περισσότερες λεπτομέρειες σχετικά με τη σκηνή.

Δεύτερο Πείραμα: Σύγκριση με Προτεινόμενη Μέθοδο

Στο δεύτερο μέρος των πειραμάτων, αξιοποιούμε το σύνολο δεδομένων για να συγκρίνουμε το αρχικό μοντέλο με τη προτεινόμενη μέθοδο, χρησιμοποιώντας διαφορετικές επιλογές για κάθε τμήμα του συστήματος. Οι διαφορετικές επιλογές φαίνονται παρακάτω:

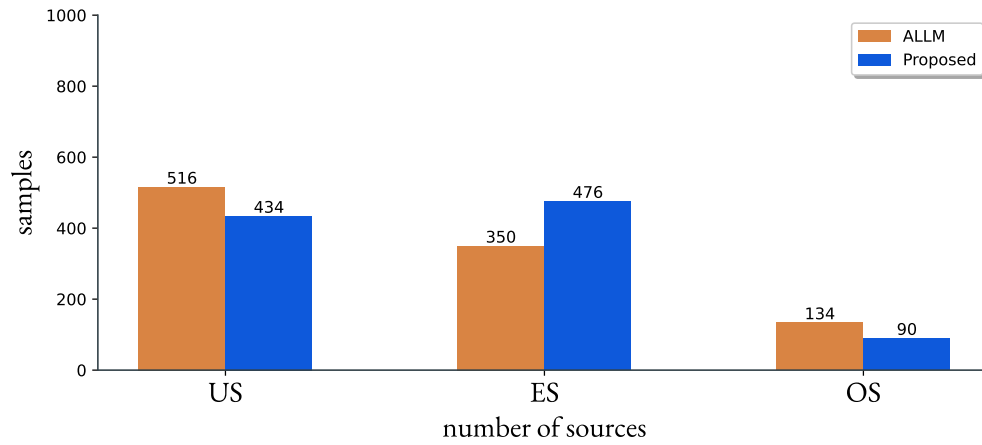
Audio-Language Model	LLM	Prompt	Size
Qwen-Audio	Qwen LLM	zero_shot prompt	15.5B
EnCLAP	Qwen LLM	CoT prompt	7.6B
EnCLAP	Qwen LLM	zero_shot prompt	7.6B
Qwen-Audio	Mistral LLM	zero_shot prompt	15.5B
EnCLAP	Mistral LLM	CoT prompt	7.6B
EnCLAP	Mistral LLM	zero_shot prompt	7.6B
Qwen-Audio	Falcon LLM	zero_shot prompt	15.5B
EnCLAP	Falcon LLM	CoT prompt	7.6B
EnCLAP	Falcon LLM	zero_shot prompt	7.6B

Πίνακας 1.0.1: Επιλογές Προτεινόμενου Συστήματος

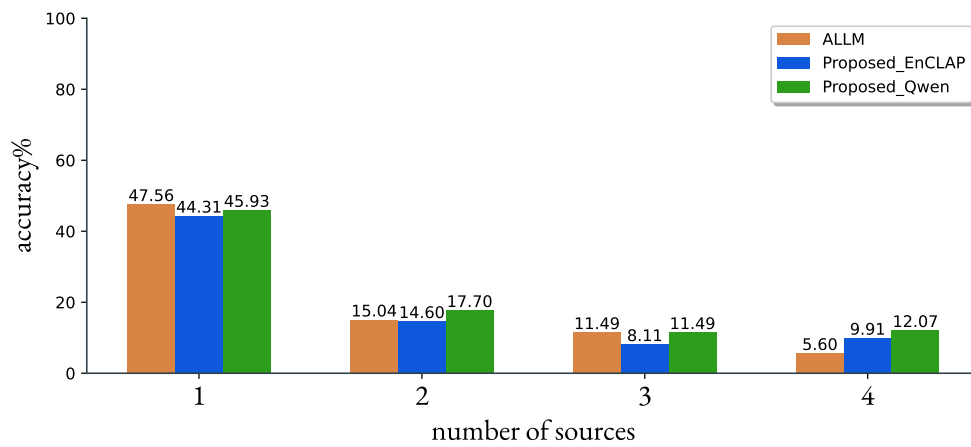
Ως μοντέλο διαχωρισμού ήχου αξιοποιούμε το μοντέλο που παρέχεται από το FUSS dataset, ως μοντέλο γλώσσας-ήχου δοκιμάζουμε το αρχικό αλλά και ένα αρκετά μικρότερο, το EnCLAP-large [43], και ως LLM, πειραματιστήκαμε με τρία διαφορετικά μεγάλα γλωσσικά μοντέλα και με δύο διαφορετικές τεχνικές προτροπής. Επομένως, συνολικά έχουμε 9 διαφορετικά αποτελέσματα. Τα αποτελέσματα για το διαχωρισμό και την αναγνώριση των ήχων φαίνεται παρακάτω:

Στο κομμάτι της αντίληψης και της αναγνώρισης του ήχου, η προτεινόμενη μέθοδος σημειώνει βελτίωση αν και όχι τόσο σημαντική. Όταν ο αριθμός των πηγών ήχου ξεπερνάει τους 3, η προτεινόμενη μέθοδος, και συγκεκριμένα με χρήση του Qwen-Audio, επιτυγχάνει τη σωστή αναγνώριση της εισόδου σε υψηλότερο ποσοστό από το αρχικό μοντέλο. Ωστόσο, σε γενικές γραμμές, τα αποτελέσματα παραμένουν συγκρίσιμα.

Όσον αφορά την αξιολόγηση των περιγραφών εικόνας, έχουμε τα εξής αποτελέσματα:



Σχήμα 1.0.31: Μετρικές Αντίληψης Αριθμού Πηγών Ήχου

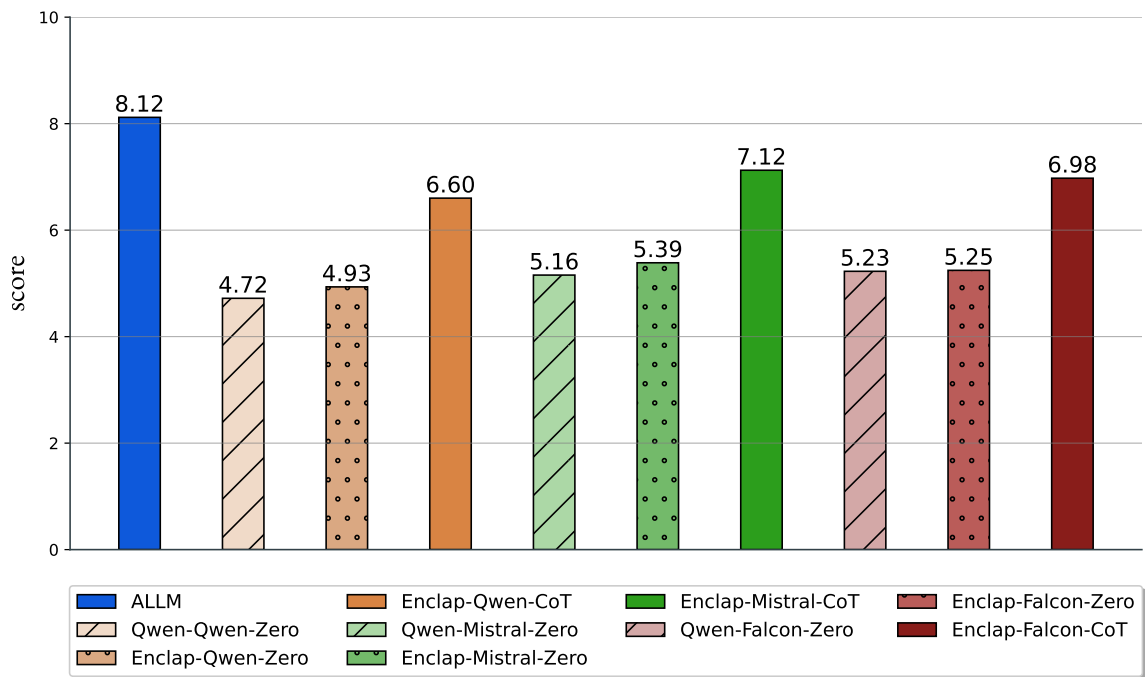


Σχήμα 1.0.32: Ακρίβεια Αναγνώρισης Πηγών Ήχου

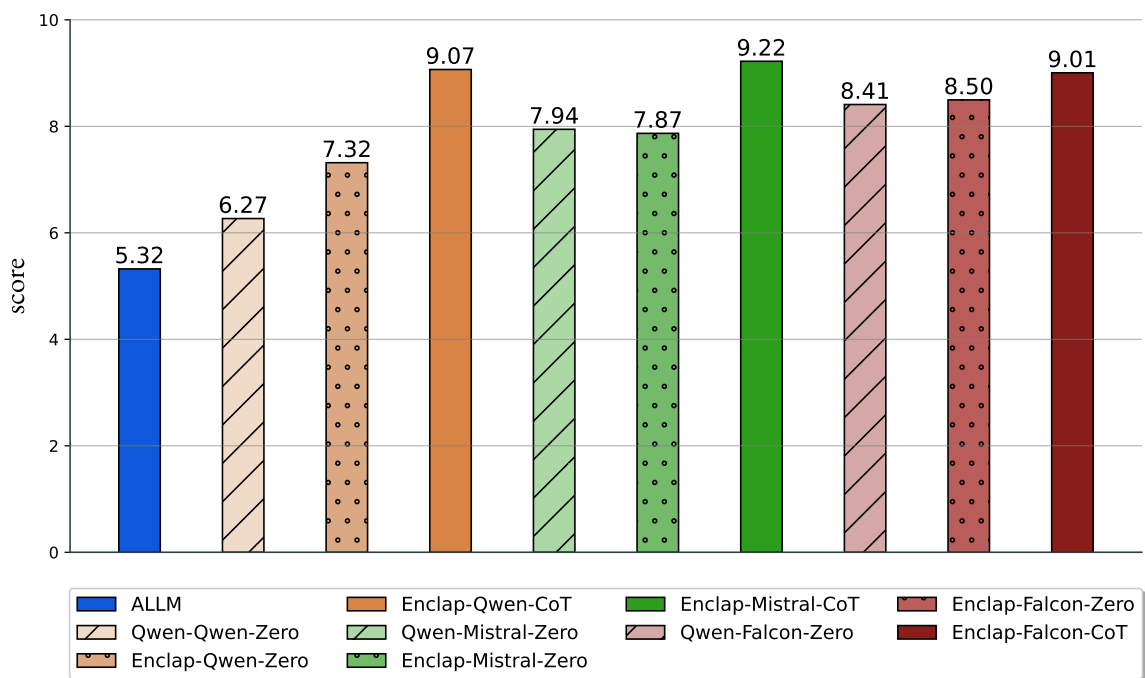
Παρατηρούμε, πως το αρχικό μοντέλο πετυχαίνει υψηλότερη βαθμολογία ως προς τη συνοχή της περιγραφής. Αυτό οφείλεται και στο γεγονός ότι η έξοδος του αρχικού μοντέλου με τη προτροπή Chain-of-Thought, είναι πιο επεξηγηματική του δοσμένου ήχου, χωρίς να εκτείνεται σε οπτικές λεπτομέρειες. Παρόλα αυτά, στις υπόλοιπες μετρικές, το προτεινόμενο σύστημα ξεπερνάει το αρχικό μοντέλο, κυρίως ως προς την οπτική λεπτομέρεια των περιγραφών. Καταλήγουμε, επίσης, στα εξής συμπεράσματα:

- Γενικά, η τεχνική Chain-of-Thought κατορθώνει το υψηλότερο σκορ και στα 3 LLM
- Χρησιμοποιώντας το EnCLAP, πετυχαίνουμε καλύτερα αποτελέσματα
- Ανάμεσα στα τρία LLM, το Mistral παρέχει περιγραφές με περισσότερη συνοχή και περισσότερες οπτικές λεπτομέρειες
- Το Falcon, είναι πιο συνεπές, έχοντας παρόμοια αποτελέσματα με διάφορες ρυθμίσεις

Τρίτο Πείραμα: Παραγωγή Εικόνων από Σκηνές Ταινιών

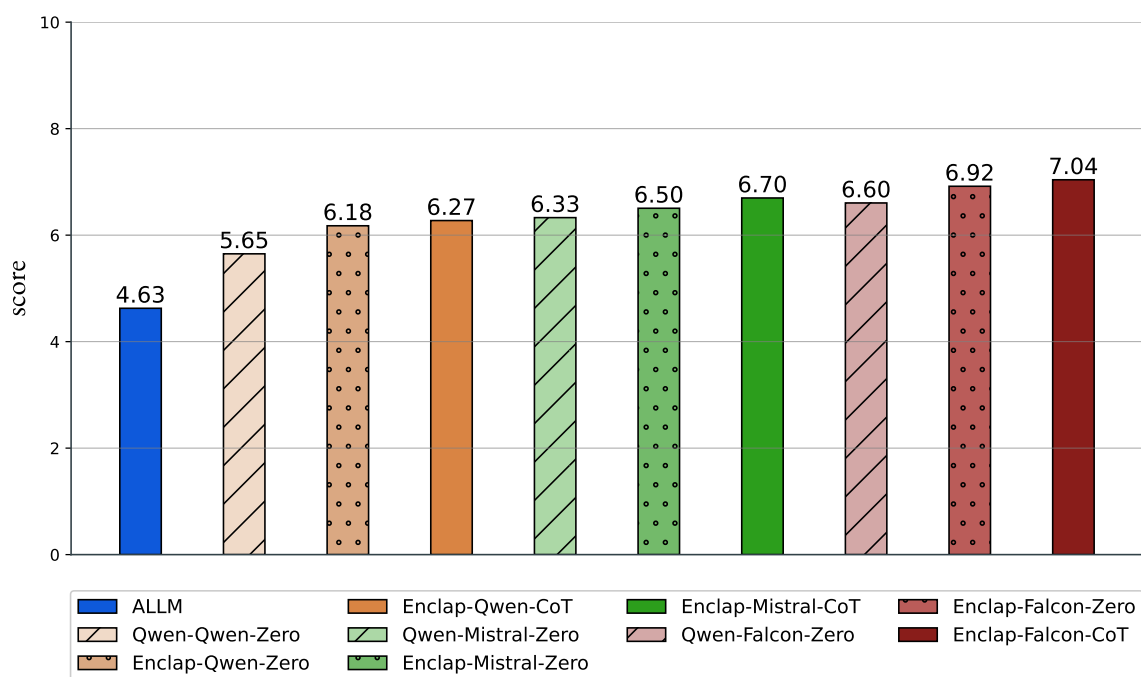


Σχήμα 1.0.33: Μέσο Σκορ Συνοχής για κάθε Μέθοδο

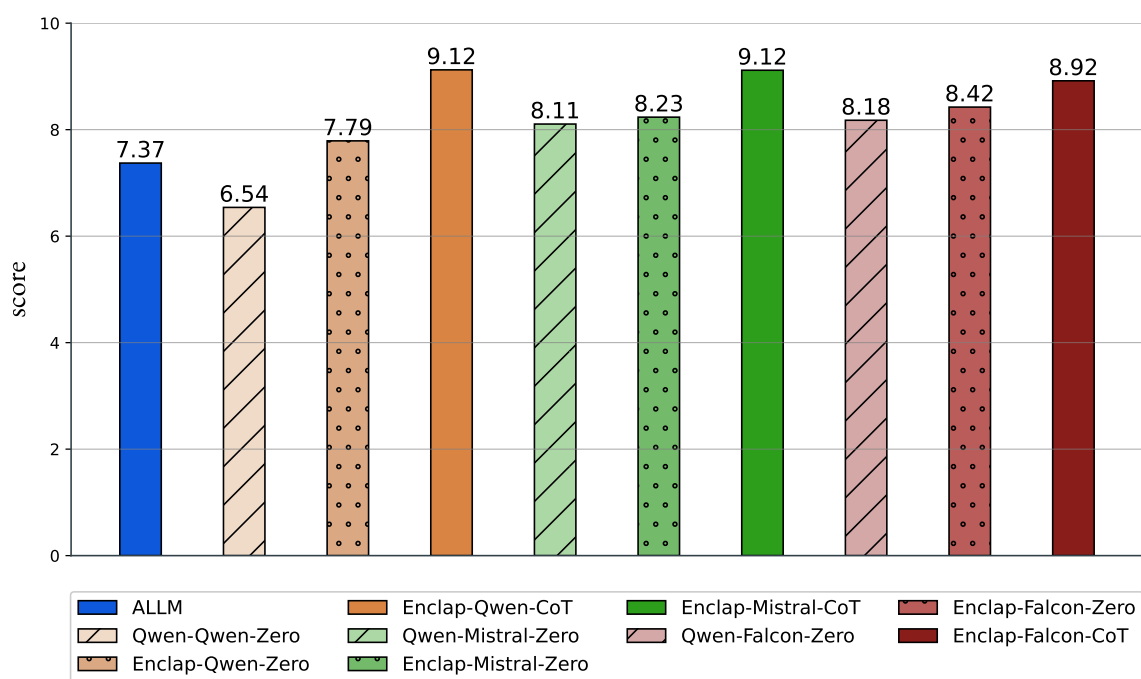


Σχήμα 1.0.34: Μέσο Σκορ Οπτικής Λεπτομέρειας για κάθε Μέθοδο

Στο στάδιο αυτό των πειραμάτων, έχοντας διαλέξει τις δύο καλύτερες διαμορφώσεις του συστήματός μας χρησιμοποιώντας τη zero-shot προτροπή μαζί με το Falcon και τη chain-of-thought προτροπή με το Mistral, συγκρίνουμε τις δύο αυτές διαμορφώσεις με το Qwen-Audio αξιοποιώντας ηχητικά κλιπ απο ταινίες.



Σχήμα 1.0.35: Μέσο Σκορ Φαντασίας για κάθε Μέθοδο



Σχήμα 1.0.36: Μέσο Σκορ Καταλληλότητας Περιγραφής Εικόνας για κάθε Μέθοδο

Σύνολο Δεδομένων

Διαλέξαμε 22 κλιπ ταινιών απο το σύνολο δεδομένων Condensed Movies [4], στοχεύοντας στην επιλογή σκηνών με ένα σύνθετο ακουστικό περιβάλλον. Απο τα κλιπ αυτά, έγινε περικοπή σε μια

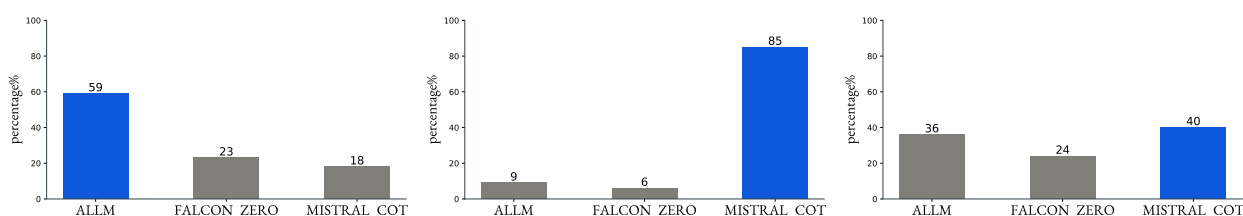
σκηνή 10 δευτερολέπτων, και εξάγαμε τον ήχο και μια αντιπροσωπευτική εικόνα.

Ανθρώπινη Αξιολόγηση

Για τη καλύτερη και πιο ποιοτική αξιολόγηση των κειμενικών περιγραφών αλλά και των παραχθέντων εικόνων, διεξήχθη μια έρευνα αποτελούμενη απο δύο μέρη. Στο πρώτο μέρος, τα άτομα που έλαβαν μέρος, κλήθηκαν να επιλέξουν ανάμεσα στις τρεις περιγραφές των μεθόδων, δοσμένου του αντίστοιχου ήχου, για κάθε μια απο τις επόμενες ερωτήσεις:

1. Ποια περιγραφή ταιριάζει καλύτερα με βάση τον δοσμένο ήχο·
2. Ποια περιγραφή είναι οπτικά πιο λεπτομερής·
3. Ποια από τις τρεις είναι συνολικά καλύτερη·

Στην έρευνα συμμετείχαν συνολικά 37 άτομα.



Σχήμα 1.0.37: Ομοιότητα με Ήχο (αριστερά), Οπτικά Λεπτομερής (κέντρο), Συνολικά Καλύτερη (δεξιά)

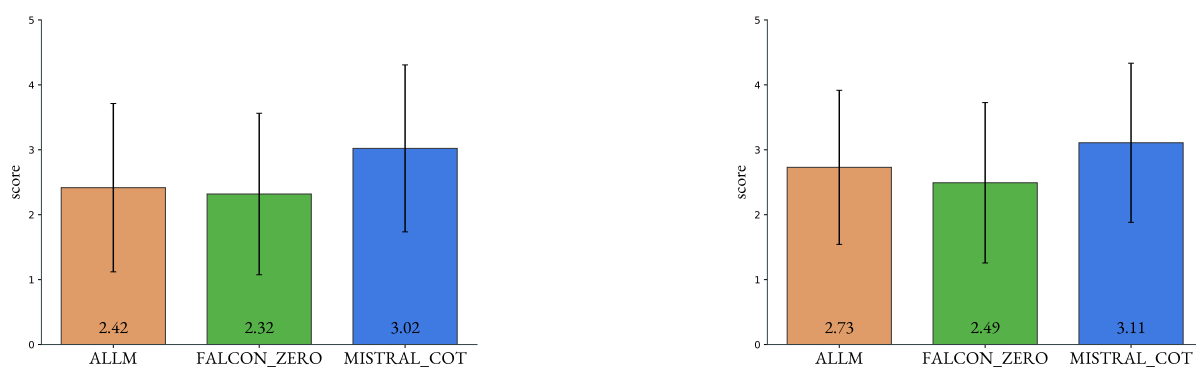
Αναφορικά με την καταλληλότητα της περιγραφής με βάση τον ήχο, η πλειοψηφία των αξιολογητών επέλεξε το αρχικό ALLM μοντέλο. Αυτό εν μέρη υποδηλώνει την απόδοση του μοντέλου, ωστόσο είναι αναγκαίο να σημειωθεί πως η επιλογή αυτή ίσως δεν έγινε αμερόληπτα, διότι οι περιγραφές του μοντέλου αυτού είναι αρκετά επεξηγηματικές και ευθείς, περιγράφοντας άμεσα τους εμπλεκόμενους ήχους, κάτι που πιθανόν να κατεύθυνε ανάλογα και τον μέσο αξιολογητή.

Στις επόμενες ερωτήσεις, και ιδιαίτερα στην περιγραφικότητα και οπτική λεπτομέρεια των περιγραφών, η προτεινόμενη μέθοδος με το Mistral σαφώς υπερέχει έναντι των υπολοίπων και συνολικά αποτελεί μια πιο ποιοτική περιγραφή εικόνας.

Στο δεύτερο μέρος της έρευνας, οι συμμετέχοντες κλήθηκαν να αξιολογήσουν τη ποιότητα των παραχθέντων εικόνων απο τις αντίστοιχες περιγραφές. Συγκεκριμένα, δοσμένου ήχου και τριών εικόνων, μια για κάθε σύστημα, τους ζητήθηκε να βαθμολογήσουν απο το 1(χαμηλό) έως το 5(υψηλό) τις εικόνες με βάση τις εξής ερωτήσεις:

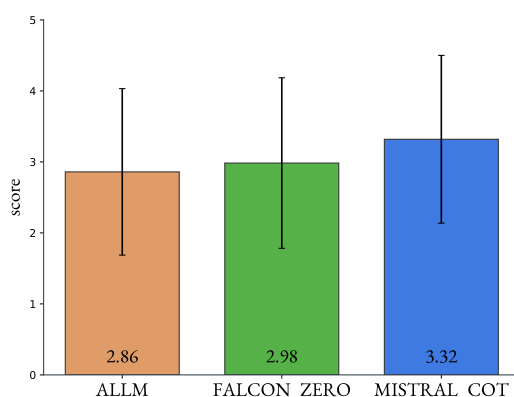
1. Ομοιότητα με τον δοσμένο ήχο
2. Ρεαλιστικότητα
3. Λεπτομέρεια

Όπως φαίνεται και στα γραφήματα παρακάτω, η προτεινόμενη μέθοδος με το Mistral μοντέλο, πετυχαίνει υψηλότερο σκορ σε όλες τις μετρικές. Παρόλα αυτά και οι τρεις αποδίδουν παρόμοια αποτελέσματα.



Σχήμα 1.0.38: Ομοιότητα με Ήχο (αριστερά), Ρεαλιστικότητα (δεξιά)

Να σημειωθεί πως η αξιολόγηση της ομοιότητας της εικόνας με τον ήχο, αλλά και οι άλλες μετρικές, διαθέτουν έναν υποκειμενικό χαρακτήρα, καθώς κάθε συμμετέχων έχει διαφορετική προοπτική και δίνει προσοχή σε διαφορετικά σημεία.



Σχήμα 1.0.39: Λεπτομέρεια

Συνολικά, το σύστημα που προτείνουμε, χρησιμοποιώντας το Mistral LLM με τεχνική Chain-of-Thought έχει τις καλύτερες αξιολογήσεις, συγκριτικά με τις υπόλοιπες μεθόδους.

CLIPscore

Για να μετρήσουμε ποσοτικά τα αποτελέσματα με ένα αντικειμενικό κριτήριο ομοιότητας ανάμεσα στις κειμενικές περιγραφές και τις αληθινές εικόνες της πραγματικής σκηνης, εφαρμόσαμε το CLIPscore. Χρησιμοποιώντας το μοντέλο CLIP εξάγαμε τα διανύσματα κειμένου και εικόνας και υπολογίσαμε τη συνημιτονική ομοιότητα ανάμεσά τους. Τα αποτελέσματα φαίνονται παρακάτω.

Τα χαμηλά νούμερα στη μετρική αυτή, υποδεικνύουν την πολυπλοκότητα της συγκεκριμένης

Method	CLIPscore \uparrow
Qwen-Audio	0.2279
Falcon-Zero_Shot	0.226
Mistral-Chain_of_Thought	0.2047

Πίνακας 1.0.2: CLIPscore

εργασίας και της ανικανότητας ενός αντικειμενικού κριτηρίου να συλλάβει όλα τα ποιοτικά χαρακτηριστικά που την απαρτίζουν. Αυτό αναδεικνύει παραπάνω και την αναγκαιότητα της ποιότητας που προσφέρει η ανθρώπινη αξιολόγηση. Ωστόσο, η σχετικά μικρή διαφορά στη βαθμολογία του αρχικού μοντέλου, μπορεί να εξηγηθεί από την αμεσότητα και σαφήνεια των κειμενικών του περιγραφών.

Συμπεράσματα και Μελλοντικές Κατευθύνσεις

Η εργασία αυτή εξερευνεί μεθόδους για τη παραγωγή εύλογων περιγραφών εικόνας από σύνθετες ακουστικές σκηνές. Πειραματιζόμενοι με διάφορες τεχνικές προτροπής και προτείνοντας ένα νέο πλαίσιο που περιλαμβάνει μοντέλο διαχωρισμού ήχου, γλωσσικό μοντέλο ήχου και μεγάλο γλωσσικό μοντέλο, καταφέραμε να βελτιώσουμε τη ποιότητα των περιγραφών, τις οποίες μετά αξιοποιούμε για τη παραγωγή εικόνων μέσω ενός μοντέλου παραγωγής εικόνας.

