National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science

Knowledge Transfer from Large Vision-Language Models for Localization and Segmentation in 2D Medical Imaging

Diploma Thesis

of

GEORGIOS TRIANTAFYLLIS

Supervisor: Athanasios Voulodimos
Assistant Professor NTUA

Athens, March 2025

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science

Knowledge Transfer from Large Vision-Language Models for Localization and Segmentation in 2D Medical Imaging

Diploma Thesis

of

GEORGIOS TRIANTAFYLLIS

Supervisor: Athanasios Voulodimos
Assistant Professor NTUA

Approved by the examination committee on 26th March 2025.

(Signature)	*(Signature)*	*(Signature)*
.....	.....	.....
Athanasios Voulodimos	Georgios Stamou	Andreas-Georgios Stafylopatis
Assistant Professor NTUA	Professor NTUA	Professor NTUA

Athens, March 2025

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism.

(Signature)

.....
Georgios Triantafyllis,
Graduate of Electrical and
Computer Engineering,
NTUA
26th March 2025

Περίληψη

Το grounded segmentation σε ιατρικές εικόνες είναι μια ιδιαίτερα απαιτητική εφαρμογή, καθώς απαιτεί σύνολα δεδομένων που αποτελούνται από εικόνες, μάσκες και κειμενικές περιγραφές για κάθε μάσκα κάτι το οποίο δε συναντάται συχνά. Για την επίλυση αυτού του προβλήματος, χρησιμοποιούμε μεγάλα μοντέλα Όρασης-Γλώσσας (LVLMs) σε συνδυασμό με ντετερμινιστικούς αλγορίθμους για την παραγωγή των κειμενικών περιγραφών για κάθε μάσκα. Όσον αφορά το grounded segmentation, αναπτύσσεται ένα σύστημα το οποίο αποτελείται από το GroundingDINO και το SAM2 ή το Med-SAM2 από τα οποία μόνο το GroundingDINO επιδέχεται περεταίρω στοχευμένη εκπαίδευση. Το σύνολο δεδομένων που θα χρησιμοποιηθεί ονομάζεται RAOS και αποτελείται από εικόνες υπολογιστικής τομογραφίας (CT) και εικόνες συνθετικής μαγνητικής τομογραφίας (synthetic MRI). Τα πειράματα που παρουσιάζονται αξιολογούν την εγκυρότητα των απαντήσεων του LVLM LLaVA-Med και τις επιδόσεις του προτεινόμενου συστήματος δοκιμάζοντας διάφορες στρατηγικές εισόδου τόσο σε παρόμοιες εικόνες όσο και σε εικόνες εκτός κατανομής συνόλου εκπαίδευσης. Τα αποτελέσματα καταδεικνύουν πως το LLaVA-Med δεν είναι αξιόπιστο για την πλήρη παραγωγή των περιγραφών των μασκών, λόγω της περιορισμένης κριτικής του ικανότητας. Επιπρόσθετα, το προτεινόμενο σύστημα παρουσιάζει ικανοποιητικές επιδόσεις σε εικόνες που ανήκουν στην ίδια κατανομή με αυτή της εκπαίδευσης, λαμβάνοντας υπόψην και τους εγγενείς περιορισμούς του.

Λέξεις Κλειδιά

Ιατρική απεικόνιση, Γειωμένη Κατάτμηση, Μεγάλα μοντέλα Όρασης-Γλώσσας, Fine-tuning, GroundingDINO, SAM2, MedSAM2, Μαγνητική τομογραφία, Υπολογιστική τομογραφία

Abstract

Grounded segmentation of medical images is a challenging task requiring expert-annotated datasets, which are scarce. To address this problem, we employ Large Vision-Language Models (LVLMs) as well as deterministic algorithms to generate the missing textual descriptions for organ masks. For the grounded segmentation task, a pipeline is developed consisting of GroundingDINO and SAM2 or Med-SAM2 with only GroundingDINO being fine-tuned. The dataset used for this study is RAOS, which includes CT scans and synthetic MRI images. Our experiments assess the accuracy of LLaVA-Med's responses and the performance of the proposed fine-tuned pipeline to various prompting strategies on both in-distribution and out-of-distribution images. The results indicate that LLaVA-Med alone cannot reliably generate the textual descriptions due to its limited reasoning ability. Additionally, our results show that the proposed pipeline performs well within the closed setting in which it was applied while acknowledging inherent limitations.

Keywords

Medical Imaging, Grounded Segmentation, Large Vision-Language Models, Fine-Tuning, GroundingDINO, SAM2, MedSAM2, MRI, CT

to my parents

Acknowledgements

I would like to express my sincere gratitude to Professor Athanasios Voulodimos and Professor Paraskevi Tzouveli for supervising this thesis and for giving me the opportunity to carry it out in the Artificial Intelligence and Learning Systems Laboratory. I would also like to extend my appreciation to Nikolaos Spanos, Paraskevi Theofilou and Iasonas Liartis for their continuous support, insightful feedback and excellent collaboration. Furthermore, I am grateful to my family for their guidance and moral support. Finally, I would like to thank my friends for their encouragement and support during this journey.

Athens, March 2025

Georgios Triantafyllis

Table of Contents

List of Figures

List of Tables

Chapter 1

Εκτεταμένη Περίληψη στα Ελληνικά

1.1 Εισαγωγή

Στην τεχνητή νοημοσύνη τα foundation μοντέλα παίζουν καθοριστικό ρόλο, καθώς βρίσκουν εφαρμογή σε πολλούς τομείς, όπως η επεξεργασία φυσικής γλώσσας (NLP) και η όραση υπολογιστών. Λόγω της ιδιότητας τους να είναι γενικεύσιμα, η δημιουργία τους είναι ακριβή και χρειάζονται αρκετοί υπολογιστικοί πόροι και μεγάλα σύνολα δεδομένων. Όμως μέσω του fine tuning μπορούν να εφαρμοστούν σε πολύ ειδικές εφαρμογές όπως η ιατρική απεικόνιση με μικρότερο κόστος από αυτό της εκπαίδευσής τους.

Στην όραση υπολογιστών πολλά μοντέλα όπως και το U-Net [6] βασίζονται σε συνελικτικά νευρωνικά δίκτυα (CNNs). Με την εμφάνιση της αρχιτεκτονικής transformers δημιουργήθηκαν επιπλέον μοντέλα με νέες δυνατότητες. Σε αυτά συγκαταλέγονται το Segment Anything Model(SAM)[7] και ο διάδοχός του SAM2[2] τα οποία πραγματοποιούν κατάτμηση εικόνων με βάση κάποιο σημείο, μάσκα ή bounding box. Ένα άλλο παράδειγμα μοντέλου της αρχιτεκτονικής transformer είναι το GroundingDINO[4] το οποίο ανιχνεύει αντικείμενα με βάση μια κειμενική περιγραφή. Μία άλλη κατηγορία μοντέλων είναι τα γνωστά μεγάλα μοντέλα Όρασης–Γλώσσας τα οποία επεξεργάζονται τόσο οπτική όσο και γλωσσική πληροφορία όπως το large language and vision assistant (LLaVA)[1] το οποίο μπορεί να απαντάει σε ερωτήσεις ανοιχτού τύπου που αφορούν το περιεχόμενο μίας εικόνας. Όλα τα παραπάνω επιδέχονται περαιτέρω εκπαίδευση και μπορούν εφαρμοστούν με ένα σχετικά μικρό κόστος σε οποιονδήποτε τομέα.

Ειδικότερα στον τομέα της ιατρικής απεικόνισης και στην grounded κατάτμηση οργάνων, το οποίο είναι αντικείμενο αυτής της εργασίας, τα σύνολα δεδομένων είναι σπάνια και οποιαδήποτε ιατρική πληροφορία θα πρέπει να είναι προϊόν ειδικού για λόγους αξιοπιστίας. Όλα τα προαναφερθέντα αποτελούν περιοριστικούς παράγοντες για την ανάπτυξη ενός μοντέλου το οποίο απαιτεί κείμενο ως είσοδο για να πραγματοποιήσει κατάτμηση οργάνων. Για την επίλυση του προβλήματος, προτείνουμε μία μέθοδο η οποία μπορεί εν μέρει να μας απαλλάξει από την ανάγκη παραγωγής των κειμενικών περιγραφών από έναν ειδικό. Γι' αυτό το σκοπό θα χρησιμοποιήσουμε το LVLM LLaVA-Med[8] το οποίο θα μεταφέρει την γνώση του στο GroundingDINO. Το σύστημα λοιπόν που θα αναπτύξουμε για το grounded segmentation αποτελείται από το GroundingDINO και το SAM2 ή Med-SAM2[3]. Όσο για το σύνολο δεδομένων θα χρησιμοποιήσουμε το RAOS[5, 9] το οποίο αποτελείται από εικόνες συνθετικής μαγνητικής και υπολογιστικής τομογραφίας με τις αντίστοιχες μάσκες οργάνων

χωρίς να υπάρχουν κειμενικές περιγραφές για αυτές. Για να τις παράξουμε θα χρησιμοποιήσουμε το LLaVA-Med και το μοντέλο που θα εκπαιδευτεί θα είναι το GroundingDINO.

Στα πειράματα που ακολουθούν θα χρησιμοποιήσουμε ως μετρικές την τομή προς την ένωση (IoU), την ακρίβεια (precision), το dice score coefficient και το wall distance. Επίσης, θα δοκιμαστούν διάφορες μορφές εισόδου ως προς το κείμενο καθώς και θα πραγματοποιηθεί μία σύγκριση μεταξύ του αρχικού GroundigDINO με το εκπαιδευμένο όπως και του SAM2 με το Med-SAM2. Τέλος θα μετρήσουμε τις επιδόσεις του εκπαιδευμένους μοντέλου σε εικόνες εκτός κατανομής συνόλου δεδομένων εκπαίδευσης.

1.2 Υπόβαθρο και σχετική Βιβλιογραφία

1.2.1 Όραση Υπολογιστών

Η όραση υπολογιστών είναι ένας κλάδος της τεχνητής νοημοσύνης που επιτρέπει στις μηχανές να αναλύουν και να επεξεργάζονται οπτική πληροφορία. Ο κλάδος αυτός αναπτύσσεται με γρήγορους ρυθμούς ιδιαίτερα μετά την εμφάνιση της βαθιάς μάθησης, λόγω μοντέλων που βασίζονται σε συνελικτικά νευρωνικά δίκτυα και στην αρχιτεκτονική transformer. Τελευταίες εξελίξεις έχουν οδηγήσει στην ανάπτυξη μοντέλων τα οποία μπορούν να επεξεργαστούν τόσο οπτική πληροφορία όσο και γλώσσα τα λεγόμενα large vision language models (LVLMs).

Μερικά από τα θεμελιώδη προβλήματα που προσπαθεί να επιλύσει η όραση υπολογιστών είναι η κατηγοριοποίηση εικόνας, η αναγνώριση αντικειμένων, η κατάτμηση εικόνας, η απάντηση ερωτήσεων που αφορούν εικόνες και το grounded segmentation.

Η κατηγοριοποίηση εικόνας είναι ένα πρόβλημα στο οποίο υπάρχει μία λίστα κλάσεων και κάθε εικόνα πρέπει να αντιστοιχηθεί σε μία κλάση. Γι' αυτό το λόγο οι εικόνες συχνά απεικονίζουν ένα μόνο αντικείμενο για να μειωθούν οι πιθανότητες να εμπεριέχονται στην εικόνα αντικείμενα δύο η περισσότερων κλάσεων.

Η αναγνώριση αντικειμένου είναι ένα άλλο πρόβλημα στο οποίο δίνονται ως είσοδοι στο μοντέλο μία εικόνα και το όνομα του αντικειμένου. Στόχος είναι η παραγωγή ενός bounding box το οποίο θα περιβάλλει το ζητούμενο αντικείμενο. Σε κάποιες περιπτώσεις μπορεί να υπάρχουν παραπάνω εμφανίσεις του ίδιου αντικειμένου στην εικόνα και τότε απαιτούνται ισάριθμα bounding boxes.

Στο πρόβλημα του Grounded Segmentation στόχος είναι ο εντοπισμός ενός αντικειμένου σε μία εικόνα με ακρίβεια πίξελ. Ως είσοδος δίνεται στο μοντέλο μία εικόνα, μία κειμενική περιγραφή και παράγεται η μάσκα του.

Η απάντηση ερωτήσεων που αφορούν εικόνες (VQA) έχει ως στόχο την παραγωγή πληροφοριών που θα απαντούν με ακρίβεια στην ερώτηση του χρήστη. Δηλαδή το μοντέλο αναμένεται να εξάγει τα χρήσιμα χαρακτηριστικά της εικόνας να κατανοήσει την ερώτηση του χρήστη και μέσω των γνώσεων που διαθέτει ή παρατηρήσεων που στηρίζονται στην εικόνα να δώσει μία σωστή απάντηση.

Μάσκα

Η μάσκα ενός αντικειμένου στην όραση υπολογιστών είναι το σύνολο των πίξελ που χρησιμοποιούνται κατά την προβολή του αντικειμένου και το ξεχωρίζουν από το φόντο. Το σχήμα μίας μάσκας είναι τυχαίο και εξαρτάται από παράγοντες όπως το σχήμα του αντικειμένου και η οπτική γωνία προβολής.

Bounding Box

Το bounding box είναι ένα ορθογώνιο παραλληλόγραμμο, το οποίο περιβάλλει κάποιο αντικείμενο. Σε αυτή τη μελέτη οι πλευρές του θα είναι παράλληλες με τον οριζόντιο και τον κάθετο άξονα της εικόνας. Ένα bounding box περιβάλλει ακριβώς το αντικείμενο με την έννοια ότι δεν εμπεριέχεται καμιά γραμμή η στήλη από πίξελ αν κανένα από αυτά δεν είναι μέρος της μάσκας του αντικειμένου.

1.2.2 Σχετικά Μοντέλα

LLaVA

Το Large Language and Vision Assistant (LLaVA)[1] είναι ένα μοντέλο το οποίο δέχεται ένα ζευγάρι εικόνας κειμένου ως είσοδο και παράγει μία κειμενική απάντηση. Το προς επίλυση πρόβλημα καθορίζεται κάθε φορά από την κειμενική είσοδο του μοντέλου και η απάντηση βασίζεται στην εικόνα.

Το LLaVA εκπαιδεύτηκε σε ένα instruction-following σύνολο δεδομένων το οποίο αποτελείται από ερωτήσεις με τις αντίστοιχες απαντήσεις. Το σύνολο δεδομένων δημιουργήθηκε με τη χρήση ενός κειμενικού GPT-4[10] στο οποίο δόθηκε μία περιγραφή της εικόνας και ένα σύνολο ερωτήσεων προς απάντηση. Οι ερωτήσεις καλύπτουν ένα ευρύ φάσμα προβλημάτων όπως απλή συζήτηση, λεπτομερής περιγραφή και σύνθετη συλλογιστική. Με αυτό τον τρόπο δημιουργήθηκε ένα ποικιλόμορφο σύνολο δεδομένων και το γλωσσικό μοντέλο GPT-4 μετέδωσε την γνώση του στο LLaVA.

Η αρχιτεκτονική του LLaVA (Figure 1.1) αποτελείται από έναν κειμενικό κι έναν οπτικό κωδικοποιητή. Ο κειμενικός κωδικοποιητής είναι ένα προ-εκπαιδευμένο γλωσσικό μοντέλο το Vicuna[11], ενώ ο οπτικός κωδικοποίητης είναι το ViT-L/14[12]. Ένα γραμμικό επίπεδο χρησιμοποιείται για να μετατρέψει τα embeddings του οπτικού κωδικοποιητή σε μία αναπαράσταση που αντιστοιχεί στα embeddings του κειμενικού κωδικοποιητή. Τα διανύσματα που προκύπτουν αποτελούν την είσοδο ενός LLM το οποίο παράγει και την απάντηση.

LLaVA-Med

Το Large Language and Vision Assistant για την βιοϊατρική (LLaVA-Med)[8] είναι ένα μοντέλο με ίδια αρχιτεκτονική με το LLaVA. Η διαφορά τους έγκειται στο γεγονός ότι το πρώτο έχει εκπαιδευτεί περαιτέρω σε ένα σύνολο δεδομένων το οποίο προέρχεται από διάφορους κλάδους της ιατρικής όπως ακτίνες X, μαγνητική τομογραφία, ιστολογία, χονδρική παθολογία και υπολογιστική τομογραφία. Η διαδικασία της εκπαίδευσης χωρίζεται σε δύο στάδια.

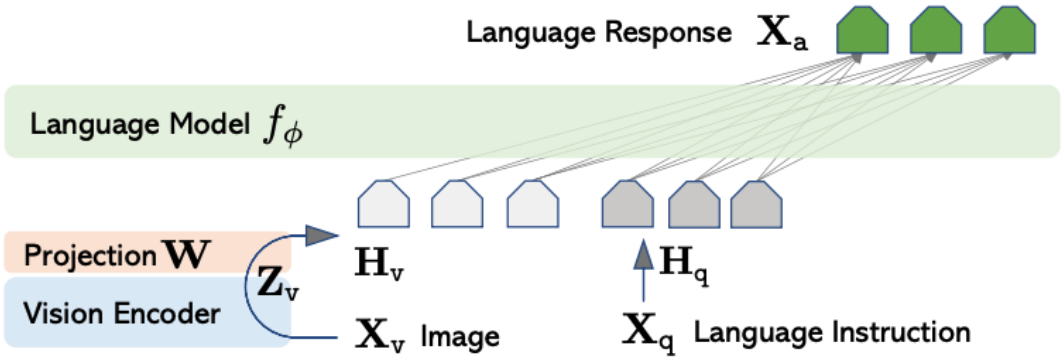


Figure 1.1. *Αρχιτεκτονική του LLaVA [1]*

Στο πρώτο στάδιο ο οπτικός κωδικοποιητής καθώς και τα βάρη του γλωσσικού μοντέλου παραμένουν σταθερά και ανανεώνονται μόνο τα βάρη του γραμμικού επιπέδου που προβάλουν τα embeddings της εικόνας στον χώρο των embeddings του κειμένου. Αυτή η διαδικασία έχει ως στόχο την ευθυγράμμιση κειμένου και εικόνας στην βιοϊατρική καθώς και την επέκταση του λεξιλογίου και της κατανόησης του μοντέλου ως προς τις εικόνες.

Στο επόμενο στάδιο χρησιμοποιείται ένα instruction-following σύνολο δεδομένων για την εκπαίδευση του γλωσσικού μοντέλου και του επιπέδου προβολής. Υποστηρίζεται ότι το μοντέλο έχει καλές επιδόσεις σε VQA σύνολα δεδομένων. Παρόλαυτά για πιο ειδικές περιπτώσεις υπάρχει ανάγκη περαιτέρω εκπαίδευσης και όπως και με άλλα large multimodal models και αυτό το μοντέλο παρουσιάζει περιορισμένη κριτική ικανότητα και hallucinations.

Segment Anything Model 2 (SAM2)

Το SAM2[2] είναι ένα θεμελιώδες μοντέλο το οποίο είναι ικανό να τμηματοποιήσει αντικείμενα σε εικόνες και βίντεο. Δέχεται μάσκες, σημεία ή bounding boxes ως είσοδο και παράγει μια μάσκα. Σε αυτή την εργασία, το μοντέλο θα δοκιμαστεί σε δισδιάστατες ιατρικές εικόνες με bounding box ως είσοδο. Η αρχιτεκτονική του μοντέλου αποτελείται από έναν κωδικοποιητή εικόνας, κωδικοποιητή εισόδου, κωδικοποιητή μνήμης, αποκωδικοποιητή μάσκας, memory attention block και memory bank (Figure 1.2).

Ο κωδικοποιητής εικόνας μετασχηματίζει τα visual data tokens σε embeddings. Χρησιμοποιεί Masked Autoencoder[13] (MAE) pre-trained Hiera[14, 15] image encoder.

Ο κωδικοποίησης εισόδου είναι ένα μέρος του μοντέλου που παράγει τα embeddings εισόδου από μάσκες, σημεία ή bounding boxes.

Ο αποκωδικοποιητης μάσκας είναι υπεύθυνος για την παραγωγή της μάσκας βασισμένος στα embeddings που λαμβάνει. Αυτό το τμήμα του συστήματος μπορεί να παράξει πολλαπλές εξόδους σε περιπτώσεις αμφισημίας, αλλά μόνο μια μάσκα θα δωθεί στην έξοδο χωρίς κάποια ακόμα είσοδο που να εξαλείψει κάθε αμφισημία. Μια σημαντική διαφορά του με το προηγούμενο μοντέλο το SAM[7] είναι ότι αυτό κάποιες φορές μπορεί να μην παράξει μάσκα αν δεν εντοπίσει έγκυρο αντικείμενο για κατάτμηση.

Σε αυτό το σημείο πρέπει να σημειωθεί πως όταν το SAM2 εφαρμόζεται σε εικόνες οι οποίες είναι ανεξάρτητες μεταξύ τους δηλαδή δεν πρόκειται για βίντεο τότε το σύστημα δεν κάνει χρήση των μηχανισμών μνήμης που διαθέτει και συμπεριφέρεται ακριβώς όπως το

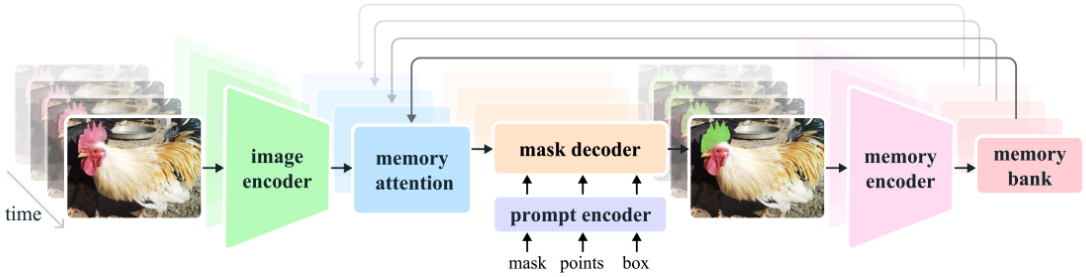


Figure 1.2. *Αρχιτεκτονική του SAM2 [2]*

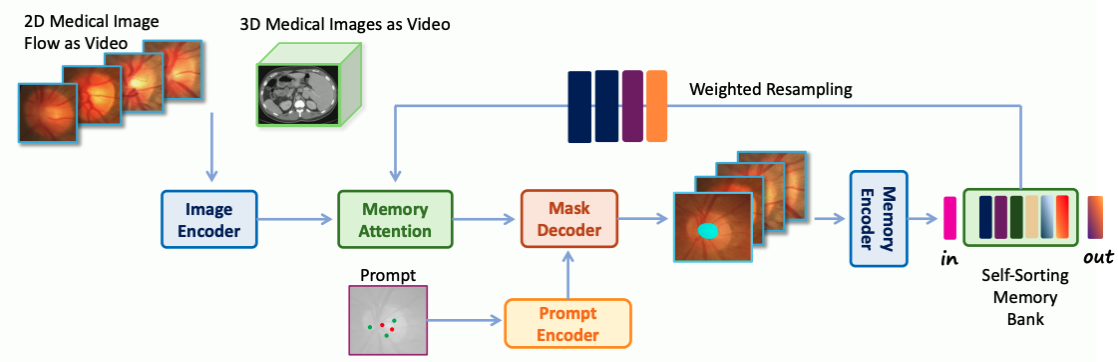


Figure 1.3. *Αρχιτεκτονική του Medical SAM 2 [3]*

SAM. Παρόλα αυτά, εμείς βασιζόμαστε στην εκπαίδευση του μοντέλου που στηρίζεται σε ένα εκτενές σύνολο δεδομένων που αποτελείται από 35.5 εκατομμύρια μάσκες σε 50.9 χιλιάδες βίντεο, το επονομαζόμενο SA-V.

Med-SAM2

Το Medical SAM 2 (MedSAM2)[3] είναι ένα μοντέλο που αναπτύχθηκε για κατάτμηση ιατρικών εικόνων τόσο σε δύο όσο και σε τρεις διαστάσεις. Χρησιμοποιεί ως θεμελιώδες μοντέλο το SAM2 κι είναι τροποποιημένο από αυτό σε αρχιτεκτονική (Figure 1.3) και fine-tuning. Ως προς την αρχιτεκτονική η διαφορά έγκειται στο memory bank το οποίο αποθηκεύει τα K τελευταία πιο χρήσιμα embeddings. Όσον αφορά το dataset που αξιοποιήθηκε για το fine-tuning αυτό ήταν το One-Prompt dataset[16] μια συλλογή από δημοσίως διαθέσιμα datasets που καλύπτουν πολλά modalities και όργανα. Όλα αυτά καθιστούν το μοντέλο γενικεύσιμο σε καινούρια προβλήματα ειδικά στον τομέα της ιατρικής.

Grounding DINO

Το Grounding DINO [4] είναι ένα open-set object detector το οποίο επεξεργάζεται ένα ζευγάρι εικόνας κειμένου ως είσοδο και παράγει στην έξοδο bounding boxes για τα αντικείμενα που περιγράφονται στην κειμενική είσοδο. Χαρακτηρίζεται ως open-set καθώς η κειμενική είσοδος μπορεί να το κατευθύνει να αναγνωρίσει αντικείμενα που δεν είναι παρόντα στο σύνολο δεδομένων εκπαίδευσής του. Αυτή η δυνατότητα καθίσταται εφικτή λόγω του modality fusion που διαθέτει το μοντέλο και των στρατηγικών grounded pre-training.