Παρόλα αυτά, ένας σημαντικός περιορισμός της συγκεκριμένης μεθοδολογίας είναι η ιεραρχική εξάρτηση των επιμέρους μοντέλων, όπου η αποτυχία ενός επηρεάζει τα επακόλουθα με επιπτώσεις στο τελικό αποτέλεσμα.

Εξετάζοντας τις μελλοντικές κατευθύνσεις, από τα ευρήματα της τρέχουσας μελέτης μπορούν να προκύψουν πολλές υποσχόμενες οδοί.

1. Δοκιμή μεγάλων γλωσσικών μοντέλων με περισσότερες παραμέτρους
2. Η βελτίωση στο διαχωρισμό καθολικών ήχων τίθεται ως καταλυτική στη συνολική απόδοση του συστήματος
3. Η παραγωγή βίντεο έναντι εικόνας, είναι πιο κοντά στον χαρακτήρα της ακουστικής πληροφορίας και καθίσταται άκρως σημαντική σε αρκετές πρακτικές εφαρμογές
4. Η αναζήτηση αν είναι εφικτή η παραγωγή ήχου από εικόνα
5. Περιορισμός σε ένα συγκεκριμένο πεδίο ήχου, όπως η μουσική, ίσως οδηγήσει σε καλύτερα αποτελέσματα
6. Βελτίωση των μετρικών για τη ποιοτική αξιολόγηση της παραγωγής εικόνας από ήχο

Chapter 2

Introduction

2.1 Motivation

Image synthesis has emerged as the prodigy of Generative AI and Computer Vision, transforming the AI landscape and reshaping industries from digital art and entertainment to medical imaging and autonomous driving. However, image generation has predominantly relied on textual descriptions to guide the desired output. Even though text is obvious and straightforward, it can be challenging to convey the intricate details of our physical world.

Audio, on the other hand, is naturally paired with images, providing a rich tapestry of information that text might overlook or find cumbersome to articulate. When we hear a dog barking, we see a dog. When we hear thunders, we expect to look at the imminence of a storm. Sound can describe image attributes in a more effortlessly way, providing a depth of context and ambiance that text descriptions may struggle to encapsulate fully.

Current audio-to-image models learn a joint representation of audio and text to leverage powerful pre-trained text-conditioned image generation models, using large audio-visual pairs of data. These models can accurately provide the image corresponding to the input audio source, but their function comes down to simply associating basic sound sources to their visual counterpart. Visual imagery from audio, however, is a manifold process that requires a different approach.

To begin with, auditory perception is an essential, complex process that is not merely designed for precise frequency encoding, but rather interprets complex acoustic motifs into distinct auditory objects reflecting our experience of the acoustic environment. Whereas sensation can be defined as the detection and processing of sensory information, perception is the act of interpreting and organizing this sensory information to produce a meaningful experience [39]. Furthermore, audio information in the real world is rather a mosaic of individual sources intertwined to complete an acoustic scene. Bregman, in his book *Auditory Scene Analysis* [8], coined the term **auditory scene** to describe the number of sounds that a human can receive simultaneously from the natural environment, investigating how we can make sense of complex auditory inputs through a top-down approach involving reasoning and prior knowledge to resolve ambiguities. It relates to the now well-known *cocktail party problem*, which describes our ability to focus on a single conversation in a noisy environment.

Therefore, audio has to be received and analyzed. It differentiates itself from visual perception, which instantly provides a great deal of information. In contrast to vision, it doesn't

require a line of sight. This suggests that we are able to associate sounds to visual stimuli not directly contributing to the audio stream. For example, if we listen to waves crashing and seagulls we immediately imagine a port or if we hear cars driving by and pedestrians we think of a city. Based on our experience, we are able to map sounds to concepts rather than isolated objects. It is no coincidence that when we listen to bagpipes we think of Scotland. This trait governs our everyday lives and it is often exploited by the marketing industry.

For our case, it also suggests that a reasoning component is crucial for the inference of visual details that cannot be directly derived solely by the audio. While constructing a comprehensive dataset covering the full spectrum of audio-visual associations could potentially render the perception module unnecessary, the curation of such a dataset poses an exceptionally challenging, if not insurmountable, task.

The emergent abilities, including basic reasoning, has been the impetus of Large Language Models over the past few years. Through text, these models have the capacity to compress the world knowledge into their weights, establishing themselves as the cognitive powerhouse of modality-specific systems. Although the development of Audio-LLMs has demonstrated great achievements in various audio-related tasks, there is much debate over their underlying audio reasoning abilities. In [16], they find that ALLMs forget some of the text-based commonsense knowledge available in the LLM. Moreover, a large number of ALLMs exhibit object hallucinations [45], where they identify a sound that does not exist in the audio. Several other issues include their limitation on deductive reasoning [20], their temporal reasoning [67] and most importantly their ability to process multiple audio inputs [14]. It remains open, whether we can utilize ALLMs to guide the image generation from complex universal sounds.

In brief, humans possess the internal competence to imagine and construct a mental image based on the receiving audio through a complex process involving high-level perception and reasoning. The question that still remains is, can machines also acquire this ability?

2.2 Goals and Contributions

Our goal is to assess the capability of ALLMs of handling complex universal audio inputs and their ability to effectively understand the acoustic environment, describing a meaningful vivid image of the scene. We further propose a pipeline that we hope can compensate the aforementioned limitations of these models and conduct a series of experiments that aim to justify or reject our hypotheses. Our contributions are:

- We investigate different prompting techniques on an Audio Large Language Model, with the aim of improving the description appropriateness of the output
- We introduce a pipeline, leveraging a source separation module and a captioning module, that improves the ability to imagine a visual scene based on the input audio
- We test an Audio Large Language Model and our proposed system on a source separation dataset containing complex acoustic clips, introducing various evaluation metrics
- We experiment generating images from short movie clips

Chapter 3

Theoretical Background

The following section lays the conceptual foundations, essential for the subsequent parts. At the core of this research lies the field of Artificial Intelligence, a subdiscipline of computer science dedicated to developing intelligent systems capable of emulating human intelligence such as decision-making, comprehension, problem-solving, creativity, and autonomy. The term "Artificial Intelligence" was formally introduced by John McCarthy in 1956 during the landmark Dartmouth Workshop, a pivotal event that established AI as an academic discipline and catalyzed decades of rapid advancement in the field.

3.1 Machine Learning

Machine learning (ML) is a subfield of artificial intelligence, concerned with the development and study of statistical algorithms that can effectively generalize and thus perform tasks without explicit instructions. The term "learning" denotes the process of determining certain model parameters from input data. Machine learning algorithms can be classified into three primary categories based on the type of learning:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

3.1.1 Supervised Learning

In supervised learning, the algorithm is explicitly provided with the "correct answers". The dataset used to train the model consists of input-output examples, where a feature vector x is paired with a corresponding label y , representing the desired output. The objective is to learn a mapping $y = f(x)$ between the input data and the desired output by minimizing a loss function, enabling the model to generalize well to unseen data. Once trained, the model can be utilized to make predictions or classify data. From a statistical point of view, the algorithm aims to approximate the underlying probability density function $p(y|x)$. Some examples of supervised learning are Naive Bayes, Support Vector Machines and Linear Regression.

3.1.2 Unsupervised Learning

In unsupervised learning, there is no external supervisor, more precisely the data is not labeled. The model is trained to discover the underlying patterns that govern the data distribution, thereby developing the capacity to formulate internal representations of the input information. Key categories of unsupervised learning algorithms include clustering, which aims to partition the data into distinct groups, and dimensionality reduction techniques that transform the initial data points into a lower-dimensional space.

A branch of unsupervised learning is self-supervised learning, where the algorithm autonomously constructs its own target variables from the input data.

Lastly, semi-supervised learning is a machine learning algorithm that works between the supervised and unsupervised learning. The goal of semi-supervised learning is to classify unlabeled data from labeled data

3.1.3 Reinforcement Learning

Reinforcement learning involves a continuous interaction with the environment, with the goal of minimizing a performance metric. The algorithm produces actions and it discovers errors, learning from this trial and error paradigm through rewards. Prominent examples of reinforcement learning methodologies include Q-Learning and Deep Q-Networks (DQN).

3.2 Deep Learning

Despite the success that machine learning brought to many areas, traditional machine learning algorithms needed manual feature extraction to facilitate the learning process, which was performed on the feature space. With the advent of deep learning, a subfield of machine learning, the idea was to let the model learn the features by itself. The term "deep learning" refers to the multilayered structure of the learning process, where each successive layer extracts increasingly abstract features, leading to improved performance in a variety of domains.

3.2.1 The neuron

The building block of neural networks is the neuron, which was first inspired by the neurons found in the human brain. The primary function of a neuron is straightforward: it takes a set of inputs \hat{x} , multiplied by learnable weights \hat{w} , adds a bias θ , and generates an output \hat{o} . This process can be expressed mathematically as $\hat{o} = \phi(\theta + \hat{x}^T \hat{w})$. The weights control the magnitude of influence of each input and the activation function introduces nonlinearity.

Activation functions

The activation function is applied on the dot product of the input signal with the neuron's weights. Common activation functions are:

1. The sigmoid function, produces outputs between 0 and 1, hence it is useful for binary classification. A main issue with sigmoid function is that it is not zero-centered.

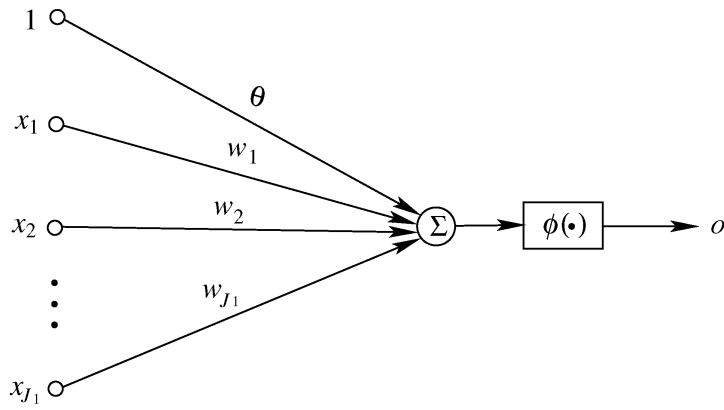


Figure 3.2.1: Perceptron Neuron [23]

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

2. The tanh function or hyperbolic tangent function, has the same shape with the sigmoid, but the output ranges from -1 to 1

$$\tanh(x) = \frac{2}{1 + e^{-x}} - 1$$

Its zero-centered output facilitates easier learning for multiple layers.

3. The Rectified Linear Unit function (ReLU) activates the neuron only when the output is positive

$$A(x) = \max(0, x)$$

A variation of ReLU is Leaky ReLU, which has a small slope for negative values instead of a flat slope. This activation function is widely used in applications where the gradients are often sparse.

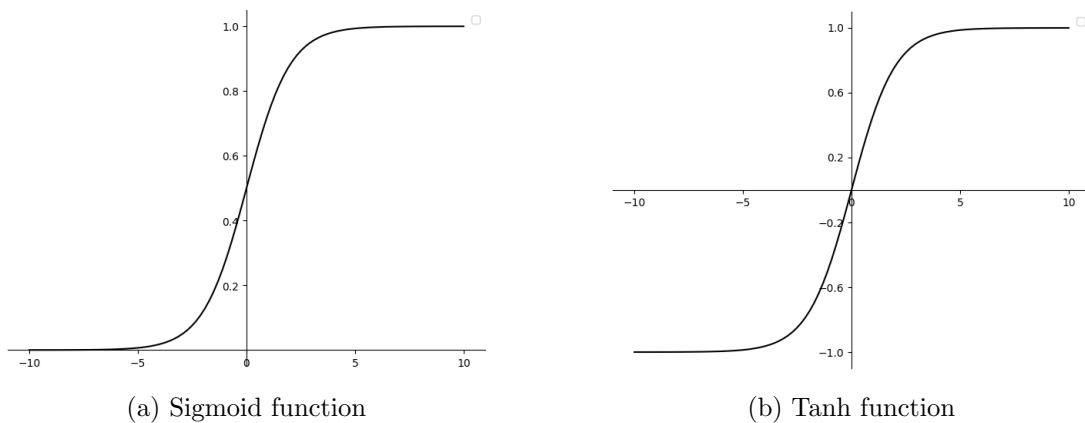


Figure 3.2.2: The tanh function extends the output range from -1 to 1

3.2.2 Feedforward Neural Networks

A simple perceptron comprises a layer of multiple neurons and yields a linear output. This architecture is advantageous when the input data exhibits linear independence. To address more complex data distributions, stacking multiple neuron layers can construct a deep structure known as a Multilayer Perceptron or feedforward neural network, a cornerstone of deep learning. Excluding the input and output layers, each intermediate layer is referred to as a hidden layer. Every neuron in a layer connects to all neurons in the preceding layer. These networks are termed feedforward due to their forward information flow, from the input layer through the hidden layers to the output layer. The objective of a feedforward network is to approximate some target function f^* [30], using the input data (curve fitting).

During training, the neural model learns the set of parameters θ that significantly reduce a cost function $J(\theta)$, which is a function of the weights of the network. The goal of the optimization is to find a minimum of the loss function using gradient descent.

Gradient Descent

The gradient of a function is a vector field, whose value at each point p indicates the direction of the steepest ascent. Therefore, gradient descent is the process of calculating the gradient of the cost function and updating the weights in the direction to which the function will be minimized, according to the rule:

$$w(i+1) = w(i) - \eta \frac{\partial J(w^T, b)}{\partial w}$$

where b is the bias term and η is the learning rate. The learning rate determines how quickly the algorithm converges to a minimum.

Backpropagation

A pivotal moment in the progression of deep learning was the introduction of the backpropagation algorithm, which allowed the model to learn. Backpropagation involves the iterative adjustment of the network's weight parameters in the appropriate direction.

3.2.3 Convolutional Networks

In the digital realm, (grayscale) images are represented as two-dimensional arrays (three-dimensional for colored images) of numerical values, where each value corresponds to the intensity or brightness of a specific pixel. This grid-like structure is a distinguishing feature of image data, as it allows multiple pixels to collectively define distinct regions within the image, independently of other areas. Traditional feedforward networks are not capable of handling this type of data, as they are computationally intensive and unable to effectively preserve the inherent structure of the image. Convolutional networks, commonly referred to as convolutional neural networks or CNNs, are a specialized type of neural network designed for processing data with a known grid-like structure. They are neural networks that take advantage of the convolution operation and have three architectural properties which distinguishes them from other types of networks: local receptive fields, shared weights and spatial or temporal subsampling [47].

Convolution operation

Discrete two-dimensional convolution entails the use of two matrices, I representing the receptive field portion of the input image and K which is the set of learnable parameters otherwise known as a kernel. The operation then is described by the following equation:

$$S(i, j) = (I * K)(i, j) = \sum_m \sum_n I(mn)K(i - m, j - n)$$

The forward propagation process involves the kernel traversing the height and width of the image, generating a spatial representation of the receptive field. This process yields a two-dimensional activation map that encodes the response of the kernel at each position within the image.

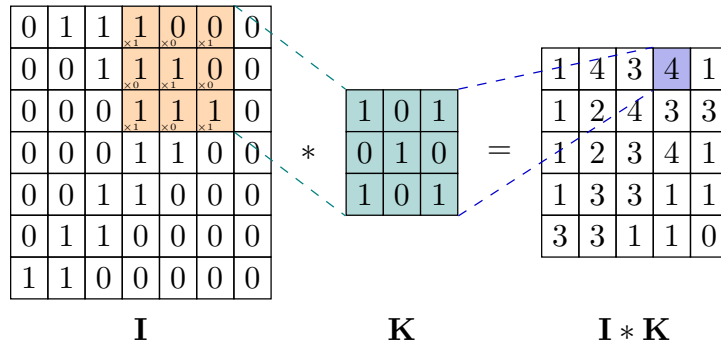


Figure 3.2.3: Convolution operation

The dimensions of the kernel defines the receptive field of the convolution layer and it is a hyperparameter of the network. Every neuron has $w \times h \times c$ number of weights, where $w \times h$ is the receptive field mentioned and c is the number of channels. The individual kernels, or filters, extract meaningful visual features from the image, such as edges, curves, and shapes. As the convolution layers are stacked, each successive layer extracts more complex, higher-level features from the input.

Pooling layer

A typical CNN consists of convolutional layers for feature extraction, pooling layers and a fully connected layer for classification.

The purpose of pooling layers is to systematically reduce the spatial dimensions of the representation, thus decreasing the number of parameters and computational requirements in the network, ultimately helping to mitigate overfitting.

Dilated Convolution

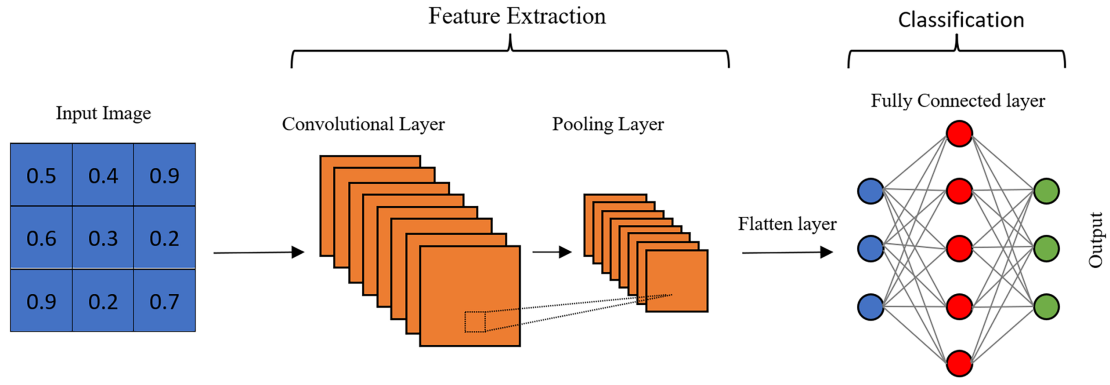


Figure 3.2.4: Architecture of a CNN from [2]

Dilated convolution is just a convolution applied to input with defined gaps. An additional parameter l (dilation rate) indicates how much the kernel is widened. There are usually $l - 1$ spaces inserted between kernel elements. The importance of the dilated convolution is that it allows a larger receptive field with same computation and memory costs while also preserving resolution.

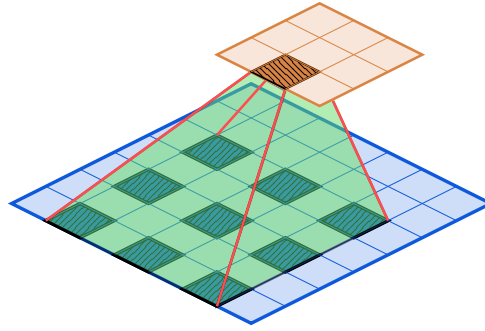


Figure 3.2.5: Dilated Convolution

3.2.4 U-Net Architecture

The U-Net architecture [59] consists of an encoder, that downsamples and extracts useful features from an input image through consecutive convolutional and max pooling layers and a decoder that is used for upsampling. Through the decoder, the field of view of the convolutional kernels is increased.

Although it was initially proposed for image segmentation, it quickly became an integral part of a large number of models specified in computer vision and audio related tasks as we will examine later.

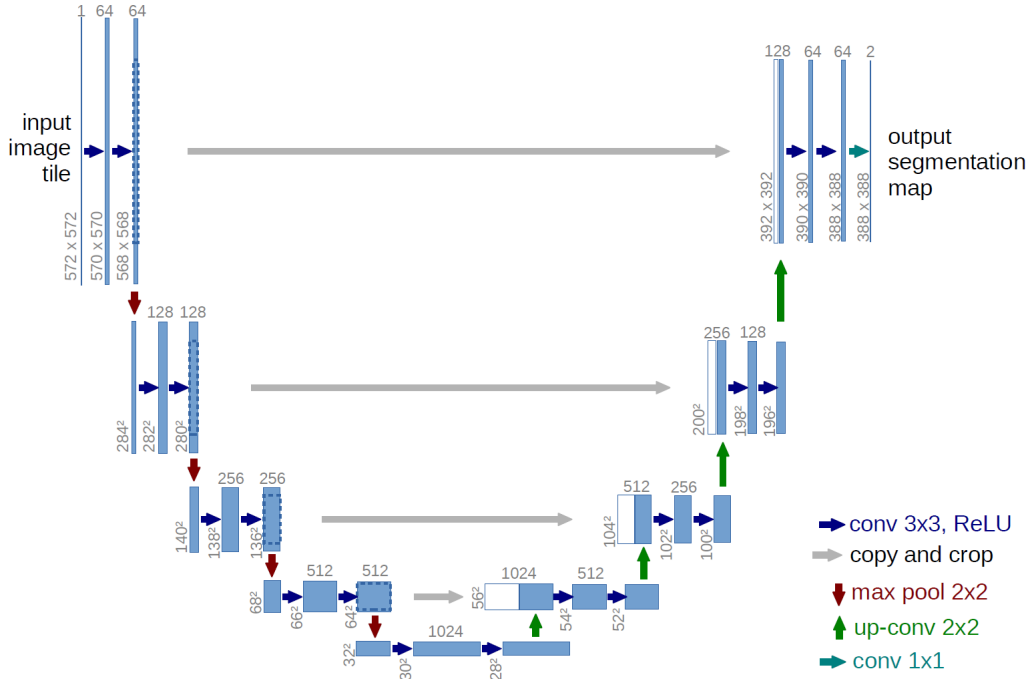


Figure 3.2.6: U-Net architecture
[59]

3.3 Transformers

Sequential data such as time series, sound, or text need a specialized network architecture that can effectively capture long-range dependencies. Recurrent Neural Networks represent a unique type of neural network that can maintain "memory" in the form of preceding states. Each neuron in an RNN accepts an input, generates an output, and recycles its previous state back into the network.

RNNs, however, face issues like exploding gradients and vanishing gradients, which were partially addressed through the introduction of a novel architecture known as Long Short-Term Memory (LSTM). LSTM uses three gates to control the flow of information that is needed to predict the output in the network.

Although LSTMs introduced advancements, maintaining a larger memory and being able to identify the most salient information are crucial for addressing longer-range dependencies. The introduction of the transformer architecture in the now seminal paper "Attention Is All You Need" [75], has shaped the current state of AI.

3.3.1 Architecture

The transformer's architecture, shown in the figure below, is comprised of an encoder and a decoder. Although it was originally intended for text-to-text generation, it can be used for a plethora of applications. The encoder consists of 6 stacked layers each one composed of a multi-head self-attention module and a position-wise feed-forward network, it maps an input sequence to a sequence of continuous representations which is fed to the decoder. The decoder is also made of the same 6 stacked layers, with an additional Masked Multi-Head Attention

module at the start of each layer.

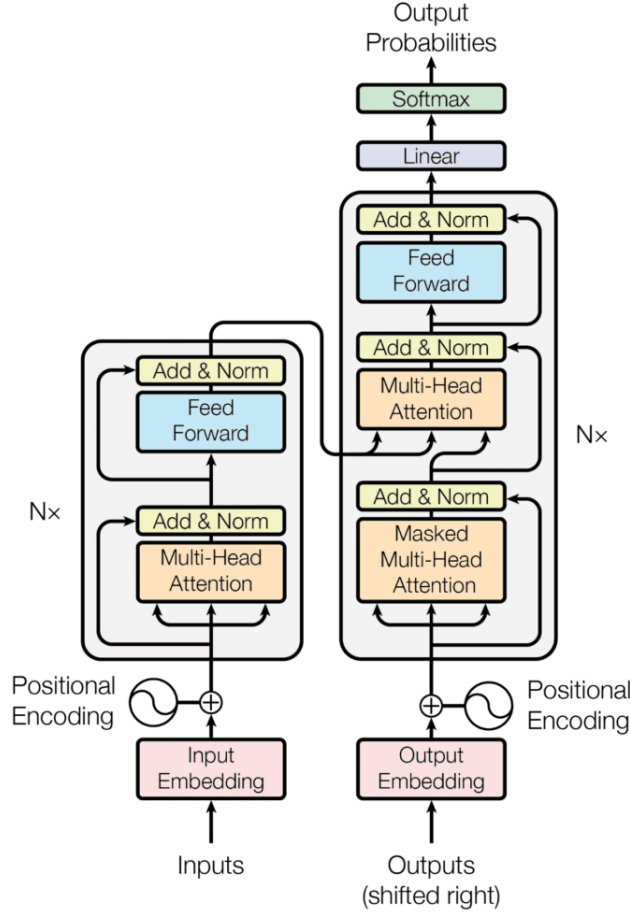


Figure 3.3.1: Transformer Architecture
from [75]

In contrast to previous architectures, the transformer receives the entire input at once and not in individual time steps. To encode the structure and the position of each piece of information however, it uses a position-aware encoding. Another improvement of this architecture is the addition of a novel operation called self-attention

3.3.2 Self-Attention

Initially, the input sequence is divided into individual elements called tokens. Each token is associated with a vector (embedding) that encodes the meaning and its position in the text. Starting with three learnable matrices, query(Q), key(K) and value(V), the embedding of each token is multiplied with the query and key matrix to produce a query and key vector. Self-attention is the process of paying attention to specific parts of the input and measuring a similarity between each element. To accomplish this, an attention score between each key and query is computed, scaled with a factor $\frac{1}{\sqrt{d_k}}$ where d_k is the dimension of the key matrix, and the result is passed through a softmax function to produce a probability distribution. This is

called attention weighting. A large dot product QK^T indicates high similarity and we say the embedding of the token "attends to" the embedding of another token. Using these weights we multiply them with the value vectors to obtain a value result

$$Attention(K, Q, V) = Softmax(\frac{QK^T}{\sqrt{d_k}})V$$

This process is what its called a single head of attention. In multi-head self-attention, there are multiple stacked self-attention modules, each with its own learnable weights, providing distinct representation subspaces and their outputs are concatenated.

3.4 Audio Source Separation

Source separation refers to the process of mixing multiple signals to form a combined signal, aiming at recovering the constituent signals from the mixture. This problem is fundamental in numerous audio signal processing tasks, as the analysis and processing of isolated sources typically yield higher accuracy compared to processing mixtures of sounds. Although the main focus of this field of study is speech enhancement, speech separation, and music instrument separation, in recent years there has been significant progress in the broader domain of universal sound separation, which will be discussed later.

3.4.1 Sound

A sound wave is the pattern of fluctuation in the air pressure over a specific distance or time, it is the pattern of changes in air pressure [1].