Η αρχιτεκτονική του μοντέλου (Figure 1.4) αποτελείται από δυο διακριτούς κωδικοποι-

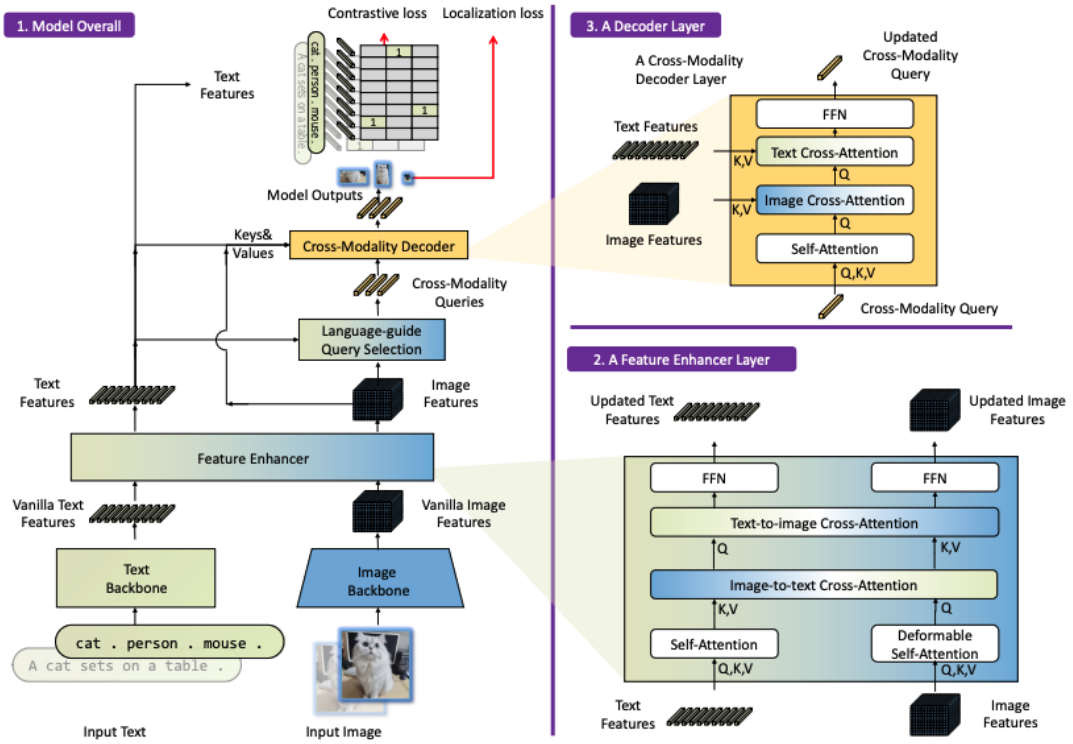


Figure 1.4. *Αρχιτεκτονική του Grounding DINO [4]*

ητές και έναν αποκωδικοποιητή. Ο image encoder χρησιμοποιεί Swin Transformer[17], ενώ ο text encoder είναι τύπου BERT[18]. Ένας feature enhancer πραγματοποιεί το fusion μεταξύ κειμένου και εικόνας με τη χρήση του cross attention. Επίσης, ο language guided query selector επιλέγει τα πιο σχετικά tokens της εικόνας υπολογίζοντας το εσωτερικό τους γινόμενο με τα tokens του κειμένου. Τέλος, ο cross modality decoder είναι υπεύθυνος για την παραγωγή των bounding boxes.

1.2.3 Ιατρική Απεικόνιση

Η ιατρική απεικόνιση έχει έναν σημαντικό ρόλο στην βιοϊατρική έρευνα με εφαρμογές όπως η κατάτμηση οργάνων, η αυτοματοποιημένη διάγνωση και ο σχεδιασμός περίθαλψης. Όσον αφορά την κατάτμηση οργάνων έχει προσελκύσει το ερευνητικό ενδιαφέρον με πολλές προσεγγίσεις βαθιάς μάθησης. Σημαντικό παράδειγμα είναι το U-Net[6] με κάποιες παραλλαγές του [19, 20, 21]. Επίσης Transformer-based αρχιτεκτονικές έχουν αναπτυχθεί για την αντιμετώπιση αυτού του προβλήματος[22, 23].

Ειδικότερα με το Segment Anything Model (SAM) όπου έχουν γίνει πολλές προσπάθειες fine-tuning [24, 25], prompting [26] και adapter-based παρεμβάσεις που ενισχύουν τις επιδόσεις κατάτμησης[27]. Με την εμφάνιση του SAM2 το ερευνητικό ενδιαφέρον επεκτάθηκε και στην τρισδιάστατη κατάτμηση με τη χρήση των μηχανισμών μνήμης που προσφέρει.

Μία άλλη κατεύθυνση ερευνών αφορά multimodal προσεγγίσεις που χρησιμοποιούν τόσο όραση όσο και γλώσσα. LVLM μοντέλα όπως το CLIP[28] έχουν χρησιμοποιηθεί γι' αυτό το σκοπό[29].

Ταυτόχρονα άλλα LVLM όπως το LLaVA[1] έχουν προσαρμοστεί στον τομέα της βιοϊατρικής (LLaVA-Med[8]) για να μπορούν να αναλύουν ιατρικές εικόνες και να απαντούν σε απλές ερωτήσεις ή ακόμα και να περιγράφουν ανωμαλίες που παρατηρούν.

Το **MIU-VL** [30] ενσωματώνει γλώσσα και όραση για τον εντοπισμό και την παραγωγή ενός bounding box στον χώρο της ιατρικής απεικόνισης. Στην είσοδο του δίνεται μόνο το όνομα του αντικειμένου προς εντοπισμό και παράγεται αυτόματα μία κειμενική περιγραφή η οποία θα περιέχει πληροφορίες όπως το χρώμα, το σχήμα και η τοποθεσία της ζητούμενης οντότητας. Το σύνολο των πληροφοριών από το κείμενο και την εικόνα τα επεξεργάζεται ένα VLM και παράγει ένα bounding box. Η έρευνα αυτή έχει ως στόχο να αξιολογήσει κατά πόσο ένα pretrained VLM μπορεί να γενικεύσει σε ιατρικές εικόνες όταν του δίνονται οι κατάλληλες είσοδοι και πόσο το fine-tuning μπορεί να βελτιώσει τις επιδόσεις εντοπισμού.

Το **MedLAM** [31] είναι ένα 3D foundation μοντέλο στον χώρο της ιατρικής απεικόνισης ικανό να εντοπίζει όργανα παράγοντας τρισδιάστατα bounding boxes. Αποτελείται από συνελικτικό κωδικοποιητή και αποκωδικοποιητή καθώς και από πολυεπίπεδο perceptron και έχει εκπαιδευτεί σε 16 σύνολα δεδομένων συμπεριλαμβανομένων 14,012 υπολογιστικών τομογραφιών. Όταν χρησιμοποιείται σε συνδυασμό με το SAM, το επονομαζόμενο MedLSAM είναι ικανό να κατάτμησει ένα όργανο παράγοντας και την αντίστοιχη τρισδιάστατη μάσκα. Η μόνη απαιτούμενη είσοδος είναι η κλάση στην οποία ανήκει το όργανο.

Στην ίδια εργασία για την παραγωγή συγκρίσιμων αποτελεσμάτων με το MedLAM στον εντοπισμό οργάνων χρησιμοποιήθηκε ένα pretrained 3D ResNet50 backbone [32] ακολουθώντας την **DetCo** [33] προσέγγιση που είναι μία μη επιβλεπόμενη contrastive learning μέθοδος για object detection. Επιπρόσθετα αξιολογήθηκαν τα **Mask R-CNN** [34] και **nnDetection** [35] στο ίδιο πρόβλημα. Το Mask R-CNN είναι ένα μοντέλο βαθιάς μάθησης για instance segmentation το οποίο βασίζεται στο Fast/Faster R-CNN [36, 37]. Αποτελείται από δύο στάδια, το region proposal network το οποίο παράγει υποψήφια bounding boxes, ενώ το δεύτερο στάδιο κατηγοριοποιεί το αντικείμενο, βελτιώνει το bounding box και ταυτόχρονα παράγει και μία μάσκα. Το nnDetection είναι μία υλοποίηση, η οποία έχει τη δυνατότητα αυτόματης ρύθμισης παραμέτρων και πραγματοποιεί object detection χρησιμοποιώντας Retina U-Net [38] backbone και ειδικά detection heads για εντοπισμό και κατηγοριοποίηση, ενώ επεκτείνει τις αρχές δόμησης του nnU-Net [21] όσον αφορά την αυτόματη ρύθμιση παραμέτρων.

1.2.4 Grounded Segmentation/Open-set object Detection

Και τα δύο προβλήματα αφορούν τη χρήση κειμενικών περιγραφών για την καθοδήγηση μοντέλων. Η διαφορά τους έγκειται στην έξοδο των μοντέλων που στην περίπτωση του grounded segmentation είναι μία μάσκα ενώ στην άλλη περίπτωση είναι ένα bounding box.

Μερικά από τα προτεινόμενα μοντέλα για την επίλυση των παραπάνω προβλημάτων είναι το GroundingDINO [4] για object detection που παράγει bounding box και το Grounded SAM [39, 7, 4] το οποίο επεκτείνει τις δυνατότητες του SAM επιτρέποντας κειμενική είσοδο.

1.3 Μεθοδολογία

1.3.1 Περιγραφή Συνόλου Δεδομένων

Τα δεδομένα που θα χρησιμοποιηθούν για την εκπαίδευση του Grounding DINO αποτελούνται από τριπλέτες (εικόνα, bounding box, κείμενο). Οι υπολογιστικές τομογραφίες και οι συνθετικές μαγνητικές τομογραφίες προέρχονται από το σύνολο δεδομένων που ονομάζεται RAOS[5, 9] και τα bounding boxes προκύπτουν από τις μάσκες που παρέχει το RAOS. Οι κειμενικές περιγραφές που αναλύονται στην υποενότητα 1.3.5 παρέχουν πληροφορίες για κάθε μάσκα που συμπεριλαμβάνουν την σχετική της θέση στην εικόνα, το σχήμα, την φωτεινότητά της και φυσικά το όνομα της οντότητας που απεικονίζεται. Ένα μέρος του αρχείου που περιέχει το σύνολο δεδομένων φαίνεται στο σχήμα 1.5.

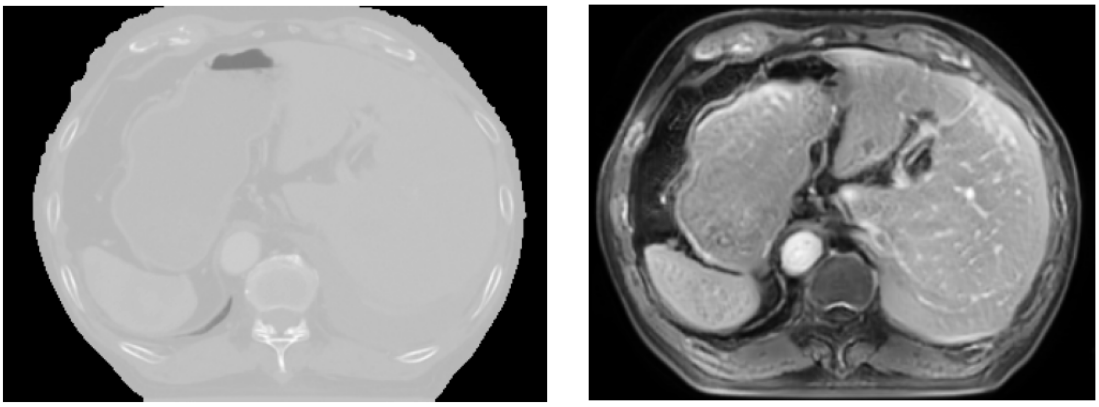

```
{
  "filename": "0_0.jpg",
  "height": 259,
  "width": 331,
  "grounding": {
    "caption": "In the image, the Colon has a slightly curved shape...",
    "regions": [
      {
        "bbox": [240, 107, 286, 157],
        "phrase": "<s> In the image, the Colon has a slightly curved shape.
        </s>Part Colon on lower middle and on right part Colon Dark Gray ."
      },
      {
        {"bbox": [44, 76, 79, 124],
        "phrase": "<s> In the image, the Intestine has a curved shape.
        </s>Part Intestine on upper middle and on left part Intestine Dark Gray ."},
      }
    ]
  }
}
```


Figure 1.5. *Παράδειγμα της δομής του συνόλου δεδομένων που αφορούν μία εικόνα με πολλά bounding boxes*

RAOS

Οι εικόνες και τα bounding boxes που χρησιμοποιούνται στο fine-tuning ανήκουν στο RAOS. Αποτελείται από 413 πραγματικές υπολογιστικές τομογραφίες, 413x9 συνθετικές μαγνητικές τομογραφίες και τις αντίστοιχες μάσκες οι οποίες έχουν παραχθεί από έναν ογκολόγο. Οι κατηγορίες που αποτελούν τα αντικείμενα προς κατάτμηση είναι liver, spleen, left and right kidneys, stomach, gallbladder, esophagus, pancreas, duodenum, colon, intestine, left and right adrenals, και left head of the femur για το synthetic MRI και επιπλέον σε αυτά rectum, bladder, right head of the femur, prostate και seminal vesicle για τα CT scans. Οι περιπτώσεις που παρουσιάζονται κατα την εκπαίδευση θεωρούνται "συνήθεις", με την έννοια πως κανένα όργανο δεν έχει αφαιρεθεί χειρουργικά. Οι ασθενείς

είναι 287 σε αριθμό σε αυτό το σύνολο. Οι 220 από αυτούς θα χρησιμοποιηθούν για την εκπαίδευση, ενώ οι υπόλοιποι 67 θα αποτελέσουν το test set. Η εκπαίδευση θα γίνει τόσο σε εικόνες CT όσο και σε MRI.



(A) CT scan (B) Delayed MRI

Figure 1.6. *Στα αριστερά (A) φαίνεται ένα CT scan. Στα δεξιά (B) απεικονίζεται η ίδια περιοχή σε μορφή MRI του τύπου "Delay". Παρατηρούμε ότι η δεξιά εικόνα αναδεικνύει με μεγαλύτερη ευκρίνεια την μορφή των οργάνων καθώς οι αντιθέσεις είναι μεγαλύτερες.*

1.3.2 Αξιολόγηση του συνόλου δεδομένων

Οι κειμενικές περιγραφές των μασκών που παράχθηκαν αξιολογήθηκαν για την ακρίβειά τους από ένα LLM και συγκεκριμένα το text-only GPT-4o-mini [40] . Αρχικά δε δόθηκαν στο LLM κατευθυντήριες οδηγίες και αυτό εστίαζε σε στοιχεία της μορφής των περιγραφών που ήταν αδύνατο να αλλάξουν, όπως η συνοχή των προτάσεων (Figure 1.7). Στη συνέχεια επιχειρήθηκε να δοθεί μια παραπάνω καθοδήγηση στο LLM για την αξιολόγηση ώστε να μην βασίζεται σε λάθος στοιχεία (Figure 1.8).

You are a dataset quality assessor. Rate the caption from 1 to 10 based on quality and accuracy.
Assessing region phrase: <s> In the image, the Liver has a slightly oval shape. </s>Liver on lower middle and on right part . Liver Light Gray .

Rating: 5

The caption describes the liver with some basic characteristics, such as its shape and color; however, it is vague and lacks clarity. The statement "Liver on lower middle and on right part" is confusing, as it does not accurately convey the liver's location in relation to the rest of the image. Overall, the caption needs improvement in precision and completeness.

Figure 1.7. *Αίτημα για αξιολόγηση της περιγραφής της μάσκας*

Στο παράδειγμα του σχήματος 1.8 το μοντέλο εστιάζει για άλλη μια φορά στην έλλειψη συνοχής και στα γραμματικά και συντακτικά λάθη. Όμως, η εξάλειψη αυτών των στοιχείων δεν είναι δυνατή με την ύπαρξη ενός φίλτρου που απομακρύνει συχνά εμφανιζόμενες λέξεις.

You are a dataset quality assessor. Rate the caption from 1 to 10 based on quality and accuracy keeping in mind that this is the dataset for grounded training of a model that detects bounding boxes. We would like to give the model the shape, relative position of the organ in the image and its brightness.
Assessing region phrase: <s> In the image, the Liver has a slightly oval shape. </s>Liver on lower middle and on right part . Liver Light Gray .

I would rate this caption a 7 out of 10.

Here's my reasoning: - **Shape**: It accurately describes the liver as having a slightly oval shape, which is relevant for the model to learn about the organ's form. - **Position**: It indicates that the liver is located in the lower middle and the right part of the image, which provides useful information about its relative position. - **Brightness**: It describes the liver's color as light gray, which is important for the model to learn about the organ's appearance.
However, the phrase structure is a bit fragmented and could be improved for clarity. Additionally, a more precise description of the liver's spatial relation to other organs or anatomical landmarks could enhance the caption's informative value. Overall, it contains the essential elements needed for training but could benefit from refinement.

Figure 1.8. *Αίτημα για αξιολόγηση της περιγραφής της μάσκας με καθοδήγηση*

Τα prompts που δόθηκαν τελικά στο LLM με βάση τα οποία έκρινε όλο το dataset φαίνονται στο σχήμα 1.9.

You are a dataset quality assessor. Rate the caption from 1 to 10 based on quality and accuracy keeping in mind that this is the dataset for grounded training of a model that detects bounding boxes in synthetic MRI images of type Delay. We would like to give the model the shape, relative position of the organ in the image and its brightness. We have filtered out words that appear too often. Thus, only important keywords remain that bind the relative words with the region. Disregard grammatical and syntactical correctness.
Assessing region phrase: <s> In the image, the Liver has a slightly oval shape. </s>Liver on lower middle and on right part . Liver Light Gray .

Rating: 8

The caption accurately describes the shape (slightly oval) and position (lower middle and right part) of the liver, as well as its brightness (light gray). It includes important keywords that are relevant for the model's understanding of the texture, location, and appearance of the organ in the synthetic MRI image. However, the inclusion of "on" before "right part" could be clearer to specify the exact position more effectively. Overall, it provides a good amount of relevant information for grounded training.

Figure 1.9. *Αίτημα για αξιολόγηση της περιγραφής της μάσκας με λεπτομερή καθοδήγηση*

Ο μέσος όρος των βαθμολογιών που παράχθηκαν από το LLM αποτελεί μια μετρική της ορθότητας των περιγραφών.

Μέση Βαθμολογία
7.25/10

Table 1.1. *Μέση βαθμολογία της ορθότητας των κειμενικών περιγραφών των μασκών*

1.3.3 Σχήματα Οργάνων που δίνονται από το LLaVA-Med

Έχουμε αναθέσει στο LLaVA-Med την παραγωγή των σχημάτων των μασκών για την κάθε κατηγορία οργάνου στο dataset. Παρακάτω παρουσιάζονται τα αναλυτικά ιστογράμματα για κάθε όργανο.