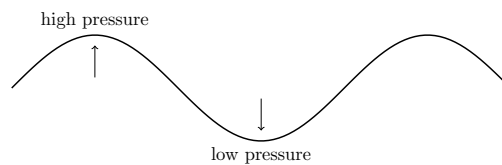


Figure 3.4.1: A sound wave is made up of compression and rarefaction

These alternations in air pressure result in the sound we perceive. The most important properties we use to describe sound are wavelength, amplitude, frequency and phase. The wavelength refers to the distance between two consecutive states of high and low pressure. The amplitude simply relates the number of air particles involved in the vibration and its directly related to the perception of loudness. To measure amplitude we use logarithmic decibels (dB). Since sound is a periodic movement it has also a frequency and a phase. The frequency describes how many cycles are completed per second, it is measured in (Hz) and it affects the perception of pitch. Natural sounds, blend frequencies with varying amplitudes, forming a spectrum.

3.4.2 Audio Representations

In its most unprocessed form, audio is stored as a waveform in the computer in a digitized form. The process of converting analog signals into digital form is called analog-to-digital (AID)

conversion [33]. The analog signal $x_a(t)$ is first sampled at regular time intervals multiples of the sampling period T_n , generating a continuous time signal $x(n)$.

$$x(n) = x_a(nT_s)$$

Afterwards the quantizer maps the continuous amplitude into a discrete set of amplitudes and the converted to a digital array in a computer.

$$\hat{x}(n) = Q[x(n)]$$

The sample rate expresses the number of samples that are taken per second and is measured in (Hz). For a signal with sample rate sr , the maximum frequency that can be reliably reconstructed is $f_N = \frac{sr}{2}$, which is called the Nyquist frequency.

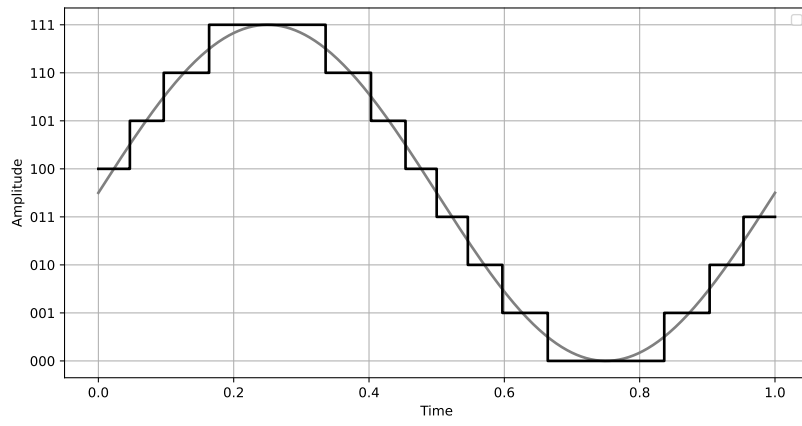


Figure 3.4.2: Analog to digital conversion

Many deep learning algorithms reduce the sampling rate of the input signal to reduce the computational load during training. A signal is monophonic (mono), if there is only one audio channel, else it is stereophonic.

Time-frequency representations

A representation in the time-frequency domain is a 2-dimensional matrix representing the frequency contents of the audio signal over time. Among numerous frequency representations, the most important are Short-Time Fourier Transformer (STFT), Log Spectrogram and Mel Spectrogram.

An STFT is the result of computing a discrete Fourier Transform (DFT) of a small moving window across the duration of the audio. The fourier transform of a signal is defined as:

$$S(f) = \int_{-\infty}^{\infty} s(t)e^{-j2\pi ft} dt$$

The absolute value of a time-frequency bin $|X(t, f)|$ determines the amount of energy of the frequency f at time t . The STFT representation of a signal is called a spectrogram.

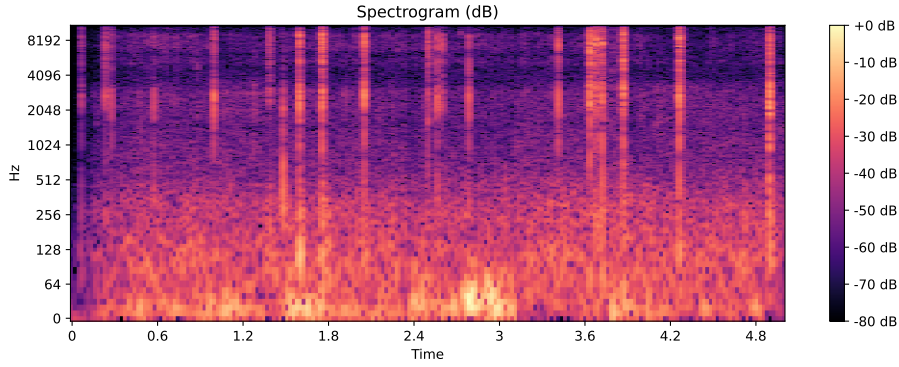


Figure 3.4.3: Spectrogram of an audio signal

The shape of the short-time window determines which frequencies get emphasized in the DFT and the window length defines the resolution of the frequency axis. A great number of machine learning algorithms involve computing the spectrogram of an audio signal as a preprocessing step to extract meaningful features.

In 1937, Stevens, Volkman, and Newmann [68] proposed a unit of pitch such that equal distances in pitch sounded equally distant to the listener. This is called the mel scale. Converting the frequencies to the mel scale and computing the DFT results in a mel spectrogram. Human hearing is also logarithmic in terms of amplitude. Thus, we can obtain a log spectrogram by calculating the log of the absolute value of each element in the STFT.

3.4.3 Traditional Methods

Before the advent of deep learning, classical machine learning and digital signal processing algorithms were utilized in source separation, where the existing approaches were predominately used as a solution to the task of blind separation. In blind separation, as introduced by [34] a set of sources, $s_1(t), \dots, s_N(t)$ are mixed together linearly by a matrix A . The term blind is coined by the fact that we do not know anything about the sources or the mixing process, all we know are the N superpositions of them, $x_1(t), \dots, x_N(t)$.

One way to categorize traditional approaches is by the number of sensors used to capture the number of sources.

Over-determined

In a determined or over-determined situation where the number of sensors is greater or equal to the number of sources, *Independent Component Analysis* (ICA) [5] is the method most commonly used. Given mixed signals $\mathbf{x} = \mathbf{A}\mathbf{s}$ where $\mathbf{x} = [x_1 x_2 \dots x_n]^T$ is an n -dimensional observable random vector, $\mathbf{s} = [s_1 s_2 \dots s_p]^T$ is a random vector with p independent source signals and $\mathbf{A} = [a_1 a_2 \dots a_p]^T$ is an $n \times p$ unknown invertible mixing matrix with linearly independent columns, ICA aims to find an unmixing matrix \mathbf{W} such that $\mathbf{u} = \mathbf{W}\mathbf{x}$ ensuring that the inferred components \mathbf{y} maintain as much statistical independence as possible. The key assumption for ICA to succeed, is the statistical independence of the source signals and that they are non-Gaussian signals. Although ICA stands out as an attractive prospect for computational auditory scene analysis, the necessary assumptions are not practical for auditory scene analysis

where there are fewer sensors than sources.

Under-determined

For an under-determined situation where the number of sensors are fewer than the number of sources, the traditional techniques involve spectral decomposition-based methods, model-based methods and *Computational Auditory Scene Analysis* (CASA) systems. In spectral decomposition, the representation of an input's mixture is dissected into fundamental elements, organized into separate sets symbolizing the distinct sources. One of those methods is *Independent Subspace Analysis* (ISA) [10]. ISA aims to find a set of projection matrices, that project the mixed signal onto independent subspaces.

The *Non-Negative Matrix Factorization* (NMF) is another decomposition strategy. First introduced in [48], NMF decomposes a non-negative nm known matrix \mathbf{V} into the product of two non-negative matrices \mathbf{W} weight matrix and \mathbf{H} the matrix of basis vectors, such that:

$$\mathbf{V} = \mathbf{WH}$$

The algorithm iteratively updates \mathbf{W} and \mathbf{H} to minimize the error between \mathbf{V} and \mathbf{WH} , converging to a local minimum representing a decomposition into meaningful, interpretable components. After training, the source matrix can be calculated as:

$$\hat{\mathbf{V}}_k = \hat{\mathbf{W}}_k \hat{\mathbf{H}}_k^T$$

Another successful approach is using CASA systems [77]. CASA systems aim to mimic the model of human auditory perception to segregate sound in complex environments, using computational models of auditory grouping principles.

Finally, model-based strategies, including *Hidden Markov Models*, model the temporal dynamics and statistical properties of individual audio sources, where each source is represented as a sequence of hidden states, with observable features generated probabilistically based on the current state.

3.4.4 Deep Learning Approaches

With the beginning of the 20th century, increased computational resources with the addition of sophisticated algorithms drove progress in the field of deep learning. Respectively, deep learning was catalytic on improving upon traditional approaches on various problems, including audio source separation. Deep learning methods can be mainly categorized into operating in the time-frequency domain and the time domain.

Time-Frequency domain

Masking is an essential part of many modern source separation approaches that operate on the frequency representation of an audio signal. A mask, $\hat{M}_i \in [0.0, 1.0]^{T \times F}$, representing the i^{th} source S_i , is a matrix containing values and is element-wise multiplied with the mixture spectrogram $|Y|$ to obtain a source estimation $S_i = \hat{M}_i \odot |Y|$. Each value of the mask determines what proportion of energy of the original mixture that a source contributes. Therefore, the objective is the estimation of a mask that approximates as close as possible the target source.

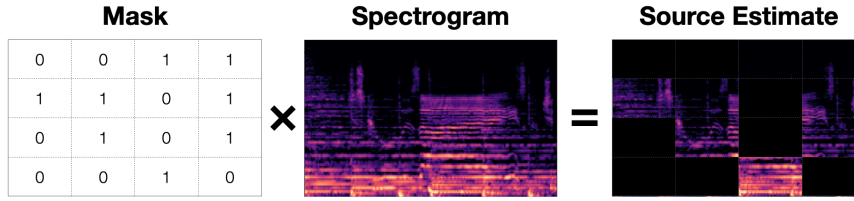


Figure 3.4.4: Example of a binary mask operation from [52]

We can recover the original source i from a mixture by inverting its mask. Depending on the values of the matrix, the mask can be a *binary mask*, containing values 0.0 or 1.0. Common masking applied for source separation is not binary but continuous. *Soft masks* are allowed to take any value within the interval $[0.0, 1.0]$. This means that the energy from a TF bin can be split between source, allowing for flexibility. Masking techniques have a maximal constraint defined by the Ideal Binary Mask (IBM).

A notable approach to source separation has been the use of Convolutional Neural Networks (CNN) due to their ability to effectively learn hierarchical representations from time-frequency representations like spectrograms. One of the earliest works, leveraging CNNs, was the influential work of Chandna et al. [11], which explicitly used CNNs to estimate time-frequency soft masks that can be applied for musical source separation. With the introduction of the U-Net for image segmentation, researchers begin using the architecture in various tasks. In [37] they adapted the U-Net architecture for the task of source separation, where they framed the source separation task as a spectrogram-to-spectrogram translation task. The model learned to map a mixed spectrogram to the spectrogram of the target source.

Time domain

A key benefit of operating in the waveform domain is the preservation of phase information, which is frequently lost in the magnitude-only representations employed by other methods. Phase is a fundamental aspect of an audio signal. Some methods assume the phase aligns with the mixed signal’s phase, while others utilize methods such as the Griffin-Lim algorithm to estimate the phase.

Several prominent works on the time domain have been established as a respectable approaches to the source separation problem, one of them being WaveNet based methods. WaveNet [73] is a deep neural autoregressive network based on PixelCNN. A crucial adaptation, is the work of Stoller et. al [69], which leverages an improved U-Net architecture for singing voice separation.

The encoder downsamples the waveform while capturing hierarchical features, and the decoder upsamples it to reconstruct the separated sources.

Another influential architecture worth mentioning is TasNet and it’s subsequent refinements. TasNet [50] employs a strategy where the input waveform is converted into a latent representation via an encoder, which is then processed by a separator network to isolate the source signals. Later on, Conv-TasNet [51], replaced the fully connected layer with a *Temporal Convolutional Network* (TCN) [46], achieving significant improvement and establishing itself as a new state-of-the-art for speech separation.

In the realm of music source separation, a considerable work is Demucs [18], which incor-

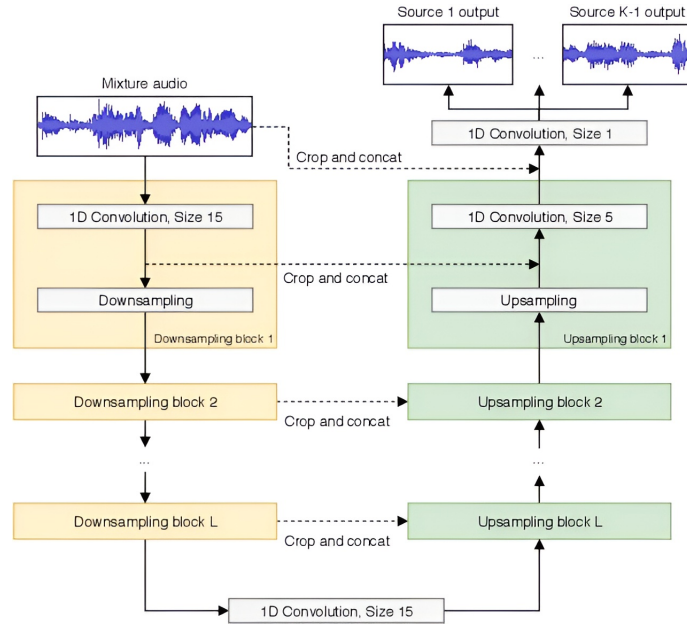


Figure 3.4.5: Wave-U-Net Architecture
[69]

porates LSTMs within the bottleneck of the U-Net, enabling the model to capture long-range temporal dependencies in the music signal.

3.4.5 Evaluation Metrics for Source Separation

As with all scientific endeavors, sound source separation relies on the systematic experimentation to verify the integrity of the results. Thus, given the complexity of the task, there is a need for quality evaluation metrics that are able to assess the outcomes. An estimate of a source \hat{s}_i assumed to be composed of four separate components, $\hat{s}_i = s_{target} + s_{interf} + e_{noise} + e_{artif}$, where s_{target} is the true source and e_{interf} , e_{noise} and e_{artif} are error terms regarding the interference, noise and added artifacts. The following are the main evaluation metrics:

Source-to-Distortion Ratio (SDR). SDR measures the overall quality of the separated signal by comparing the desired signal to the unwanted components and it is defined as:

$$SDR = 10 \log_{10} \left(\frac{\|s_{target}\|^2}{\|e_{interf} + e_{noise} + e_{artif}\|^2} \right)$$

Source-to-Artifact Ratio (SAR). It is the amount of unwanted artifacts a source estimate has with relation to the true source:

$$SAR = 10 \log_{10} \left(\frac{\|s_{target} + e_{interf} + e_{noise}\|^2}{\|e_{artif}\|^2} \right)$$

Source-to-Interference Ratio (SIR). Measures the "leakage" of the estimation, meaning how much of other sources can be heard in the estimate.

$$SIR = 10 \log_{10} \left(\frac{\|s_{target}\|^2}{\|e_{interf}\|^2} \right)$$

Source-to-Noise Ratio (SNR).

$$SNR = 10 \log_{10} \left(\frac{\|s_{target}\|^2}{\|s_{target} - \hat{s}\|^2} \right)$$

Scale-Invariant Source-to-Distortion Ratio (SI-SDR). The SI-SDR measure removes the SDR's dependency on the amplitude scaling of the signal.

3.5 Diffusion Models**3.5.1 Generative Models**

Deep learning drove progress and research in image classification, speech recognition, natural language processing and other tasks. Neural networks in this era were typically trained as discriminative models, due to the adversity of generative modeling. The groundbreaking work of Diederik P Kingma and Ian J. Goodfellow, led the way to a new era for generative modeling, producing the first practical deep neural networks capable of learning generative, rather than discriminative, models of complex data such as images.

The purpose of deep generative modeling is the development of statistical models that not only recognize patterns in data, but also have the capacity to learn the underlying data distribution and generate novel samples based on the learned distribution. Generative modeling can take two forms:

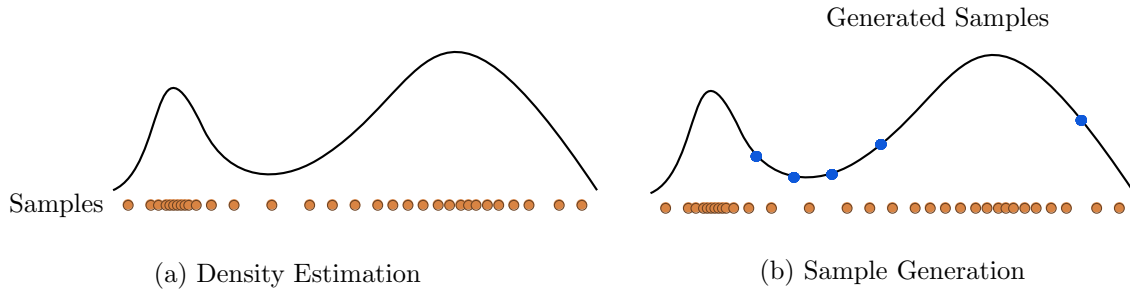


Figure 3.5.1: Generative modeling can estimate the underlying distribution and generate new samples

Autoencoders

Autoencoders are a class of generative models consisting of an encoder and a decoder. The encoder compresses the input data into a low-dimensional latent space called "bottleneck". This captures the salient features of the data while simultaneously reducing the computational cost. Afterwards, the decoder learns to reconstruct the latent representation minimizing the reconstruction loss between the output and the ground truth.

$$\mathcal{L}(x, \hat{x}) = \|x - \hat{x}\|^2$$

To alleviate the deterministic nature of vanilla autoencoders, variational autoencoders [44] introduce an element of randomness. This ensures that the sampling of novel samples resemble the initial distribution, without being a mere copy. For every latent variable we define an average value μ and a standard deviation σ which determine a probability distribution, typically Gaussian, over this variable. This design choice encourages the encoder to place the latent variables around a central point, rendering the memorization of extreme deviations challenging. The loss function employed during training combines the reconstruction loss and a regularization term. The regularization term provides:

- Continuity: Points closer in latent space correspond to similar decoded content.
- Completeness: Sampling on the latent space should provide meaningful results.

Generative Adversarial Networks

When the modeling of the data distribution is intractable, *Generative Adversarial Networks* (GANs) [31] provide a solution by taking samples from the easiest distribution possible, random noise. A generator network is trained to translate random noise samples into meaningful samples from an input distribution and a discriminator network is trained to discriminate between real samples and generated samples. Together, the two networks operate in an adversarial manner, competing against each other. During training, the generator learns to produce faithful samples that will effectively deceive the discriminator and the discriminator develops the ability to better distinguish real from fake samples.

Their generative strength in image synthesis, creating photo-realistic images, placed GANs in the forefront of image generation for years to come. Despite their performance however, VAEs and GANs have some limitations such as mode collapse and difficulty in training.

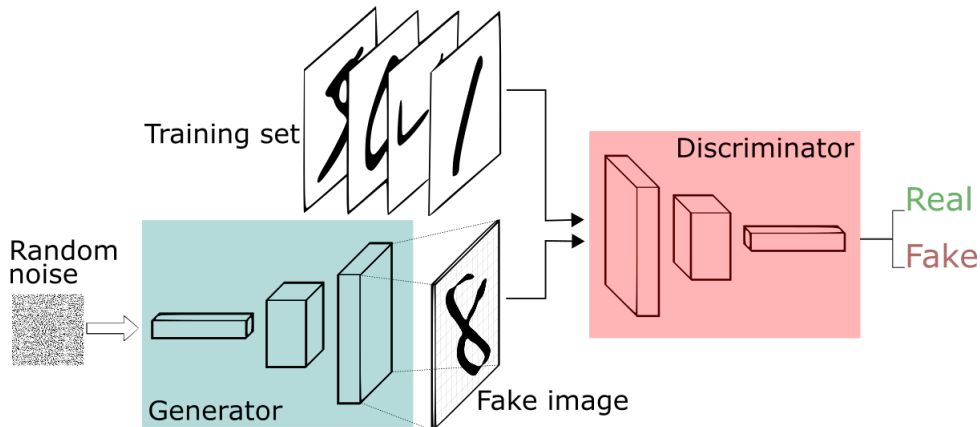


Figure 3.5.2: GAN Pipeline from [62]

3.5.2 Diffusion Process

Diffusion models have emerged as the new state-of-the-art family of generative models, gaining much popularity in recent years with the development of models specified in image

generation. They managed to outperform other generative models that were at the spotlight in many areas. It all began in 2015, when Sohl-Dickstein et al. [63], inspired by non-equilibrium thermodynamics, pioneered the diffusion process for novel samples generation. Diffusion models have three predominant formulations: denoising diffusion probabilistic models (DDPMs), score-based generative models (SGMs) and stochastic differential equations (Score SDEs)

Denoising Diffusion Probabilistic Models (DDPMs)

The main idea of this method is the progressive perturbation of the data distribution through an iterative process and the learning to reverse this process, recovering the original distribution. This results in a generative model capable of creating new samples, similar to those of the training data. Their framework consisted of two processes that are modeled by two Markov chains.

Forward Process The initial phase is characterized by a procedure called the forward process. Starting from a data distribution $q(\mathbf{x}^{(0)})$, noise is systematically introduced in incremental steps according to a predetermined schedule until any structure is completely eradicated.

$$q(\mathbf{x}^{(0...T)}) = q(\mathbf{x}^{(0)}) \prod_{t=1}^T q(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})$$

Reverse Process The reverse process involves reversing the forward process by learning transition kernels parameterized by a neural network.

$$p(\mathbf{x}^{(T)}) = \pi(\mathbf{x}^{(T)})$$

$$p(\mathbf{x}^{(0...T)}) = p(\mathbf{x}^{(0)}) \prod_{t=1}^T p(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)})$$

During learning, the mean and the covariance of the Gaussian diffusion kernel are needed to be estimated, therefore the computational cost of this algorithm solely depends on the cost of the functions defining them, times the number of time-steps.

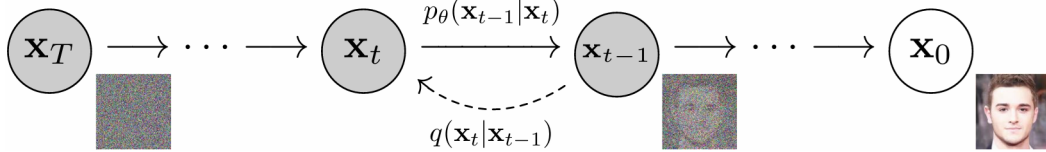


Figure 3.5.3: Forward and Reverse processes from [35]

Score-based Generative Models

Later on, in 2019 [65], introduced a new generative model that depends on the concept of *Stein* score of the logarithmic density of data, namely the gradient of the log likelihood of the given distribution, $\nabla_x \log p(x)$, which is essentially a vector field pointing on the direction where the log data density grows the most. Their main idea was to perturb the data with increasing

various levels of Gaussian noise and with the aid of a neural network, which they define as *Noise Conditional Score Network* (NCSN), to estimate the scores for all noise distributions. For generating new samples, they proposed the use of Langevin dynamics starting at the noisiest score annealing down the noise level.

Noise Conditional Network (NCSN) is a deep neural net $\mathbf{s} : \mathbb{R}^D \rightarrow \mathbb{R}^D$ parameterized with θ and is trained to have the ability of estimating the score of the data distribution $p_{data}(\mathbf{x})$. Hence, the purpose of score-based generative modeling is to learn a model that can generate new samples from the underlying data distribution.

Stochastic Differential Equations (Score SDEs)

This approach [66], generalized the previous concepts in the case of infinite time steps or noise levels, where diffusion and denoising processes are solutions to Stochastic Differential Equations (SDEs). Instead of a finite number of noise distributions, they considered a continuum of distributions evolving in time based on a stochastic differential equation. This process progressively diffuses data with noise and is derived by a stochastic differential equation

$$dx = f(x, t)dt + g(t)dw$$

that does not depend on data and does not have trainable parameters. Respectively, for sample generation, a reverse process that satisfies a reverse-time SDE

$$dx = [f(x, t) - g(t)^2 \nabla_x \log p_t(x)]dt + g(t)d\bar{w},$$

can be estimated by training a neural network to estimate scores. New samples are then obtained by using SDE solvers.

3.5.3 Improvements

In the following years, a plethora of significant research papers further refined diffusion models with several improvements. The most important among them is the simplification of the variational bound presented in [35], the acceleration of the sampling procedure with the *Denoising Diffusion Implicit Model* [64] and the introduction of the cosine schedule that aids in a more gradual destruction of data [54].



Figure 3.5.4: Linear(*top*) and Cosine schedule(*bottom*). With the Cosine Schedule data is perturbed more gradually [54]

Advancements regarding the guidance of sample generation will be discussed in the forthcoming chapter.

Chapter 4

Related Work

This section provides details about the specifications of the models related to this work.

4.1 Language Models

A language model is a computational model, capable of capturing the structure, patterns and statistical properties of human text, allowing it to predict or generate novel text samples. Specifically, a language model is trained on a large amount of text, to conduct a probability distribution over predefined words called the vocabulary. Although language models are able to perform all sorts of tasks, it's basic function comes down to estimating the probability of a token or sequence of tokens occurring within a longer sequence of tokens.

$$P(w_1, w_2, \dots, w_n) = P(w_1) \cdot P(w_2|w_1) \cdot P(w_3|w_1, w_2) \dots P(w_n|w_1, w_2, \dots, w_{n-1})$$

A *token* is defined as a word or a chunk of a word. This partition is a preprocessing step to numerous natural language processing (NLP) tasks.

4.1.1 Large Language Models

Predicting the next sentence is quite challenging even for a human. But how about predicting something that seems much easier, like the next word someone is going to say? What word, for example, is likely to follow [\[40\]](#)

The water of Walden Pond is so beautifully ...,

the next word might be **green** or **calm**, but the probability of the word **refrigerator** appearing in the end of this sentence is very low.

In the 1950s, the dream of effortless communication across languages fueled the birth of NLP. The primary driving force was machine translation and the initial approach centered around rule-based systems. However, rule-based systems struggled with the nuances of the human language with all its idioms and variations. Around the time machine learning was getting much progress, there was a shift to a more statistical approach. The N-Gram model, a statistical model, assumes that the probability of a word depends only on a limited history (the Markov assumption). Thus, the model is able to estimate conditional probabilities from observed frequency counts in a text corpus. A paradigm shift was introduced in 2003, when

the n-gram model was outperformed by a Multilayer Perceptron (MLP) [6], paving the way into a neural network era. Several advancements were made in language modeling using Recurrent Neural Networks (RNN) and LSTMs and particular in the representation of words with the proposal of Word2Vec in 2013 [53]. The model utilizes either the continuous bag-of-words (CBOW) or the skip-gram architecture to predict contextual information, thereby representing words in a continuous vector space where semantic and syntactic similarities are captured as geometric proximity. This innovative approach transformed natural language processing by enabling efficient and scalable learning of word embeddings, which have since become essential components in a wide range of NLP applications.

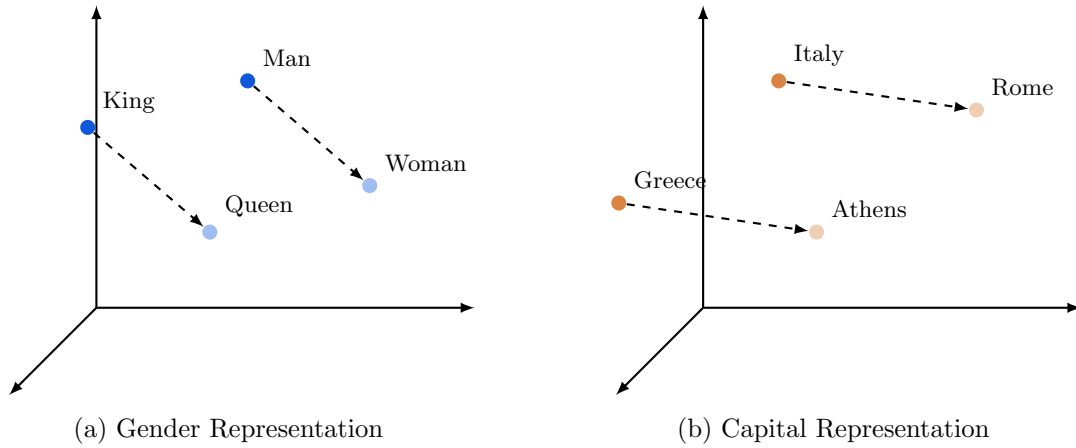


Figure 4.1.1: The word embeddings capture meaningful relations between words. As an example, a specific direction might encode the gender information or the capital of a country.

The establishment of the transformer architecture as a staple, transformed the landscape of language modeling. Their parallelization allowed the analysis of a larger input and their long-range receptive field facilitated more contextual and human-like generation. In recent years, *Large Language Models* (LLMs) have emerged as a state-of-the-art in the field of natural language processing (NLP). These models, built upon the transformer architecture and trained on a vast corpora of text data, are capable of capturing intricate statistical and semantic relationships within language.

Categories

Broadly, LLMs can be grouped into three distinct categories:

Auto-regressive models like the *Generative Pre-Trained Transformers* GPT-series, where the model is trained in a next-token prediction fashion. These models utilize only the decoder of the transformer architecture and through training they learn to efficiently compress the textual knowledge of the world into the model’s weights.

Auto-encoding like Bert, where the model is trained using masked language modeling

Sequence-to-sequence models like T5, are designed for tasks where the input and output are sequences, such as translation, summarization, and text generation.

Key Characteristics

The current success of LLMs is perhaps predominantly attributed to scaling. As the number of model parameters increase, the performance of these models improve and such has been the case with LLMs. In the past few years, sophisticated algorithms coupled with increased computational resources, enabled the models to be trained from billions to trillion of tokens encompassing nearly the entire *World Wide Web*. This brings us to the motif of large language models and the reason why they have been the main focus of the research community.

One of the most remarkable aspects of current LLMs is their ability to exhibit behaviors that they were not explicitly trained to have. These emergent abilities include performing novel tasks with little to no additional training (in-context learning), showcasing a form of adaptability that has significant implications.

The ability that this research has been relied upon is the emergence of reasoning in current LLMs. As these models have scaled in size, they have demonstrated a capacity for performing tasks that require logical inference, multi-step problem solving, and even rudimentary forms of reasoning through in-context learning.

4.1.2 Prompting

Artificial Intelligence has seen great progress from its initial notion. From explicitly programming a system with strict rules, to gathering the necessary training data and currently, harnessing the knowledge distilled into a pre-trained model for novel tasks. *Prompt engineering* requires composing natural language instructions called prompts to elicit knowledge from LLMs in a structured way [76]. We define as a *prompt*, the input text query into the LLM.

At its core, prompt engineering is based on the premise that the performance of an LLM on a given task can be substantially influenced by the manner in which the input is presented. In this work we examine some of the prompt engineering techniques:

Zero-Shot

The zero-shot prompt directly instructs the model to perform a task without any additional examples to steer it. It is the most basic form of guiding the text generation with a straightforward instruction.

Few-Shot

Few-Shot prompting [9] refers to the process of providing one (one-shot) or more (few-shot) examples inside the prompt, steering the model into the proper direction.

As we can see in the below figure, as the model size increases, it benefits from the inclusion of examples in the prompt. Thus, instead of finetuning a separate language model checkpoint for each new task, one can simply “prompt” the model with a few input–output exemplars demonstrating the task.

For example, for translation, we could provide the model with few translation examples to the desired language.

Chain-of-Thought A chain of thought is a series of intermediate natural language reasoning

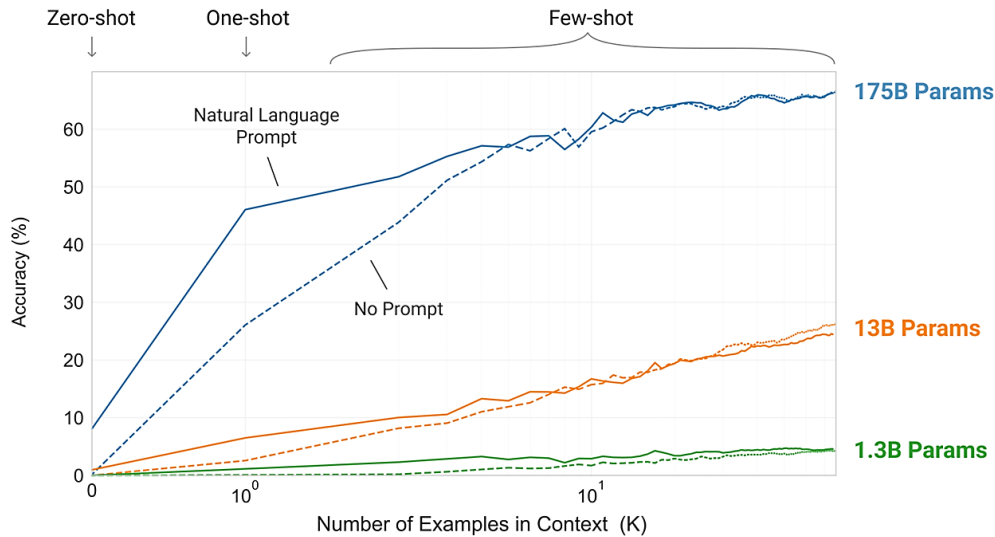


Figure 4.1.2: Larger models make increasingly efficient use of in-context information [9]

steps that lead to the final output, this approach is referred to as *Chain-of-Thought* prompting [79]. This method guides the model to explain its response in a step-by-step manner. There is evidence that chain-of-thought prompting elicits reasoning abilities in large language models, forcing the model to decompose a complex problem into intermediate steps.

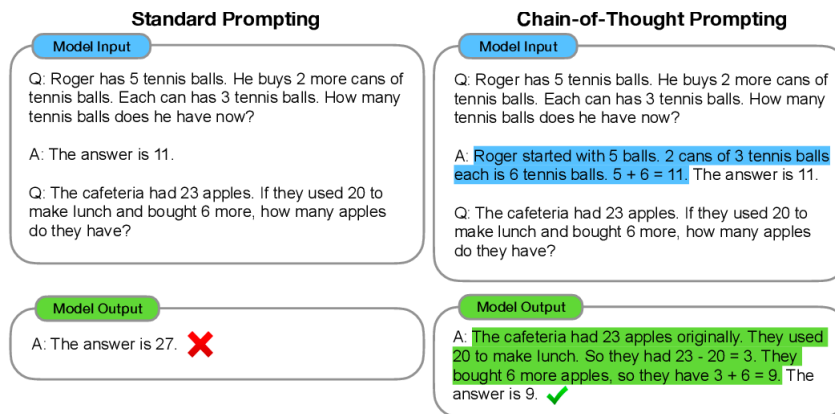


Figure 4.1.3: Standard (zero-shot) prompting and Chain-of-Thought prompting [79]

Role Prompting Another interesting prompting technique is the assignment of a particular role to the model. This encourages the AI to approach a question by assuming the specified role, leading to a more focused, creative, or empathetic response depending on the given role. For example, if we expect a helpful, patient and friendly tone we might assign the role of a helpful assistant.

4.2 Guided Image Generation

Guided image generation refers to a class of generative modeling techniques in which auxiliary information is used to steer the output of an image synthesis model. The progress of GANs and the development of Diffusion models later on, promoted rapid advancements in the domain of image generation and computer vision in general. Over time, the field of image synthesis has evolved from producing simple facial renderings to creating visual masterpieces inspired by renowned artists, establishing itself as a valuable tool in many areas such as the entertainment industry, architecture, gaming development and arts. This progress would not have been feasible without several improvements on the diffusion model architecture and the introduction of textual guidance.

4.2.1 Text-to-Image

The training of a diffusion network that operated on the pixel space demanded many time steps, inducing slow inference and escalating computational cost. The authors of Stable Diffusion [58] provided a solution by introducing a new class of diffusion models called *Latent Diffusion Models*. These models employ an autoencoder to encode the information into a low-dimensional latent space. In the latent space, images are encoded with fewer dimensions, reducing significantly the computational cost and accelerating the image generation process.

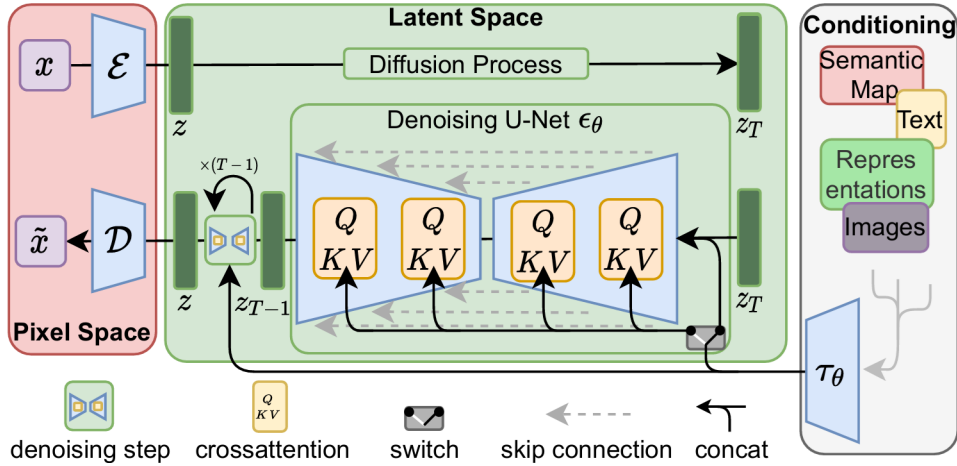


Figure 4.2.1: Stable Diffusion Architecture [58]

The introduction of text-guided diffusion models completely revolutionized the field by enabling the synthesis of images that are closely aligned with detailed natural language prompts. To guide the image generation through text prompts, the textual input first is encoded into a meaningful representation that aligns with the representation of a similar image. This is achieved with a CLIP (*Contrastive Language Image Pretraining*) model [56], which is trained using contrastive learning to encode similar text-image pairs closer into the latent space.

These embeddings are then incorporated into the diffusion model through *cross-attention* layers of the U-Net. A cross-attention layer is different than the self-attention, in a sense that the query refers to a different modality than the key-value vectors. Therefore, the cross-attention

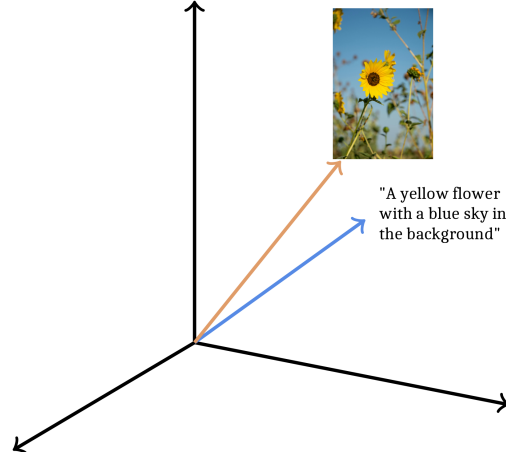


Figure 4.2.2: Image-Text Embeddings Alignment Illustration

layers can influence the image features that are most closely aligned to the text features of the input prompt.

4.2.2 Audio-to-Image

While text-guided image generation is straightforward, textual descriptions are not naturally paired with the image. On the other hand, audio and visual information are naturally paired and sound can capture more subtle nuances that are difficult to be effectively expressed through text. These details can be the intensity of a thunder, the soothing tranquility of a summer morning breeze or the intricate and complex soundscape of a bustling metropolitan street. Several works have tried to incorporate audio to guide the image synthesis process, which rely on the learning of a joint representation.

Sound2Scene

In this work, they learn to align audio to visual latent space [70]. In contrast to previous works that were limited to a small dataset where images and sounds were closely related, they use VGGSound [12]. They first train a GAN to generate images from the visual embeddings of the image encoder and then train an audio encoder to translate an input sound to its corresponding visual feature by aligning the audio to the visual space. After training, the audio embedding can be directly used to guide the image generation, as it is related to a corresponding visual embedding.

Specifically, using the audio-visual data pairs $\mathcal{D} = \{V_i, A_i\}_{i=1}^N$, where V_i is a video frame selected from the dataset clip and A_i is audio, the objective is to learn the audio encoder to extract informative audio features $\mathbf{z}^{\mathbf{A}} = f_A(A)$ that are aligned well with anchored visual features $\mathbf{z}^{\mathbf{V}} = f_V(V)$. These representations are aligned minimizing the InfoNCE loss [74]. This work is also possibly the only one that experiments with combining multiple audio waveforms, something that leads to a more complex audio.

AudioToken



Figure 4.2.3: Generated images by mixing two different audios in the waveform space [70]

AudioToken [81] is a model that leverages the high-quality image generation capabilities of Stable Diffusion together with an audio representation model, BEATs [13]. The system consists of a text tokenizer that encodes the initial prompt "A photo of a" and an audio embedder which produces an audio embedding e_{audio}

$$\text{"A photo of a"} \rightarrow e_{text} \in \mathbb{R}^{4 \times d_a},$$

where d_a is the embedding dimension of the text input. The concatenation $[e_{text}, e_{audio}]$ is then used as input to the text-conditioned Stable Diffusion, as if it was a simple prompt. The overall architecture of the model is shown below.

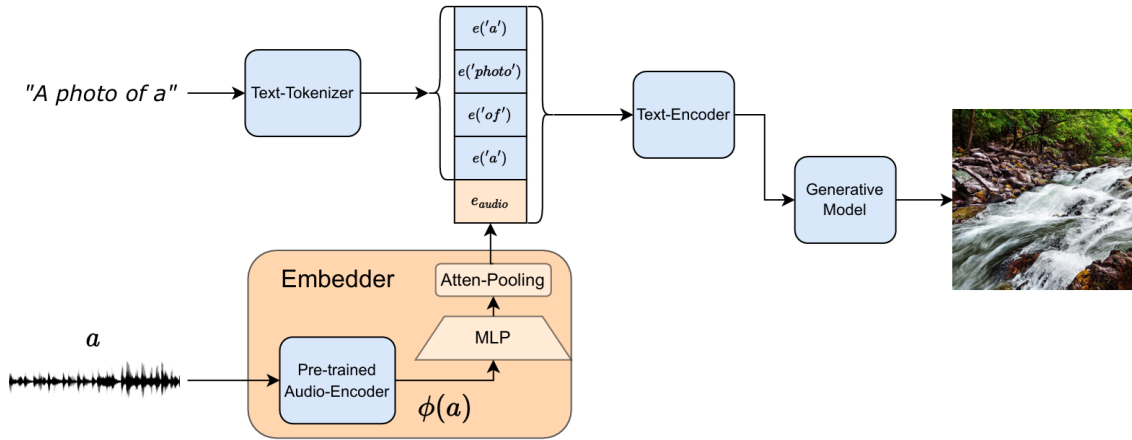


Figure 4.2.4: AudioToken Architecture [81]

This work paved the way for similar publications which leveraged a pre-trained text guided image generation model and focused on aligning the audio to the textual space.

SonicDiffusion

SonicDiffusion [7], except from audio-text alignment, they incorporate cross-attention layers through which the audio tokens are infused and guide the image generation. Thus, this process enhances the semantic and visual alignment between the synthesized images and the input audio.

Limitations

The aforementioned models, although they achieved impressive results in the audio guided image generation, they come with certain limitations.

- Their generative strength is constrained on simple sounds that come from distinctive sources
- They lack the reasoning required to infer details of complex audio inputs that are essential for a detailed image

4.3 Audio Language Models

The emergence of LLMs as a powerful tool in language processing tasks raised an important question concerning other modalities. Consequently, a surge of works contributed to extending the capabilities of language models beyond the confines of text into the realm of vision, acoustics, video, and other modalities. Before delving into the intricacies of their internal mechanics, it is crucial to revisit and evaluate the role of audition within the broader field of artificial intelligence concerning around which main tasks the research community revolves.

4.3.1 Computer Audition

Computer audition (CA) or machine listening is the general field of study of algorithms and systems for audio interpretation by machines [78]. Since audio signals are interpreted by the human ear-brain system, that complex perceptual mechanism should be simulated somehow in software for "machine listening". We use the term "audio" in the broader sense, encompassing all natural sounds such as speech, music, and environmental sounds. Although the focus of most research works is on speech or music related tasks, we are interested in *universal audio*. The most important universal audio related tasks are:

Acoustic Scene Classification and Sound Event Classification

This tasks refers to the classification of an input audio into a predefined set of labels, treating the audio as a whole (acoustic scene) or classifying a single source present in the audio.

Sound Event Detection (SEC)

Sound Event Detection is the task of detecting individual sound events in the input audio and their temporal activity.

Automated Audio Captioning (AAC)

The automated audio captioning problem can be defined as the task of automatically generating a textual description (i.e. caption) for an audio signal, where the caption closely approximates a human-generated one for the same audio signal. This differs from sound event detection (SED),

because the audio captioning method does not predict sound events and their start and end times [21].

Audio Question Answering (AQA)

This task can include either the generation of a complete sentence from an audio or a sequence of audio-question answers from subsequent audio clips and corresponding questions.

4.3.2 Audio Large Language Models (ALLMs)

Until recently, the main template for addressing the aforementioned audio tasks was the construction of a task-specific neural network architecture. However, as the scale of models and datasets continues to expand, traditional *Multi-Modal* MM models incur substantial computational costs, particularly when trained from scratch [82].

Considering these models operate at the intersection of multiple modalities and text, it was only logical to capitalize on readily available text-conditioned foundation models, Large Language Models (LLMs). This junction between modalities and the linguistic knowledge capacity of LLMs led to the emergence of a novel field, *Multi-Modal Large Language Models* (MM-LLMs), specifically for the audio modality they are known as Audio Large Language Models (ALLMs). These models harness LLM’s advantageous properties such as language generation, emergent reasoning abilities, and In-Context Learning (ICL), integrating an LLM as the cognitive powerhouse to empower various multimodal tasks.

ALLMs are trained on millions of audio-text pairs using either Contrastive Learning or Next-Token Prediction.

Contrastive ALMs

These models utilize an audio and a text encoder in a tower-like structure to learn a joint audio-text multimodal latent space. They are trained using contrastive learning and after training, they can be used for zero-shot inferences such as classification and retrieval. An example of such model is the *Contrastive Language-Audio Pretraining* CLAP [24].

Next-Token Prediction

These models frame all audio tasks as audio question-answering tasks and generate free-form text. Specifically, given a pair of data (\mathbf{a}, \mathbf{x}) , where \mathbf{a} denotes the audio sequence and \mathbf{x} the text sequence (input prompt), the training objective is to maximize the next-token probability, denoted as

$$\mathcal{P}_{\theta}(x_t | \mathbf{x}_{<t}, \text{Encoder}_{\phi}(\mathbf{a})),$$

conditioning on audio representations and previous text sequences $\mathbf{x}_{<t}$, where θ and ϕ denote the trainable parameters of the LLM and audio encoder respectively [15].

The general architecture of ALLMs as illustrated in the figure 4.3.1, consists of five fundamental components:

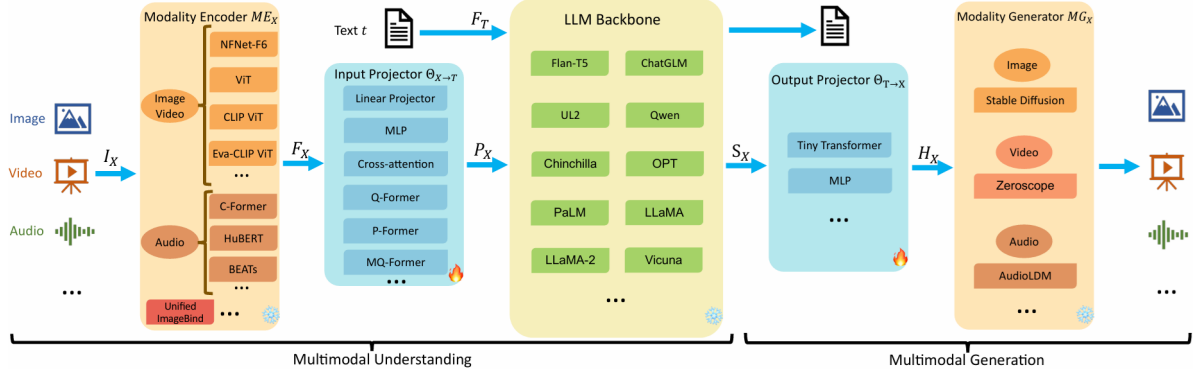


Figure 4.3.1: The general model architecture of MM-LLMs [82]

- Modality Encoder
- Input Projector
- LLM Backbone
- Output Projector
- Modality Generator

Depending on the model’s purpose it may include or not a modality generator for an end-to-end intermodality generation. Although over the years there have been many variations, most of the models follow the same architectural blueprint, improving upon prior works over the modality representation, the latent alignment and the LLM fine-tuning.

The *Input Projector* module is tasked with aligning the encoded features of the input modality with the text feature space T of the LLM. The aligned features are then fed as prompts P_x into the LLM alongside with the textual features. The majority of models incorporate a simple linear projection (MLP) to project the audio embeddings into the input space of the language model, but other choices may include cross-attention layers, Q-Former, P-Former and MQ-Former.

Afterwards, the LLM backbone processes the input embeddings engaging in semantic understanding and decision making regarding the inputs. It produces either simple text, if the model is modality-to-text, or modality-specific signal tokens S_x for any-to-any generation. The aligned representations of other modalities can be considered as *soft prompts* for the language model. Soft prompts are learnable tensors concatenated with the input embeddings.

One of the main reasons MM-LLMs utilize a foundational model such as *LLMs*, is that they have been pre-trained on massive amount of data and have the inherent ability to adapt to modality specific tasks without further training. Because training even a fraction of the model’s weights, can lead to major implications both on computational cost and issues such as *catastrophic forgetting*, where the model “forgets” some of the knowledge gained during the training process. Therefore, a number of works have included the LLM without further retraining.

But to make the most out of the language model’s capacity and even extend its abilities, several have adopted a *Parameter-Efficient Fine-Tuning* (PEFT) strategy. PEFT methods train only a portion of the model’s weights, or attach new layer of weights (adapters) and

only train those weights. By training a small set of parameters, PEFT preserves most of the large pretrained model’s structure and saves time and computational resources. The most used approach is *Low-Rank Adaptation* (LoRA) [36], but others include Prefix-Tuning and Layer Norm tuning.

One of the first models encompassing transfer learning to approach all audio tasks as text generation tasks was *Pengi* [19]. The resulted system is able to take an audio clip and text as input and generate free-form text. A great leap forward was achieved by *LTU* model [29], which focuses on improving the ability to perceive and comprehend audio signals in a deeper level than previous efforts. It was a significant achievement on the field of audio perception and it stood as a pillar for the emergence of audio reasoning. It utilized AST [28] as the audio encoder and an open-source LLaMA [72] with LoRA adapters. Qwen-Audio [15], is one of the largest and most potent audio large language models. It scales the audio-language tasks to over 30 tasks, enabling universal audio understanding abilities. It employs Whisper-large-v2 [57], which is a 32-layer Transformer model, as the audio encoder and Qwen-7B [3] as the LLM-backbone. They also propose a multi-task training framework, used to perform co-training on a wide range of tasks, eliminating the need for laborious model switching when dealing with different tasks.

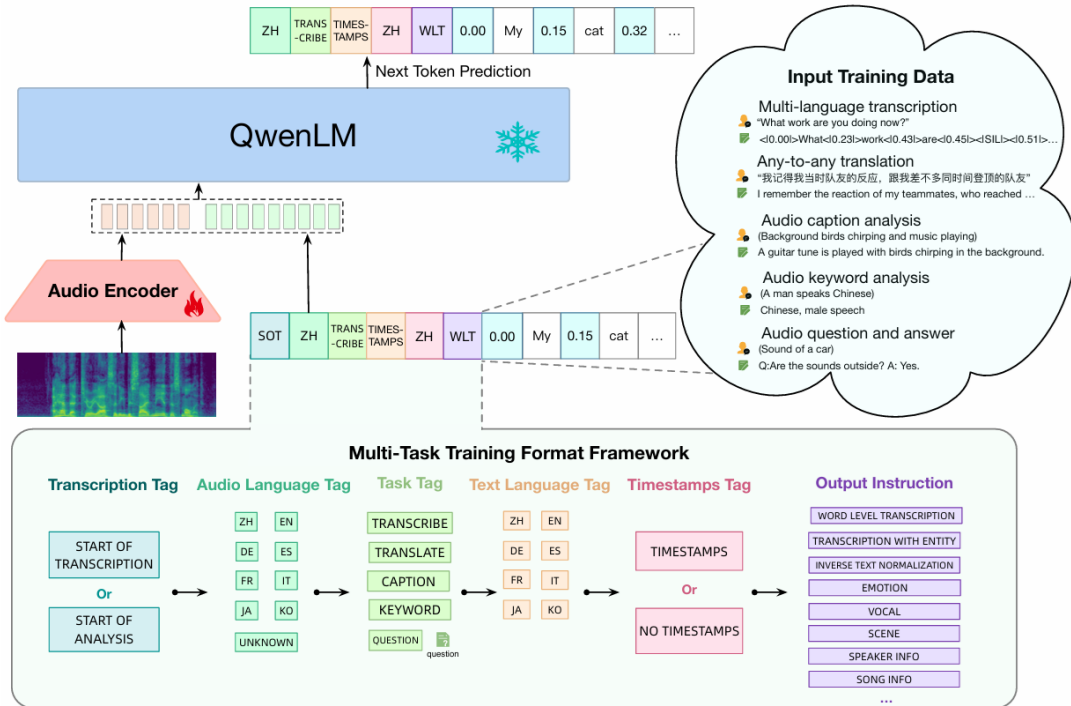


Figure 4.3.2: Qwen-Audio Architecture [15]

Although the emergent Audio-LLMs have demonstrated great potential in various audio-related tasks, as we discussed in the introduction, there is much debate concerning their underlying audio reasoning abilities that retrogrades audio understanding.

4.4 Universal Sound Separation

Universal sound separation consists on building source-agnostic models that are not constrained to a specific domain and targets at separating an unknown number of sources given an arbitrary mix. Unlike other forms of source separation where the dataset is constrained on a specific task such as speech separation or musical instrument separation, universal sound encompasses a wide range of sound sources.