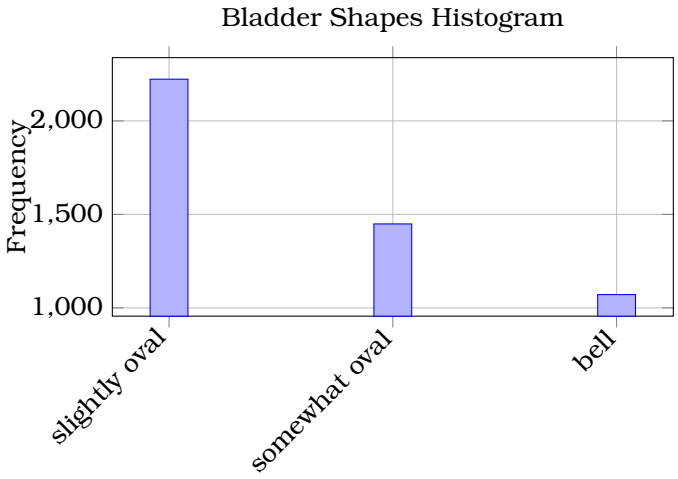


Figure 1.10. *Ιστόγραμμα σχημάτων Bladder*

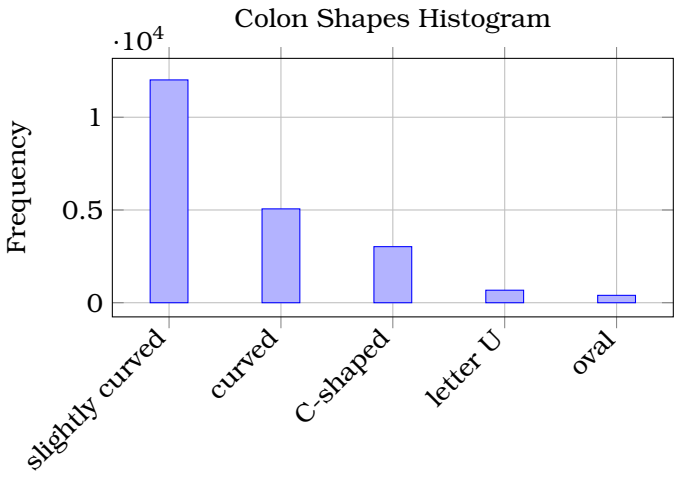


Figure 1.11. *Ιστόγραμμα σχημάτων Colon*

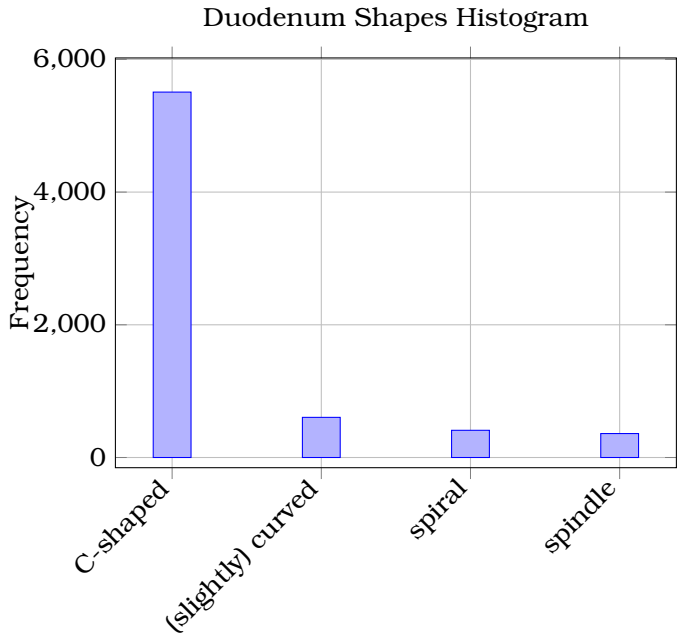


Figure 1.12. *Ιστόγραμμα σχημάτων Duodenum*

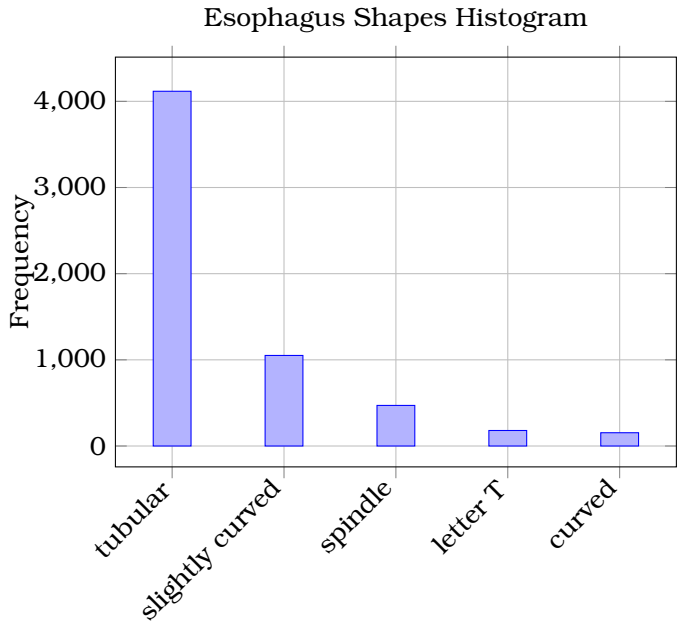


Figure 1.13. *Ιστόγραμμα σχημάτων Esophagus*

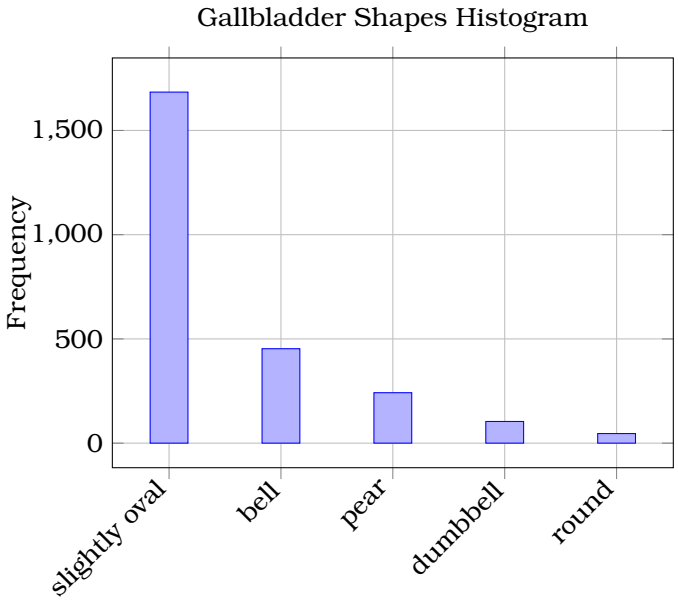


Figure 1.14. *Ιστόγραμμα σχημάτων Gallbladder*

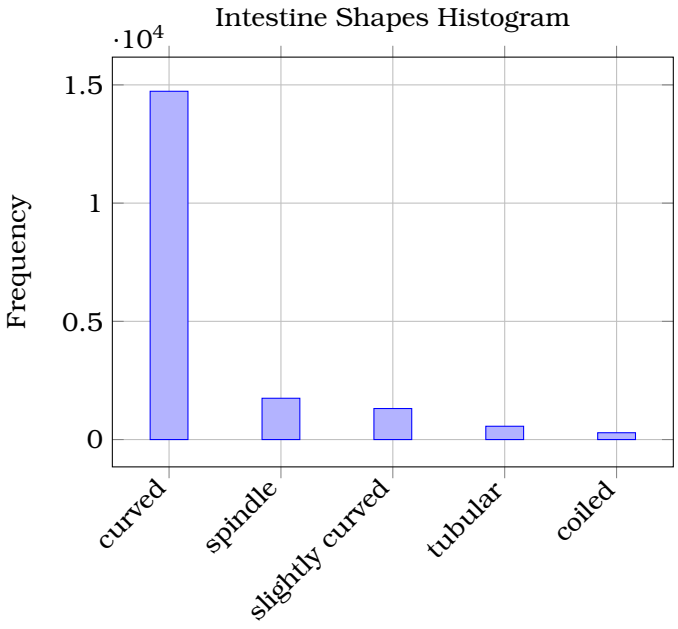


Figure 1.15. *Ιστόγραμμα σχημάτων Intestine*

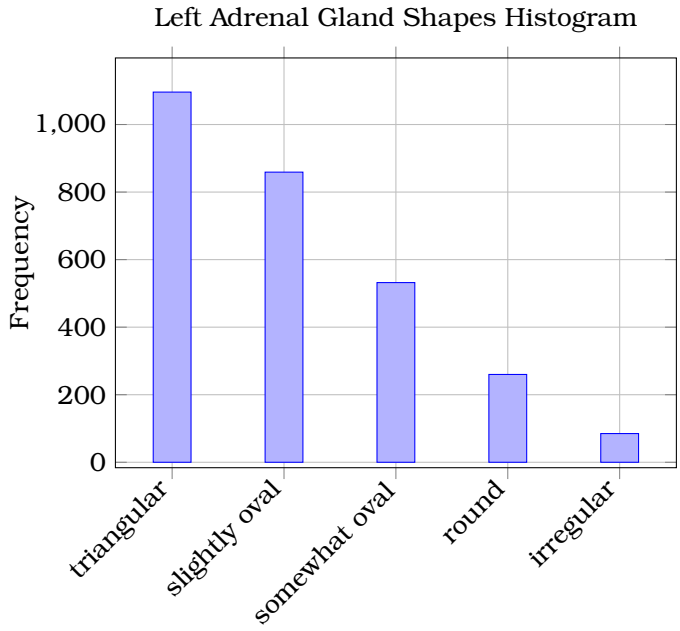


Figure 1.16. *Ιστόγραμμα σχημάτων Left Adrenal Gland*

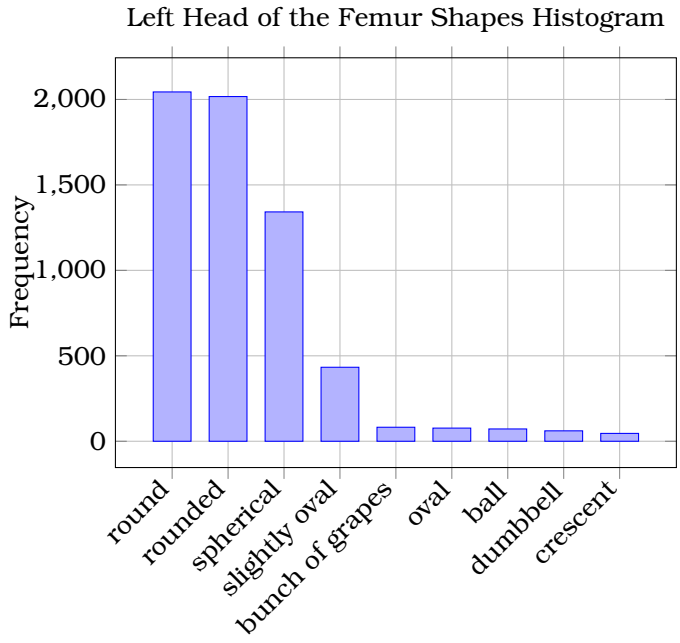


Figure 1.17. *Ιστόγραμμα σχημάτων Left Head of the Femur*

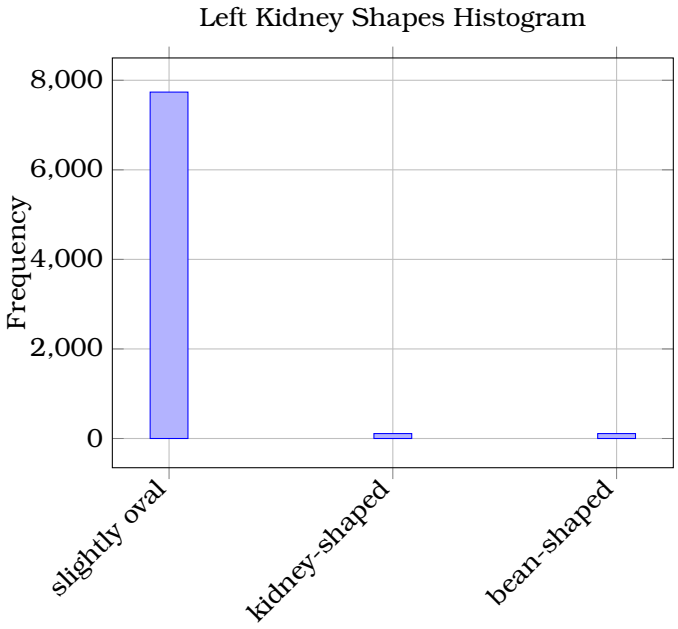


Figure 1.18. *Ιστόγραμμα σχημάτων Left Kidney*

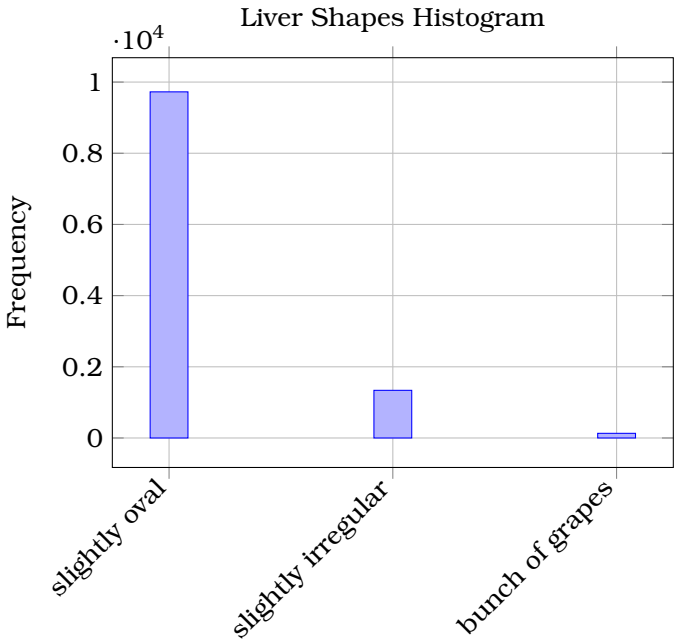


Figure 1.19. *Ιστόγραμμα σχημάτων Liver*

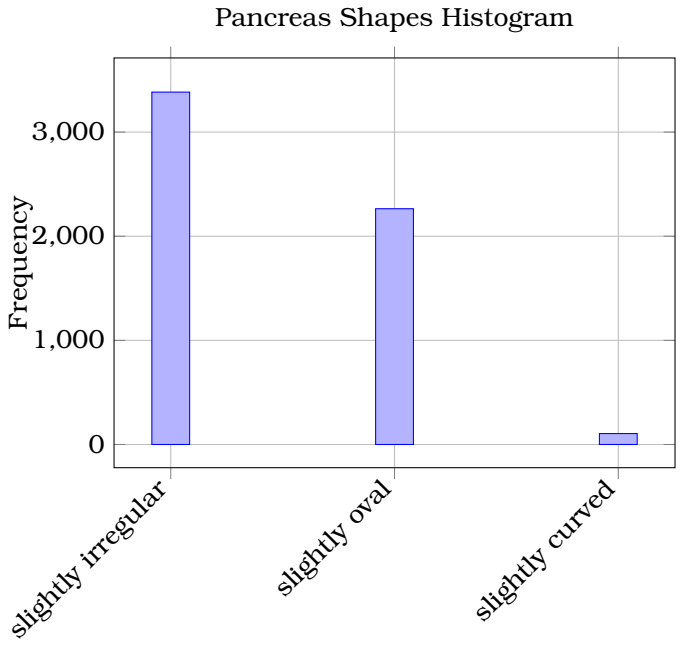


Figure 1.20. *Ιστόγραμμα σχημάτων Pancreas*

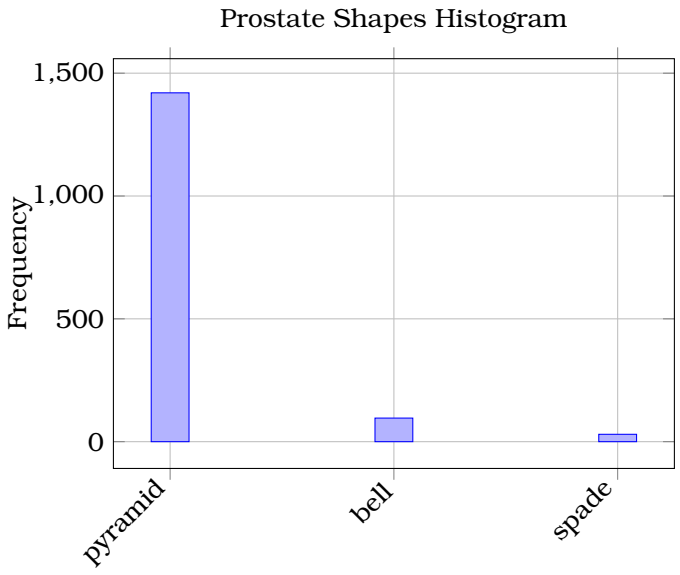


Figure 1.21. *Ιστόγραμμα σχημάτων Prostate*

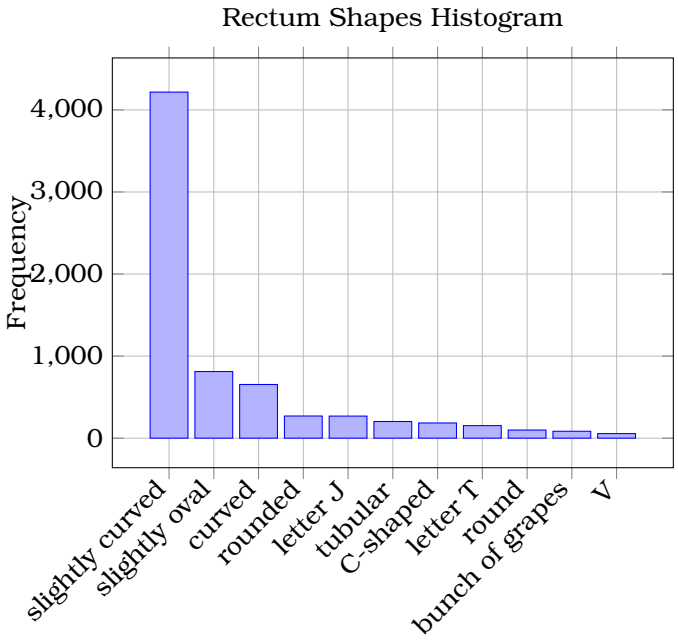


Figure 1.22. *Ιστόγραμμα σχημάτων Rectum*

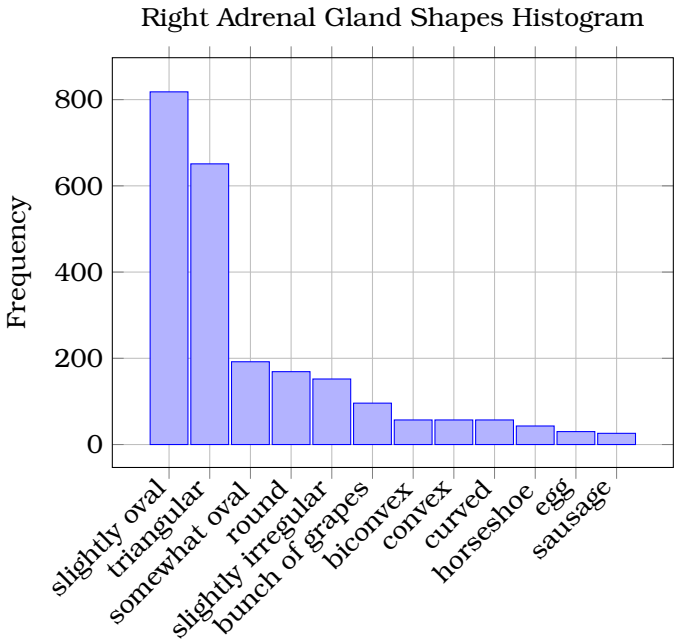


Figure 1.23. *Ιστόγραμμα σχημάτων Right Adrenal Gland*

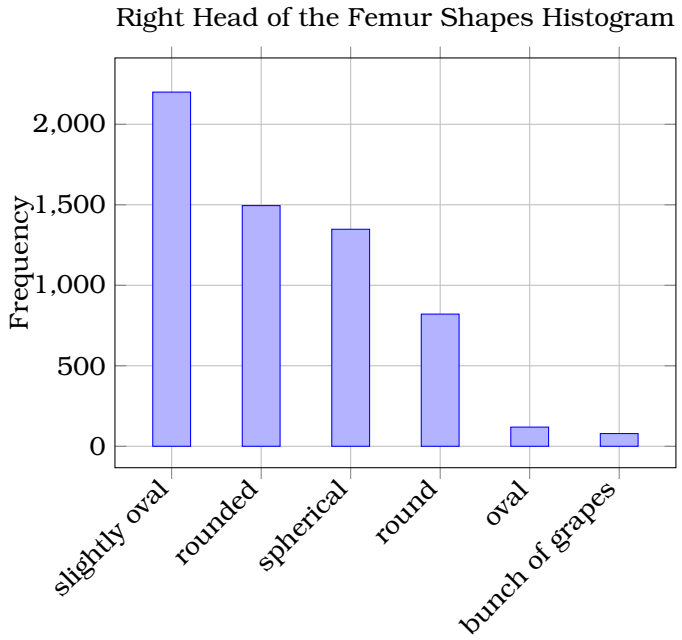


Figure 1.24. *Ιστόγραμμα σχημάτων Right Head of the Femur*

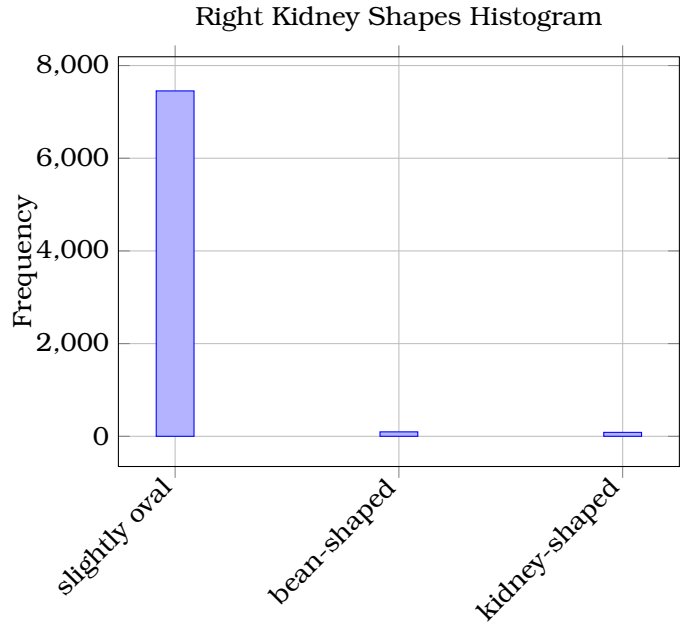


Figure 1.25. *Ιστόγραμμα σχημάτων Right Kidney*

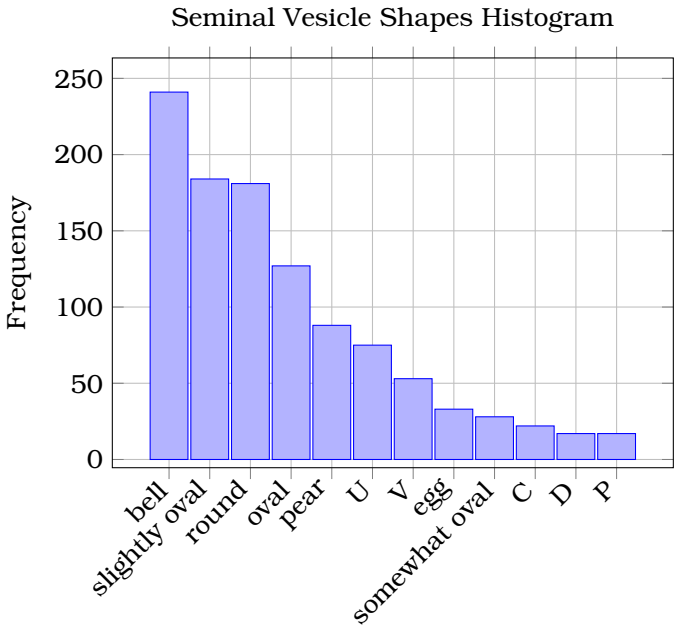


Figure 1.26. *Ιστόγραμμα σχημάτων Seminal Vesicle*

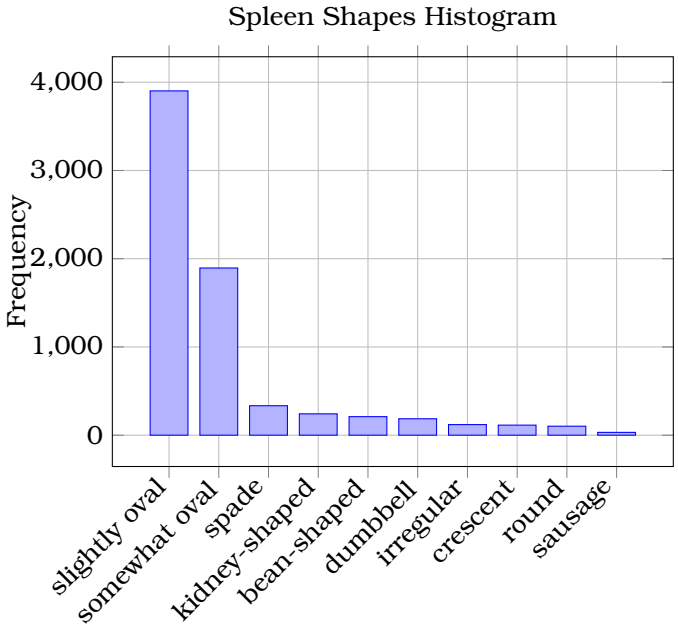


Figure 1.27. *Ιστόγραμμα σχημάτων Spleen*

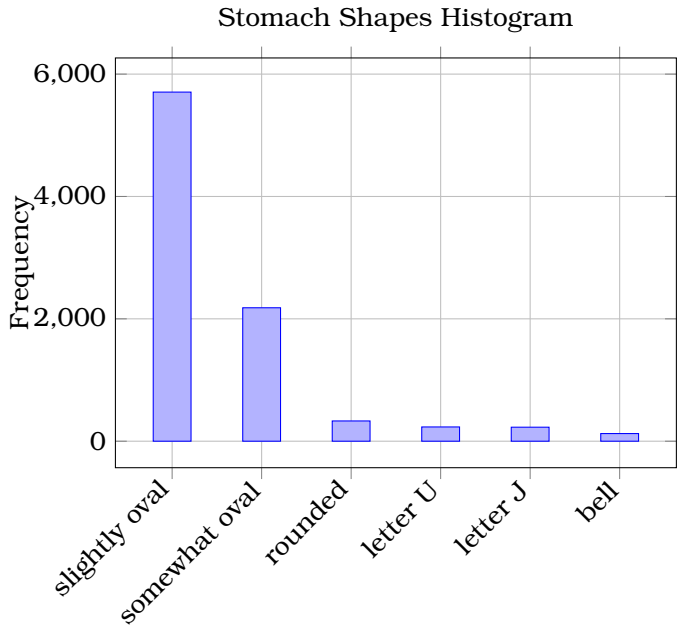


Figure 1.28. *Ιστόγραμμα σχημάτων Stomach*

1.3.4 Επαύξηση Δεδομένων

Για να γίνει το μοντέλο πιο εύρωστο ενσωματώθηκαν στο training dataset οι ίδιες εικόνες αλλά περιστραμμένες κατά 90, 180 και 270 μοίρες. Επίσης, προστέθηκαν και αναποδογυρισμένα αντίγραφα κατα τον κάθετο άξονα που διέρχεται από το κέντρο των εικόνων. Το μοντέλο διαθέτει επίσης αυτόματο cropping και αλλαγή των διαστάσεων της εικόνας. Όλα τα παραπάνω αποσκοπούν στην βελτίωση των επιδόσεων του μοντέλου.

1.3.5 Στάδια

Παραγωγή των Κειμενικών Περιγραφών

Στο πρώτο στάδιο των πειραμάτων χρειάστηκε να δοκιμάσουμε πολλές μορφές ερωτήσεων προς το LLaVA-Med ώστε να μας παράξει ακριβείς περιγραφές για κάθε μάσκα. Στόχος είναι να μεταφερθεί η γνώση του fine-tuned LVLM στο Grounding DINO ώστε το σύστημα που θα δημιουργηθεί να έχει τη δυνατότητα να κατανοήσει την περιγραφή και να την ταιριάξει στην σωστή μάσκα.

Για να το πετύχουμε αυτό επιλέξαμε κάποια χαρακτηριστικά των μασκών που τις ξεχωρίζουν από τις υπόλοιπες και μπορούν να περιγραφούν λεκτικά. Αυτά είναι αναλυτικά το σχήμα, η σχετική θέση, η φωτεινότητα και η υφή. Επομένως αυτά θα ζητηθούν από το LLaVA-Med να αναγνωρίσει και να περιγράψει. Βέβαια, αποφασίστηκε πως η υφή δεν θα χρησιμοποιηθεί ως χαρακτηριστικό και λόγω περιορισμών στην κριτική ικανότητα του LLaVA-Med επίσης η θέση και η φωτεινότητα θα εξαχθούν με τη χρήση ντετερμινιστικών αλγορίθμων. Ένα σχήμα που παρουσιάζει την προαναφερθείσα διαδικασία είναι το 1.29 . Οι αναλυτικοί λόγοι των παραπάνω επιλογών αναλύονται στην υποενότητα 1.4.1.

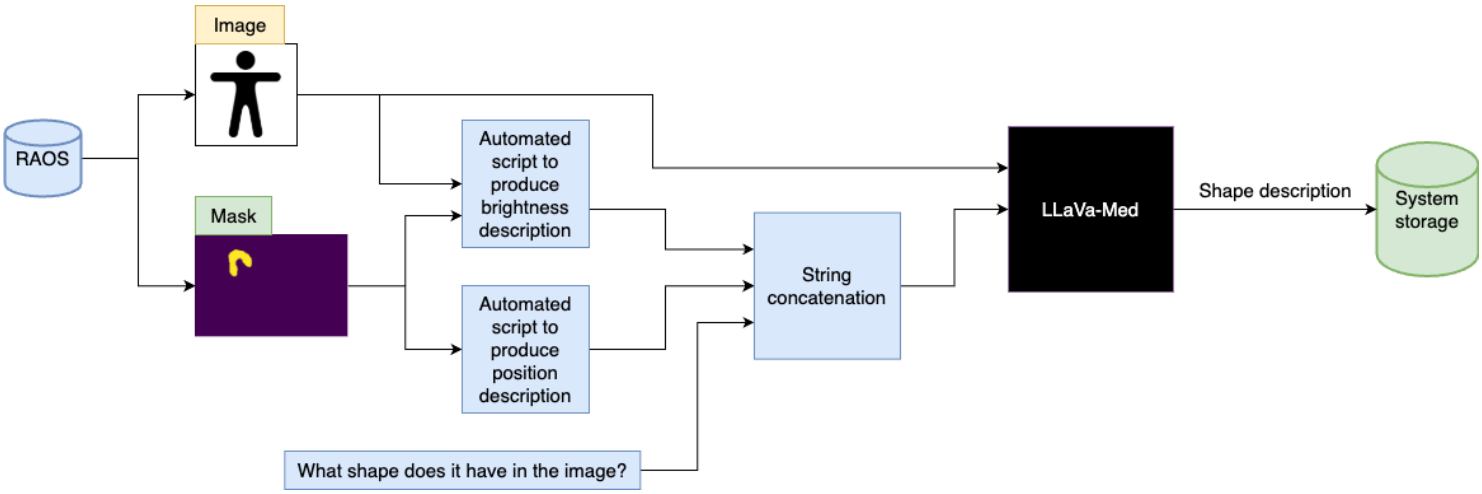
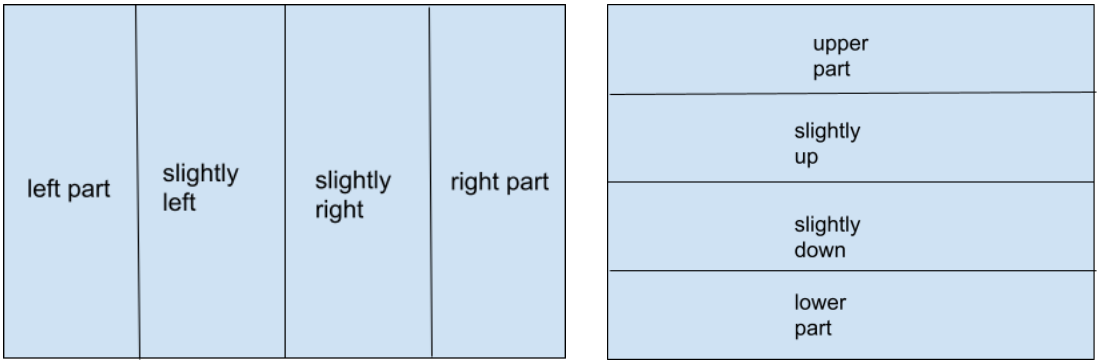


Figure 1.29. *Εξαγωγή των κειμενικών περιγραφών του σχήματος κάθε μάσκας*

Σχετική Θέση

Οι σχετικές θέσεις που επιστρέφει το LLaVA-Med είναι λανθασμένες και ακατάλληλες για χρήση στην εκπαίδευση του Grounding DINO. Επομένως δημιουργήθηκε ένας ντετερμινιστικός αλγόριθμος για την εξαγωγή της θέσης της κάθε μάσκας μέσα στην εικόνα. Συγκεκριμένα, για κάθε μάσκα υπολογίζεται το κέντρο μάζας της και ανάλογα με την περιοχή στην οποία ανήκει εξάγεται ένας τοπικός προσδιορισμός για τον κάθετο άξονα και ενας για τον οριζόντιο άξονα. Οι υποδιαιρέσεις της εικόνας φαίνονται στο ακόλουθο σχήμα 1.30 .



(A) Horizontal division (B) Vertical division

Figure 1.30. *Η εικόνα χωρίζεται σε τέσσερα ίσα μέρη ως προς του οριζόντιο και του κάθετο άξονα. Σε κάθε μέρος αντιστοιχίζεται μία ταμπέλα. Το κέντρο μάζας κάθε μάσκας χρησιμοποιείται για τον προσδιορισμό της θέσης της μάσκας και λαμβάνει έναν προσδιορισμό από την οριζόντια διαίρεση και έναν από την κάθετη διαίρεση. Ένα παράδειγμα πρότασης είναι : "The Left Kidney is located slightly up and slightly left in the image."*

Φωτεινότητα

Υλοποιήθηκε επίσης μία αυτοματοποιημένη μέθοδος για την παραγωγή των περιγραφών που αφορούν την φωτεινότητα. Καθώς γνωρίζουμε όλες τις μάσκες των οντοτήτων που επιθυμούμε να κατατμήσουμε μπορούμε να τις χρησιμοποιήσουμε για να υπολογίσουμε την μέση τιμή της φωτεινότητας των πίξελ που την αποτελούν. Για την κατηγοριοποίηση της μάσκας σε κάποια από τις κλάσεις της φωτεινότητας δημιουργούμε την ακόλουθη κλίμακα 1.31.