Only recently was universal sound separation shown to be feasible. The stagnation of the progress on this particular field can be attributed on two major impediments:

- The available datasets were scarce and limited on the variety of sounds
- Universal sound separation is an overall challenging task

Even the process of separating audio streams by humans is an ongoing research field, so one can imagine how challenging it can be to replicate this ability in machines. One great question that arises is what constitutes an individual sound source, and this can sometimes be ambiguous depending on the task. For example, if the audio contains people taking in a restaurant, the available sources can be the speech of the people, the sound of cutlery and crowd noise, but if there is background music it may be unnecessary to separate each single instrument.

One of the first to tackle the task of universal sound separation was the work of Kavalerov et al. [41], curating a new constrained dataset and using it to investigate the space of mask-based separation architectures. They experiment with different network architectures and analysis-synthesis bases, and their findings suggest that for universal sound separation, the STFT representation is the appropriate option and the optimal window sizes are shorter in general compared to speech/non-speech separation. One of the architectures they experiment with is an improved version of ConvTasNet’s (TDCN) [51] masking network, which they refer to as *TDCN++*. This architecture is the one used also to serve as the baseline separation model on the *Free Universal Sound Separation* (FUSS) dataset [80], and the one we used in this work as well.

The ConvTasNet model, as we describe in previous chapter, functions in the time-domain, directly processing raw waveforms, using a deep convolutional encoder-decoder framework. It features three stages, *encoder*, *separation* and *decoder*. The encoder first converts the input waveform into a learned feature representation, this representation is then used to estimate the masks for each source through a *Temporal Convolutional Network* and the source waveforms are reconstructed with the decoder module.

The improved version of this architecture (TDCN++) incorporates feature-wise layer normalization over frames instead of global normalization, longer-range skip-residual connections, and a learnable scaling parameter after each dense layer initialized to an exponentially decaying scalar equal to 0.9^ℓ , where ℓ is the layer or block index.

The FUSS baseline model, is able to separate mixtures with a variable number of sources. Initially, the input audio is transformed into a spectrogram with $32ms$ window and $8ms$ hop, and the magnitude of this representation is fed to the TDCN++ network, that M masks. These masks are multiplied with complex spectrogram input, and initial separated sources \mathbf{s} are produced by applying an inverse STFT. Finally, a mixture consistency layer is applied to these initial separated source waveforms:

$$\hat{s}_m = s_m + \frac{1}{M}(x - \sum_{m'} s_{m'})$$

, which projects the separated sources such that they sum up to the original input mixture x . Since the FUSS mixtures contain up to 4 sources (as we will see later), they choose $M = 4$.

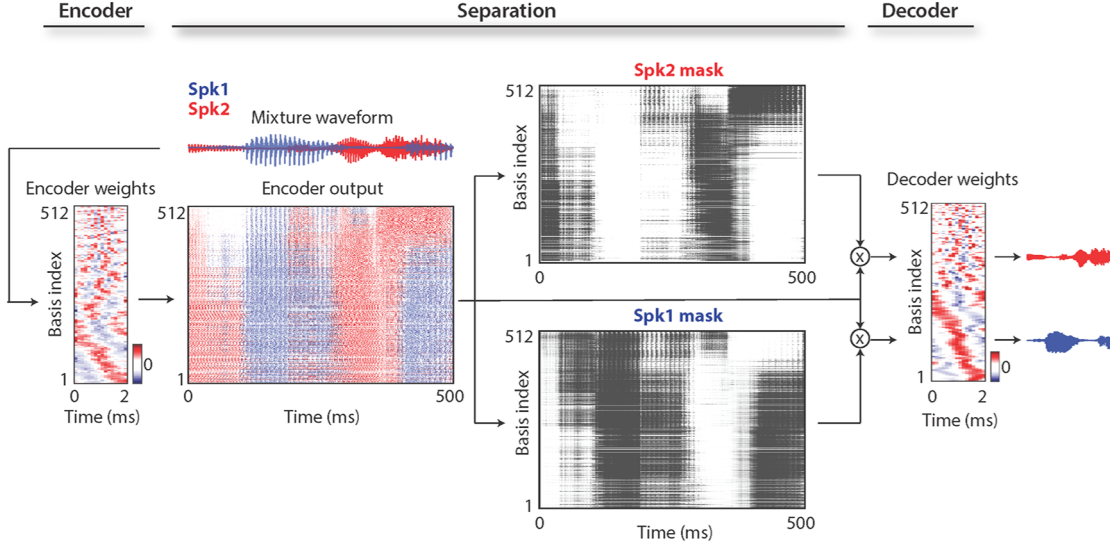


Figure 4.4.1: ConvTasNet Pipeline [51]

The loss function they employ for variable source separation is:

$$\mathcal{L}(\mathbf{s}, \hat{\mathbf{s}}) = \min_{\pi \in \Pi} [\sum_{m_a=1}^{M_a} \mathcal{L}_{SNR}(s_{m_a}, \hat{s}_{\pi(m_a)}) + \sum_{m_0=M_a+1}^M \mathcal{L}_0(x, \hat{s}_{\pi(m_0)})],$$

assuming the training mixture x consists of M_a reference sources which can be less than the output sources M , where the active per-source loss is $\mathcal{L}_{SNR}(y, \hat{y}) = 10 \log_{10}(\|y - \hat{y}\|^2 + \tau \|y\|^2)$ and the loss for inactive separated source is $\mathcal{L}_{SNR}(x, \hat{y}) = 10 \log_{10}(\|\hat{y}\|^2 + \tau \|x\|^2)$. The SNR loss, as used in ConvTasNet, measures the quality of the separated signals. Thus, to compensate the ignorance of the source order during training, they utilize the permutation invariant training technique that finds the best possible matching between the estimated sources and ground-truth sources to minimize the loss.

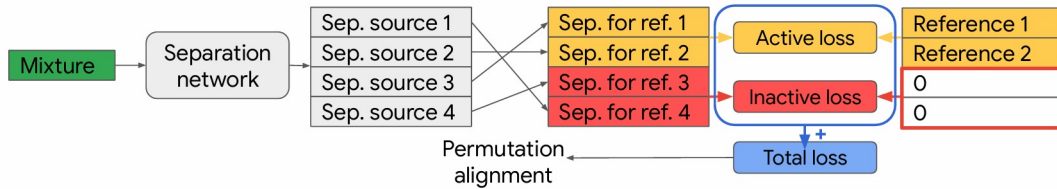


Figure 4.4.2: Variable source separation for a separation model with $M = 4$ outputs and input mixture with $M_a = 2$ active references from [80]

Chapter 5

Proposal

The process of image generation based solely from acoustic information is merely explored and the end-to-end models as we previously examined, do not have the capacity to undergo a reasoning step that we believe is necessary when facing a complex audio scene. Visualizing an image based on audio, comes down to three basic stages, as seen in the figure below.

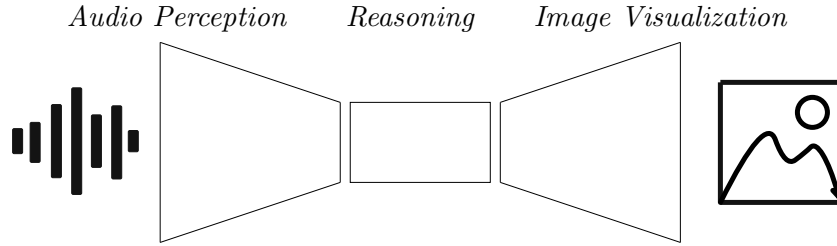


Figure 5.0.1: Process of Acoustic Imagery. Input audio (*left*) and generated image (*right*).

First, the audio is perceived and analyzed to infer the sources that are present. Then, through a reasoning process, the individual sounds as well as the relationships between them are mapped to visual concepts, describing a vivid and detailed picture that is then visualized. The visual details present in the image description are not part of the audio information, but are logically deduced from our inherent knowledge of the world. For example, if we hear a *cow moo*, it is rational to assume the audio takes place close to a farm, even though such information is not present in the audio (the physical entity of a farm does not make sounds).

In the landscape of audio-language models, there has been a significant progress in audio understanding and reasoning. However, current models have a number of limitations concerning the ability to handle complex acoustic scenes with multiple and possibly overlapping sounds, and their capacity to fully utilize the linguistic reasoning abilities of their LLM component to infer visual details from the input audio.

To this end, we aim to first investigate the ability of ALLMs to generate a meaningful and visually compelling image description based on the input audio. Then, we propose a novel pipeline, with the aim of enhancing the image description and, consequently, the image generation based on audio. In order to properly compare the two methods, we seek to find an optimal version of ALLM, comparing different prompting techniques that could possibly yield better results.

The proposed pipeline is illustrated below.

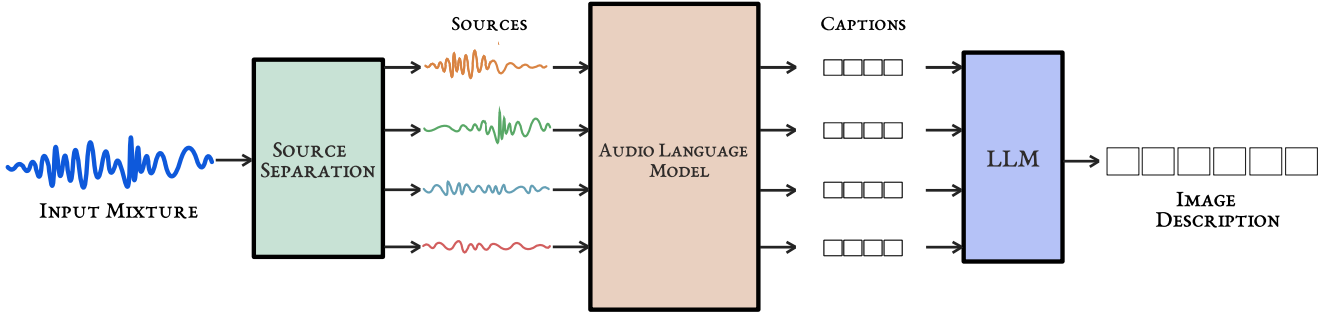


Figure 5.0.2: Proposed Pipeline

Initially, the input audio mixture is passed through an audio separation component to separate the individual sources that are present. This step, facilitates the understanding of the audio by the subsequent audio-language model, simplifying the process of the recognition and description of the audio via text. Finally, the respective captions for each of the audio sources are used as input to a Large Language Model, that will merge them into a coherent and visually rich image description, which can then be used as input prompt to an image generation model.

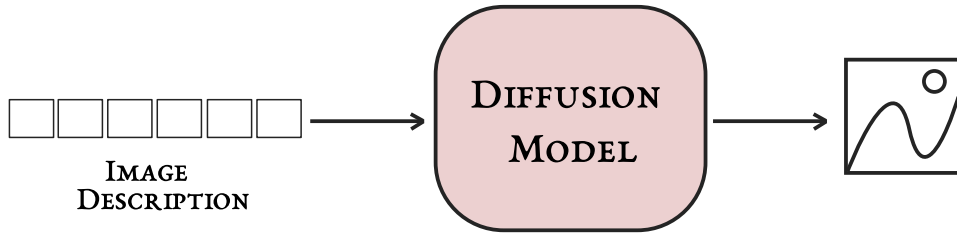


Figure 5.0.3: Image Generation Pipeline

Our hypothesis, is that by separating the individual sound sources, it will be easier for the language model to understand and perceive the content of more composite acoustic scenes, leading to a better overall result. Then, to utilize the linguistic reasoning abilities of the LLM to it's fullest, we separately include an LLM component that will analyze the text input. We hypothesize that providing a textual input to the language model, rather than an audio input, may allow the model to better focus on the nuanced instruction of visual description.

To support our hypotheses, we conducted experiments that compare the results obtained by using only the Audio Large Language Model, and using our proposed method. For the ALLM, we chose the Qwen-Audio-Chat model [15], that achieves considerable results in the MMAU benchmark [60].

The experiments are three fold:

- First, we experiment with different prompting techniques for the ALLM, with the goal of selecting the optimal candidate for comparison.

- Second, we generate text descriptions with the two methods from a dataset of complex audio samples, comparing the quality of the image descriptions.
- Third, we generate images, comparing the two methodologies with ground truth frames.

We probed various options for each individual component of the pipeline, and given the intricacy of the task, we introduced various evaluation metrics, with the purpose of adequately capturing the characteristic elements that comprise a proper image description.

Chapter 6

Experiments

This chapter exhibits the experimental process and the respective results are demonstrated. The experiments are divided into three segments:

- A Initially, we investigate different prompting techniques for the Qwen-Audio, with the purpose of finding the optimal option.
- B Next, we compare the selected option of the Qwen-Audio with our proposed pipeline, using different settings. We demonstrate the evaluation results.
- C Finally, we generate images with the selected methods and demonstrate the results.

6.1 Dataset Preparation

For the purpose of our research, we needed a dataset that contained audio mixtures of universal audio. Datasets that contain universal audio, particularly those with multiple and potentially overlapping sounds, are exceedingly limited. The samples needed to contain multiple audio sources and possibly overlapping each other to create a composite soundscape. Subsequently, we used a source separation dataset, FUSS dataset [80].

The Free Universal Sound Separation (FUSS) Dataset is a database of arbitrary sound mixtures and source-level references, for use in experiments on arbitrary sound separation. The audio data is sourced from a prerelease of *Freesound Dataset* known as (FSD50K) [26], a sound event dataset composed of Freesound content annotated with labels from the AudioSet Ontology [27].

To create the mixtures, they used Scaper [61] to generate the audio data. 10 second clips of sources are convolved with simulated room impulse responses and added together. Each 10 second mixture contains between 1 and 4 sources. Source files longer than 10 seconds are considered "background" sources. Every mixture contains one background source, which is active for the entire duration and the timing of the sound events are random. Room impulse responses are simulated using the image method with frequency-dependent walls. Each impulse corresponds to a rectangular room of random size with random wall materials, where a single microphone and up to 4 sources are placed at random spatial locations. The data is reverberated using a different room simulation for each mixture. In this simulation, each source has its own reverberation corresponding to a different spatial location. Reverberated mixtures are created by summing over the reverberated sources.

The room simulation is based on the image method with frequency-dependent wall filters. A simulated room with width between 3-7 meters, length between 4-8 meters, and height between 2.13-3.05 meters is sampled for each mixture, with a random microphone location, and the sources in the clip are each convolved with an impulse response from a different randomly sampled location within the simulated room.

In this work, we use the evaluation set of FUSS containing 1000 samples. To allow the assessment of the audio labels, we changed the script to include the labels of the FSD50K dataset, which are weak labels, meaning that they have been created using heuristic rules and they have a lower level of certainty.

Therefore, our dataset consists of 1000 samples of 10-second audio clips with $16kHz$ sample rate encoded with *Pulse-Code Modulation* on a bit depth of 16-bits and overall it contains 2,77 hours of audio. Each row of the dataset, contains the filename, the audio array, the source labels and the number of sources composed.

	filename	mixed_audio_array	source_labels	number_of_sources
0	example0000.wav	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	[Rain,Water]	1
1	example0001.wav	[0.0, 0.0, 0.0, -3.0517578e-05, -3.0517578e-05...	[Wind, Glockenspiel,Mallet_percussion,Percussi...	2
2	example0002.wav	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, -3.05...	[Harmonica,Musical_instrument,Music]	1
3	example0003.wav	[0.0, 0.0, -3.0517578e-05, 0.0, -3.0517578e-05...	[Mechanical_fan,Mechanisms, Alarm, Female_spee...	4
4	example0004.wav	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, ...	[Hiss, Wind_chime,Chime,Bell, Tick-tock,Clock,...	4

Figure 6.1.1: Dataset

There are 187 total unique labels, and in 19 instances the labels were not found. The most frequent label is:

`['Electric_guitar,Guitar,Plucked_string_instrument,Musical_instrument,Music']`,

with frequency 66 and the least frequent label is `[Bicycle,Vehicle]` with frequency 1.

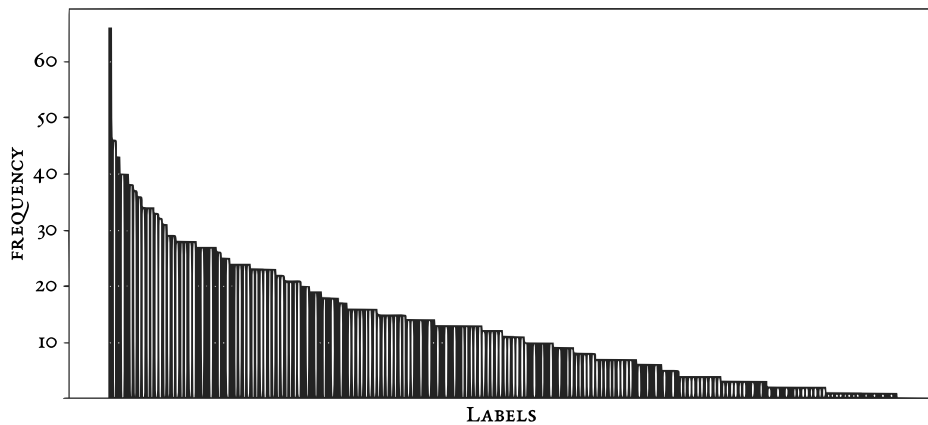


Figure 6.1.2: Label Distribution

Below, we can see the distribution of the number of sources in the audio samples.

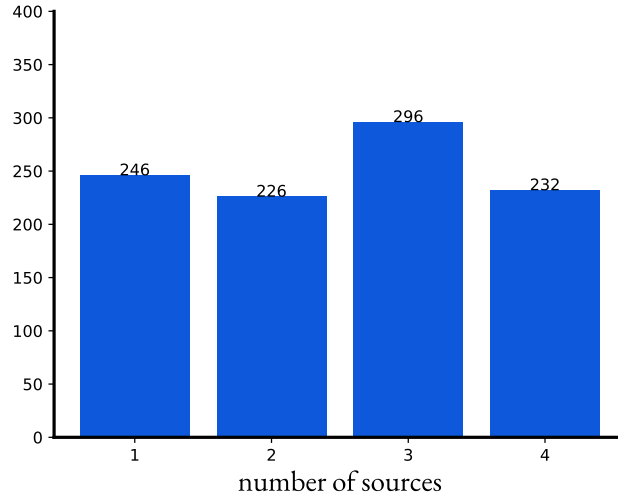


Figure 6.1.3: Number of Sources Distribution

The figure illustrates that there are 246 samples with one source, 226 samples with two sources and so on.

Dataset	Number
Samples	1000
Duration	2 hours 46 minutes
Unique Labels	187

Table 6.1.1: Dataset Statistics

6.2 Experiment A: Prompting Techniques Comparison

The first part of the experiments, is devoted on the textual description of an image through audio.

In this experimental set, we generated image descriptions through the Qwen-Audio model using different prompting techniques. Our goal is to appoint the strategy that yield superior results but without further fine-tuning. The ALLM that we use in our experiments is Qwen-Audio-Chat, which is built upon Qwen-Audio with further instruction-tuning and has around 8.4B parameters. We load the model using 4bit quantization, which does not have a major impingement on the performance and mitigates high memory consumption.

Three different prompting strategies were used, namely **Zero-Shot**, **Chain-of-Thought** and **Role**. We also experimented with **Few-Shot**, but providing examples as text audio input to the prompt mislead the model into focusing on the examples and distracted it from the audio

input. The model draws its audio input from the prompt template and completely disregards the actual audio output, leading to invalid results.

Technique	Prompt Template
Zero-Shot	Generate a visual description of a scene based on the input audio.
Chain-of-Thought	Imagine you are in a scene where you hear the input audio. Think step-by-step: 1. What are the individual sound sources present in the audio? 2. Where might these sounds typically be heard together? 3. What visual elements would be present in such a location? 4. What are the spatial relationships between the sound sources? Describe the scene you would see, based on your reasoning, including visual details and spatial relationships between any sound sources.
Role	Imagine you are a filmmaker who needs to create a visual scene that matches the input audio. Write a concise scene description that captures the atmosphere and visual elements that correspond to the audio.

Table 6.2.1: Prompt Templates

For this experiment only, we selected 100 random samples from the original dataset in order to compare the different strategies, keeping the label and number of sources distribution approximately the same.

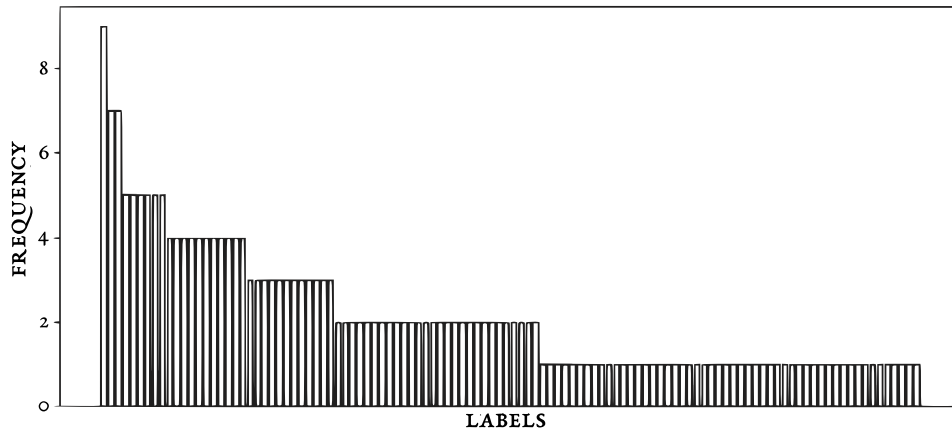


Figure 6.2.1: Label Distribution for the small dataset

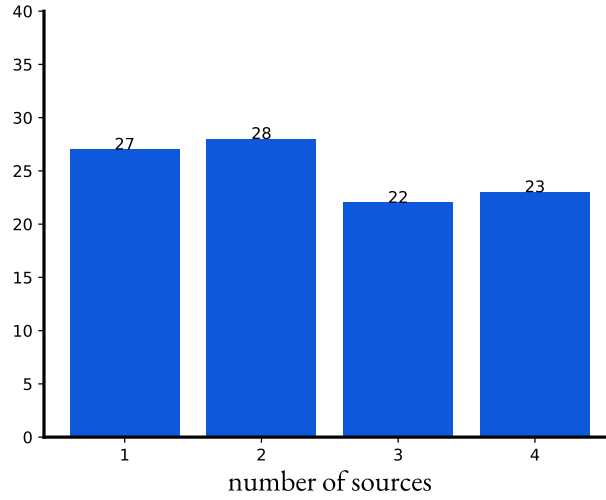


Figure 6.2.2: Number of Sources Distribution for the small dataset

6.2.1 Qualitative Results

In the table 6.1.1 below, is an example of the responses of each strategy given an audio input with the following source labels:

```
['Giggle,Laughter,Human_voice']
['Acoustic_guitar,Guitar,Plucked_string_instrument,Musical_instrument,Music']
['Cowbell,Percussion,Musical_instrument,Music,Bell']
```

Attention should be given to the response of the Chain of Thought approach. The model generated a detailed explanation and not a concrete description. On one hand, this thorough explanation is necessary at explaining each interaction and detail in the audio, on the other hand it is very rigid, lacking the imaginative qualities needed. Another thing worth mentioning is the CoT technique recognized the guitar sound, where the other two methods failed.

6.2.2 Evaluation Methods

Below, we introduce the evaluation methods that were used to assess the results throughout this experiment and the subsequent experiment B. The task of describing a visual image based on the input audio is a complex process, encompassing a gamut of constituent parts. One must first perceive each sound in the mixture, recognize the source of each audio stream, reason about the intricate acoustic properties, translate these acoustic insights into visual scenes and aptly describe the image. Thus, the evaluation of the text descriptions should reflect the complexity and the nuance of the task and take into account each important factor.

CLAP score

For this experiment, one of the evaluation metric used is CLAPscore. Similar to CLIP, the CLAP model [24] consists of an audio and a text encoder. Through contrastive learning, the

Technique	Response
Zero-Shot	A baby is laughing while a fork clatters on a plate.
Chain-of-Thought	The input audio contains two separate sound sources: a baby laughing and a guitar strumming. The baby laughing is the primary sound source, while the guitar strumming is a supporting sound source. These two sound sources are likely to be heard together in a music room or a classroom where a baby is playing with a guitar. The visual elements present in such a location could be a baby sitting on a guitar case, a teacher playing the guitar, or other babies and children playing with musical instruments nearby. The spatial relationships between the sound sources could be close, with the baby’s laughter and the guitar strumming occurring simultaneously or in close proximity to each other.
Role	A baby laughs as a metal object strikes a surface.

Table 6.2.2: Responses of each prompting technique

model learns to correlate in the latent space the audio-text embeddings that are similar, and push away the audio-text embeddings that are dissociated. Therefore, we extracted the input audio embedding and the text response of the ALLM and we measured the cosine similarity. The cosine similarity indicates the alignment between two vectors and consequently how much their context match, without being affected by the magnitude of the vector’s magnitude like dot product. The range of the values is $[-1,1]$. Cosine similarity closer to 1 indicates high similarity between the audio and the text.

$$\text{cosine similarity} = S_C(A, B) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

Note that, while we use the CLAPscore for this experiment to compare the different prompting techniques, the score is probably not adequate to encapsulate the similarity between the audio and a proper image description, mainly for two reasons:

- First, the model is not trained to handle very complex acoustic scenes
- Second, an acceptable visual description may not serve as an audio caption, deviating from the strict description of the audio alone.

We used the *clap-htsat-unfused* model.

LLM-as-a-Judge

Recently, LLMs have seen extraordinary progress. Naturally, people started using them as a way to evaluate results in aspects that cannot be evaluated otherwise [32]. It is a cost-effective alternative to human evaluation, which establishes the golden standard of evaluating responses close to human perception. For the LLM judge we used Mistral-7B-Instruct [38], a highly used model for this purpose.

We introduce the following evaluation metrics:

- **Audio Understanding**

- **Source Perception:** This metric measures the ability of the model to perceive the sounds in the audio input. Due to the vague nature of the response, we employ an LLM to assess the number of different sounds that are present in the text description, which indicates the number of sounds the model perceived. Comparing this number to the true number of sources, we can infer the model’s capacity to handle complex audio inputs with overlapping sounds. We draw inspiration from the source separation task, and we measure the percentage of under separated samples *US*, samples that the model perceived fewer than the actual sounds, equal separated *ES* (same number of sounds) and over separated *OS* (greater than the true number of sounds).
- **Source Recognition:** With source recognition, we aim to evaluate the capability of the model to not only perceive the individual sounds, but to also recognize the source. The reason for the two metrics concerning the perception of audio, is because the model might be able to perceive a sound but misjudge the source, something that naturally can happen even to humans. We employ the LLM to assign a binary label (Yes/No) to the model’s responses, as to whether it incorporates the source labels in the description. Then, we measure the accuracy.

$$Accuracy = \frac{\text{number of correctly recognized}}{\text{total number of samples}}$$

- **Image Description**

We draw a distinction between the competence of the model to perceive and understand audio, and it’s quality of describing a visual scene. Portraying an image through text consists of an intricate mixture of properties.

- **Coherence:** The image description should make logical sense and be coherent. Given the recognized sources, the scene described by the model should have a reasonable connection to the sounds involved. We utilize the LLM-as-a-Judge to assign a score (1 – 10) with 10 being highest, to each response.
- **Visual Details:** Furthermore, it is natural to assume that a proper image description should be visually rich, painting the whole picture as vivid as possible. We follow the same scoring as before.
- **Imagination:** This metric aims at evaluating the ”truthful verbosity” of the responses. The process of imagining an image based solely from sound, incorporates some level of imagination, or hallucination, but an image to be valid, it should be grounded on the acoustic information. Consequently, the LLM-as-a-Judge assigns high score (close to 10) if the model ”hallucinated” details that are not present in the

audio but can be logically deduced from the audio, and low score if the details are fictional.

- **Image Appropriateness:** Finally, we measure the overall quality of the textual response as an image description and its suitability as input to an image generation model. Again, we use the $(1 - 10)$ measure scale.

In our experiments, we define as *level of difficulty* the number of sources in the audio, assuming that larger number of sources becomes an increasing challenge for the model.

6.2.3 Quantitative Results

CLAPscore

The figure below shows the cosine similarity values for each prompt strategy per number of sources in the audio.

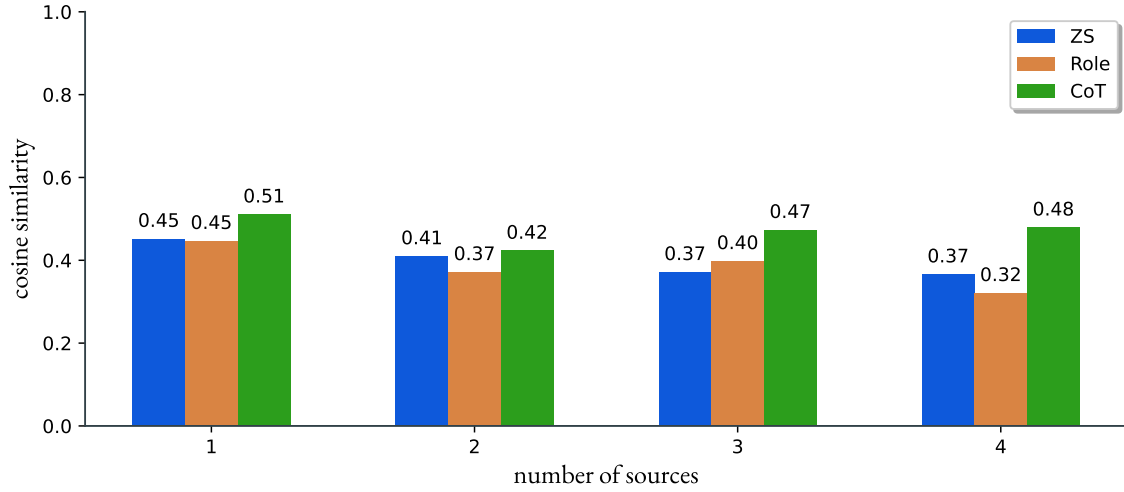


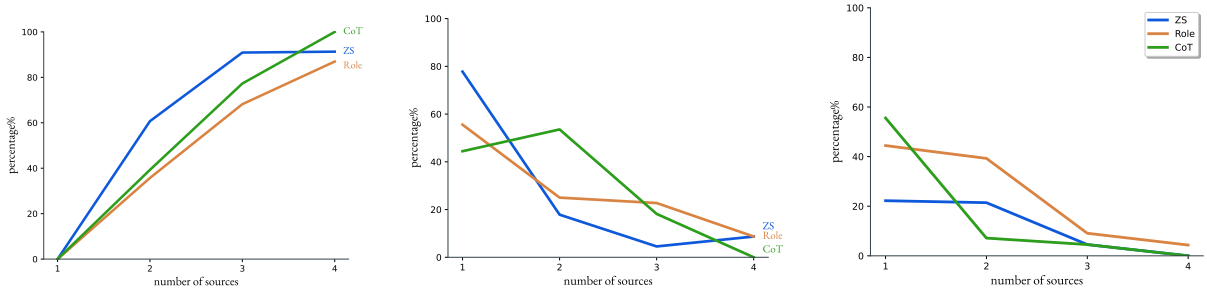
Figure 6.2.3: CLAPscore

As we can see, all strategies have similar results with the CoT method performing slightly better. However, all techniques have low similarity score which is a result of both the inadequacy of the model to generate valid responses, the inability of the CLAP model to encode the complex audio and also the intricacy of the task. The responses are not audio captions but (supposedly) image descriptions based on audio. Therefore, despite the fact that we demonstrate the results, it should be noted that CLAPscore may not be a valid evaluation approach. It indicates, nonetheless, the variations in each prompting technique.

Source Perception

We plot the percentage of US, ES and OS for each number of sources.

As we can see, as the number of sources in the input audio increases the model struggles to identify the exact number of sources in the audio (ES) and it perceives fewer and fewer

Figure 6.2.4: Under Separation (*left*), Equal Separation (*center*), Over Separation (*right*)

sources than there are actually in the audio (US). When there are four different sounds in the input mixture, the percentage of the correctly perceived samples for all prompting methods are minuscule.

Figure 6.2.5, shows the overall performance in source perception of the model for each prompting strategy.

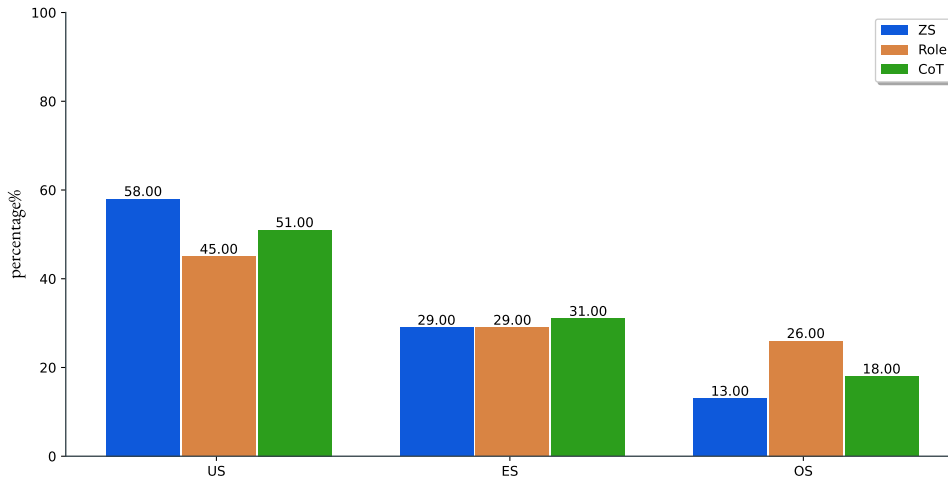


Figure 6.2.5: Source Perception for Each Prompting Technique

The US values, indicate the inability of the model to perceive the correct amount of sounds and the ES values the exact opposite. The OS values on the other hand, indicate the extent to which the model engages in hallucination, as it generates more sources than actually exist. The findings reveal that the model exhibits limited hallucination, which is a desirable outcome.

Despite evaluating the prompting strategies on the task of perceiving the sounds in the audio, is safe to say that we do not expect great discrepancies between the different prompting strategies, because different prompts can not have a large impact on the audio perception which is performed by the audio encoder. This is illustrated also by the uniform distribution of the equal separation, where all three of the prompts yield consistent results.

Source Recognition

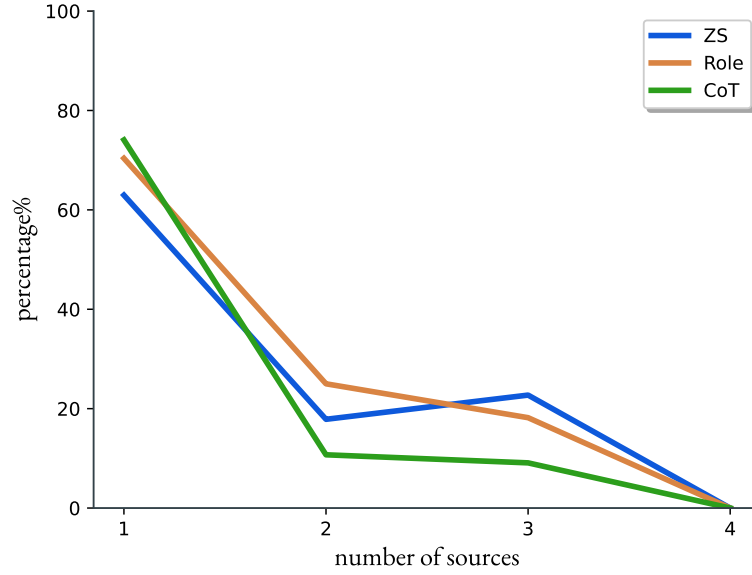


Figure 6.2.6: Label Accuracy

Similar to source perception, the model struggles to correctly identify the sound sources as the number of sources increases, but overall the prompting methods perform approximately the same.

In order to test the ability of the LLM judge to infer the number of sources and the label inclusion from the text description alone, we manually went through each sample annotated each information from the text description only. Due to the large amount of data, we chose the CoT method. We found that the LLM-as-a-Judge was **89%** accurate at judging the number of sources from the text response of the ALLM alone, which is pretty precise. We also found that it was **84%** accurate at discerning the inclusion or not of all source labels. This is a somewhat high score, given that the labels provided by FSD50K are weak and don't paint the whole picture of the sound. Some sounds are pure sound effects that are hard to categorized even by humans. Therefore, we concluded that the LLM-as-a-Judge is a valid process of evaluating the responses of ALLM. Keep in mind that the value of the LLM-as-a-Judge's accuracy at scoring the inclusion of all labels is subject to noise, which is the human error itself, as we could possibly be wrong with some of our answers.

Image Description

In the coherence metric, the *Chain-of-Thought* technique dominates with a mean score of 8.27 over the other methods. Especially, the *Role* method yields the lowest score. A reason for this might be because the role we assigned to the model is of a filmmaker and the responses might be too artistic.

It is also clear, that the responses of the model do not entail the visual vividness appropriate to an image description. The highest score achieved with a prompting technique is 5.37, signifying the incompetence of the ALLM to generate a visually rich description.

The *Role* strategy has a slight superior imagination score, but all methods perform similarly.

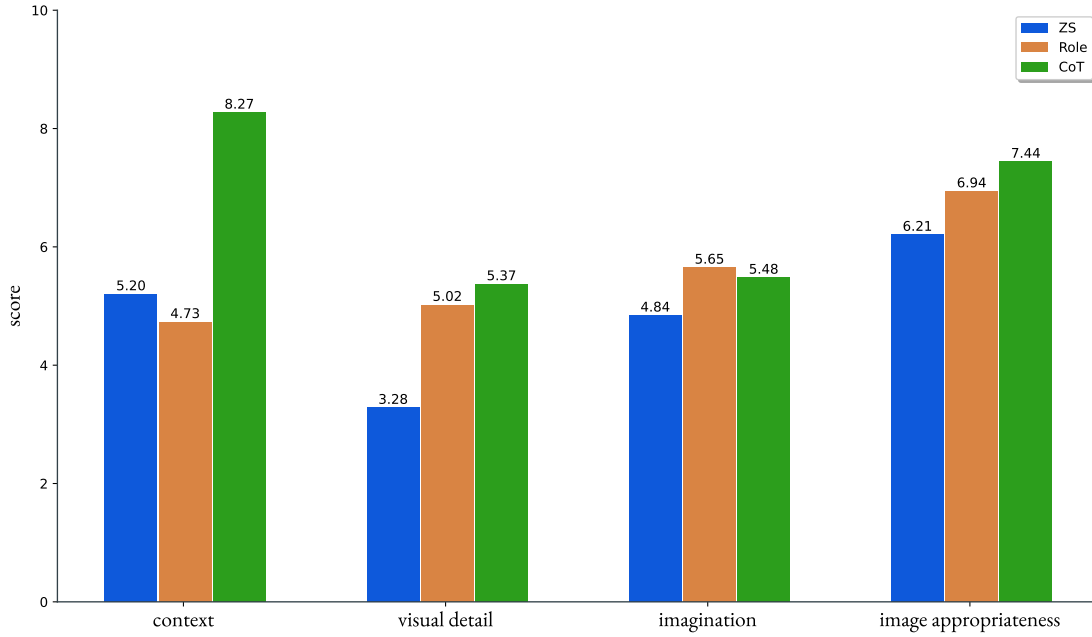


Figure 6.2.7: Image Description Metrics

Overall, the *Chain-of-Thought* prompting technique, achieves the best results. Therefore, we choose to use the ALLM with the CoT prompt for the next set of experiments. Notably, the *Zero-Shot* prompt yield the lowest score in the majority of the metrics, emphasizing the improvement caused through prompting techniques. This elucidates the fact that describing the visual scene based on audio can benefit from a step-by-step reasoning process, where the model decomposes the problem into intermediate components focusing on each individual aspect.

6.3 Experiment B: Comparison with Proposed Method

So far, we have used only the ALLM as a means of describing an image based on audio. In this set of experiments, we employ our proposed pipeline to generate visual narratives, and compare different settings for each component of the proposed pipeline with the ALLM.

We use the same evaluation metrics used as before for the LLM-as-a-Judge but we skip the CLAPscore due to the discrepancies explained earlier.

We partition the evaluation and comparison of the proposed pipeline to the ALLM in two sections.

Audio Understanding

The first section, is concerned with audio understanding, which consists of the audio perception and recognition. We contemplate the two approaches, showcasing any improvements that the source separation component may have induced in approaching complex acoustic scenes.

As we explained in the previous section, our proposed pipeline 5.0.2 consists of an audio source separation component, an audio-language component and an LLM. The source separation component and the audio-language model are responsible for the perception and understanding

of the acoustic insights of the input. For the source separation we used the baseline model of the FUSS dataset [80]. This model, based on TDCN++, generates four audio waveforms regardless of the true number of sources of the audio. For example, if the audio contains two sources, the model will generate four estimates, two of which will be zero (silence) or nearly zero (might have artifacts). To eliminate the waveforms that are zero or practically zero and only pass the active sources through the audio-language model, we propose a filtering method.

To filter out the non-active separations we compute the sum of the audio and it's max value and return the product. If the product is above a threshold we preserve the audio, otherwise we discard it. From experimentation, we found that a good threshold is 0.1. Therefore, the sum value rejects short high intensity audio abnormalities and the max value rejects long, low intensity constant noise. We also compute the *root mean square* (RMS) value of the signal and compare it with the RMS value of the original audio. The RMS value of an audio signal, is a measure of its "loudness".

$$RMS = \sqrt{\frac{1}{N} \sum_{i=1}^N x_i^2}$$

RMS is an important measure in signal theory, as it describes the dynamic range of the audio. High rms value indicates an active signal. Because every sample has different rms value, we compare the rms value of the estimated source and the original sample and if the estimate is within 10% of the original rms, it is accepted. It is important to note that with this method, in 2 out of 1000 samples (0.2%) none of the sources were accepted, but we did not lower the threshold because other non-active sources would then pass through.

For the audio-language component, we experimented with two different models. The first one, is the same we used in the experiment A, namely Qwen-Audio-Chat, which has the ability to meticulously describe the audio. We used the prompt "*Describe the audio in detail*". However, this model is quite large in size, having 8.5B parameters, therefore we investigated whether the performance would not deteriorate by utilizing a smaller audio captioning model. We performed the same experiments, using the EnCLAP-large model [43] which has around 620M parameters.

Automated Audio Captioning (AAC) is an emerging audio perception task that involves recognizing environments, sound events and then describing them in fluent sentences. EnCLAP used two acoustic representation models, EnCodec [17] and CLAP [24], together with a pre-trained language model, BART [49], as a captioning decoder. EnCLAP is trained using masked codec modeling, guiding BART to learn the relationships between the encoded audio, using AudioCaps [42] and Clotho [22] datasets During training, they employ cross-entropy loss between the correct caption and the generated caption.

To measure the number of sources in the description, we counted the number of descriptions of each sample. The figure 6.3.1 below, compares the performance of audio perception between the proposed pipeline and the ALLM.

The results, even though comparable, justify the inclusion of a source separation component. As we can see, the proposed pipeline achieves better results in all three of the metrics. It yields 36% improvement on equal separation and even diminishes the hallucination incidents. It is evident, that the audio-language model will benefit from a component that can preprocess a complex audio, into its constituent sounds.

The improvement will also be highlighted plotting the equal separation across the number of sources.

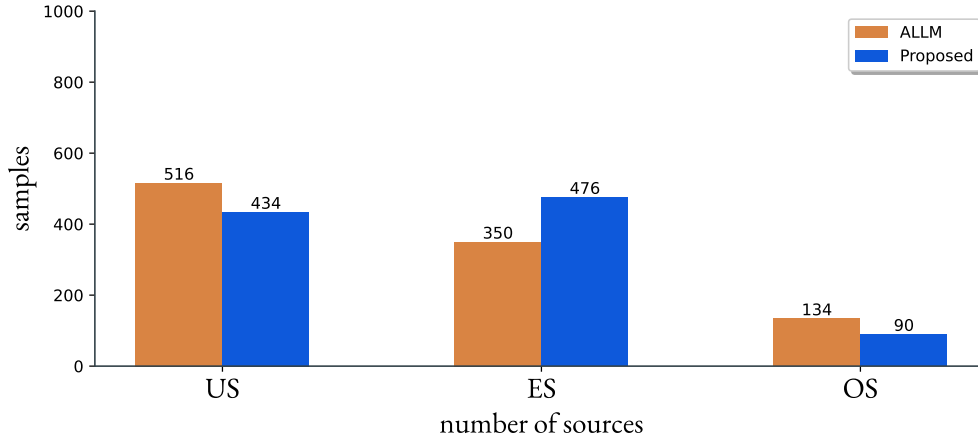


Figure 6.3.1: Source Separation Metrics

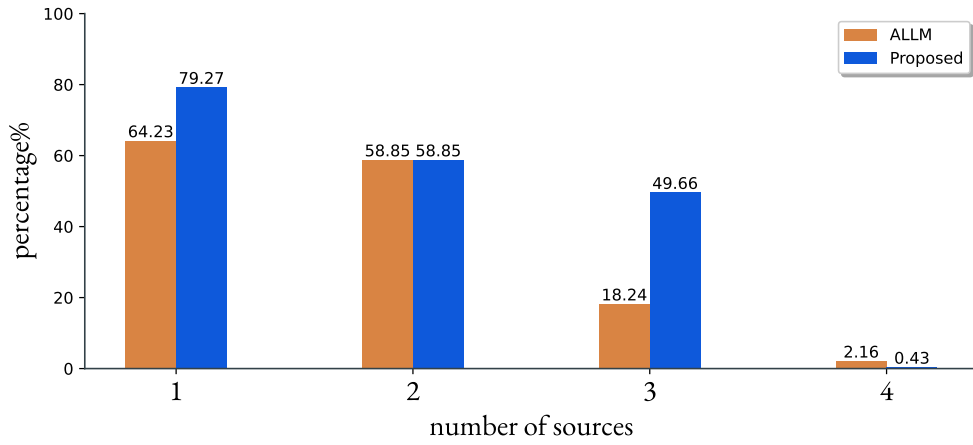


Figure 6.3.2: Equal Separation

When the number of sources are three, the proposed pipeline manages to perceive far more sounds than the ALLM alone. Overall, however, the results are proportional.

To evaluate the recognition of each sound based on the weak labels, we opted for a slightly different approach than the ALLM. While the ALLM responses were direct and easy to decipher the number of different elements, the descriptions of our proposed method are more in line with an image description, thus they impose a challenge to identify the number of acoustic sources. Consequently, we evaluated each individual description of the audio-language component, and considered a sample as "valid", if all the descriptions of the sample had correctly predicted the label. We evaluate the results of the proposed pipeline with the EnCLAP model and the Qwen-Audio-Chat, figure 6.3.3.

Again, the results between the proposed pipeline and the ALLM are comparable, with improvement starting to appear above two number of sources, where Qwen-Audio-Chat manages to recognize accurately more samples than the ALLM, and at four number of sources, both EnCLAP and Qwen-Audio-Chat are able to identify slightly more samples correctly. The EnCLAP

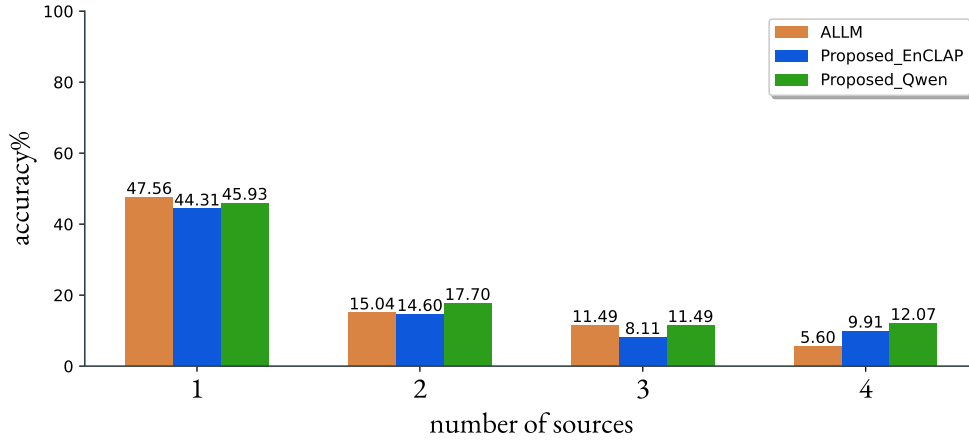


Figure 6.3.3: Label Accuracy per Number of Sources (We define as *Proposed_EnCLAP* the proposed pipeline using the EnCLAP model and *Proposed_Qwen* the proposed pipeline using the Qwen model)

model despite having 7.3% of the Qwen-Audio-Chat parameters, has a solid performance.

Image Description

The individual captions of the audio-language model are then passed through the LLM to merge them into a unified description of an image where the sounds could be heard. For the choice of the LLM, we experiment with three different LLMs and two different prompting techniques, a zero-shot prompt and a chain-of-thought prompt. We use the Qwen 2.5 7B-Instruct [55], the Mistral 7B-Instruct [38] and the Falcon3 7B-Instruct [71] models. Together, in total we have a set of nine different results.

Audio-Language Model	LLM	Prompt	Size
Qwen-Audio	Qwen LLM	zero_shot prompt	15.5B
EnCLAP	Qwen LLM	CoT prompt	7.6B
EnCLAP	Qwen LLM	zero_shot prompt	7.6B
Qwen-Audio	Mistral LLM	zero_shot prompt	15.5B
EnCLAP	Mistral LLM	CoT prompt	7.6B
EnCLAP	Mistral LLM	zero_shot prompt	7.6B
Qwen-Audio	Falcon LLM	zero_shot prompt	15.5B
EnCLAP	Falcon LLM	CoT prompt	7.6B
EnCLAP	Falcon LLM	zero_shot prompt	7.6B

Table 6.3.1: Proposed Pipeline Settings

We try the zero shot prompt with both audio-language models but experiment with the

chain-of-thought prompt using the EnCLAP model, which has significantly smaller size. Given that the separation module has around $9M$ parameters, the overall size of the proposed pipeline with the EnCLAP audio-language model is around $7.6B$, $1B$ smaller than the initial Qwen-Audio-Chat model.

In the figure below, we show the results for each evaluation metric of image description.

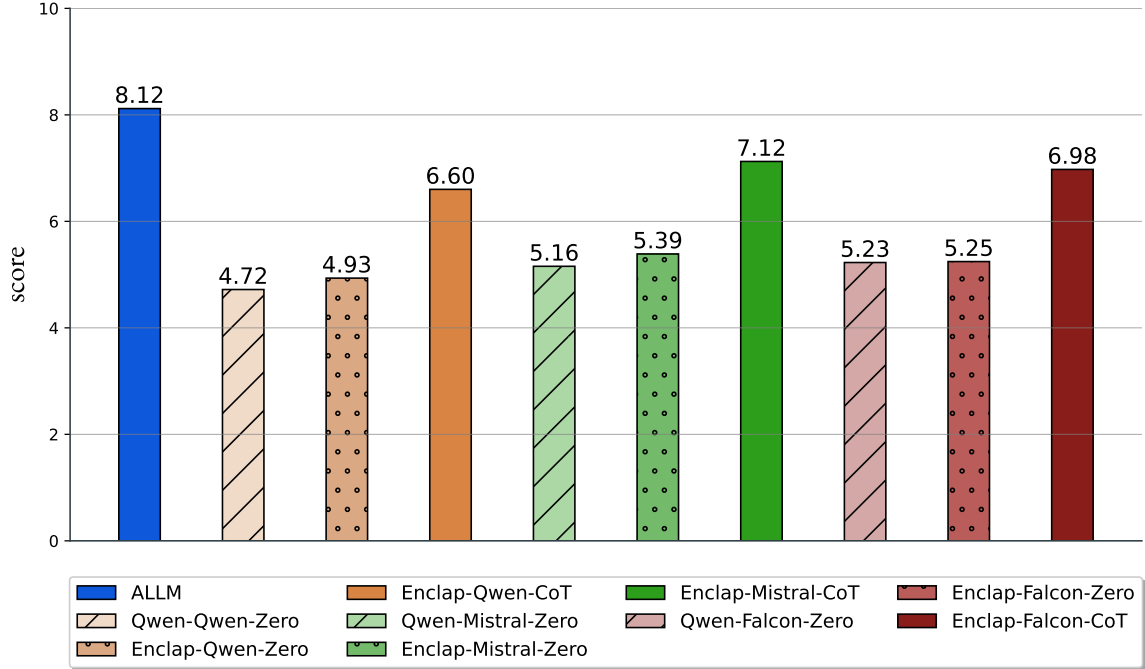


Figure 6.3.4: Coherence Mean Score for Each Method

In terms of consistency, the ALLM alone manages to surpass our proposed method. The output of the ALLM is more coherent and grounded to the audio, and its structure provides as it seems a more coherent answer than the other methods. We assume, this is also caused by the fact that the generated text of the ALLM is more like a thorough explanation than a visual description, as it can be seen in the examples in the appendix.

Furthermore, the generated descriptions of our proposed methodology offer a detailed and nuanced representation of the visual elements rather than a superficial and rigid audio-visual mapping. Mistral achieves the highest score using EnCLAP and Chain-of-Thought prompting 6.3.5.

Falcon with EnCLAP and CoT prompt, seems to outpace the other methods in terms of imagination, managing to hallucinate reasonably enough visual details that can be inferred from the audio. Again, our proposed methods accomplish highest score compared to the ALLM 6.3.6.

Finally, the figure 6.3.7 shows that our proposed method’s descriptions offer a more suitable image description in total, contributing to a more suitable portrayal of a visual scene than can aptly guide an image generation model.