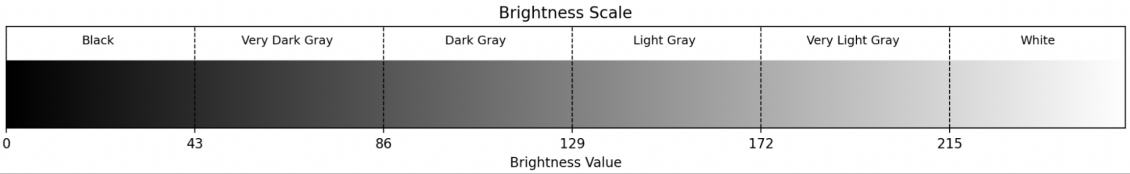
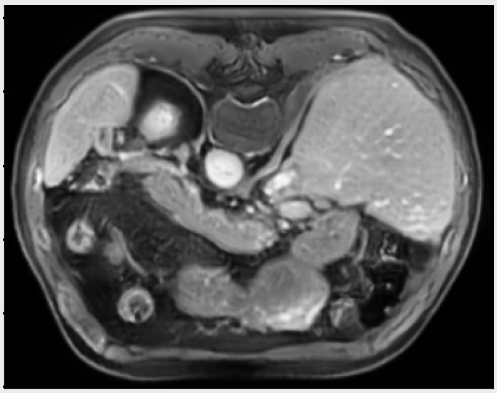


Figure 1.31. *Κλίμακα για την κατηγοριοποίηση της φωτεινότητας κάθε μάσκας σε μια κλάση με βάση την μέση τιμή της*

Σχήμα

Αποφασίσαμε να παρέχουμε στο LLaVA-Med την πληροφορία για την θέση και την φωτεινότητα της οντότητας έτσι ώστε να βοηθηθεί στην διαδικασία εντοπισμού της στην εικόνα. Τα τελικά prompts που δόθηκαν στο LVLM καταδεικνύονται στο επόμενο παράδειγμα 1.32.



User: The Stomach is located on the lower part and slightly right in the image and appears to be Dark Gray. What shape does it have in the image?

LLaVA-Med: The stomach in the image has a round shape.

Figure 1.32. *Τελικά prompts, χρησιμοποιώντας την πληροφορία για την θέση και την φωτεινότητα, η οποία θα βοηθήσει το μουτέλο να εντοπίσει το ζητούμενο όργανο.*

Εκπαίδευση του Grounding DINO

Σε αυτή τη φάση χρησιμοποιήθηκε το Open Grounding DINO [41] μια υλοποίηση του Grounding DINO με δυνατότητα fine-tuning. Στο διάγραμμα που ακολουθεί φαίνεται η διαδικασία εκπαίδευσης 1.33.

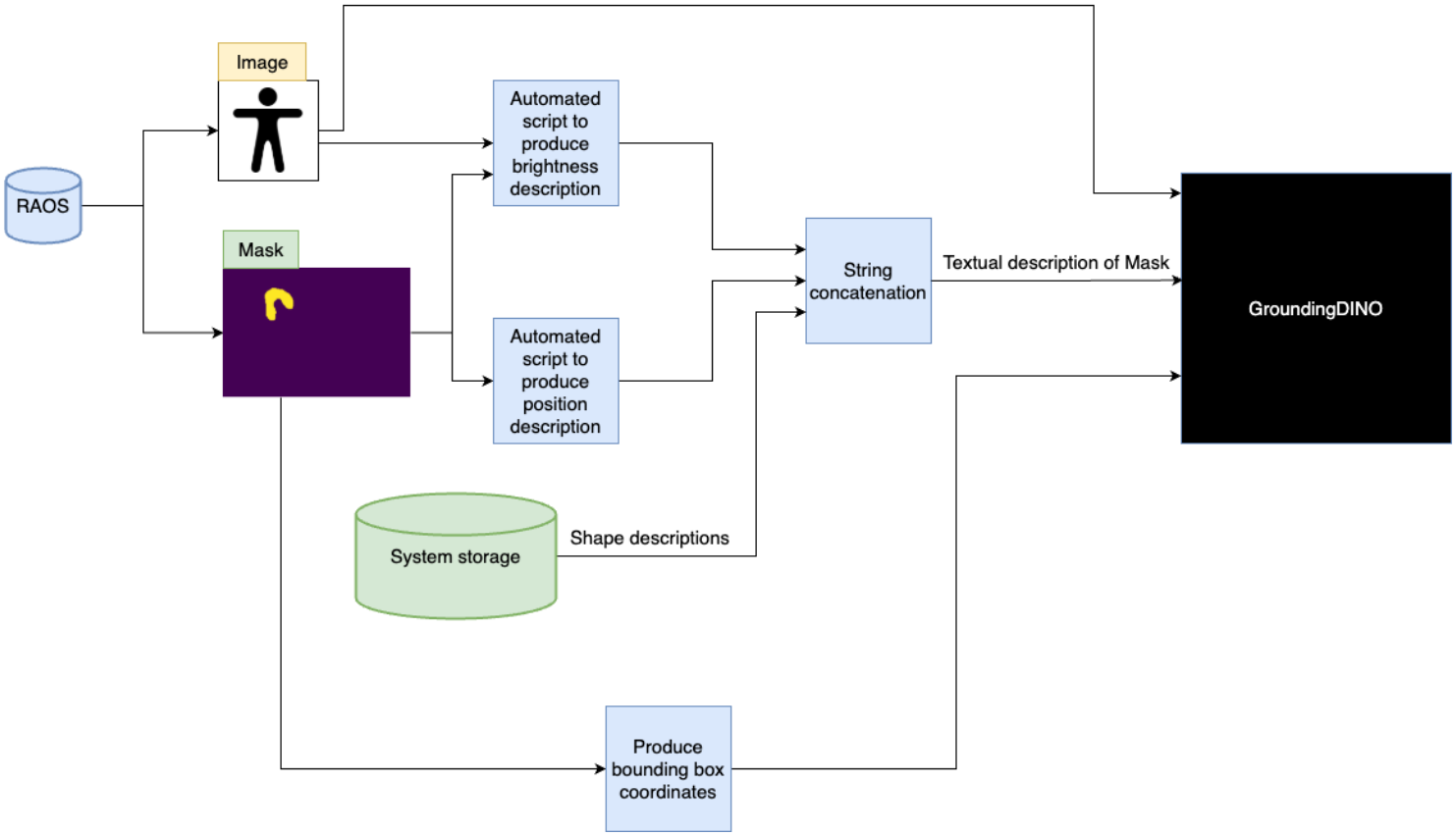
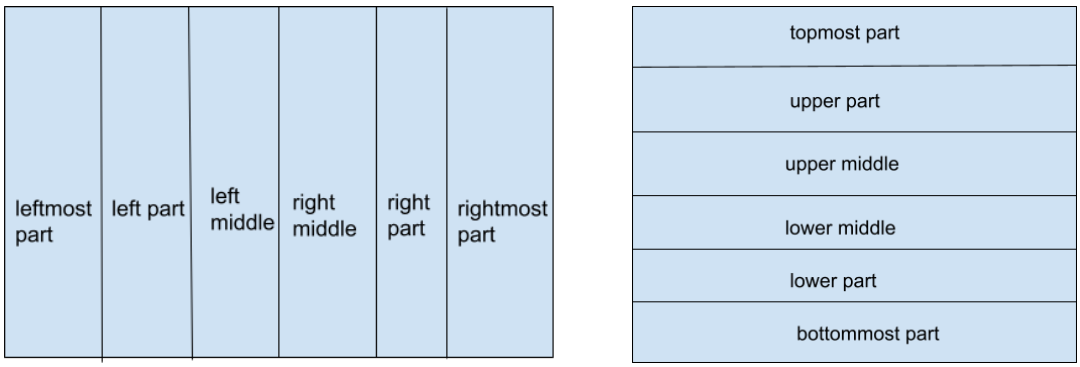


Figure 1.33. *Fine-tuning του GroundingDINO*

Το dataset που χρησιμοποιήθηκε ήταν το RAOS εμπλουτισμένο με τις περιγραφές των μασκών που περιέχουν πληροφορία για τη σχετική θέση, φωτεινότητα και σχήμα της κάθε μάσκας. Η φωτεινότητα υπολογίζεται με τον τροπο που παρουσιάστηκε προηγουμένως, ενώ η θέση υπολογίζεται όπως πριν με την διαφορά ότι αυξήθηκαν οι υποδιαιρέσεις της εικόνας για μεγαλύτερη ακρίβεια (Σχήμα 1.34). Για το σχήμα των μασκών χρησιμοποιήθηκε το LLaVA-Med.



(A) Horizontal division (B) Vertical division

Figure 1.34. *Η εικόνα χωρίζεται σε έξι μέρη ως προς του κάθετο και οριζόντιο άξονα. Υπολογίζεται το κέντρο βάρους της μάσκας και από την θέση του προσδίδονται δυο τοπικοί προσδιορισμοί στην μάσκα. Ένα παράδειγμα είναι: "The Left Kidney is located on the topmost part and on the left part in the image."*

Πριν πραγματοποιηθεί η συνένωση των τριών περιγραφών εφαρμόστηκε ένα φίλτρο για την αφαίρεση των συχνά εμφανιζόμενων λέξεων που αφορούν τις προτάσεις για φωτεινότητα και θέση. Αυτό αποσκοπούσε στην διευκόλυνση του μοντέλου να εστιάσει σε σημαντικές πληροφορίες. Αντιθέτως, κανένα φίλτρο δεν εφαρμόστηκε στην περιγραφή του σχήματος για τις κανονικές εικόνες αλλά στα επαυξημένα δεδομένα φιλτραριστήκαν όλες οι προτάσεις.

Υπερπαράμετροι

Το Grounding DINO εκπαιδεύτηκε χρησιμοποιώντας το Swin-B [17] (swin_B_384_22k) ως backbone εικόνας και το BERT-base-uncased [18] ως κωδικοποιητή κειμένου. Το μοντέλο λειτουργεί με 900 queries και υποστηρίζει μέγιστο μήκος κειμενικού token 256. Διαθέτει έξι επίπεδα encoder και έξι επίπεδα decoder, με διάσταση κρυφού επιπέδου 256 και διάσταση feedforward 2048. Ο μετασχηματιστής χρησιμοποιεί 8 κεφαλές προσοχής (attention heads) με ReLU ενεργοποίηση. Η εκπαίδευση πραγματοποιείται με μέγεθος παρτίδας (batch size) 4, βασικό ρυθμό μάθησης 0.0001 και weight decay 0.0001. Εφαρμόζεται data augmentation με κλίμακες από 480 έως 800 και μέγιστο μέγεθος εικόνας 1333. Ο βελτιστοποιητής (optimizer) μειώνει τον ρυθμό μάθησης στις εποχές (epochs) 4 και 8. Για τον υπολογισμό της απώλειας (loss computation), χρησιμοποιείται το Hungarian matching με classification, L1 και GIoU losses. Το μοντέλο βελτιώνεται περαιτέρω μέσω text cross-attention και ενός επιπέδου σύντηξης (fusion layer).

Περιβάλλον

Το περιβάλλον εκπαίδευσης αποτελούνταν από έναν NVIDIA A10G GPU έκδοση driver 550.127.05 και CUDA 12.4. Η διαδικασία εκπαίδευσης χρειάστηκε περίπου 44 ώρες για τις MRI εικόνες, ενώ το LLaVA-Med χρειάστηκε περίπου 24 ώρες για την παραγωγή των περιγραφών των σχημάτων.

Αξιλόγηση της εκπαίδευσης

Η απόδοση του μοντέλου ποσοτικοποιείθηκε μέσω δυο μετρικών συγκεκριμένα την τομή προς ένωση (IoU) και την ακρίβεια. Για την σύγκριση με άλλα μοντέλα χρησιμοποιήθηκαν οι μετρικές dice score coefficient και wall distance.

Τομή προς Ένωση (IoU)

Η μετρική αυτή παίρνει ως είσοδο τις δυο μάσκες, την πραγματική και την παραγόμενη από το μοντέλο. Υπολογίζεται ως ο λόγος της τομής προς την ένωση των δυο μασκών:

$$\text{IoU} = \frac{\text{Εμβαδόν Τομής}}{\text{Εμβαδόν Ένωσης}} = \frac{|A \cap B|}{|A \cup B|} \tag{1.1}$$

Ακρίβεια

Η μετρική αυτή ορίζεται ως ο λόγος των true positives προς το άθροισμα των true positives με τα false positives.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{1.2}$$

Dice Similarity Coefficient

Το Dice Similarity Coefficient εκφράζεται μαθηματικά:

$$\mathrm{DSC}(A,B) = \frac{2|A \cap B|}{|A| + |B|}. \tag{1.3}$$

Wall Distance

Το wall distance εκφράζει την μέση απόλυτη διαφορά μεταξύ των ορίων του προβλεπόμενου και του πραγματικού bounding box. Στις δύο διαστάσεις τα όρια είναι τέσσερα ενώ στις τρεις είναι έξι. Δίνεται το wall distance για ένα 2D bounding box:

$$\mathrm{WD} = \frac{1}{4}\sum_{i=1}^{4} |\hat{d}_i - d_i| \tag{1.4}$$

όπου τα d_i και $\hat{d}_i$ είναι τα όρια του προβλεπόμενου και του πραγματικού bounding box αντίστοιχα και ανήκουν στην ίδια πλευρά.

Αξιολόγηση

Το σύστημα που προτείνεται αξιολογείται σε ένα σύνολο από εικόνες MRI και CT της κοιλιακής χώρας από 67 ασθενείς. Αναλυτικότερα, παρέχονται στο GroundingDINO μια εικόνα με μια περιγραφή της οντότητας προς κατάτμηση. Έπειτα το παραγόμενο bounding box δίνεται σε ένα μοντέλο όπως το SAM2 ή το Med-SAM2 το οποίο παράγει και την προβλεπόμενη μάσκα. Η διαδικασία φαίνεται στο διάγραμμα 1.35.

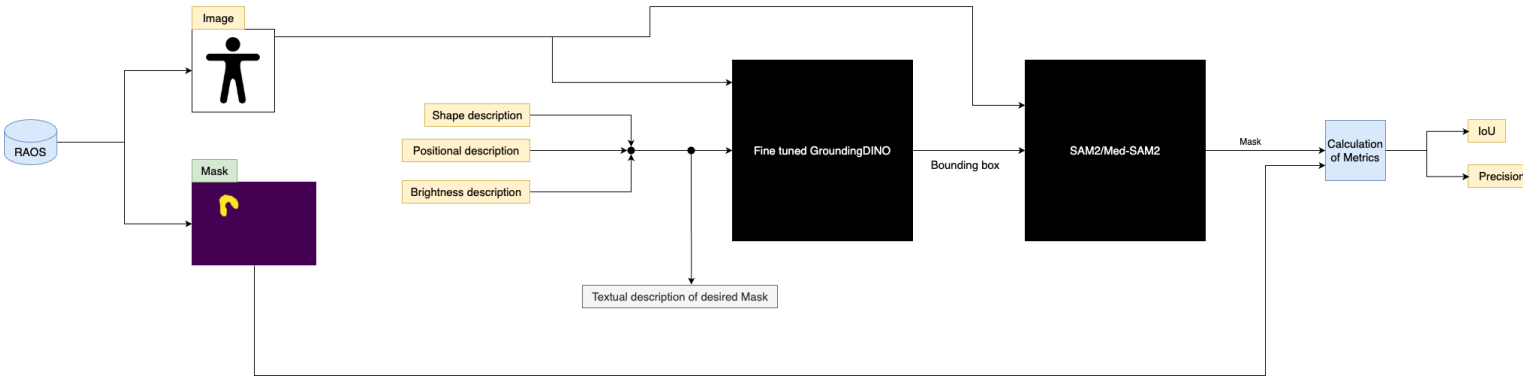


Figure 1.35. *Αξιολόγηση του συστήματος*

1.4 Αποτελέσματα

1.4.1 Prompting LLaVA-Med

Αρχικά εξετάστηκε η ικανότητα του LLaVA-Med να αναγνωρίζει τα όργανα που παρουσιάζονται σε μία εικόνα. Οι απαντήσεις του δεν ήταν ακριβείς, καθώς όργανα που δεν ανήκουν στο σύνολο δεδομένων αναγνωρίζονταν και τα σχετικά όργανα τα αγνοούσε. Για την βελτίωση των απαντήσεων αποφασίστηκε να αναφέρονται ρητώς τα ονόματα των σχετικών οργάνων πριν την διατύπωση περεταίρω ερωτήσεων όπως φαίνεται στο παράδειγμα 1.36.

Στο επόμενο παράδειγμα του σχήματος 1.37 η εικόνα έχει αναστραφεί και αναφέρουμε ρητώς τα ονόματα των σχετικών οργάνων πριν την διατύπωση της ερώτησης. Η απάντηση που δίνεται καταδεικνύει ότι το μοντέλο δεν είναι εύρωστο σε αναστραμμένες εικόνες, καθώς επιστρέφει την γενική ανατομική θέση του liver. Όσον αφορά το σχήμα του kidney, το

LLaVA-Med σωστά το αναγνωρίζει ως “bean-shaped”. Παρόλαυτά, παρέχει άχρηστες γενικές πληροφορίες όπως είναι η λειτουργία των νεφρών που δε συνεισφέρουν στην ικανότητα του GroundingDINO να παράγει ακριβή bounding boxes. Από αυτό το παράδειγμα καταλαβαίνουμε ότι το μοντέλο απαντά με βάση πληροφορίες που έχει μάθει και δεν τις εξάγει από την εικόνα.

Στο επόμενο πείραμα (Σχήμα 1.38) περιστρέφουμε την εικόνα 180 μοίρες για να εξετάσουμε αν οι περιγραφές των θέσεων των οργάνων θα είναι πιο ακριβείς από προηγουμένως. Παρόλο που κάποιες περιγραφές θέσεων βελτιώθηκαν, υποθέτουμε ότι οι απαντήσεις δεν είναι αποτέλεσμα κατανόησης της εικόνας αλλά ευθυγραμμίζονται με την γνώση που έχει αποκτήσει το μοντέλο κατά την εκπαίδευση του. Επίσης το LLaVA-Med ξεχνάει να απαντήσει για την θέση και το σχήμα κάποιον οργάνων αποτυγχάνοντας με αυτό τον τρόπο να φέρει εις πέρας όλα όσα του έχουν ανατεθεί.

Επίσης δοκιμάστηκε και μία άλλη μορφή ερωτήσεων για τον εντοπισμό των θέσεων που έχει να κάνει με τα τεταρτημόρια (Παράδειγμα 1.39). Φάνηκε ότι το μοντέλο δίνει απαντήσεις με μία συγκεκριμένη μορφή σε αυτή την ερώτηση χωρίς να εμπεριέχει γενικές πληροφορίες. Για την βελτίωση των επιδόσεων του μοντέλου αποφασίστηκε ακόμη κάθε ερώτηση να φορά ένα μόνο όργανο και ένα μόνο χαρακτηριστικό.

Προσπάθειες για εξαγωγή των μασκών με κριτική σκέψη του μοτέλου

Από τις απαντήσεις του μοντέλου είναι εμφανές πως γνωρίζει τα γενικά σχήματα των οργάνων, παρόλαυτα σε πολλές περιπτώσεις οι απαντήσεις αφορούν το τρισδιάστατο σχήμα και όχι αυτό που αποτυπώνεται στις δύο διαστάσεις της εικόνας. Για παράδειγμα μπορεί το duodenum να περιγραφεί ως “C-shaped“ ενώ στην εικόνα είναι ελλειψοειδές.

Για την επίλυση αυτού του προβλήματος αποφασίσαμε να παρουσιάσουμε στο μοντέλο το ερώτημα αυτό ως multiple choice μεταξύ των δισδιάστατων περιγραφών των οργάνων. Αυτή η στρατηγική αποδείχτηκε μη αποτελεσματική 1.40.

Όσον αφορά το σχήμα και την υφή ακολουθήθηκε ακόμη μία στρατηγική κατά την οποία δίνονταν στο μοντέλο τα χαρακτηριστικά ορισμένων κατηγοριών με στόχο το LVLM να αναγνωρίσει από την εικόνα ποια από αυτά ταιριάζουν στο όργανο και να το κατηγοριοποιήσει σε κάποια υφή ή σχήμα. Παρόλαυτά η στρατηγική δεν έφερε τα επιθυμητά αποτελέσματα (Παραδείγματα 1.40, 1.41).

Περιγραφές Υφής

Όσον αφορά την υφή μία εξέταση του συνόλου δεδομένων εκπαίδευσης του LVLM οδήγησε στην αναγνώριση δυο όρων και συγκεκριμένα “ομογενές”, “ετερογενές” ως γνωστούς όσον αφορά την υφή. Παρόλα αυτά το μοντέλο δυσκολευόταν να χαρακτηρίσει ένα όργανο σωστά ως ομογενές η ετερογενές και γι’ αυτό το λόγο η υφή τελικά δε χρησιμοποιήθηκε ως χαρακτηριστικό.

1.4.2 RAOS synthetic MRI τύπου "Delay"

Το GroundingDINO εκπαιδεύτηκε για 11 εποχές σε εικόνες συνθετικής μαγνητικής τομογραφίας. Οι επιδόσεις ανά όργανο διαφέρουν αρκετά καθώς κάποια όργανα είναι πιο ορατά από άλλα λόγω της φωτεινότητας. Στα πειράματα δοκιμάστηκαν prompts με διάφορα επίπεδα πληροφορίας όπως θέση, σχήμα και φωτεινότητα και όπως φαίνεται όσο πιο περι-

γραφικά είναι τα prompts τόσο πιο καλές είναι οι επιδόσεις. Επίσης δοκιμάστηκε το SAM2 σε σύγκριση με το Med-SAM2 και φάνηκε πως το Med-SAM2 παρόλο που είναι fine-tuned δεν βελτιώνει τα αποτελέσματα. Ακόμη δοκιμάστηκαν διαφορετικά είδη prompts, όπως μια κανονική πρόταση για το σχήμα με λέξεις κλειδιά για φωτεινότητα και θέση ή κανονικές προτάσεις χωρίς λέξεις κλειδιά. Αποδείχθηκε ότι η πληροφορία όταν δίνεται σε συνοπτική μορφή με λέξεις κλειδιά, περνάει πιο καλά στο μοντέλο και έχει καλύτερες επιδόσεις.

Fine-tuned Grounding DINO (Epoch 11) + SAM2	IoU	Precision
Liver	0.7491	0.7996
Spleen	0.7869	0.8399
Left Kidney	0.8344	0.8770
Right Kidney	0.8153	0.8582
Stomach	0.6430	0.7549
Gallbladder	0.4746	0.5584
Esophagus	0.5132	0.6663
Pancreas	0.3246	0.3831
Duodenum	0.3641	0.4523
Colon	0.4988	0.6183
Intestine	0.3063	0.3541
Left Adrenal	0.2499	0.3186
Right Adrenal	0.3187	0.4171
Segmentation Overall Average	0.5230	0.6010

Table 1.2. *Performance results*

Fine-tuned Grounding DINO (Epoch 11) + Med-SAM2	IoU	Precision
Liver	0.7281	0.8194
Spleen	0.7815	0.8763
Left Kidney	0.7989	0.9235
Right Kidney	0.7915	0.9036
Stomach	0.6312	0.7738
Gallbladder	0.4637	0.5387
Esophagus	0.5432	0.6648
Pancreas	0.3055	0.3916
Duodenum	0.3642	0.4635
Colon	0.4967	0.6139
Intestine	0.2965	0.378
Left Adrenal	0.233	0.2896
Right Adrenal	0.3334	0.4088
Segmentation Overall Average	0.5133	0.6157

Table 1.3. *Performance results*

Fine-tuned Grounding DINO (Epoch 11) + SAM2 (Position only)	IoU	Precision
Liver	0.7557	0.8051
Spleen	0.7677	0.8160
Left Kidney	0.8205	0.8608
Right Kidney	0.7996	0.8388
Stomach	0.5759	0.6674
Gallbladder	0.3927	0.4663
Esophagus	0.4758	0.6228
Pancreas	0.2641	0.3090
Duodenum	0.2509	0.3022
Colon	0.2906	0.3613
Intestine	0.1744	0.1980
Left Adrenal	0.1670	0.2223
Right Adrenal	0.2199	0.2991
Segmentation Overall Average	0.4232	0.4785

Table 1.4. *Performance results*

Fine-tuned Grounding DINO (Epoch 11) + SAM2 (Position and shape)	IoU	Precision
Liver	0.7509	0.8010
Spleen	0.7830	0.8353
Left Kidney	0.8295	0.8705
Right Kidney	0.8106	0.8517
Stomach	0.6419	0.7539
Gallbladder	0.439	0.5173
Esophagus	0.4916	0.6321
Pancreas	0.3175	0.3718
Duodenum	0.3254	0.3987
Colon	0.4431	0.5517
Intestine	0.2859	0.3333
Left Adrenal	0.2103	0.2794
Right Adrenal	0.2916	0.3869
Segmentation Overall Average	0.4993	0.5726

Table 1.5. *Performance results*

Fine-tuned Grounding DINO (Epoch 11) + SAM2 (Unfiltered prompts natural language shortened version with 2 sentences)	IoU	Precision
Liver	0.7461	0.7957
Spleen	0.7432	0.7956
Left Kidney	0.8304	0.8712
Right Kidney	0.8138	0.8553
Stomach	0.6350	0.7652
Gallbladder	0.4557	0.5473
Esophagus	0.4915	0.6363
Pancreas	0.3088	0.3659
Duodenum	0.2570	0.3207
Colon	0.4616	0.5756
Intestine	0.237	0.2786
Left Adrenal	0.2043	0.2657
Right Adrenal	0.2828	0.3710
Segmentation Overall Average	0.4855	0.5597

Table 1.6. *Example of input: "In the image, the Liver has a slightly oval shape. The Liver is located on the lower middle and on the right part of the image and appears to be Light Gray."*

Fine-tuned Grounding DINO (Epoch 11) + SAM2 (Unfiltered prompts natural language with 3 sentences)	IoU	Precision
Liver	0.7369	0.7863
Spleen	0.7814	0.8351
Left Kidney	0.8323	0.8739
Right Kidney	0.8146	0.8575
Stomach	0.6319	0.7619
Gallbladder	0.4647	0.5561
Esophagus	0.4981	0.6475
Pancreas	0.2968	0.3533
Duodenum	0.2962	0.3685
Colon	0.4357	0.5450
Intestine	0.2377	0.2818
Left Adrenal	0.2292	0.2974
Right Adrenal	0.3048	0.3981
Segmentation Overall Average	0.4842	0.5590