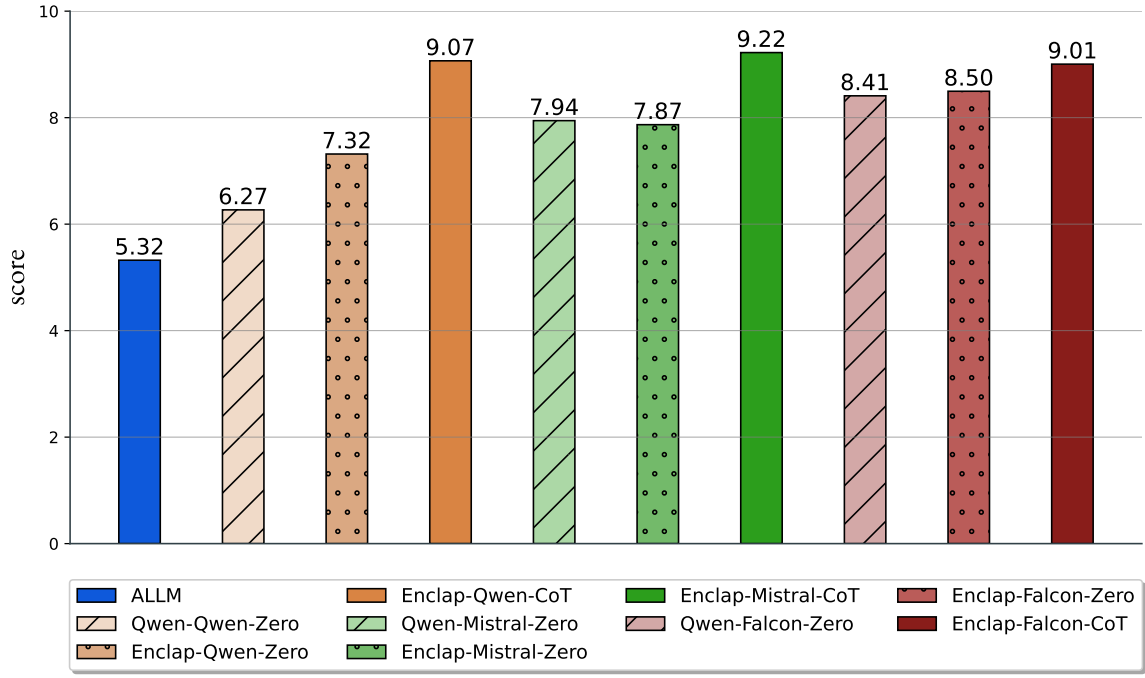


Figure 6.3.5: Visual Detail Mean Score for Each Method

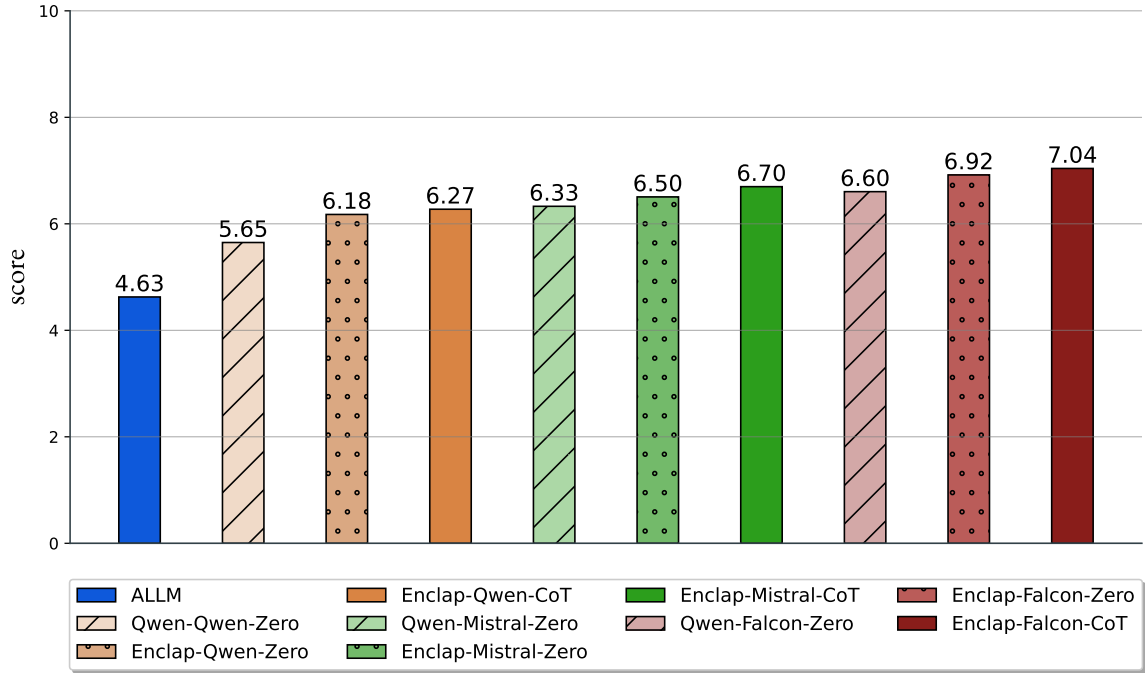


Figure 6.3.6: Imagination Mean Score for Each Method

Overall, LLM-as-a-Judge has interpreted the descriptions of our proposed pipeline as visually more compelling and suitable than the end-to-end audio large language model. Though the ALLM has the capacity to understand sounds and their spatial relationships, it lacks the capability to leverage the semantic knowledge of the language model component to organize this

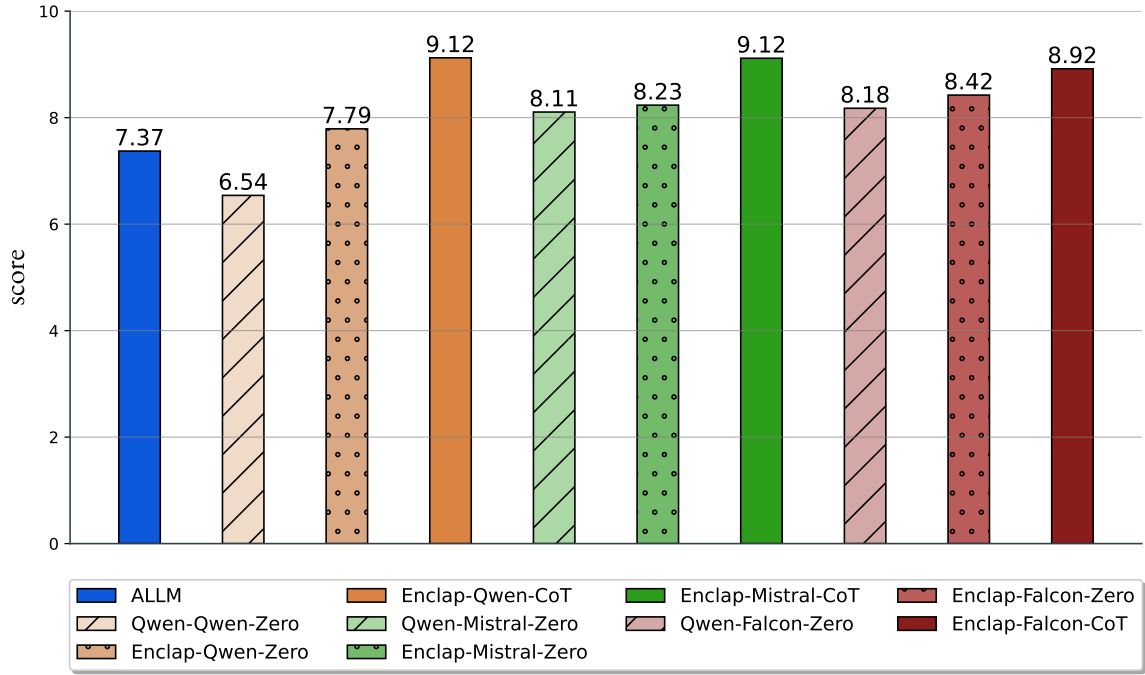


Figure 6.3.7: Image Appropriateness Mean Score for Each Method

auditory perception into a meaningful visual representation. Analyzing the different choices of our pipeline, we can draw the following conclusions:

- In general, the chain-of-thought prompt achieves the highest results among all LLMs.
- Leveraging the EnCLAP model as an intermediate audio-language step yield superior results.
- Among LLMs, Mistral provides the more coherent and visually detailed description.
- Falcon is a more consistent and robust LLM, producing uniform results across different settings.

So far, we have analyzed the capacity of the audio large language model and our proposed pipeline to provide a visually adequate description of a scene based solely on a complex audio input, being able to translate the auditory insights into valid visual representations that are reasonable according to our commonsense. In the following section, we are going to implement this process in a somewhat real-world experiment, generating images from auditory scenes.

6.4 Experiment C: Generating Images from Movie Scenes

In the final stage of the experimental process, we implemented the Qwen-Audio-Chat model and two of the most effective configurations from our proposed approach, in an image generation assessment involving complex acoustic environments.

6.4.1 Dataset

We drew samples from *Condensed Movies* dataset [4], which contains video clips from popular movie scenes. We manually went through examples and selected clips that contain a composite acoustic scene, with a variety of overlapping environmental and speech sounds. We opted for scenes that are amenable to mental imagery, but still challenging.

We handpicked 22 example clips and trimmed the video clip in the appropriate time frame of 10 seconds. Then, we extracted the audio and selected a representative frame that depicts most of the acoustic elements for further evaluation.

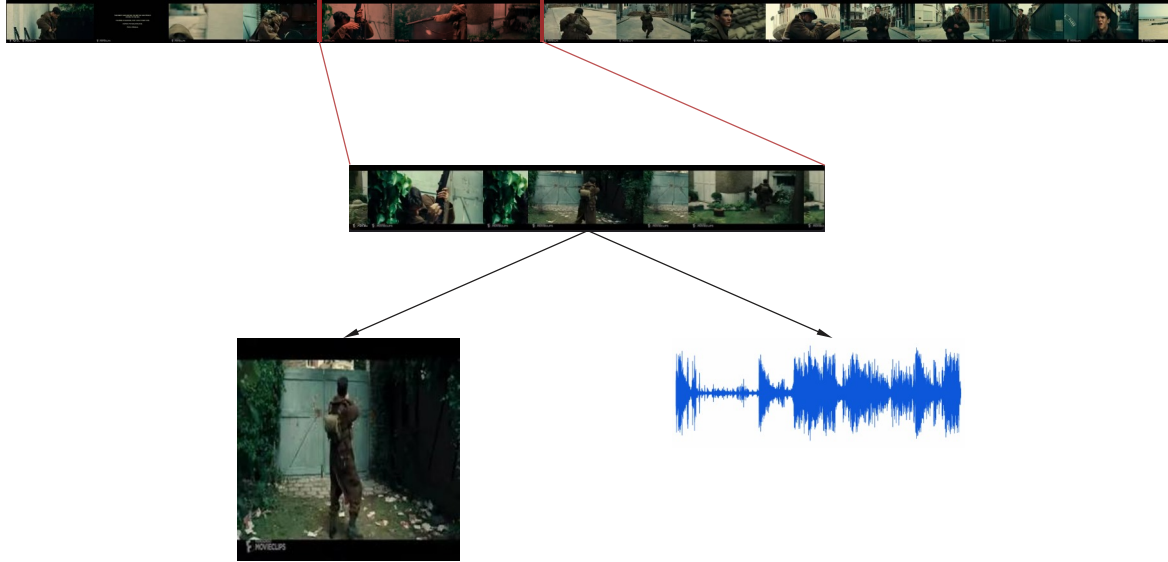


Figure 6.4.1: Samples Preparation

6.4.2 Human Evaluation

We chose as our method’s configurations the *EnCLAP-Falcon-Zero_Shot* and the *EnCLAP-Mistral-Chain_of_Thought*, which yield the highest rating from the previous experiments among the zero-shot and chain-of-thought prompt respectively. Together, with the initial Qwen-Audio model, we generated textual descriptions through the extracted audio of the movie samples. Afterwards, using the textual outputs of the models, we generated images leveraging *Stable Diffusion 3.5 Medium* [25]. To effectively and qualitatively evaluate the text descriptions, which describe the visual scene of the audio, and the generated images as well, we conducted a human evaluation.

First Part

The first part of the human evaluation aims at assessing the text descriptions among the three methods. The evaluators were given three examples of input audio and three text descriptions, in random method order, and were asked to select one of three methods for each of the three below questions:

1. *Which description fits better given the input audio?*
2. *Which description is visually more detailed?*
3. *Which one of the three is overall better?*

The total participants were 37. The figures below show the percentage of the evaluators option between each method, for each question.

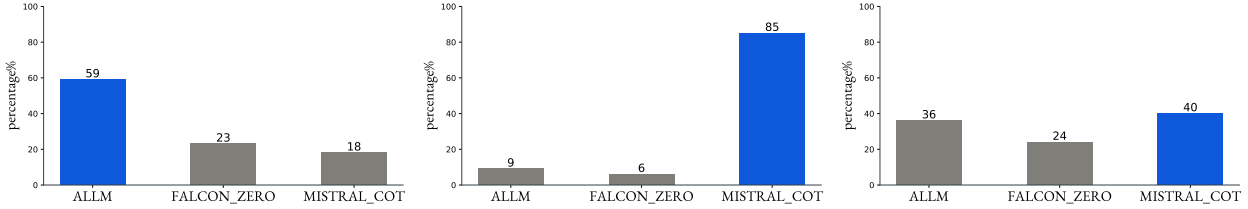


Figure 6.4.2: Suitability According to Audio (*left*), Visual Detail (*center*), Overall Quality (*right*)

Regarding the appropriateness of the textual description to the input audio, the majority of the evaluators chose the ALLM model. Although this result may reflect the true performance between the models, it is important to mention a possible bias concerning this specific question. The output of the Qwen-Audio, in contrast to our proposed method’s outputs, is more straightforward and more explanatory. At first, it states the sounds prevalent in the audio and then proceeds at explaining other attributes. This trait, may elicit a bias to the evaluators towards this description, despite the correctness of the other descriptions.

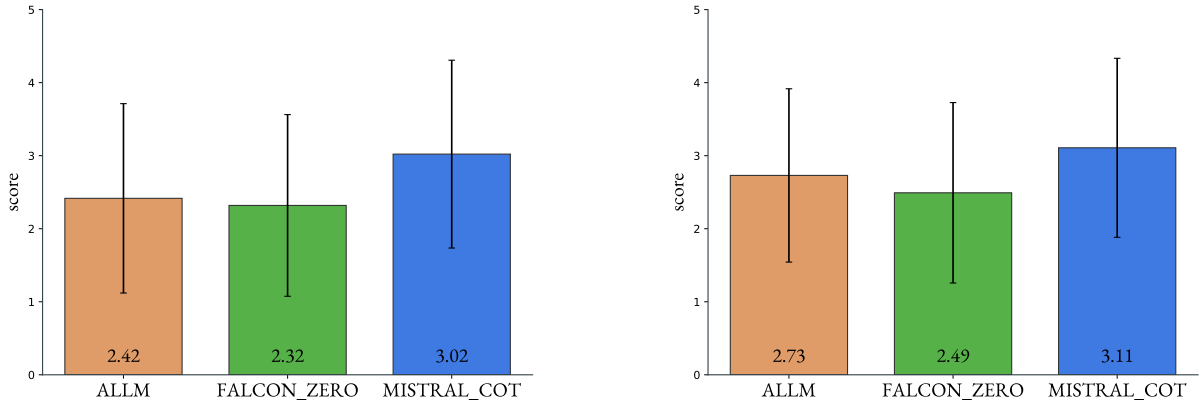
For the next two questions, especially for the visual detail, the mistral-cot configuration is undoubtedly the most favorable. Overall, this method seems to yield better descriptions of the visual scene rather than explicitly reciting the audio.

Second Part

In the second stage of the user study, the evaluators were given five examples of an input audio and three generated images from the textual descriptions of each method. For each image, they were asked to rate from 1(lowest) to 5(highest), for each of the metrics below:

1. *Similarity to the input audio*
2. *Realism*
3. *Detail*

Concerning the similarity to the input audio, our method with mistral and chain-of-thought achieves higher score among the others, however all three do not manage to accomplish high rating.

Figure 6.4.3: Similarity to Audio (*left*), Realism (*right*)

It is worth mentioning that, for this metric as well as the following, the evaluation of such task is prone to subjectivity. Each individual is different and may pay attention to distinct elements of the audio-image pair. Also, each person could have followed a different process of evaluation, affecting the results.

Regarding the realism of the images, the results are similar to the previous metric. Our method is rated with a highest overall score with a mean value of 3.11, but in general all three have comparable scores. Finally, for the level of the detail of the images we have analogous results. In total, our proposed pipeline, using Mistral and Chain-of-Thought prompt, yields the best results.

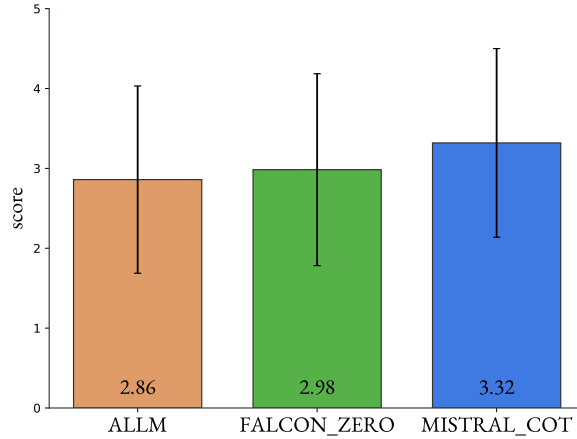


Figure 6.4.4: Level of Detail Mean Score for Each Method

6.4.3 CLIPscore

To quantitatively measure the alignment of the generated textual descriptions to the true image frames from the scene, we computed the CLIPscore for each method using CLIP [56]. We extracted the text embedding of the textual descriptions and the image embeddings of the ground truth frames, and computes the cosine similarity

$$\text{cosine similarity} = S_C(A, B) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|}$$

Method	CLIPscore \uparrow
Qwen-Audio	0.2279
Falcon-Zero_Shot	0.226
Mistral-Chain_of_Thought	0.2047

Table 6.4.1: CLIPscore

The low level of CLIPscore for all three of the methods, indicates the intricacy of the task and the incompetence of quantitative measures and the objective evaluation of the process in general. The visual scene from the film is subjective based on the audio and the selection of the representative frame may not be ideal. This, further emphasizes the importance of the qualitative evaluation provided by human assessment, contributing to a more nuance understanding of the performance.

However, this slight improvement of the ALLM CLIPscore, could be attributed to the forwardness of the textual descriptions, which does not deviates from the input audio with details. The model, as we mentioned, does not have the capacity to imagine in depth a visual scene from the audio, which explains its advantage over the other methods in objective metrics like CLIPscore.

Qwen-Audio

The input audio contains two sound sources: sizzling and a woman speaking. The sizzling sound is coming from a pan or a frying pan, and the woman is speaking. The woman is speaking in a foreign language, which makes it difficult to understand what she is saying. The sizzling sound is a common sound that is often associated with cooking, especially frying food. The woman's voice is faint and can be heard in the background. The visual elements that would be present in such a location would likely include a kitchen with a stove, a frying pan, and possibly a person cooking or frying food. The spatial relationship between the sound sources is that the sizzling sound is coming from the stove, and the woman's voice is coming from somewhere nearby."



Falcon_Zero

"Imagine a dynamic scene where the sound of sizzling and crackling fills the air, suggesting a stovetop or a campfire, while the unmistakable scrape of metal against metal adds a touch of urgency or struggle. In the foreground, a woman's voice is heard, perhaps offering instructions or commentary, followed by the clank of dishes or a metal container, which is then answered by a man's voice, possibly providing a contrasting perspective or additional details. This auditory tableau sets the stage for a lively, perhaps bustling or action-packed environment, ripe for a vivid visual representation."



Mistral_CoT

"In a bustling, open-air market, a vibrant scene unfolds. A woman in traditional attire, her voice lively and inviting, engages with a nearby vendor. The vendor, a man with a hearty laugh, responds with a clanking sound as he rummages through a large, metal barrel filled with various kitchenware. Nearby, a food stall sizzles and crackles with the aroma of freshly cooked dishes wafting through the air. A large, flat metal griddle rests atop a roaring fire, its surface glistening with oil and the remnants of recently cooked meals. The sound of metal scraping against metal echoes as the cook stirs the food, skillfully flipping and seasoning the dishes. The market is bustling with activity, with the clanking and scraping of metal mixed with the lively chatter of vendors and patrons. The scene is bathed in warm, golden sunlight filtering through the market's colorful canopies, casting dappled shadows on the cobblestone ground. The market, with its lively atmosphere, tantalizing aromas, and the rhythmic sounds of cooking and conversation, is a visual feast for the senses."



"The input audio contains two sources of sound: a man speaking and a crowd cheering. The man is speaking into a microphone, and the crowd is cheering. This audio could be coming from a live event, such as a concert or sports game, where the man is giving a speech or announcement, and the crowd is responding in excitement. The music in the background could be playing live, and the sound of the crowd cheering could be heard through the speakers. The spatial relationships between the sound sources are unclear, as the audio is not clear enough to determine the exact location of each sound source."



"Imagine a dimly lit room where a man, with a gentle, focused expression, is seated on a worn wooden stool, his fingers dancing across the strings of a guitar. The room is filled with an ethereal ambiance as the guitar's melodies intertwine with the hauntingly beautiful sounds of a synthesizer, creating a tapestry of sound that seems to weave through the air. In the background, muffled voices echo softly, adding an intimate layer to the scene, as if they are part of a conversation that is both



"In a spacious, dimly lit room, a man is engrossed in playing a classic acoustic guitar, his fingers gracefully strumming the strings, creating a melody that fills the air. The guitar's sound, warm and resonant, echoes off the walls, blending harmoniously with the rhythmic beats of a distant synthesizer. The synthesizer, a modern and electronic counterpoint to the guitar, adds a layer of futuristic, pulsating notes to the room. Its music, while different, complements the guitar's melody, creating a unique symphony that resonates through the space. In the background, the man's voice can be heard, a deep, soothing baritone that carries a sense of wisdom and passion. His words, spoken with a rhythm that mirrors the music, add another layer to the room's ambiance, creating a captivating, dynamic atmosphere. In the distance, muffled voices can be heard, a reminder of the world outside the room. These voices, though faint, add a sense of life and activity to the scene, creating a contrast that enhances the serenity and focus of the man playing the guitar and the synthesizer's rhythmic beats. The room, bathed in a soft, warm light, is filled with a sense of peace and creativity. The guitar, the synthesizer, the man's voice, and the distant voices all coexist in harmony, creating a visual scene that is as dynamic and engaging as the symphony of sounds they produce."

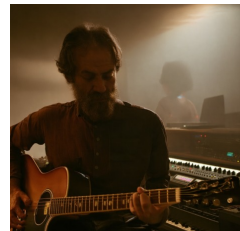


Figure 6.4.5: Example Outputs for each Method

Chapter 7

Conclusion

7.0.1 Discussion

This thesis explores methodologies for generating plausible visual descriptions from a complex audio input. We employ a pretrained Audio Large Language Model (ALLM) and experiment with different prompting techniques aimed at improving the quality of the results. Furthermore, we explore the feasibility of leveraging audio source separation with a large language model to enhance the comprehension of the audio and construct a text description of an image, which can then be used to guide the image generation process. Our experimental results, indicate an improvement upon the quality of the textual descriptions and the generated images.

In our implementation, leveraging an intermediate audio captioning model (EnCLAP) yields improved performance compared to a substantially larger ALLM. Through experimentation with various LLMs and prompts, we have determined that Mistral 7B-Instruct with a Chain-of-Thought prompt generates a visually compelling description, having the capacity to infer numerous visual details that are coherent with the audio input while also demonstrating a degree of imaginative interpretation.

7.0.2 Limitations

Despite the advantages of our proposed framework, we should also mention several limitations that it imposes. While the modularity of the system facilitates the interchangeability of components and enables experimentation with different configurations, it also introduces a considerable number of potential failure points in a hierarchical structure. If the audio separation module is unable to adequately differentiate the individual audio sources, it impedes the proper recognition by the audio language model. Consequently, if the audio language model fails to interpret the audio content, the resulting description becomes meaningless. Another limitation, in part intentional, is the elimination of the time dimension of the audio, which is a characteristic trait of auditory information. Audio is a sequential data and the task of generating a single frame from audio collapses the chronological order into one moment.

7.0.3 Future Work

In future work, there are several research paths that can offer substantial contributions.

1. Experimenting with different, larger models. Larger LLMs tend to exhibit more emerging abilities
2. Improving universal sound separation will effectively enhance the overall quality of the results
3. Generating video output will be more in line with the auditory perception and will establish a new frontier in many industries
4. Cross-modality reasoning is an intriguing topic and further investigating the opposite, image-to-audio will be of great interest
5. Restricting to a specific domain, like music visualization, leveraging musical instrument separation may lead to better results than the more general universal audio.
6. Contributing to better evaluation metrics for the quality assessment of audio to image generation

Appendix A

Additional Details

A.1 Prompts

LLM-as-a-Judge prompts

Evaluation	Prompt Template
Number of Sources	<p>< <i>s</i> > [INST] You are a helpful assistant acting as a judge. You will be given an LLM response, which is a text description of a scene based on a given audio input. Your task is to identify the number of different sound descriptions that are present in the response. The number should be from 1 to 4 different sounds. Sounds can be environmental sounds, onomatopoeias like roar or chirp, music, speech. For example: LLM response: "The audio is of a child speaking followed by the sound of rustling and then a click." The output should be: [/INST]</p> <p>3 < /<i>s</i> ></p> <p>[INST] Output only a number from 1 to 4 and nothing else.</p> <p>LLM response: {<i>prediction</i>} [/INST]</p>
Labels Accuracy	<p>< <i>s</i> > [INST] You are a helpful assistant acting as a judge. You will be given an LLM response, which is a text description of a scene based on a given audio input and the corresponding weak labels of the input audio. Your task is to assign a binary label (Yes/No) to each description, whether it includes the sounds that are described by the weak labels. Careful, the weak labels are of poor quality so they can be generic. For example if the description is "The audio is of a child speaking followed by the sound of rustling and then a click." and the weak labels are: "[child laughter, children voice', 'crumpling_and.crinkling']" then the output should be: [/INST]</p> <p>"Yes" < /<i>s</i> ></p> <p>[INST] Output only 'Yes' or 'No' and nothing else. Keep in mind that the labels are weak. Only output yes if all sounds are included. If the description describes a sound that can not be inferred from the weak labels output 'No'.</p> <p>LLM response:{<i>prediction</i>} weak labels: {<i>source_labels</i>} [/INST]</p>

Table A.1.1: Prompt Templates for the Audio Perception Evaluation

Evaluation	Prompt Template
Coherence	<p>[<i>INST</i>] You are a helpful assistant acting as a judge. You will be given an LLM response, which is a text description of a scene based on a given audio input. Your task is to output a score from 1 to 10 with 1 being lowest and 10 being highest, on how much the description makes logical sense given the identified sound sources. For example if the identified sources are "children playing" and "splashing water" and the response is "children playing in a pool" the output score should be high. Output only the score and nothing else.</p> <p>LLM response: {<i>prediction</i>} [<i>INST</i>]</p>
Visual Detail	<p>[/<i>INST</i>] You are a helpful assistant acting as a judge. You will be given an LLM response, which is a text description of a scene based on a given audio input. Your task is to output a score from 1 to 10 with 1 being lowest and 10 being highest, on how much visually detailed is the generated description, how rich are the visual details. Output only the score and nothing else. LLM response: {<i>prediction</i>} [<i>INST</i>]</p>
Imagination	<p>[/<i>INST</i>] You are a helpful assistant acting as a judge. You will be given an LLM response, which is a text description of a scene based on a given audio input and the weak labels of the audio. Your task is to output a score from 1 to 10 with 1 being lowest and 10 being highest, on how much the LLM hallucinated details that are not present in the audio weak labels but make sense from the audio weak labels. For example, if the description is: "A man breathing heavily wearing a blue shirt" and the weak labels are: ['breathing'], the output score should be high because the LLM imagined something that is not present in the audio but makes sense. If the description is: "A man is walking holding a cranky cat in his hands" and the weak labels are: ['walking'], the output score should be low because there is no cat present in the audio.</p> <p>LLM response: {<i>prediction</i>} weak labels: {<i>source_labels</i>} [<i>INST</i>]</p>
Image Appropriate- ness	<p>[<i>INST</i>] You are a helpful assistant acting as a judge. You will be given an LLM response, which is a text description of a visual scene based on a given audio input. Your task is to output a score from 1 to 10 with 1 being lowest and 10 being highest, on how good it is as an image description and how much suitable it is as an input to an image generation model. Output only the score and nothing else.</p> <p>LLM response: {<i>prediction</i>} [<i>INST</i>]</p>

Table A.1.2: Prompt Templates for Image Description Evaluation

Text merging prompts

These prompts were used for the final text merging of our proposed pipeline for each LLM we used. Each model follows a slightly different format.