Table 1.7. *Example of input: "In the image, the Liver has a slightly oval shape. The Liver is located on the lower middle and on the right part of the image. The Liver in the image appears to be Light Gray."*

Fine-tuned Grounding DINO (Epoch 7) + SAM2	IoU	Precision
Liver	0.7548	0.8057
Spleen	0.7876	0.840
Left Kidney	0.8345	0.8775
Right Kidney	0.8196	0.8626
Stomach	0.6311	0.7402
Gallbladder	0.4402	0.5185
Esophagus	0.486	0.6211
Pancreas	0.3184	0.3753
Duodenum	0.3259	0.4006
Colon	0.5085	0.6311
Intestine	0.3052	0.3528
Left Adrenal	0.2262	0.2871
Right Adrenal	0.263	0.3391
Segmentation Overall Average	0.5188	0.5951

Table 1.8. *Performance results*

Original Grounding DINO + SAM2	IoU	Precision
Liver	0.1059	0.1078
Spleen	0.0289	0.0291
Left Kidney	0.11306	0.1171
Right Kidney	0.2639	0.2709
Stomach	0.0225	0.0250
Gallbladder	0.0085	0.0085
Esophagus	0.00161	0.0016
Pancreas	0.0177	0.0177
Duodenum	0.00262	0.0026
Colon	0.0230	0.0233
Intestine	0.02708	0.0271
Left Adrenal	0.00088	0.0009
Right Adrenal	0.0018	0.0018
Segmentation Overall Average	0.0487	0.0498

Table 1.9. *Performance results*

Original Grounding DINO + Med-SAM2	IoU	Precision
Liver	0.0612	0.0683
Spleen	0.021	0.0223
Left Kidney	0.1054	0.1207
Right Kidney	0.2504	0.2724
Stomach	0.0204	0.02414
Gallbladder	0.0042	0.0042
Esophagus	0.00012	0.00012
Pancreas	0.0085	0.0087
Duodenum	0.00055	0.00056
Colon	0.028	0.0298
Intestine	0.01681	0.01773
Left Adrenal	0.000409	0.000423
Right Adrenal	0.00123	0.00124
Segmentation Overall Average	0.0407	0.0445

Table 1.10. *Performance results*

Pipeline	IoU	Precision
Fine-tuned Grounding DINO (Epoch 11) + SAM2	0.5230	0.6010
Fine-tuned Grounding DINO (Epoch 11) + Med-SAM2	0.5133	0.6157
Original Grounding DINO + SAM2	0.0487	0.0498
Original Grounding DINO + Med-SAM2	0.0407	0.0445

Table 1.11. *Results based on model combination*

Prompt information	IoU	Precision
Position	0.4232	0.4785
Position + shape	0.4993	0.5726
Position + shape + brightness	0.5230	0.6010

Table 1.12. *Results based on prompt information for the combination of the Fine-tuned Grounding DINO (Epoch 11) with SAM2*

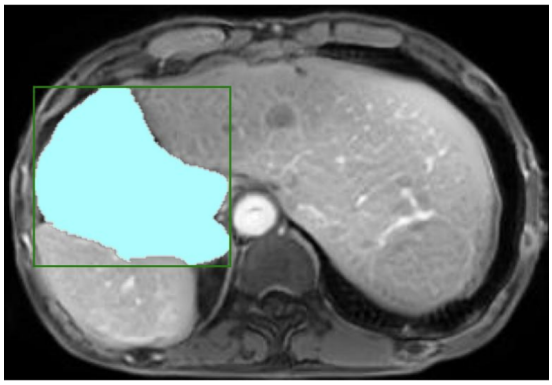


Figure 1.42. *Ground truth bounding box and mask [5]*

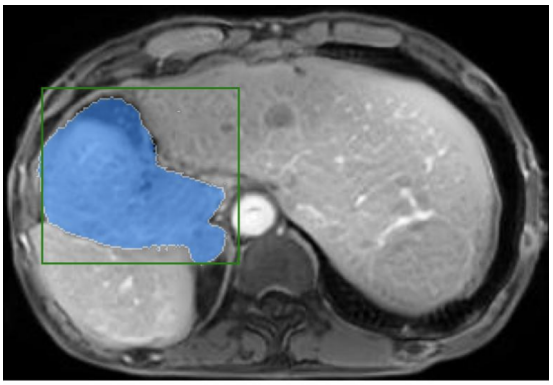


Figure 1.43. *Predicted bounding box and mask. IoU = 0.85 Precision = 0.997 [5]*

Figure 1.44. *Example of stomach grounded segmentation with input: "In the image, the Stomach has a slightly oval shape. Stomach on upper middle and on left part . Stomach Light Gray ."*

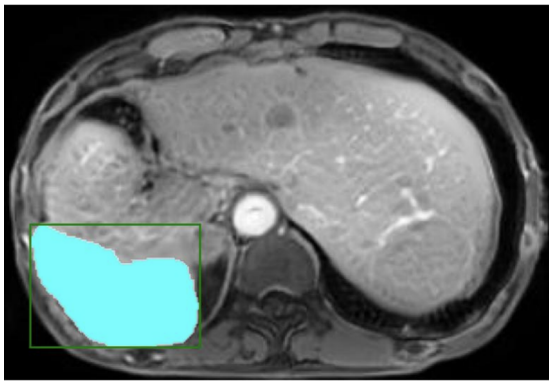


Figure 1.45. *Ground truth bounding box and mask [5]*

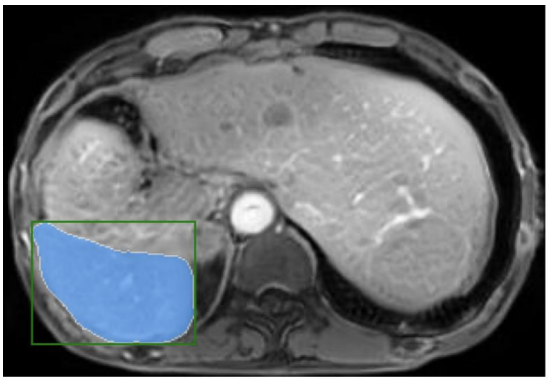


Figure 1.46. *Predicted bounding box and mask. IoU = 0.927 Precision = 0.969 [5]*

Figure 1.47. *Example of spleen grounded segmentation with input: "In the image, the Spleen has a shape that resembles a crescent. Spleen on lower part and on left part . Spleen Very Light Gray ."*

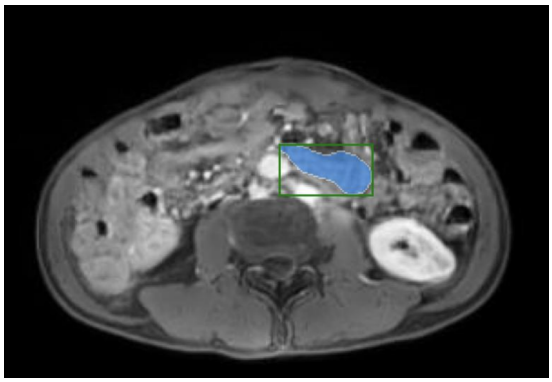


Figure 1.48. *Ground truth bounding box and mask [5]*

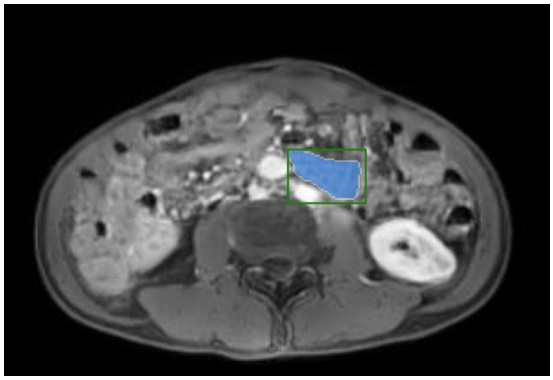


Figure 1.49. *Predicted bounding box and mask. IoU = 0.595 Precision = 0.784 [5]*

Figure 1.50. *Example of duodenum grounded segmentation with input: "In the image, the Duodenum has a C-shaped appearance. Duodenum on upper middle and on right middle . Duodenum Light Gray ."*

1.4.3 RAOS CT Scans

Το σύστημα που αναπτύχθηκε δοκιμάστηκε και σε εικόνες υπολογιστικής τομογραφίας. Έγινε αξιολόγηση στα ίδια πειράματα με εικόνες της κοιλιακής χώρας από τους ίδιους 67 ασθενείς.

Fine-tuned Grounding DINO (Epoch 11) + SAM2	IoU	Precision
Liver	0.7844	0.8200
Spleen	0.8311	0.8829
Left Kidney	0.8777	0.9288
Right Kidney	0.8689	0.9250
Stomach	0.7219	0.8462
Gallbladder	0.6155	0.7125
Esophagus	0.6553	0.8081
Pancreas	0.5152	0.5899
Duodenum	0.4490	0.5535
Colon	0.5822	0.7347
Intestine	0.3308	0.3770
Left Adrenal	0.5283	0.6468
Right Adrenal	0.4800	0.5863
Rectum	0.7242	0.8564
Bladder	0.8437	0.9023
Left Head of the Femur	0.7773	0.9074
Right Head of the Femur	0.7586	0.9000
Prostate	0.6632	0.7589
Seminal Vescicle	0.6205	0.7329
Segmentation Overall Average	0.5779	0.6702

Table 1.13. *Performance Results*

Fine-tuned Grounding DINO (Epoch 11) + Med-SAM2	IoU	Precision
Liver	0.7879	0.8395
Spleen	0.8435	0.9160
Left Kidney	0.8773	0.9448
Right Kidney	0.8563	0.9257
Stomach	0.7288	0.8368
Gallbladder	0.6108	0.6874
Esophagus	0.6800	0.7888
Pancreas	0.5056	0.5790
Duodenum	0.4601	0.5525
Colon	0.6030	0.6937
Intestine	0.3338	0.3717
Left Adrenal	0.4803	0.5725
Right Adrenal	0.4929	0.5709
Rectum	0.7418	0.8546
Bladder	0.8113	0.8952
Left Head of the Femur	0.6695	0.9011
Right Head of the Femur	0.6967	0.8859
Prostate	0.6445	0.7631
Seminal Vescicle	0.6246	0.7240
Segmentation Overall Average	0.5783	0.6588

Table 1.14. *Performance Results*

Fine-tuned Grounding DINO (Epoch 11) + SAM2 (Position only)	IoU	Precision
Liver	0.7358	0.7682
Spleen	0.8230	0.8729
Left Kidney	0.8698	0.9223
Right Kidney	0.8629	0.9194
Stomach	0.6222	0.7217
Gallbladder	0.5396	0.6216
Esophagus	0.6500	0.7945
Pancreas	0.4272	0.4870
Duodenum	0.2862	0.3516
Colon	0.3965	0.4958
Intestine	0.1449	0.1591
Left Adrenal	0.5025	0.6145
Right Adrenal	0.4445	0.5456
Rectum	0.7143	0.8545
Bladder	0.8160	0.8684
Left Head of the Femur	0.7210	0.8487
Right Head of the Femur	0.7292	0.8670
Prostate	0.5699	0.6492
Seminal Vescicle	0.5197	0.5937
Segmentation Overall Average	0.4592	0.5262

Table 1.15. *Performance Results*

Fine-tuned Grounding DINO (Epoch 11) + SAM2 (Position and shape)	IoU	Precision
Liver	0.7661	0.7994
Spleen	0.8295	0.8814
Left Kidney	0.8775	0.9286
Right Kidney	0.8687	0.9248
Stomach	0.7137	0.8386
Gallbladder	0.6166	0.7148
Esophagus	0.6566	0.8055
Pancreas	0.4937	0.5646
Duodenum	0.4272	0.5254
Colon	0.5584	0.7064
Intestine	0.2892	0.3274
Left Adrenal	0.5267	0.6476
Right Adrenal	0.4739	0.5780
Rectum	0.7237	0.8530
Bladder	0.8417	0.9003
Left Head of the Femur	0.7164	0.8457
Right Head of the Femur	0.7176	0.8570
Prostate	0.6508	0.7441
Seminal Vescicle	0.5970	0.7023
Segmentation Overall Average	0.5540	0.6422

Table 1.16. *Performance Results*

Fine-tuned Grounding DINO (Epoch 11) + SAM2 (Unfiltered prompts natural language shortened version with 2 sentences)	IoU	Precision
Liver	0.7272	0.7585
Spleen	0.8280	0.8787
Left Kidney	0.8785	0.9300
Right Kidney	0.8687	0.9246
Stomach	0.6909	0.8130
Gallbladder	0.6049	0.7077
Esophagus	0.6565	0.8047
Pancreas	0.4506	0.5212
Duodenum	0.3806	0.4719
Colon	0.4993	0.6360
Intestine	0.2564	0.2956
Left Adrenal	0.5239	0.6512
Right Adrenal	0.4765	0.5859
Rectum	0.7250	0.8639
Bladder	0.8159	0.8725
Left Head of the Femur	0.7102	0.8360
Right Head of the Femur	0.7117	0.8514
Prostate	0.6481	0.7375
Seminal Vescicle	0.5236	0.6220
Segmentation Overall Average	0.5231	0.6088

Table 1.17. *Example of input: "In the image, the Liver has a slightly oval shape. The Liver is located on the lower middle and on the right part of the image and appears to be Light Gray."*

Fine-tuned Grounding DINO (Epoch 11) + SAM2 (Unfiltered prompts natural language with 3 sentences)	IoU	Precision
Liver	0.7754	0.8111
Spleen	0.8279	0.8789
Left Kidney	0.8781	0.9297
Right Kidney	0.8683	0.9242
Stomach	0.7117	0.8364
Gallbladder	0.6090	0.7156
Esophagus	0.6504	0.8054
Pancreas	0.4918	0.5695
Duodenum	0.3982	0.4941
Colon	0.5427	0.6858
Intestine	0.2702	0.3092
Left Adrenal	0.5287	0.6541
Right Adrenal	0.4851	0.5996
Rectum	0.7223	0.8650
Bladder	0.8304	0.8876
Left Head of the Femur	0.7159	0.8393
Right Head of the Femur	0.7368	0.8768
Prostate	0.6615	0.7587
Seminal Vescicle	0.5489	0.6511
Segmentation Overall Average	0.5443	0.6326

Table 1.18. *Example of input: "In the image, the Liver has a slightly oval shape. The Liver is located on the lower middle and on the right part of the image. The Liver in the image appears to be Light Gray."*

Fine-tuned Grounding DINO (Epoch 7) + SAM2	IoU	Precision
Liver	0.7859	0.8223
Spleen	0.8323	0.8849
Left Kidney	0.8764	0.9285
Right Kidney	0.8675	0.9251
Stomach	0.7187	0.8392
Gallbladder	0.6019	0.7010
Esophagus	0.6523	0.8040
Pancreas	0.5080	0.5837
Duodenum	0.4260	0.5174
Colon	0.5810	0.7358
Intestine	0.3264	0.3731
Left Adrenal	0.5305	0.6542
Right Adrenal	0.4763	0.5861
Rectum	0.7111	0.8459
Bladder	0.8463	0.9062
Left Head of the Femur	0.7717	0.9033
Right Head of the Femur	0.7581	0.8995
Prostate	0.6520	0.7521
Seminal Vescicle	0.6196	0.7277
Segmentation Overall Average	0.5741	0.6668

Table 1.19. *Performance Results*

Original Grounding DINO + SAM2	IoU	Precision
Liver	0.0906	0.0910
Spleen	0.0385	0.0385
Left Kidney	0.0433	0.0441
Right Kidney	0.0425	0.0430
Stomach	0.0321	0.0431
Gallbladder	0.0066	0.0066
Esophagus	0.0018	0.0018
Pancreas	0.0184	0.0184
Duodenum	0.0038	0.0044
Colon	0.0113	0.0131
Intestine	0.0259	0.0260
Left Adrenal	0.0014	0.0015
Right Adrenal	0.0015	0.0015
Rectum	0.0130	0.0130
Bladder	0.0717	0.0718
Left Head of the Femur	0.0299	0.0318
Right Head of the Femur	0.1072	0.1155
Prostate	0.0201	0.0201
Seminal Vescicle	0.0051	0.0051
Segmentation Overall Average	0.0293	0.0306

Table 1.20. *Performance Results*

Original Grounding DINO + Med-SAM2	IoU	Precision
Liver	0.1060	0.1098
Spleen	0.0422	0.0422
Left Kidney	0.0427	0.0442
Right Kidney	0.0326	0.0333
Stomach	0.0346	0.0433
Gallbladder	0.0077	0.0077
Esophagus	0.0011	0.0011
Pancreas	0.0205	0.0206
Duodenum	0.0038	0.0043
Colon	0.0102	0.0117
Intestine	0.0266	0.0268
Left Adrenal	0.0012	0.0012
Right Adrenal	0.0015	0.0015
Rectum	0.0123	0.0124
Bladder	0.0776	0.0777
Left Head of the Femur	0.0253	0.0295
Right Head of the Femur	0.0883	0.1053
Prostate	0.0215	0.0215
Seminal Vescicle	0.0062	0.0062
Segmentation Overall Average	0.02949	0.03134

Table 1.21. *Performance Results*

Pipeline	IoU	Precision
Fine-tuned Grounding DINO (Epoch 11) + SAM2	0.5779	0.6702
Fine-tuned Grounding DINO (Epoch 11) + Med-SAM2	0.5783	0.6588
Original Grounding DINO + SAM2	0.0293	0.0306
Original Grounding DINO + Med-SAM2	0.02949	0.03134

Table 1.22. *Results based on model combination*

Prompt information	IoU	Precision
Position	0.4592	0.5262
Position + shape	0.5540	0.6422
Position + shape + brightness	0.5779	0.6702

Table 1.23. *Results based on prompt information for the combination of the Fine-tuned Grounding DINO (Epoch 11) with SAM2*

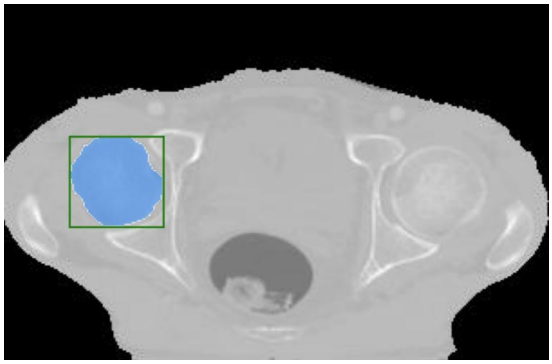


Figure 1.51. *Ground truth bounding box and mask [5]*

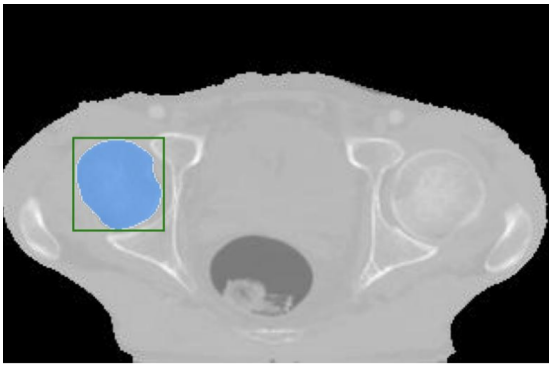


Figure 1.52. *Predicted bounding box and mask. IoU = 0.91 Precision = 0.98 [5]*

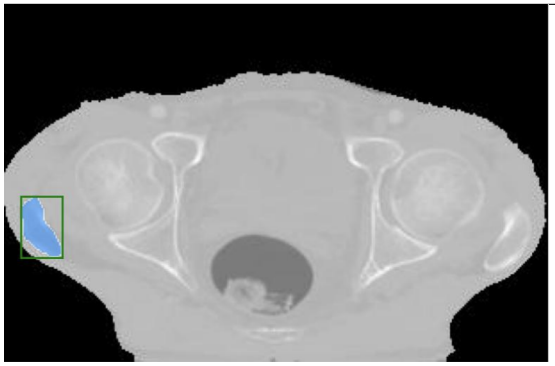


Figure 1.53. *Ground truth bounding box and mask [5]*

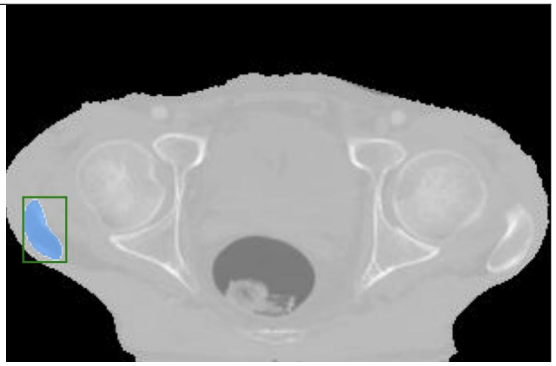


Figure 1.54. *Predicted bounding box and mask. IoU = 0.89 Precision = 0.95 [5]*

Figure 1.55. *Example of left head of the femur grounded segmentation with input: "In the image, the Left Head of the Femur has a rounded shape. Part Left Head Femur on lower middle and on **left part** Left Head Femur Very Light Gray ." for the first pair and "In the image, the Left Head of the Femur has a rounded shape. Part Left Head Femur on lower middle and on **leftmost part** Left Head Femur Very Light Gray ." for the second pair*

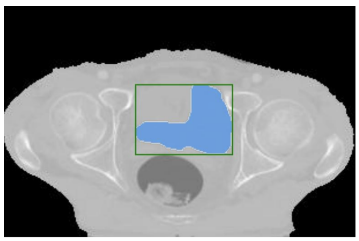


Figure 1.56. *Ground truth bounding box and mask [5]*

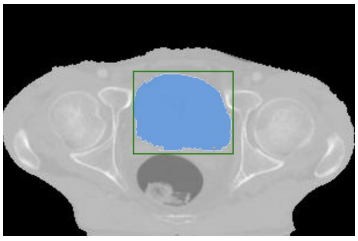


Figure 1.57. *Predicted bounding box and mask. IoU = 0.57 Precision = 0.58 [5]*

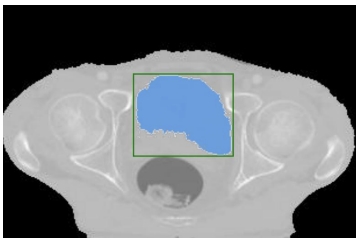


Figure 1.58. *Predicted bounding box and mask. IoU = 0.54 Precision = 0.46 [5]*

Figure 1.59. *Example of bladder grounded segmentation with SAM2 on the left and MedSAM2 on the right*

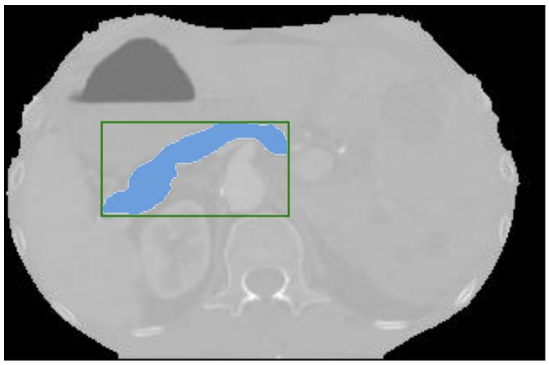


Figure 1.60. *Ground truth bounding box and mask [5]*

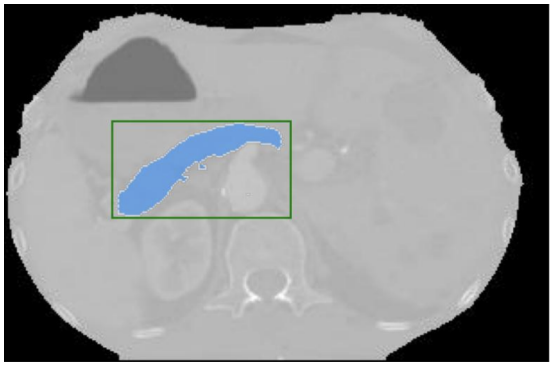


Figure 1.61. *Predicted bounding box and mask. IoU = 0.73 Precision = 0.89 [5]*

Figure 1.62. *Example of pancreas grounded segmentation with input: "In the image, the Pancreas has a slightly irregular shape. Pancreas on upper middle and on left middle . Pancreas Very Light Gray."*

1.4.4 RAOS αξιολόγηση σε άλλα Synthetic MRI

Μέχρι αυτό το σημείο, το μοντέλο έχει αξιολογηθεί σε σύνολα δεδομένων στα οποία εκπαιδεύτηκε. Σε αυτό το πείραμα, αξιολογούμε την απόδοση σε σύνολα δεδομένων εκτός κατανομής. Στις νέες συνθετικές εικόνες MRI, απεικονίζονται οι ίδιοι ασθενείς με σημαντικές διαφοροποιήσεις στη φωτεινότητα των οργάνων σε σύγκριση με τον τύπο "Delayed". Αυτό σημαίνει ότι οι απεικονίσεις των οργάνων διαφέρουν, γεγονός που αντικατοπτρίζεται και στις κειμενικές περιγραφές σχετικά με τη φωτεινότητα. Υπό αυτές τις νέες συνθήκες, δοκιμάζουμε εάν το μοντέλο κατανοεί την αλλαγή και προσαρμόζεται, διατηρώντας παρόμοια απόδοση.

Pre Artery type of Synthetic MRI	IoU	Precision
Liver	0.7166	0.7903
Spleen	0.6833	0.7694
Left Kidney	0.6740	0.7274
Right Kidney	0.5578	0.6058
Stomach	0.5818	0.6977
Gallbladder	0.3575	0.4267
Esophagus	0.4236	0.5426
Pancreas	0.1910	0.2356
Duodenum	0.2007	0.2658
Colon	0.4410	0.5885
Intestine	0.2908	0.3617
Left Adrenal	0.0197	0.0284
Right Adrenal	0.0996	0.1343
Segmentation Overall Average	0.4352	0.5235

Table 1.24. *Fine-tuned Grounding DINO (Epoch 11) + SAM2*

PV type of Synthetic MRI	IoU	Precision
Liver	0.7402	0.8076
Spleen	0.7547	0.8254
Left Kidney	0.8173	0.8739
Right Kidney	0.8172	0.8729
Stomach	0.6395	0.7526
Gallbladder	0.4103	0.4979
Esophagus	0.4859	0.6572
Pancreas	0.3341	0.4071
Duodenum	0.3439	0.4404
Colon	0.4703	0.6039
Intestine	0.2985	0.3569
Left Adrenal	0.2556	0.3367
Right Adrenal	0.2742	0.3706
Segmentation Overall Average	0.5071	0.5972

Table 1.25. *Fine-tuned Grounding DINO (Epoch 11) + SAM2*

Type of Synthetic MRI	IoU	Precision
Delay	0.5230	0.6010
Pre Artery	0.4352	0.5235
PV	0.5071	0.5972

Table 1.26. *Results across different types of Synthetic MRI*

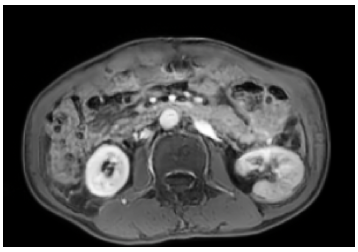


Figure 1.63. *Delay [5]*

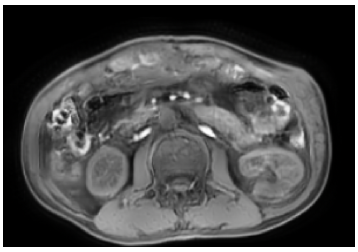


Figure 1.64. *Pre Artery [5]*

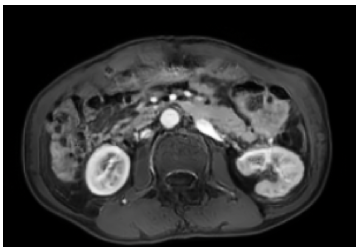


Figure 1.65. *PV [5]*

Figure 1.66. *Examples of different synthetic MRI types*

1.5 Συμπεράσματα

1.5.1 Σύγκριση με άλλα μοντέλα

Σε αυτή την ενότητα οι επιδόσεις του προτεινόμενου συστήματος θα συγκριθούν με άλλα μοντέλα. Η σύγκριση θα γίνει με το MedLAM [31], ένα foundation μοντέλο για εντοπισμό οργάνων που σε συνδυασμό με το SAM [7] ή το MedSAM [24] πραγματοποιεί κατάτμηση. Επίσης παρουσιάζονται οι επιδόσεις του SAM και του MedSAM όταν δοθούν τα πραγματικά bounding boxes [31] (manual prompting) και τέλος παρουσιάζεται η επίδοση του nnU-Net [21], ένα πλήρως επιβλεπόμενο μοντέλο βαθιάς μάθησης.

Είναι σημαντικό να σημειωθεί ότι τα αποτελέσματα που αφορούν άλλα μοντέλα προέρχονται από το σύνολο δεδομένων WORD [9] το οποίο περιέχει εικόνες υπολογιστικής τομογραφίας από την κοιλιακή χώρα, ενώ η δουλειά που παρουσιάστηκε με αυτή την εργασία έχει αξιολογηθεί στο σύνολο δεδομένων RAOS. Τα δύο αυτά σύνολα δεδομένων μοιράζονται τις εικόνες υπολογιστικής τομογραφίας κάποιων ασθενών. Συγκεκριμένα, οι επισημειώσεις του WORD που αφορούν τους κοινούς ασθενείς έπρεπε να επεκταθούν έτσι ώστε να δημιουργηθούν 19 κλάσεις όπως στο RAOS αντί για τις 16 που υπήρχαν.

Table 1.27. *Dice scores (%) for region segmentation. Prior work is evaluated on the WORD dataset. Our framework is evaluated on RAOS, a continuation of WORD, and thus, "-" refers to values missing from prior work.*

Regions	MedLAM		Man. Prompt		Fully Supervised nnU-Net		Our pipeline	
	SAM	MedSAM	SAM	MedSAM	5-shot	full	SAM2	MedSAM2
Liver	66.0	23.8	84.2	46.6	94.3	96.3	85.2	85.7
Spleen	61.7	36.3	85.3	65.0	90.9	95.7	89.4	90.4
Kidney L	82.1	70.7	92.1	84.1	83.4	94.7	93.0	93.0
Kidney R	88.3	77.3	92.9	86.4	86.0	95.2	92.4	91.7
Stomach	44.6	37.2	77.1	80.3	83.2	93.1	80.8	81.3
Gallbladder	13.1	10.2	72.7	68.8	54.7	77.0	70.7	70.5
Esophagus	36.6	27.8	67.0	63.1	72.0	81.5	77.2	79.0
Pancreas	29.7	21.4	64.4	46.9	72.0	84.7	61.6	60.9
Duodenum	26.0	21.1	54.1	51.0	53.7	77.3	53.9	54.9
Colon	25.6	26.6	41.8	44.1	74.7	86.1	67.4	68.7
Intestine	37.5	34.1	61.4	52.5	77.6	88.0	41.2	41.7
Adrenal	3.3	10.0	17.4	26.5	58.3	72.8	63.5	62.0
Left Adrenal	–	–	–	–	–	–	65.6	61.7
Right Adrenal	–	–	–	–	–	–	60.9	62.4
Prostate	–	–	–	–	–	–	75.3	74.3
Seminal Vesicle	–	–	–	–	–	–	72.9	73.2
Rectum	50.1	46.0	75.5	80.0	70.2	80.3	82.2	83.6
Bladder	65.3	59.1	83.0	82.9	84.0	91.8	89.9	87.9
Head of Femur L	81.7	71.5	90.5	80.3	43.1	31.2	84.9	77.3
Head of Femur R	80.1	74.3	89.1	83.2	47.8	40.98	83.3	79.1
Average	49.5	40.5	71.8	65.1	71.6	80.4	66.1	66.3

Από το πινακάκι 1.28 παρατηρούμε ότι η προτεινόμενη μέθοδος παράγει τα καλύτερα

bounding boxes. Η σύγκριση έχει γίνει με το **MedLAM**, το **DetCo** [33], μία μη επιβλεπόμενη μέθοδο contrastive learning για object detection, το **Mask R-CNN** [13] και το **MIU-VL** [30], ένα μοντέλο όρασης–γλώσσας για κατανόηση ιατρικής εικόνας. Το τελευταίο μοντέλο δεν έχει εκπαιδευτεί στο WORD το οποίο καταδεικνύει ότι περαιτέρω εκπαίδευση θα μπορούσε να βελτιώσει τις επιδόσεις.

Table 1.28. *Wall Distance (WD) results for each region and method on the RAOS dataset for our work and WORD for prior work. The best performance for each region is noted in bold*

Regions	**MedLAM**	**DetCo**	**Mask R-CNN**	**Our pipeline**	**MIU-VL**
Liver	10.5	73.8	7.1	**4.5**	80.4
Spleen	5.7	37.9	6.0	**1.7**	142.5
Kidney L	5.0	39.4	5.7	**0.9**	164.3
Kidney R	3.6	61.4	4.1	**1.0**	166.1
Stomach	17.7	49.3	8.8	**4.6**	102.8
Gallbladder	16.9	71.4	5.1	**2.5**	153.2
Esophagus	6.8	47.3	4.1	**1.5**	153.8
Pancreas	12.2	42.0	7.0	**5.1**	135.2
Duodenum	12.7	56.6	8.1	**5.9**	139.8
Colon	15.6	49.0	12.3	**9.8**	83.3
Intestine	15.4	50.7	**11.3**	25.3	106.0
Adrenal	7.6	52.3	6.2	**2.1**	160.2
Rectum	8.6	54.7	5.8	**2.2**	203.1
Bladder	8.1	65.1	3.7	**1.9**	167.7
Head of Femur L	5.3	55.8	3.3	**4.2**	175.1
Head of Femur R	5.9	52.1	4.1	**4.0**	163.4
Average	9.9	53.9	6.5	4.7	143.6

1.5.2 Επίλογος

Τα αποτελέσματα των πειραμάτων του LLaVA-Med αποδεικνύουν ότι έχει περιορισμένη κριτική ικανότητα. Παρόλο που χρησιμοποιήθηκαν διάφορες τεχνικές prompting όπως ερωτήσεις πολλαπλής επιλογής και rule based inference το μοντέλο φαίνεται να μην έχει κατανοήσει τις σχέσεις μεταξύ των κειμενικών περιγραφών και των χαρακτηριστικών της εικόνας. Γι' αυτό το λόγο δεν θεωρείται έμπιστο να παράξει το σύνολο των κειμενικών περιγραφών όπως αρχικά θεωρήθηκε.

Σε αυτό το σημείο πρέπει επίσης να τονίσουμε ότι το LLaVA-Med όπως τα LLMs δεν είναι ντετερμινιστικό και δεν παράγει την ίδια κειμενική έξοδο με το ίδιο prompt, το οποίο εισάγει έναν βαθμό βεβαιότητας. Στην περίπτωση των περιγραφών σχήματος αυτός ο μη ντετερμινισμός δεν αποτελεί σοβαρό πρόβλημα καθώς τα παραγόμενα σχήματα είναι αποδεκτά. Παρόλαυτά στην περίπτωση της υφής που δοκιμάστηκε και ζητήθηκε από το το μοντέλο να κατηγοριοποιήσει τα όργανα σε ομογενή και ετερογενή οι απαντήσεις δεν ήταν πάντα ίδιες με ίδια είσοδο.

Όπως φαίνεται και από τον πίνακα ο εντοπισμός που προσφέρει το προτεινόμενο σύστημα είναι αποδοτικός και είναι επίσης το μόνο πρόβλημα για το οποίο εκπαιδεύτηκε το GroundingDINO. το επόμενο μέρος του συστήματος, το οποίο παράγει την μάσκα, επιτυγχάνει ανταγωνιστικά αποτελέσματα. Παρόλα αυτά, υπάρχουν περιπτώσεις στις οποίες περαιτέρω εκπαίδευση του segmentation μοντέλο είναι απαραίτητη, καθώς τον bounding box είναι το ιδανικό.

Συγκρίνοντας το αρχικό GroundingDINO με την fine-tuned εκδοχή του παρατηρούμε ότι

το αρχικό δεν κατανοεί τα ονόματα των οργάνων που του παρουσιάζονται και απαιτεί παραπάνω εκπαίδευση στον τομέα της βιοϊατρικής καθώς επιστρέφει bounding boxes γύρω από το σύνολο του σώματος. Το MedSAM2 από τη μεριά του δεν αποτελεί σημαντική βελτίωση στο foundation μοντέλο SAM2, καθώς τα αποτελέσματα τους είναι συγκρίσιμα.

Όσον αφορά τις κειμενικές περιγραφές που δίνονται στο σύστημα αυτές θα πρέπει να είναι όσο πιο περιγραφικές γίνεται εμπεριέχοντας πληροφορία για την σχετική θέση του οργάνου στην εικόνα, την φωτεινότητα του και το σχήμα του έτσι ώστε να επιτυγχάνεται η μέγιστη απόδοση. Το μοντέλο είναι επίσης ικανό να εντοπίζει όργανα και να τα κατατμήσει ακόμα κι αν τα prompts τους διαφέρουν από αυτά του training set.

Επίσης το μοντέλο δοκιμάστηκε και σε καινούργια σύνολα δεδομένων στα οποία μόνο η φωτεινότητα των οργάνων άλλαζε, εισάγοντας νέα οπτικά χαρακτηριστικά. Το μοντέλο παρουσίασε καλές αποδόσεις για όργανα τα οποία παραμένουν ορατά. Οι αλλαγές στην φωτεινότητα των οργάνων ενσωματώθηκαν και στις κειμενικές περιγραφές και το σύστημα κατάλαβε αυτή την αλλαγή σε πολλές περιπτώσεις. Για εικόνες στις οποίες τα όργανα ήταν θολά η παρουσίαζαν μεγάλες διαφοροποιήσεις στη φωτεινότητα, η απόδοση μειώθηκε. Συνολικά εικόνες με καθαρή απεικόνιση των οργάνων είχαν τα καλύτερα σκορ.

Το προτεινόμενο σύστημα επέδειξε μία προσαρμοστικότητα στο κλειστό περιβάλλον στο οποίο εκπαιδεύτηκε. Με την έννοια κλειστό περιβάλλον εννοούνται εικόνες οι οποίες προέρχονται από την ίδια κατανομή για παράδειγμα έχουν παραχθεί από το ίδιο μηχάνημα. Αυτό το σύστημα θα μπορούσε να χρησιμοποιηθεί ως ένα βοηθητικό εργαλείο για επαγγελματίες στον τομέα σε ένα τέτοιο κλειστό περιβάλλον.

1.5.3 Περιορισμοί

Ένας βασικός παράγοντας που περιορίζει τις βαθμολογίες σε όλες τις μετρικές είναι η αδυναμία prompting γειτονικών μη συνεχών περιοχών του ίδιου οργάνου. Για παράδειγμα, όπως φαίνεται στο Σχήμα 1.67, διαφορετικά τμήματα του παχέος εντέρου μπορεί να μην είναι διακριτά βάσει της θέσης τους, καθώς καταλαμβάνουν την ίδια γενική περιοχή και θα αποκτήσουν την ίδια περιγραφή θέσης.

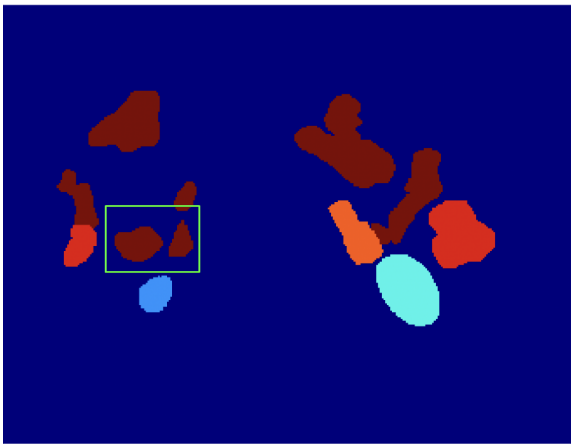


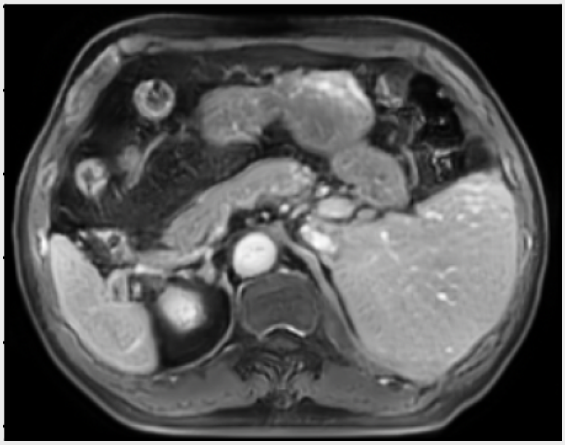
Figure 1.67. *Για τα δύο μέρη του ίδιου οργάνου που βρίσκονται μέσα στο πράσινο πλαίσιο θα πραγματοποιηθούν δύο prompts στο σύστημα. Όμως υπάρχει πιθανότητα τα δύο prompts να είναι ακριβώς τα ίδια λόγω της εγγύτητάς τους.*

1.5.4 Μελλοντικές Επεκτάσεις

Σε αυτήν την εργασία, το μοντέλο εκπαιδεύτηκε ξεχωριστά σε δύο διαφορετικά περιβάλλοντα (υπολογιστικές τομογραφίες και συνθετικές μαγνητικές τομογραφίες) και αξιολογήθηκε σε αυτά τα συγκεκριμένα κλειστά περιβάλλοντα και σε κάποια ακόμα καινούρια δεδομένα. Μία πιθανή μελλοντική βελτίωση θα ήταν η περαιτέρω προσαρμογή του μοντέλου σε ένα ευρύ σύνολο δεδομένων που καλύπτει ολόκληρο τον τομέα της ιατρικής απεικόνισης, συμπεριλαμβάνοντας πολλαπλές περιοχές του ανθρώπινου σώματος και εικόνες από διάφορες μεθόδους απεικόνισης (ακτινογραφίες, πολλαπλοί τύποι μαγνητικής τομογραφίας, υπολογιστικές τομογραφίες, ιστολογία, μακροσκοπική παθολογία). Αυτό θα οδηγούσε στη δημιουργία ενός μοντέλου με εκτεταμένη γνώση στον ιατρικό τομέα, το οποίο θα μπορούσε να χρησιμοποιηθεί σε οποιοδήποτε περιβάλλον, καθώς θα βασίζεται σε ένα μεγάλο σύνολο δεδομένων και θα είναι γενικεύσιμο.

Μία άλλη κατεύθυνση για μελλοντική έρευνα θα μπορούσε να είναι η συνεχής ενημέρωση των βαρών του LVLM μοντέλου με τις κειμενικές περιγραφές που παράγονται από τους ντετερμινιστικούς αλγορίθμους. Αυτό θα μπορούσε να έχει ως αποτέλεσμα την βελτίωση της κριτικής ικανότητας του LVLM ώστε να είναι πιο αξιόπιστο και να δίνει ακριβείς απαντήσεις.

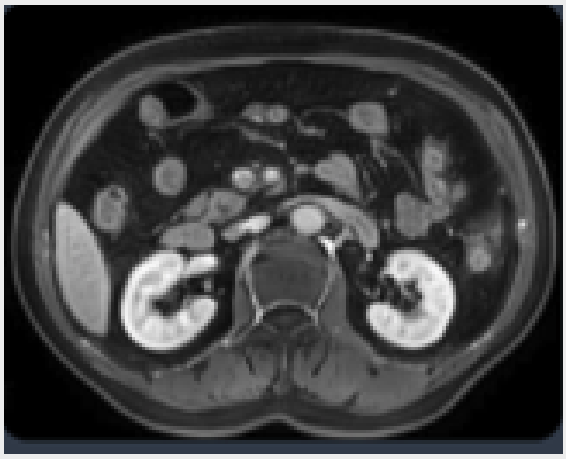
Τέλος, θα μπορούσε να ερευνηθεί η εφαρμογή του προτεινόμενου συστήματος σε 3D ιατρικές εικόνες παράγοντας τρισδιάστατα bounding boxes is και χρησιμοποιώντας το MedSAM2, ώστε να επεξεργαστεί και να κατατμήσει όγκους. Αυτό θα μπορούσε να οδηγήσει σε βελτιωμένα αποτελέσματα τα οποία θα επωφελούνται από την φύση των δεδομένων στις τρεις διαστάσεις, όπου όλα τα slices είναι διαθέσιμα σε ένα prompt το οποίο θα έχει ως αποτέλεσμα μία πιο αποδοτική χρήση του μηχανισμού μνήμης του segmentation μοντέλου.



User: Describe the organs in this image in detail in regards to their position and shape.
LLaVA-Med: In this image, the organs are positioned in a way that is typical for a female. The liver is located on the right side, and the spleen is on the left side. The liver appears to be enlarged, with a shape that resembles a "bunch of grapes." The spleen is also enlarged, with a shape that resembles a “sausage."

User: The visible organs in this image are the liver, spleen, duodenum, colon, stomach, pancreas, left and right adrenal. Describe them in detail in regards to their position and shape.
LLaVA-Med: In the image, the liver is located in the right upper quadrant, and it appears to be enlarged. The spleen is situated in the left upper quadrant, and it also appears to be enlarged. The duodenum is the first part of the small intestine and is located in the upper right quadrant. The colon is the large intestine and is situated in the lower right quadrant. The stomach is located in the upper left quadrant, and the pancreas is in the upper left quadrant, just below the stomach. The left and right adrenal glands are small, triangular-shaped glands located on top of each kidney.

Figure 1.36. *Το LLaVA-Med αγνοεί κάποια όργανα. Γι’ αυτό το λόγο αποφασίστηκε σε κάθε prompt να αναφέρονται τα ονόματα όλων των οργάνων που μας ενδιαφέρουν.*



User: In this image the organs are the two kidneys, the liver, the colon, the intestine, the pancreas, and the duodenum. Describe them as they appear in the image.
LLaVA-Med: In the image, the two kidneys are visible as well as the liver, colon, intestine, pancreas, and the duodenum. The kidneys are shown in their normal positions, with the liver located in the upper right part of the abdomen, the colon and intestine in the lower right part, the pancreas in the upper left part, and the duodenum in the upper middle part of the abdomen. The organs appear to be in their usual anatomical positions, and there are no visual abnormalities or pathologies mentioned in the context.

User: Describe the shape of the kidneys.
LLaVA-Med: The kidneys in the image are described as having a normal shape. They are bean-shaped organs with the right kidney being slightly higher than the left one. The shape of the kidneys is important for their proper functioning as it allows them to efficiently filter waste products and excess water from the blood.

User: Where is the liver here?
LLaVA-Med: In the image the liver is located in the upper right part of the abdomen.

User: Here the spinal cord is located in the lower part of the image in the middle. With that in mind, where is the liver?
LLaVA-Med: In the image the liver is located in the upper right part of the abdomen.

Figure 1.37. *Το μουτέλο επικεντρώνεται πιο πολύ σε πληροφορίες που έχει μάθει και δεν τις εξάγει με κριτική σκέψη από την εικόνα*

User: The visible organs in this image are the liver, spleen, duodenum, colon, stomach, pancreas, left and right adrenal. Describe them in detail in regards to their position and shape.

LLaVA-Med: In the image, the liver is located in the right upper quadrant of the abdomen. It appears to be enlarged and has a nodular contour. The spleen is situated in the left upper quadrant, and it also appears to be enlarged. The duodenum, which is the first part of the small intestine, is visible in the image. The colon is located in the lower right quadrant, and it appears to be dilated. The stomach is situated in the upper left quadrant, and it appears to be enlarged. The pancreas is visible in the image, and it appears to be enlarged as well. The left and right adrenal glands are also visible in the image, and they appear to be enlarged.

Figure 1.38. *Η εικόνα περιστράφηκε κατά 180 μοίρες μήπως αυτό βοηθήσει το μοντέλο για την παραγωγή σωστών περιγραφών.*

User: The visible organs in this image are the stomach, the liver, the spleen, the left kidney, the right kidney, the colon, the intestine, the right adrenal, the pancreas and the duodenum. Describe them in detail in regards to their position and say in which quadrant they belong or center of the image.

LLaVA-Med: In this image, the stomach is located in the left upper quadrant, the liver is in the right upper quadrant, the spleen is in the left upper quadrant, the left kidney is in the left lower quadrant, the right kidney is in the right lower quadrant, the colon is in the left lower quadrant, the intestine is in the right lower quadrant, the right adrenal is in the right lower quadrant, the pancreas is in the left lower quadrant, and the duodenum is in the right upper quadrant.

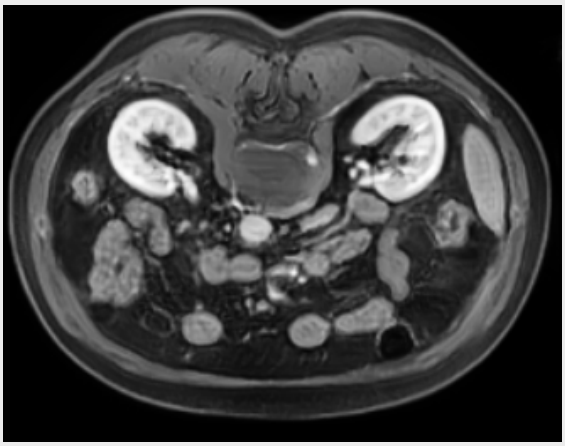
User: The visible organs in this image are the stomach, the liver, the spleen, the left kidney, the right kidney, the colon, the intestine, the right adrenal, the pancreas and the duodenum. Describe their shape in the image in detail.

LLaVA-Med: In this image, the stomach appears to be oval-shaped, the liver is oval-shaped, the spleen is oval-shaped, the left kidney is bean-shaped, the right kidney is bean-shaped, the colon is oval-shaped, the intestine is oval-shaped, the right adrenal is oval-shaped, the pancreas is oval-shaped, and the duodenum is oval-shaped.

User: The visible organs in this image are the stomach, the liver, the spleen, the left kidney, the right kidney, the colon, the intestine, the right adrenal, the pancreas and the duodenum. Tell me about the brightness of each organ if it is white, gray or dark closer to black.

LLaVA-Med: In this image, the stomach, the liver, the spleen, the left kidney, the right kidney, the colon, the intestine, the right adrenal, the pancreas, and the duodenum appear to be white. This means that they are well-visualized and easily distinguishable from the surrounding tissues.

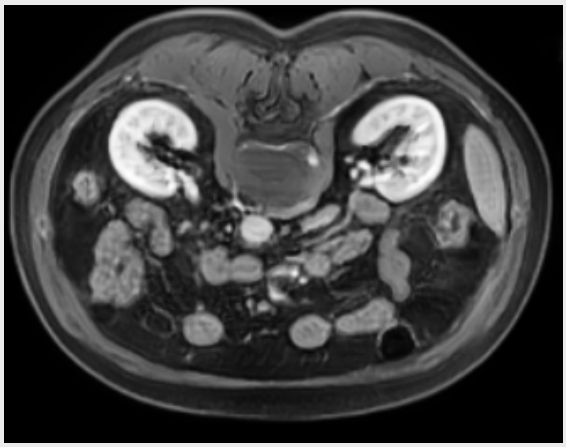
Figure 1.39. *Αποφασίστηκε κάθε ερώτηση να αφορά ένα μόνο χαρακτηριστικό*



User: The Liver is located slightly up and on the right part of the image and appears to be Light Gray. Describe its shape in detail. If the liver appears to be elongated and small that means it has the shape of a sausage but if it is large it has its normal shape. What shape does it have here?
LLaVA-Med: In this image, the liver has a large shape, which is its normal shape.
User: Can you locate the liver in this image?
LLaVA-Med: Yes, the liver is located slightly up and on the right part of the image.
User: But it has the shape of a sausage because it is small and elongated.
LLaVA-Med: No, in this image, the liver has its normal shape, which is large.
User: The liver lies on the right of the right kidney. What shape does it have?
LLaVA-Med: The liver has a large shape, which is its normal shape.
User: No, it has the shape of a sausage.
LLaVA-Med: No, in this image, the liver has its normal large shape.
User: The liver here appears to cover a small area in regards to the area of the image, so it must be small, and a small liver typically has the shape of a sausage. So what shape does it have here?
LLaVA-Med: In this image, the liver has the shape of a sausage because it is small.

Another example for the same image
User: The Liver is located slightly up and on the right part of the image and appears to be Light Gray. If the liver appears to cover a small area in regards to the area of the image it must be small and a small liver typically has the shape of a sausage. Otherwise it has an irregular shape and is wedge-shaped. So what shape does it have here?
LLaVA-Med: In this image, the liver appears to be wedge-shaped.

Figure 1.40. *Πειράματα με ερωτήσεις πολλαπλών επιλογών*



User: The Liver is located slightly up and on the right part of the image and appears to be Light Gray. If the liver appears to cover a small area in regards to the area of the image it must be small and a small liver typically has the shape of a sausage. Otherwise it has an irregular shape and is wedge-shaped. So what shape does it have here?
LLaVA-Med: In this image, the liver appears to be wedge-shaped.