Evaluation	Prompt Template
Labels Accuracy	<p>< <i>s</i> > [INST] You are a helpful assistant acting as a judge. You will be given an LLM response, which is a text caption of an audio input and the corresponding weak labels of the input audio. Your task is to assign a binary label (Yes/No) to the text description, whether it includes the sounds that are described by the weak labels. Careful, the weak labels are of poor quality so they can be generic. [/INST] < /<i>s</i> ></p> <p>[INST] Output only 'Yes' or 'No' and nothing else. Keep in mind that the labels are weak. If the description describes a sound that can not be inferred from any of the weak labels, output 'No'.</p> <p>LLM response: {<i>prediction</i>} weak labels: {<i>source_labels</i>}/[/INST]</p>

Table A.1.3: Prompt Template for Label Accuracy of Intermediate Captions

Method	Prompt Template
Zero-Shot	<p>[SYSTEM]: "You are a creative assistant that transforms audio descriptions into vivid visual scenes. Your task is to analyze multiple audio captions, combine their details, and create a cohesive, richly detailed description of a visual scene. This description will be used to generate an image, so ensure it is clear, imaginative, and visually evocative." "Prompt: Combine the following audio descriptions into a single, vivid image description suitable for an image generation model:</p> <p>{<i>descriptions</i>}</p> <p>The output must be concise "</p>
Chain-of-Thought	<p>[SYSTEM]: "You are a creative assistant that transforms audio descriptions into vivid visual scenes. Your task is to analyze multiple audio captions, combine their details, and create a cohesive, richly detailed description of a visual scene. This description will be used to generate an image, so ensure it is clear, imaginative, and visually evocative." Prompt: "Think step-by-step: 1. Read each individual audio description and identify key elements such as sound sources, 2. Imagine a visual scene that can naturally incorporate all the identified audio elements. 3. Consider spatial relationships, 4. Convert the auditory clues into visual descriptors, 5. Merge all these elements into a single, flowing narrative that paints a vivid picture of the scene, 6. Ensure the description is coherent, contextually consistent, and naturally integrates all auditory details. Produce one unified image description suitable for an image generation model, that captures a scene where all the individual audio elements coexist in a coherent, visually engaging manner.</p> <p>Audio Descriptions: {<i>descriptions</i>}"</p>

Table A.1.4: Prompt Template for Qwen 2.5 Text Merging

Method	Prompt Template
Zero-Shot	"User: Combine the following audio descriptions into a single, vivid image description suitable for an image generation model. Audio Descriptions: $\{descriptions\}$. Assistant:
Chain-of-Thought	"Think step-by-step: 1. Read each individual audio description and identify key elements such as sound sources, 2. Imagine a visual scene that can naturally incorporate all the identified audio elements. 3. Consider spatial relationships, 4. Convert the auditory clues into visual descriptors, 5. Merge all these elements into a single, flowing narrative that paints a vivid picture of the scene, 6. Ensure the description is coherent, contextually consistent, and naturally integrates all auditory details. Produce one unified image description suitable for an image generation model, that captures a scene where all the individual audio elements coexist in a coherent, visually engaging manner. Audio Descriptions: $\{descriptions\}$

Table A.1.5: Prompt Template for Falcon Text Merging

Method	Prompt Template
Zero-Shot	"Combine the following audio descriptions into a single, vivid image description suitable for an image generation model: descriptions The output must be concise"
Chain-of-Thought	"Think step-by-step: 1. Read each individual audio description and identify key elements such as sound sources, 2. Imagine a visual scene that can naturally incorporate all the identified audio elements. 3. Consider spatial relationships, 4. Convert the auditory clues into visual descriptors, 5. Merge all these elements into a single, flowing narrative that paints a vivid picture of the scene, 6. Ensure the description is coherent, contextually consistent, and naturally integrates all auditory details. Produce one unified image description suitable for an image generation model, that captures a scene where all the individual audio elements coexist in a coherent, visually engaging manner. Audio Descriptions: $\{descriptions\}$

Table A.1.6: Prompt Template for Mistral Text Merging

A.2 Example Results

Visual Descriptions examples

In the tables below we show example text descriptions from the Experiment B

Model	Text Output
Qwen-Falcon-Zero_Shot	"Imagine a dimly lit room where the gentle rhythm of water droplets creates a soothing background melody. The droplets fall steadily from a high point, like a soft, continuous rain on unseen surfaces. The sound is calming, almost meditative, as it fills the air with a tranquil hum."
EnCLAP-Falcon-Zero_-Shot	"Imagine a dimly lit room where the gentle rhythm of water droplets creates a soothing, almost hypnotic background melody. The droplets fall steadily, each one echoing softly as they hit a surface, casting a series of tiny, shimmering reflections on the walls and ceiling. This auditory symphony is accompanied by the subtle sheen of water, which seems to dance in the dim light, transforming the space into a tranquil oasis of sound and sight."
EnCLAP-Falcon-Chain_-of_Thought	"In a serene courtyard, a faucet drips rhythmically, echoing the soft tolling of a distant church bell. Overhead, a lively tree is home to chirping birds, their cheerful songs blending with the intermittent barks of a playful dog nearby. In the center, a person is typing away at a typewriter, the clack-clack of keys punctuating the scene. In the background, the joyful laughter of children fills the air, creating a vibrant tapestry of sounds that bring the peaceful courtyard to life."

Table A.2.1: Example Results from Experiment B for an audio clip with weak labels: [Rain, Water]

Model	Text Output
ALLM	"The input audio contains the sound of rain falling on different surfaces. The rain is falling on the ground, roof, and water. The audio is a bit noisy and in mono. This audio can be heard in a rainy day outside."
Qwen-Qwen-Zero_Shot	"Water drips steadily from a faucet, creating a gentle cascade of droplets in the quiet room."
EnCLAP-Qwen-Zero_-Shot	"Water drips steadily from a faucet, creating a gentle pitter-patter on the sink below."
EnCLAP-Qwen-Chain_-of_Thought	"In a dimly lit bathroom, a dripping faucet hangs above a sink, casting a soft, diffused light on the tiles below. Water droplets form at the tip of the faucet and fall in a gentle, rhythmic pattern, creating a soothing, ambient sound. Each drop splits the air as it descends, leaving a faint trail of moisture behind. The background features a few scattered toiletries and a towel hanging on the side, adding to the cozy, domestic atmosphere. The sound of the dripping water blends seamlessly with the quiet hum of the room, creating a serene and peaceful ambiance."
Qwen-Mistral-Zero_Shot	"A tranquil scene unfolds, featuring a small, moss-covered waterfall cascading into a serene, crystal-clear pool below. Droplets of water are continuously falling, creating a soothing, rhythmic sound as they splash into the pool, rippling its surface."
EnCLAP-Mistral-Zero_-Shot	"A tranquil scene features a small, moss-covered stone bridge arching over a gently flowing stream. The water cascades from the bridge above, creating a rhythmic dripping sound as it lands in the clear, pebble-bottomed stream below. The surrounding area is lush with verdant ferns and tall, swaying trees, casting dappled sunlight onto the water's surface. A soft mist hangs in the air, adding an ethereal quality to the serene landscape."
EnCLAP-Mistral-Chain_-of_Thought	"Imagine a bustling city street at dusk. The sound of water dripping from a leaky roof above creates a rhythmic patter on the ground below, mingling with the crunch of footsteps on gravel. In the distance, a dog barks, its sound echoing off the surrounding buildings, creating an eerie, almost haunting effect. A car horn blares in the distance, its sharp, discordant note cutting through the otherwise peaceful ambiance. The soft, melodic tune of a nearby radio can be heard, the music adding a touch of serenity to the chaotic scene. A bird chirps, its song sweet and clear against the backdrop of the city sounds. Rain falls gently, the drops hitting the ground with a soft plop, creating a soothing symphony that seems to calm the city's heartbeat. The scene is bathed in the warm, golden glow of the setting sun, casting long, dancing shadows on the cobblestone street. The city is alive with sound and movement, yet there is a sense of peace and tranquility that permeates the air."

Table A.2.2: Example Results from Experiment B for an audio clip with weak labels: [Rain, Water]

A.3 Human Evaluation

Participants Statistics

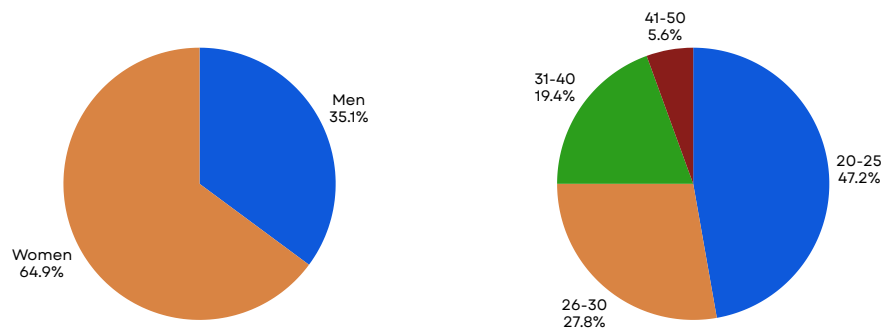


Figure A.3.1: Participants Gender (left) and Age (right)

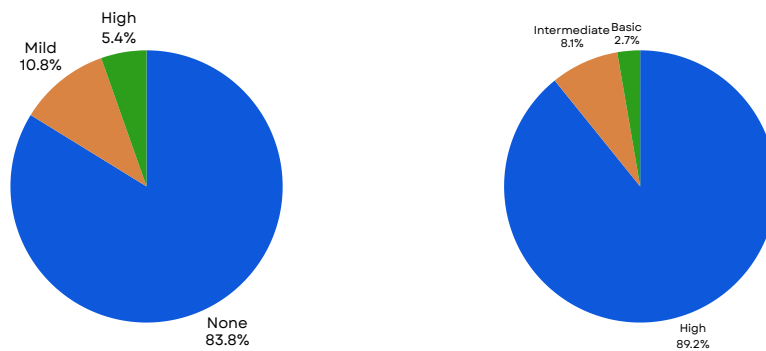


Figure A.3.2: Participants Hearing Difficulty (left) and Level of English (right)

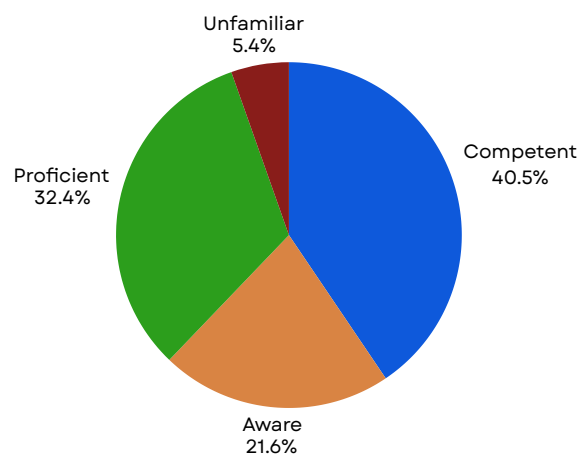


Figure A.3.3: Participants Familiarity with AI

The majority of the participants are young adults, under the age of 30, and do not have a

hearing impairment. Given the importance of comprehending the English text descriptions, the majority of the participants possess a high level of English understanding. Also, most of the evaluators were familiar with Artificial Intelligence.

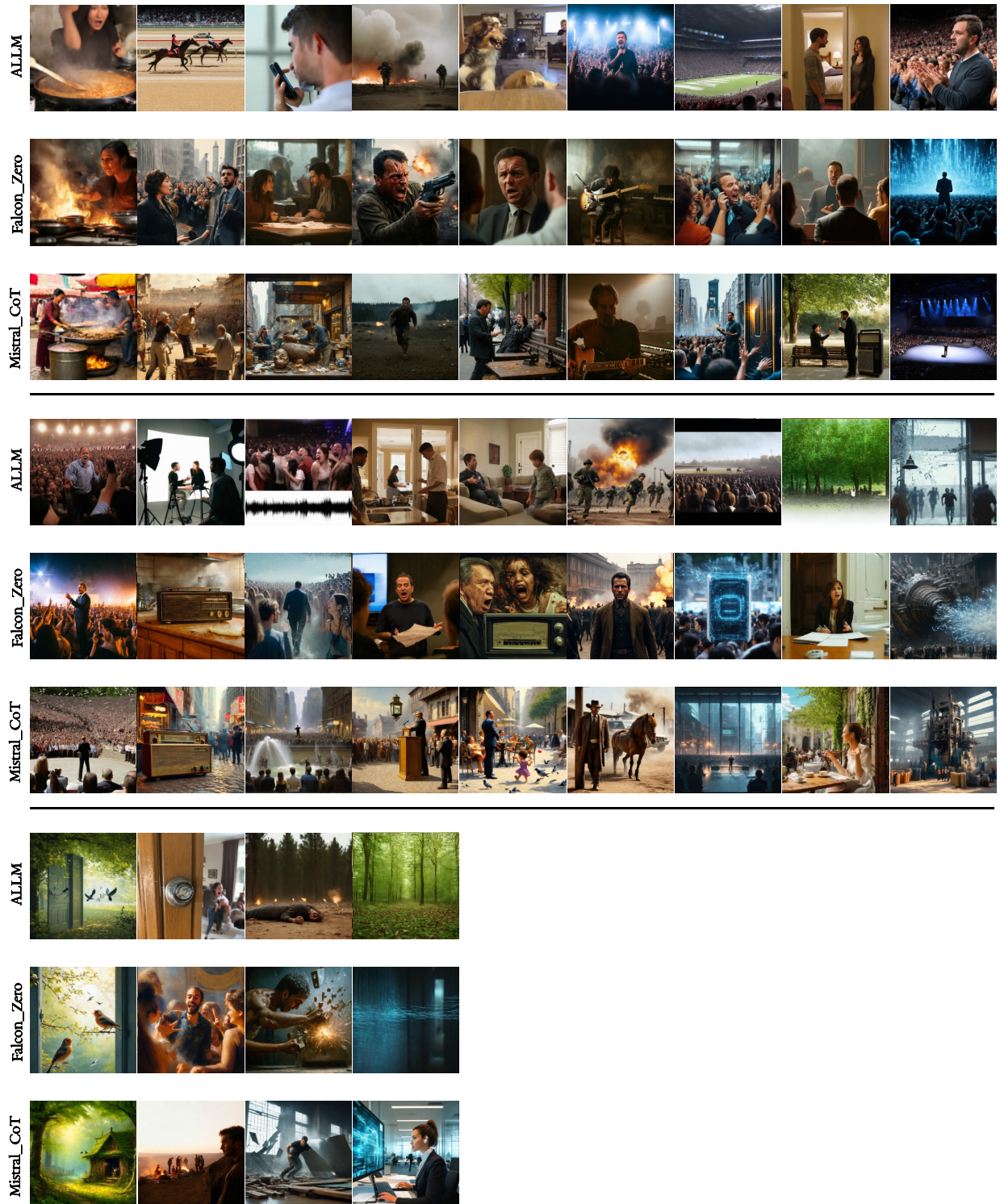
Text Descriptions

Model	Captions	Descriptions
Falcon-Zero_Shot	"Birds are chirping", "Birds chirp, music plays, and a door slams", "Clanking and rustling"	"A scene unfolds with birds cheerfully chirping, their songs harmonizing with a lively melody of music playing in the background. Suddenly, the tranquility is disrupted by the sharp sound of a door slamming, adding an element of surprise and movement to the otherwise serene auditory environment."
Mistral-Chain_of_Thought	"Birds are chirping", "Birds chirp, music plays, and a door slams", "Clanking and rustling"	"In a picturesque, tranquil landscape, a vibrant, verdant forest stretches out, filled with the harmonious symphony of nature. The canopy above is a lively tapestry of leaves, swaying gently in the breeze, while the forest floor is a mosaic of ferns, wildflowers, and fallen leaves. Nestled within this verdant haven, a variety of birds are perched, their colorful plumage standing out against the greenery. They chirp melodiously, their voices blending with the soft rustling of leaves and the distant, soothing melody of a lullaby, as if the forest itself is lulled into a peaceful slumber. In the distance, a wooden cabin comes into view. Its aged wooden walls are weathered by time, and the roof is adorned with moss and small, colorful flowers. The cabin door swings open, and a clanking sound echoes through the forest, as if a blacksmith is at work within. The clanking sound is interspersed with the soft rustling of fabric, hinting at the presence of a person moving about inside the cabin. The scene is bathed in the soft, warm glow of the setting sun, casting long, dappled shadows across the forest floor. The overall atmosphere is one of peace and tranquility, where the beauty of nature and the comfort of human presence coexist in perfect harmony."

Table A.3.1: Example Description Results from the Corresponding Captions

Generated Images

Figure A.3.4: Image Results from Experiment C: ALLM Images (top row), Falcon-Zero_Shot Images (middle row), Mistral-Chain_of_Thought Images (bottom row)



Bibliography

- [1] 1 - physics of sound. In J. T. Talty, editor, *Industrial Hygiene Engineering (Second Edition)*, pages 372–389. William Andrew Publishing, Park Ridge, NJ, second edition edition, 1998.
- [2] P. Alsaleh A. A space and time efficient convolutional neural network for age group estimation from facial images, 2023. Accessed: 2025-1-31.
- [3] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, and T. Zhu. Qwen technical report, 2023.
- [4] M. Bain, A. Nagrani, A. Brown, and A. Zisserman. Condensed movies: Story based retrieval with contextual embeddings, 2020.
- [5] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.
- [6] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. In *Journal of machine learning research*, 2003.
- [7] B. C. Biner, F. M. Sofian, U. B. Karakaş, D. Ceylan, E. Erdem, and A. Erdem. Son-icdiffusion: Audio-driven image generation and editing with pretrained diffusion models, 2024.
- [8] A. S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press, 05 1990.
- [9] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020.
- [10] M. A. Casey and A. Westner. Separation of mixed audio sources by independent subspace analysis. In *International Conference on Mathematics and Computing*, 2000.
- [11] P. Chandna, M. Miron, J. Janer, and E. Gómez. Monoaural audio source separation using deep convolutional neural networks. In *Latent Variable Analysis and Signal Separation*, 2017.

- [12] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman. Vggsound: A large-scale audio-visual dataset, 2020.
- [13] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei. Beats: Audio pre-training with acoustic tokenizers, 2022.
- [14] Y. Chen, X. Yue, X. Gao, C. Zhang, L. F. D’Haro, R. T. Tan, and H. Li. Beyond single-audio: Advancing multi-audio processing in audio large language models. In *Conference on Empirical Methods in Natural Language Processing*, 2024.
- [15] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models, 2023.
- [16] E. B. Çoban, M. I. Mandel, and J. Devaney. What do MLLMs hear? examining the interaction between LLM and audio encoder components in multimodal large language models. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024.
- [17] A. D’efosse, J. Copet, G. Synnaeve, and Y. Adi. High fidelity neural audio compression. *ArXiv*, abs/2210.13438, 2022.
- [18] A. Défossez, N. Usunier, L. Bottou, and F. R. Bach. Demucs: Deep extractor for music sources with extra unlabeled data remixed. *ArXiv*, abs/1909.01174, 2019.
- [19] S. Deshmukh, B. Elizalde, R. Singh, and H. Wang. Pengi: An audio language model for audio tasks, 2024.
- [20] S. Deshmukh, S. Han, H. Bukhari, B. Elizalde, H. Gamper, R. Singh, and B. Raj. Audio entailment: Assessing deductive reasoning for audio understanding, 2024.
- [21] K. Drossos, S. Adavanne, and T. Virtanen. Automated audio captioning with recurrent neural networks. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, New York, U.S.A., Oct. 2017.
- [22] K. Drossos, S. Lipping, and T. Virtanen. Clotho: An audio captioning dataset, 2019.
- [23] K.-L. Du, C.-S. Leung, W. H. Mow, and M. N. S. Swamy. Perceptron: Learning, generalization, model selection, fault tolerance, and role in the deep learning era. *Mathematics*, 10(24), 2022.
- [24] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang. Clap: Learning audio concepts from natural language supervision, 2022.
- [25] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, K. Lacey, A. Goodwin, Y. Marek, and R. Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024.

- [26] E. Fonseca, J. Pons, X. Favory, F. Font, D. Bogdanov, A. Ferraro, S. Oramas, A. Porter, and X. Serra. Freesound datasets: A platform for the creation of open audio datasets. In *International Society for Music Information Retrieval Conference*, 2017.
- [27] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.
- [28] Y. Gong, Y.-A. Chung, and J. Glass. Ast: Audio spectrogram transformer, 2021.
- [29] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. Glass. Listen, think, and understand, 2024.
- [30] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. <http://www.deeplearningbook.org>.
- [31] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks, 2014.
- [32] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Y. Wang, W. Gao, L. Ni, and J. Guo. A survey on llm-as-a-judge, 2025.
- [33] M. Hayes. *Schaum's Outline of Digital Signal Processing*. Schaum's Outline Series. McGraw-Hill Companies, Incorporated, 1999.
- [34] J. Hérault and C. Jutten. Space or time adaptive signal processing by neural network models. 1987.
- [35] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models, 2020.
- [36] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models, 2021.
- [37] A. Jansson, E. J. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde. Singing voice separation with deep u-net convolutional networks. In *International Society for Music Information Retrieval Conference*, 2017.
- [38] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b, 2023.
- [39] K. Joos, A. Gilles, P. Van de Heyning, D. De Ridder, and S. Vanneste. From sensation to percept: The neural signature of auditory event-related potentials. *Neuroscience Biobehavioral Reviews*, 42:148–156, 2014.
- [40] D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd edition, 2025. Online manuscript released January 12, 2025.
- [41] I. Kavalero, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. L. Roux, and J. R. Hershey. Universal sound separation, 2019.

- [42] C. D. Kim, B. Kim, H. Lee, and G. Kim. Audiocaps: Generating captions for audios in the wild. In *North American Chapter of the Association for Computational Linguistics*, 2019.
- [43] J. Kim, J. Jung, J. Lee, and S. H. Woo. Enclap: Combining neural audio codec and audio-text joint embedding for automated audio captioning, 2024.
- [44] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.
- [45] C.-Y. Kuan, W.-P. Huang, and H. yi Lee. Understanding sounds, missing the questions: The challenge of object hallucination in large audio-language models, 2024.
- [46] C. Lea, R. Vidal, A. Reiter, and G. D. Hager. Temporal convolutional networks: A unified approach to action segmentation, 2016.
- [47] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [48] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.
- [49] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. rahman Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics*, 2019.
- [50] Y. Luo and N. Mesgarani. Tasnet: Time-domain audio separation network for real-time, single-channel speech separation. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 696–700, 2017.
- [51] Y. Luo and N. Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27:1256–1266, 2018.
- [52] E. Manilow, P. Seetharman, and J. Salamon. *Open Source Tools & Data for Music Source Separation*. <https://source-separation.github.io/tutorial>, Oct. 2020.
- [53] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013.
- [54] A. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models, 2021.
- [55] Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu, K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang, T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu. Qwen2.5 technical report, 2025.
- [56] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021.

- [57] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision, 2022.
- [58] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models, 2022.
- [59] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.
- [60] S. Sakshi, U. Tyagi, S. Kumar, A. Seth, R. Selvakumar, O. Nieto, R. Duraiswami, S. Ghosh, and D. Manocha. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [61] J. Salamon, D. MacConnell, M. Cartwright, P. Q. Li, and J. P. Bello. Scaper: A library for soundscape synthesis and augmentation. *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 344–348, 2017.
- [62] T. S. Silva. A short introduction to generative adversarial networks. <https://sthalles.github.io>, 2017.
- [63] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015.
- [64] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models, 2022.
- [65] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution, 2020.
- [66] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations, 2021.
- [67] A. K. Sridhar, Y. Guo, and E. Visser. Enhancing temporal understanding in audio question answering for large audio language models. *ArXiv*, abs/2409.06223, 2024.
- [68] S. S. Stevens, J. E. Volkman, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8:185–190, 1937.
- [69] D. Stoller, S. Ewert, and S. Dixon. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *ArXiv*, abs/1806.03185, 2018.
- [70] K. Sung-Bin, A. Senocak, H. Ha, A. Owens, and T.-H. Oh. Sound to visual scene generation by audio-to-visual latent alignment, 2023.
- [71] T. Team. The falcon 3 family of open models, December 2024.
- [72] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models, 2023.

-
- [73] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu. Wavenet: A generative model for raw audio. In *Speech Synthesis Workshop*, 2016.
 - [74] A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding, 2019.
 - [75] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.
 - [76] S. Vatsal and H. Dubey. A survey of prompt engineering methods in large language models for different nlp tasks, 2024.
 - [77] T. Virtanen and A. Klapuri. Separation of harmonic sound sources using sinusoidal modeling. In *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*, volume 2, pages II765–II768 vol.2, 2000.
 - [78] W. Wang. Machine audition: Principles, algorithms and systems. 2010.
 - [79] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.
 - [80] S. Wisdom, H. Erdogan, D. Ellis, R. Serizel, N. Turpault, E. Fonseca, J. Salamon, P. Seetharaman, and J. Hershey. What’s all the fuss about free universal sound separation data?, 2020.
 - [81] G. Yariv, I. Gat, L. Wolf, Y. Adi, and I. Schwartz. Audiotoken: Adaptation of text-conditioned diffusion models for audio-to-image generation, 2023.
 - [82] D. Zhang, Y. Yu, J. Dong, C. Li, D. Su, C. Chu, and D. Yu. Mm-llms: Recent advances in multimodal large language models, 2024.