Figure 1.41. *Παράδειγμα κριτικής σκέψης του μοντέλου για το σχήμα του liver*

Chapter 2

Introduction

2.1 Scope of Work

In Artificial Intelligence (AI), foundation models play a significant role and can be utilized in various domains such as natural language processing (NLP) and computer vision (CV). Due to the need for generalizability, they are expensive to create, requiring vast computational resources and rely on large-scale datasets. Any improvement regarding their architecture, size, training dataset and optimization method can significantly enhance performance, bringing new opportunities for their application. A major advantage of foundation models lies in the fact that their use can be extended to more domains than they were trained for at a relatively low cost. Notable foundation models in NLP are BERT [18] and GPT-4 [10].

In computer vision, models such as the U-Net [6], which is a task-specific model, used to rely on convolutional neural networks(CNNs). With the advent of the transformer architecture, new models emerged with broader capabilities. These include the Segment Anything Model(SAM) [7] and its successor SAM2 [2], which are capable of segmenting images based on a bounding box, point, or mask. Another example is the GroundingDINO [4], which can detect objects based on a textual description. Beyond these, a new category of models known as large vision-language models (LVLMs) integrate vision and language modalities and can be utilized in multimodal tasks. One such model is the large language and vision assistant (LLaVA) [1] that can be used in the context of visual question answering (VQA), a task where an image is provided and the model has to answer an open-ended question. All of the aforementioned foundation models have learned through their training general knowledge and can be applied with slight modifications and low cost to any specific domain.

Especially in the field of medical imaging and more specifically grounded organ segmentation, which is the application of this work, datasets are often either scarce or unannotated and the creation of ground truth data can only be the work of experts for reliability purposes. These constitute limiting factors for the development of models, which specialize in medical imaging, as the dataset is the backbone of the training process. To address this issue, in this work, we propose a method that is able to partly replace the need for expert annotations by utilizing the domain-specific knowledge of a fine-tuned LVLM, the LLaVA-Med [8] and transferring it to GroundingDINO. For the grounded segmentation

of the organs (or parts of them if they are non-contiguous), we develop a pipeline consisting of the GroundingDINO for the production of a bounding box from the (2D image, text) input and SAM2 or Med-SAM2 [3] to generate the predicted mask of the organ. As for the dataset it comprises both CT scans and Synthetic MRI images along with their respective labelmaps all of which belong to the RAOS [5, 9] dataset. The only missing part for grounded segmentation training are the organ textual descriptions which will be generated by LLaVA-Med and deterministic algorithms. From the utilized models, only GroundingDINO will be fine-tuned in the combined data of the RAOS and the produced textual descriptions, while the segmentation models will remain unaltered.

With the proposed method we proceed on fine-tuning the GroundingDINO and conduct a series of experiments to analyze the performance of the pipeline on grounded segmentation. As metrics, the intersection over union (IoU), the precision, the dice score coefficient and wall distance are used to quantify the performance. During the testing phase, different types of prompting strategies are initially assessed both for the whole pipeline and LLaVA-Med. Additionally, a comparison is presented between the fine-tuned version of the GroundingDINO and its original counterpart. Furthermore, the pipeline performance is assessed with the use of SAM2 and then the fine-tuned Med-SAM2 to understand if the fine-tuning of the second component is imperative for improved results. Lastly, the pipeline will be evaluated on unseen synthetic MRI images of unknown type for the system.

2.2 Thesis Organization

This work is organized into the following chapters:

- **Chapter 2: Introduction** - This section provides a brief overview of the work, the motives and the purposes.
- **Chapter 3: Background and Related Work** - In this chapter a small introduction to computer vision and relevant concepts are presented. Additionally, the utilized models are analyzed in more detail and related work is also mentioned.
- **Chapter 4: Methodology** - Here, all of the decisions regarding the architecture of the pipeline, the hyperparameters of the model, the dataset and a dataset evaluation are provided.
- **Chapter 5: Results** - In this section, the results are presented in both quantitative and qualitative form with the use of example images and segmentation masks.
- **Chapter 6: Discussion** - In the last chapter, we include a comparison with other models, an analysis of the findings, a discussion about the limitations of the project and a suggestion about future work.

Chapter 3

Background and Related Work

3.1 Overview of Computer Vision

Computer vision (CV) is a subfield of Artificial Intelligence (AI) that focuses on enabling machines to analyze and interpret visual information in a meaningful way. In the past computer vision relied on handcrafted features and rule-based approaches for tasks such as edge detection and feature matching. However, the introduction of large-scale data-driven models, availability of information and improvement of computational methods and hardware have all resulted in more robust and generalizable models. The field has progressed rapidly with the advent of deep learning, particularly with convolutional neural networks and the more recent transformer-based architectures.

Recent developments have enabled the use of large vision language models (LVLMs) that bridge the gap between visual and textual modalities performing applications such as image captioning, visual question answering and grounded segmentation. Other tasks that fall under CV are image classification, object detection and image segmentation.

In the image classification task, there is a list of classes and the objective is to match each image to one class. To this end, images usually depict a single object to reduce the chance of containing objects from two different classes.

Object detection is another task in which an image and a class name are given as prompts and the output should be the bounding box surrounding the requested object in the image. In some cases, there might exist multiple instances of the object in a single image and multiple bounding boxes are required.

Grounded image segmentation is a task in which the object is detected with pixel precision and a mask is formed. The input for this task should be the image along with a textual description of the requested object.

Visual question answering (VQA) is a task in which an image along with a text-based question regarding its content is provided as input. The model in turn is expected to extract image features, understand the given text and produce an accurate response to the question.

Mask

A mask of an object in computer vision is the set of pixels that are used to delineate it, segmenting it from the background. The shape of a mask is arbitrary, depending on

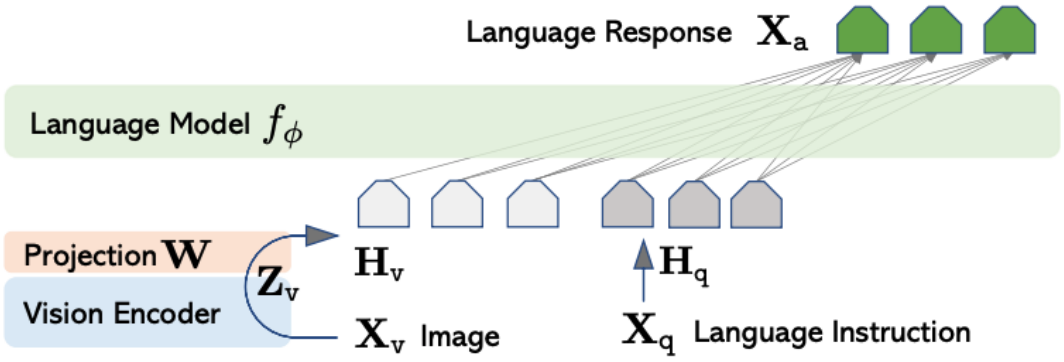


Figure 3.1. *Architecture of LLaVA [1]*

factors such as the structure of the object, occlusion and the camera's capture angle.

Bounding box

A bounding box around the object of interest is a rectangular region that perfectly fits the object. In this study, the edges of the bounding box are parallel to the horizontal and vertical axes of the images. Notably, a bounding box does not include any row or column of pixels unless at least one of them is part of the object's mask.

3.2 Relevant Models

3.2.1 LLaVA

The Large Language and Vision Assistant (LLaVA) is a large-scale multimodal model [1] that accepts image text pairs as input and generates a textual response. The tasks it can handle vary as each task can be described in the textual input and the answer is based on the input image.

The model was trained using an instruction-following dataset, consisting of tasks and their corresponding responses. The dataset was curated using text-only GPT-4, which was prompted with the caption for each image along with a set of questions to answer. The questions covered a wide variety of tasks and difficulties ranging from simple conversation and detailed description to complex reasoning. This process resulted in an elaborate set of questions and answers with diversity. As a result, the GPT-4 worked as a teacher model that distilled its knowledge to the LLaVA model.

The architecture of LLaVA (Figure 3.1) consists of both text and vision encoders. The text encoder is a pre-trained large language model, the Vicuna [11], while the vision encoder is the ViT-L/14 [12], the same model used in CLIP [28]. A linear layer is employed to transform the embeddings of the image encoder into a representation in the vector space of their textual counterparts acting as a form of fusion. The resulting embeddings of the linear layer along with the text embeddings are then processed by the LLM that generates a textual response.

3.2.2 LLAVA-Med

The Large Language and Vision Assistant for BioMedicine (LLaVA-Med) [8] is a conversational generative artificial intelligence model. Similarly to LLaVA it takes a multimodal input (text-image) and produces a textual output. The difference between LLaVA and LLaVA-Med is that the latter is a fine-tuned variant of the former, retaining the same architecture. The fine-tuning process is based on a dataset that includes images, captions and question-answer pairs regarding multiple fields of medicine such as X-rays, MRI, Histology, Gross pathology and CT scans. The training procedure consists of two stages one entailing a concept feature alignment and an end-to-end instruction tuning stage.

In the first stage, the image encoder and the language model (LM) weights remain frozen, while only the projection matrix that transforms the visual input embeddings into a vector in the text domain is updated. This training process leads to an improved alignment between the two modalities in the biomedical domain, thus expanding the vocabulary and image understanding of the model.

The subsequent stage involved training the LM and the projection layer weights using an instruction-following dataset in the same way the original LLaVA used one. The model is claimed to achieve good zero-shot task transfer performance on several VQA datasets. However, for more specialized scenarios the model needs further fine-tuning and as with other large multimodal models (LMMs), the model presents hallucinations and has a limited reasoning capability.

3.2.3 Segment Anything Model 2 (SAM2)

SAM2 [2] is a foundation model capable of segmenting objects in both images and videos. It accepts bounding boxes, points, or masks as prompts and outputs a segmentation mask. In this study, it will be tested solely on medical 2D images with a bounding box acting as input. The model's architecture comprises an image encoder, prompt encoder, memory encoder, mask decoder, a memory attention block and a memory bank (Figure 3.2).

The **image encoder** transforms visual data tokens into embeddings, leveraging hierarchical feature extraction. It employs a Masked Autoencoder [13] (MAE) pre-trained Hiera [14, 15] image encoder and after producing the image embedding, multiple prompts and segmentation masks can be inferred.

The **prompt encoder** is a part of the model that produces the input embeddings from masks, points, or bounding boxes.

The **mask decoder** is responsible for generating a mask based on the embeddings it receives. This component can produce multiple outputs in cases of ambiguity, but only one mask will be given on the output without a refinement of the prompts. A notable difference between its predecessor in SAM [7] is that this part can sometimes output no mask in case the model does not detect any valid objects that require segmentation.

At this point, it should be noted that when the SAM2 model is used to segment images instead of videos, the memory mechanism is not utilized and remains empty. This leads SAM2 to behave as SAM. Despite this, in our work, we employ SAM2 due to its training

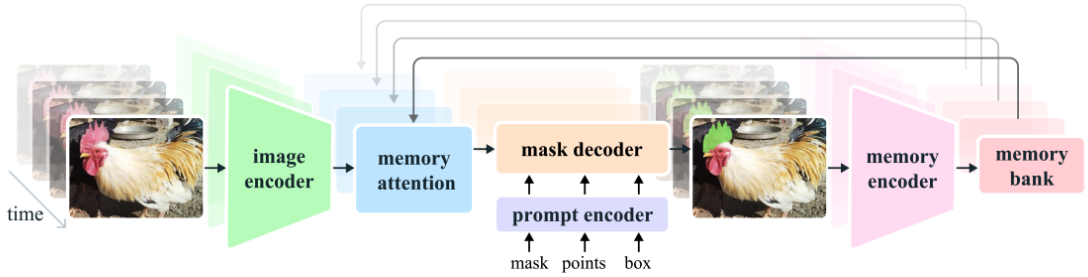


Figure 3.2. *Architecture of SAM2 [2]*

that is based on a larger dataset.

This dataset known as SA-V was created with the use of a data engine. It includes 35.5M masks across 50.9K videos. The masks are not limited to a predefined list of objects but rather regard any visible object or even occluded.

The dataset generation was completed in three phases. Initially, the original SAM model was used in videos to segment objects in each frame and human annotators adjusted the predicted masks. In the subsequent stage, both SAM and SAM2 were utilized. SAM generated a mask for the first frame of the video and SAM2 was tasked with predicting the mask of the same object in the next frames using its memory mechanism. At any frame, annotators may have adjusted the masks with pixel precision. In the last phase, only SAM2 was used in its final structural form accepting all kinds of prompts such as points or masks. During this phase, only slight modifications to the masks were made.

3.2.4 Med-SAM2

The Medical SAM 2 (MedSAM2)[3] is a model developed to handle medical image segmentation tasks both for 2D and 3D applications. It utilizes the SAM2 model as a foundation and builds upon it in terms of architecture (Figure 3.3) and fine-tuning enabling a more diverse set of functions such as the One prompt image segmentation [16]. In more detail, the memory bank of the model has been altered and does not keep the last K frames, but only the most relevant and informative embeddings. In this way, the model is more robust as the images presented are not sorted and the tasks may vary from image to image. Another action in the direction of enhancing the model is the fine-tuning. The dataset used for this was the One-Prompt dataset [16] a compilation of publicly available datasets that span many modalities and organs. All of these actions make the model more generalizable to unseen tasks directed to the medical domain and enhance its capabilities for 3D input, video segmentation and One prompt segmentation.

3.2.5 Grounding DINO

Grounding DINO [4] is an open-set object detector that processes an image-text pair as input to generate bounding boxes for objects relevant to the textual description within the image. It is characterized as open-set because it can be guided from the textual input to detect objects outside of the training set. This capability is enabled through modality fusion (image-text) and grounded pre-training strategies.

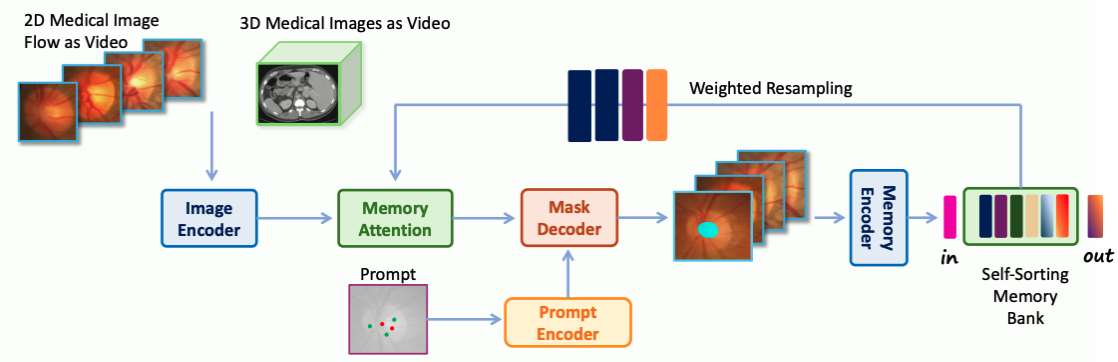


Figure 3.3. *Medical SAM 2 Architecture [3]*

The architecture (Figure 3.4) comprises two distinct encoders and a single decoder. The image encoder utilizes the Swin Transformer [17], a deformable attention-based framework, while the text encoder employs BERT [18]. A feature enhancer facilitates a fusion between the two modalities as it combines cross attention to align the two modalities. Subsequently, a language guided query selection stage selects the most relevant image tokens by evaluating their inner product with text tokens. Lastly, the cross modality decoder is responsible for the production of the bounding boxes.

A key innovation of the Grounding DINO lies in its ability to process each text input on a sub-sentence level (Figure 3.5) in the process of grounded training. This approach enables the model to analyze each sentence word by word, independently of other sentences within the same text prompt. By doing so, it mitigates irrelevant dependencies and ensures more precise alignment between textual and visual features while maintaining the fine grained details of each sentence regarding a single class.

3.3 Related Work

Medical Imaging. Medical imaging plays a fundamental role in healthcare and biomedical research, enabling the delineation of anatomical structures, automated diagnosis and treatment planning. Among its applications, organ segmentation has received significant attention with various deep learning based approaches. A notable example is the U-Net [6] and some variations [19, 20, 21]. Transformer based models have also been developed to address this problem [22, 23], leveraging self-attention mechanisms.

In addition, the advent of the Segment Anything Model (SAM) has further fueled research interest, with numerous efforts being made to tackle this problem by fine-tuning [24, 25], prompting [26] or developing adapter-based enhancements to improve segmentation efficiency [27]. The introduction of Segment Anything 2 (SAM2) [2] has expanded the scope of segmentation research, particularly in the realm of 3D segmentation, with the use of the memory bank architecture.

Another prominent direction of research interest involves multimodal approaches that utilize both image and language modalities. Vision language models such as CLIP [28] have been adapted to address this issue [29]. They utilize text as a means to make

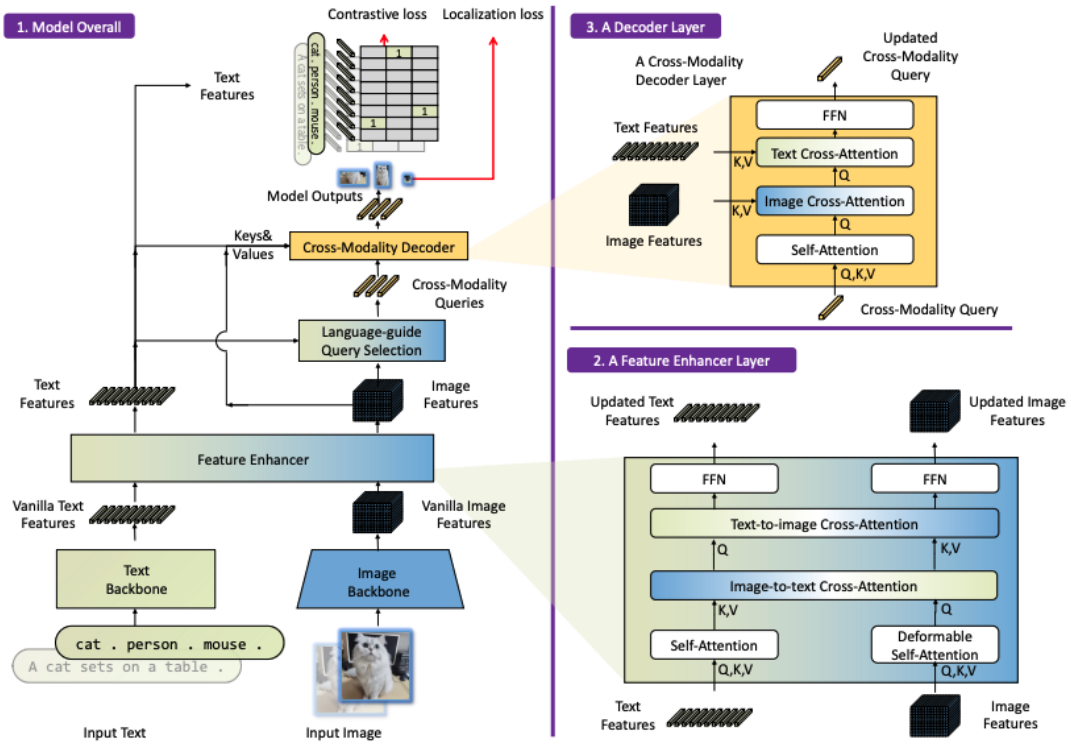


Figure 3.4. *Grounding DINO Architecture [4]*

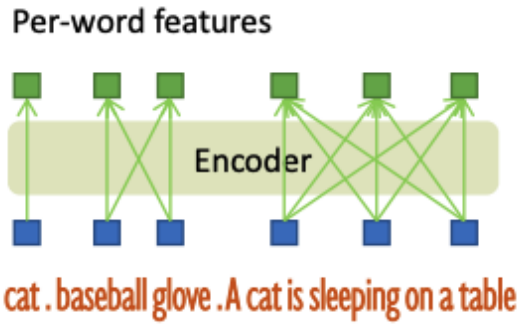


Figure 3.5. *Sub-sentence level [4]*

segmentation more robust and efficient across diverse datasets.

More specialized LVLM's, such as LLaVA [1] and especially LLaVA-Med [8] for the biomedical domain, have been trained to analyze images describe them in detail often detecting if there is an anomaly and are able to answer simple questions.

Grounded Segmentation/Open-set object Detection. Traditional segmentation approaches rely solely on visual features and predefined categories or prompts in the form of points and bounding boxes. Unlike those, the problem of grounded segmentation refers to the use of textual prompts to guide a model in segmenting objects within an image. This method leads to an inherently more generalizable segmentation that utilizes learned features to detect even unseen objects. Similarly, open-set object detection is a task in which a textual input describes the requested entity. The key difference between those two tasks lies in the output, which is a mask in the first case and a bounding box in object detection.

Several implementations have been proposed, including Grounding DINO [4], which integrates transformer-based object detection with textual grounding mechanisms and Grounded SAM [39, 7, 4], which extends the capabilities of SAM by incorporating text prompts.

MIU-VL [30] is a framework that integrates both language and vision to localize and produce a bounding box in the medical domain. It is prompted with the name of the object to be localized and a textual prompt is automatically generated containing information about the color, shape and location of the requested entity. The combined information from the text and image modalities is assessed by a VLM and a bounding box is produced. This study aims to evaluate the generalization of pretrained VLMs on medical images with the appropriate prompts and proceed on fine-tuning models to showcase the performance improvement.

MedLAM [31] is a 3D medical foundation model that is capable of localizing organs, producing three dimensional bounding boxes. It consists of CNN based encoder and decoder along with a multilayer perceptron and has been trained on 16 datasets including 14,012 CT scans. When integrated with SAM, MedLSAM is able to segment an organ, generating the respective 3D mask. The only required input is the class of the organ.

In the same work, to create comparable data to MedLAM for organ localization, they utilized a pretrained 3D ResNet50 backbone [32] following the **DetCo** [33] approach, an unsupervised contrastive learning method for object detection. Additionally, they evaluated the **Mask R-CNN** [34] framework in the same task. Mask R-CNN is a deep learning model for instance segmentation and builds upon Fast/Faster R-CNN [36, 37]. It comprises two stages, the region proposal network which produces candidate bounding boxes and a second stage where it classifies the object and refines the bounding box, while simultaneously predicting a mask.

Chapter 4

Methodology

4.1 Dataset Description

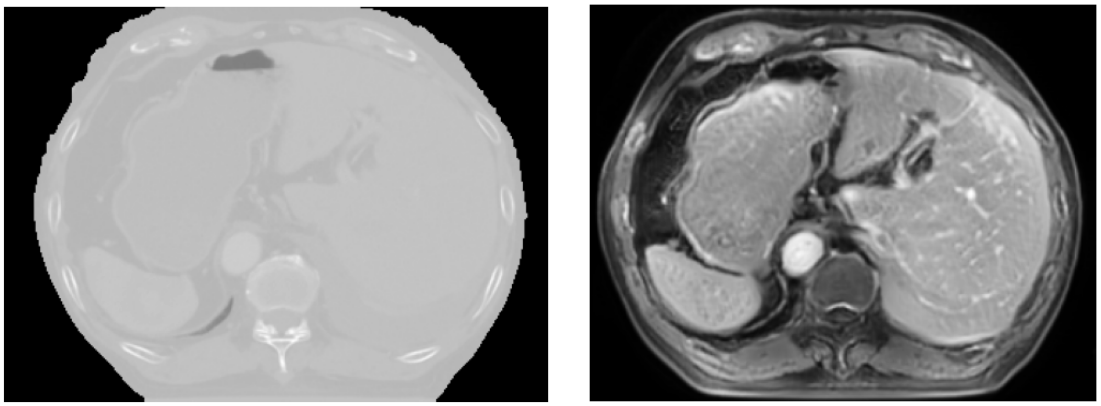
The dataset used for fine-tuning the Grounding DINO is multimodal, consisting of (image, bounding box, text) triplets. The CT scans and Synthetic MR images originate from the RAOS [5, 9] dataset, with the bounding boxes deriving from the annotated label maps within the same dataset. The textual input analyzed in section 4.6.1 provides a detailed description of each annotated entity, including its relative position in the image, shape, brightness and its corresponding name. A snippet of the dataset file can be viewed in Figure 4.1.

```
{
  "filename": "0_0.jpg",
  "height": 259,
  "width": 331,
  "grounding": {
    "caption": "In the image, the Colon has a slightly curved shape...",
    "regions": [
      {
        "bbox": [240, 107, 286, 157],
        "phrase": "<s> In the image, the Colon has a slightly curved shape.
        </s>Part Colon on lower middle and on right part Colon Dark Gray ."
      },
      {
        {"bbox": [44, 76, 79, 124],
        "phrase": "<s> In the image, the Intestine has a curved shape.
        </s>Part Intestine on upper middle and on left part Intestine Dark Gray ."},
      }
    ]
  }
}
```

Figure 4.1. *Example of dataset structure regarding a single image with multiple bounding boxes*

4.1.1 RAOS

The images and bounding boxes employed for fine-tuning the groundingDINO model derive from the RAOS [5, 9] dataset. It consists of 413 real clinical CT scans and 413x9 MR scans from the human abdominal area. It is also worth noting that this dataset contains mask annotations by a senior oncologist, which enables supervised fine-tuning. From the nine available types of Synthetic MRI scans, we have selected to train the model on the category “Delayed” as various filters and imaging techniques have been applied to increase the contrast between the organs and thus create more suitable conditions for computer vision tasks. Among the annotated organs which will be the categories of interest are the liver, spleen, left and right kidneys, stomach, gallbladder, esophagus, pancreas, duodenum, colon, intestine, left and right adrenals and left head of the femur. The rest of the organs in the dataset, such as the rectum, bladder, right head of the femur, prostate and seminal vesicle, are not present in this type of Synthetic MRI but exist in the CT scans. The dataset is also separated into sets A, B and C. Set A has slices considered as the average case, while sets B and C consist of corner cases because of missing organs or parts of them. For this reason, we chose to fine-tune and evaluate the model on the images of Set A, which contains CT scans and MRI slices from 287 patients. It should be noted that the fine-tuning is separate for the CT scans and the Synthetic MRI of type "Delay".



(A) CT scan (B) Synthetic MRI of type "Delayed"

Figure 4.2. *On the left (A) is an original CT scan that displays the organs of the abdomen. On the right (B) the same area is displayed but it appears to be an enhanced version as the organs are more distinguishable, their boundaries are clearer and the contrast is higher.*

The images have been rotated so that the spinal cord is positioned at the lower part of the image, as illustrated in Figure 4.2. Each slice is presented as a top-down view to maintain consistency across the dataset. In total, there are 220 patients for the training set and 67 patients for the test set.

4.2 Dataset Evaluation

The textual descriptions of the masks were evaluated for their accuracy and reliability using a large language model (LLM) and specifically the GPT-4o-mini model [40]. The LLM was initially prompted to assess the descriptions without any further instructions. However, it was quickly evident that it focused on aspects that cannot be changed, such as sentence cohesiveness (Figure 4.3). Subsequently, an explanation regarding the evaluation task was provided to guide the model and ensure that it focused on relevant points (Figure 4.4).

You are a dataset quality assessor. Rate the caption from 1 to 10 based on quality and accuracy.
Assessing region phrase: <s> In the image, the Liver has a slightly oval shape. </s>Liver on lower middle and on right part . Liver Light Gray .

Rating: 5

The caption describes the liver with some basic characteristics, such as its shape and color; however, it is vague and lacks clarity. The statement "Liver on lower middle and on right part" is confusing, as it does not accurately convey the liver's location in relation to the rest of the image. Overall, the caption needs improvement in precision and completeness.

Figure 4.3. *Request to assess the mask description*

In the next example illustrated in Figure 4.4, the model focused on the fragmented nature of the sentence, penalizing the lack of cohesiveness, grammatical and syntactical correctness. Admittedly, the cohesiveness cannot be improved as the sentences pass through a filtering function that omits all the frequently occurring words such as “and” and "the". This filtering aims to mitigate the redundancy by preventing the repetition of such words during the training process, thereby allowing the model to focus on relevant information. For this reason, further refinement of the prompt was necessary. The final version, which serves as the basis for evaluating all textual descriptions is presented in Figure 4.5.

You are a dataset quality assessor. Rate the caption from 1 to 10 based on quality and accuracy keeping in mind that this is the dataset for grounded training of a model that detects bounding boxes. We would like to give the model the shape, relative position of the organ in the image and its brightness.
Assessing region phrase: <s> In the image, the Liver has a slightly oval shape. </s>Liver on lower middle and on right part . Liver Light Gray .

I would rate this caption a 7 out of 10.

Here's my reasoning: - **Shape**: It accurately describes the liver as having a slightly oval shape, which is relevant for the model to learn about the organ's form. - **Position**: It indicates that the liver is located in the lower middle and the right part of the image, which provides useful information about its relative position. - **Brightness**: It describes the liver's color as light gray, which is important for the model to learn about the organ's appearance.
However, the phrase structure is a bit fragmented and could be improved for clarity. Additionally, a more precise description of the liver's spatial relation to other organs or anatomical landmarks could enhance the caption's informative value. Overall, it contains the essential elements needed for training but could benefit from refinement.

Figure 4.4. *Request to assess the mask description with some guidance*

You are a dataset quality assessor. Rate the caption from 1 to 10 based on quality and accuracy keeping in mind that this is the dataset for grounded training of a model that detects bounding boxes in synthetic MRI images of type Delay. We would like to give the model the shape, relative position of the organ in the image and its brightness. We have filtered out words that appear too often. Thus, only important keywords remain that bind the relative words with the region. Disregard grammatical and syntactical correctness.
Assessing region phrase: <s> In the image, the Liver has a slightly oval shape. </s>Liver on lower middle and on right part . Liver Light Gray .

Rating: 8

The caption accurately describes the shape (slightly oval) and position (lower middle and right part) of the liver, as well as its brightness (light gray). It includes important keywords that are relevant for the model's understanding of the texture, location and appearance of the organ in the synthetic MRI image. However, the inclusion of "on" before "right part" could be clearer to specify the exact position more effectively. Overall, it provides a good amount of relevant information for grounded training.

Figure 4.5. *Request to assess the mask description with detailed guidance*

Finally, the average rating will be a metric of the accuracy and quality of the dataset.

Average Rating
7.25/10

Table 4.1. *Average rating of the accuracy and quality of dataset's textual modality*

4.3 Shape Descriptions by LLaVA-Med

In this work, we have assigned the task of finding the shape for each mask to LLaVA-Med. In this section the descriptions are presented analytically per category.

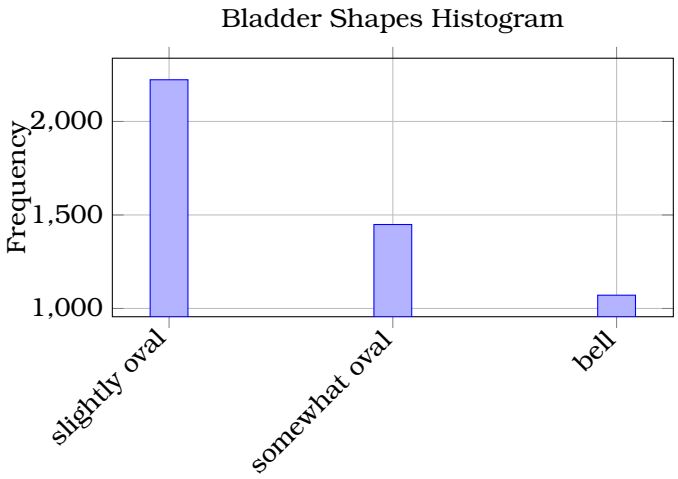


Figure 4.6. *Histogram of Bladder Shapes*

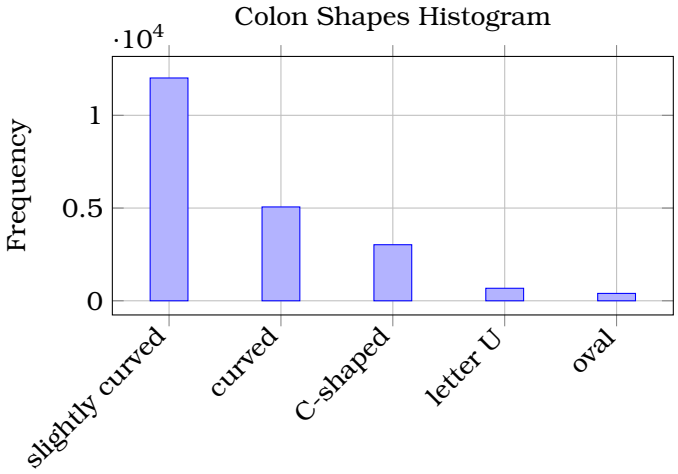


Figure 4.7. *Histogram of Colon Shapes*

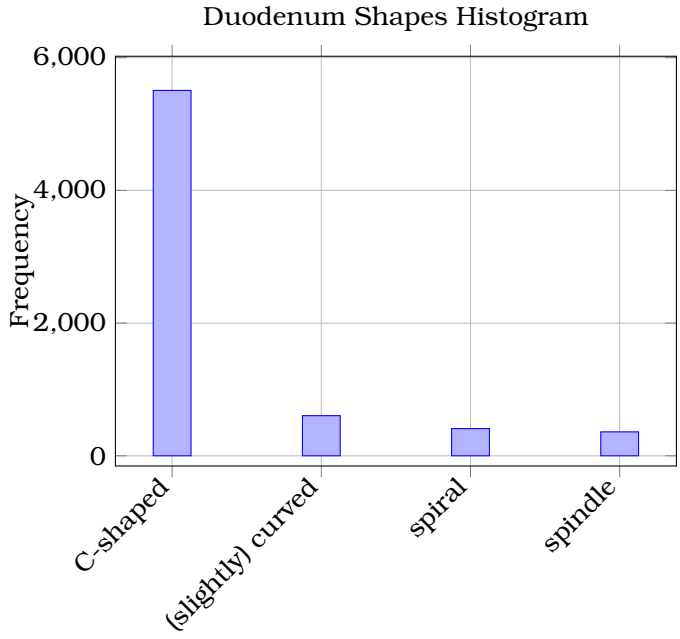


Figure 4.8. *Histogram of Duodenum Shapes*

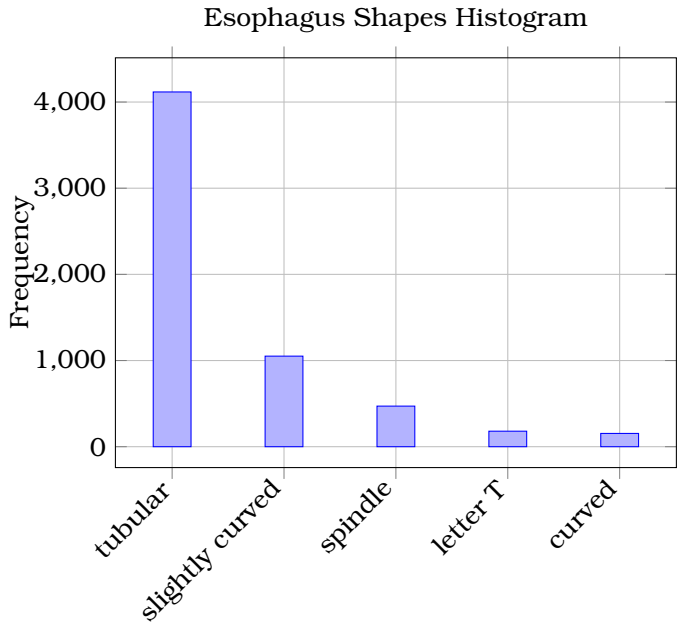


Figure 4.9. *Histogram of Esophagus Shapes*

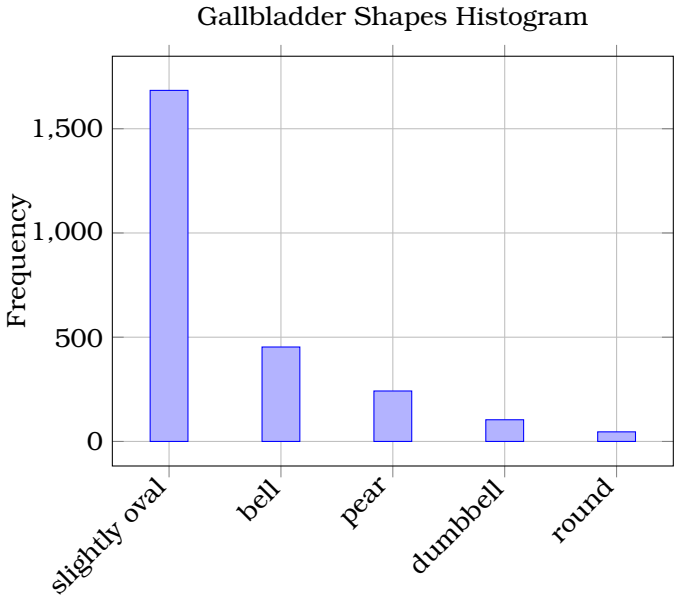


Figure 4.10. *Histogram of Gallbladder Shapes*

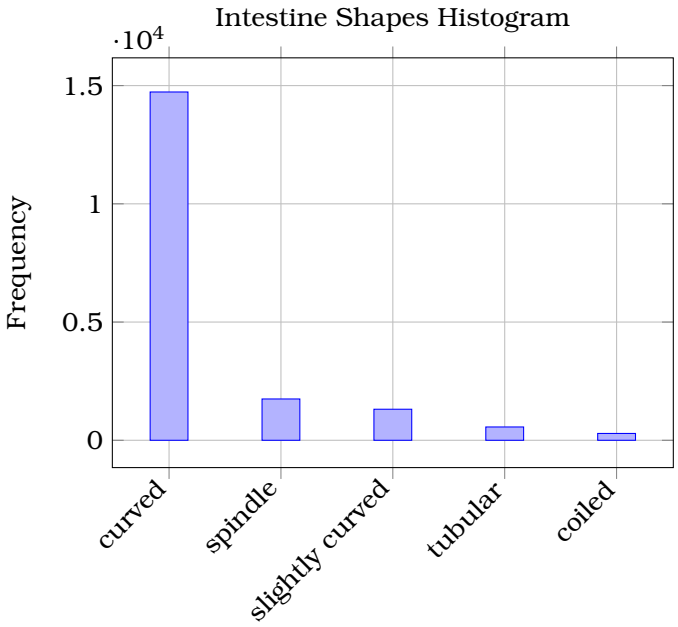


Figure 4.11. *Histogram of Intestine Shapes*

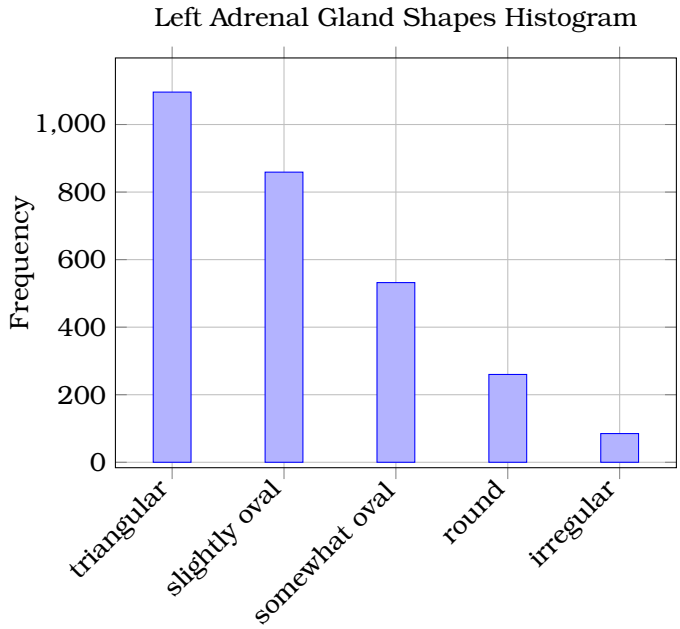


Figure 4.12. *Histogram of Left Adrenal Gland Shapes*

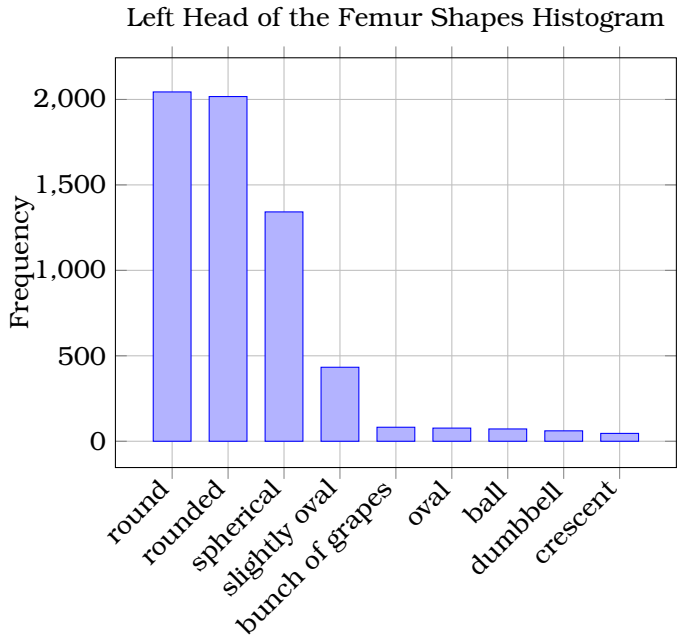


Figure 4.13. *Histogram of Left Head of the Femur Shapes*

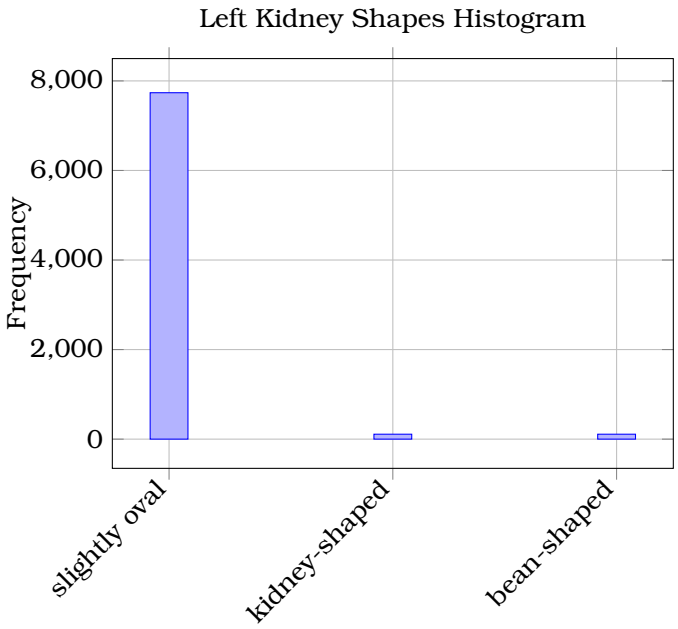


Figure 4.14. *Histogram of Left Kidney Shapes*

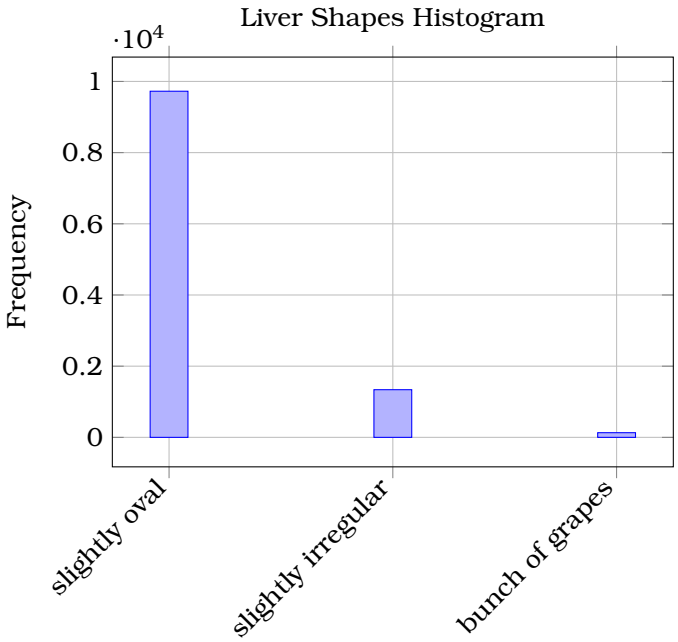


Figure 4.15. *Histogram of Liver Shapes*

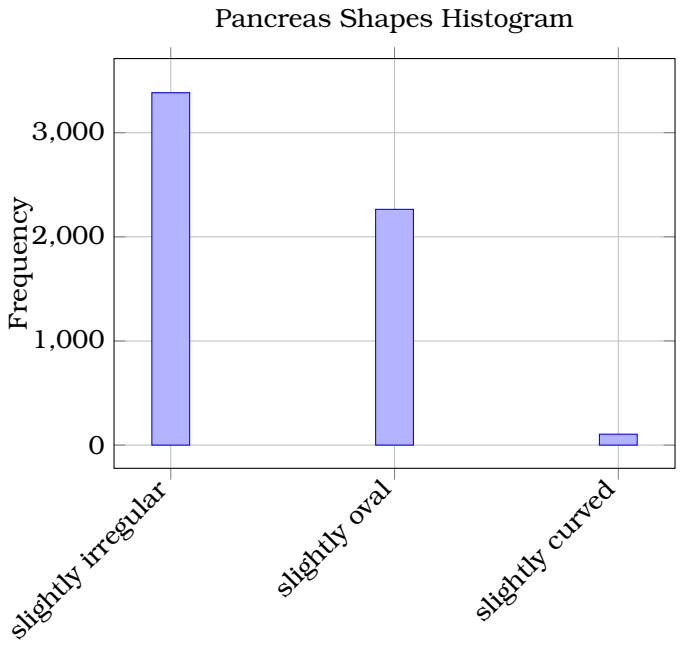


Figure 4.16. *Histogram of Pancreas Shapes*

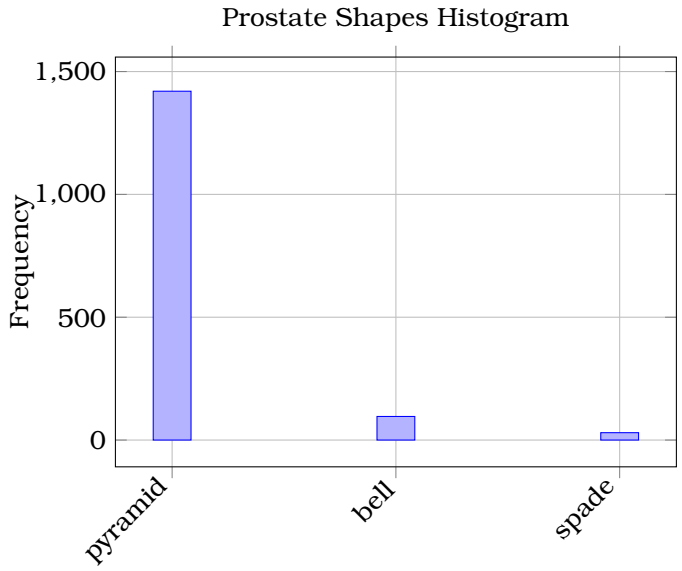


Figure 4.17. *Histogram of Prostate Shapes*

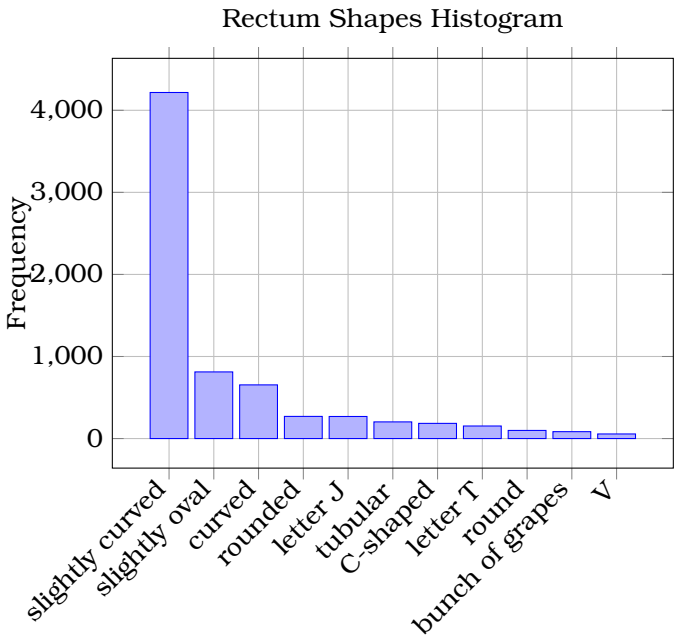


Figure 4.18. *Histogram of Rectum Shapes*

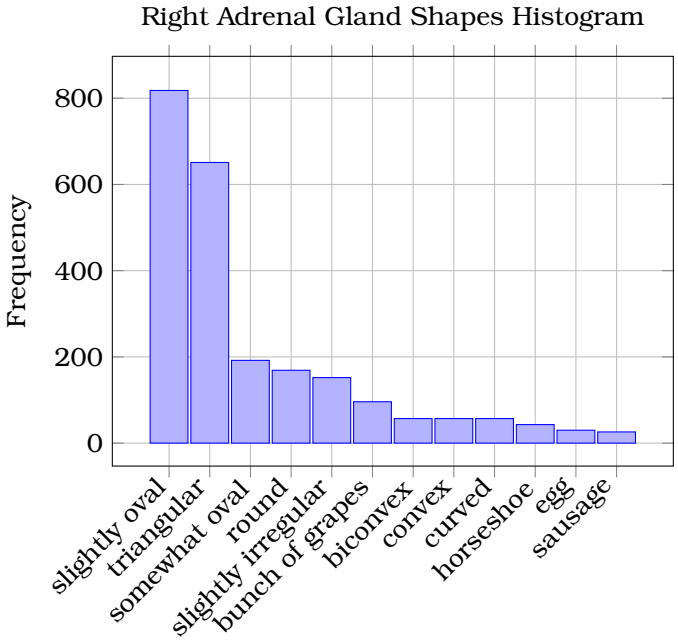


Figure 4.19. *Histogram of Right Adrenal Gland Shapes*

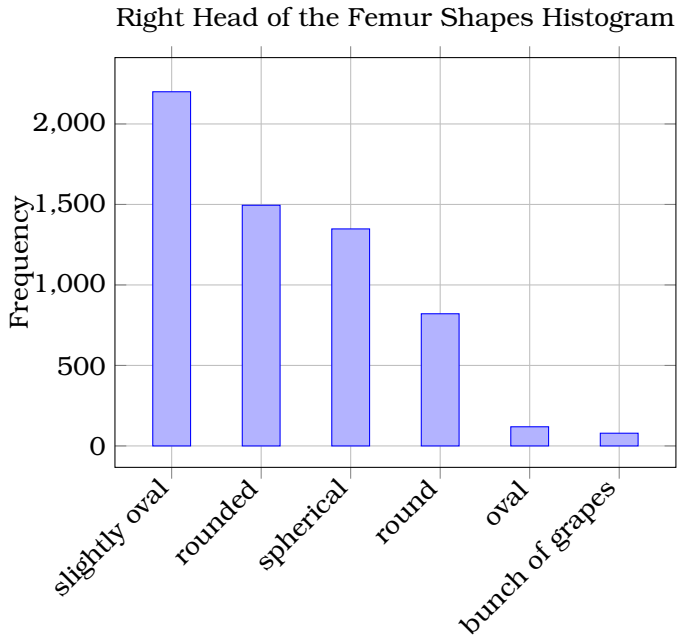


Figure 4.20. *Histogram of Right Head of the Femur Shapes*

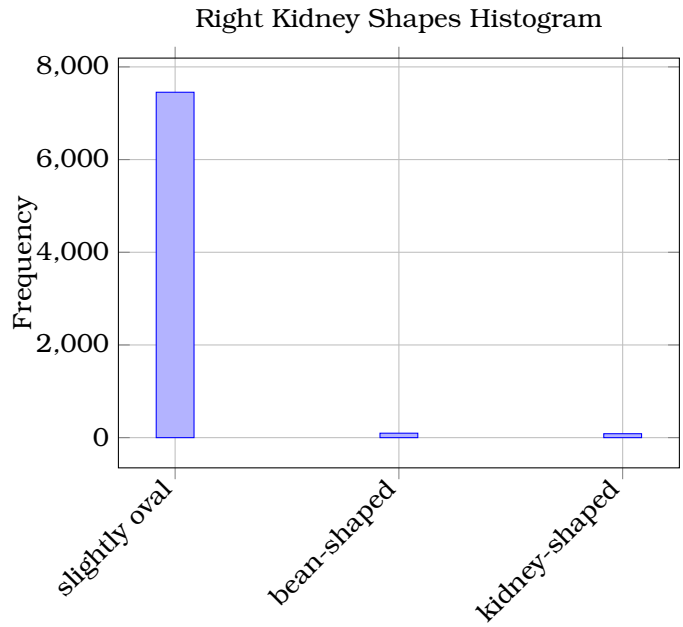


Figure 4.21. *Histogram of Right Kidney Shapes*

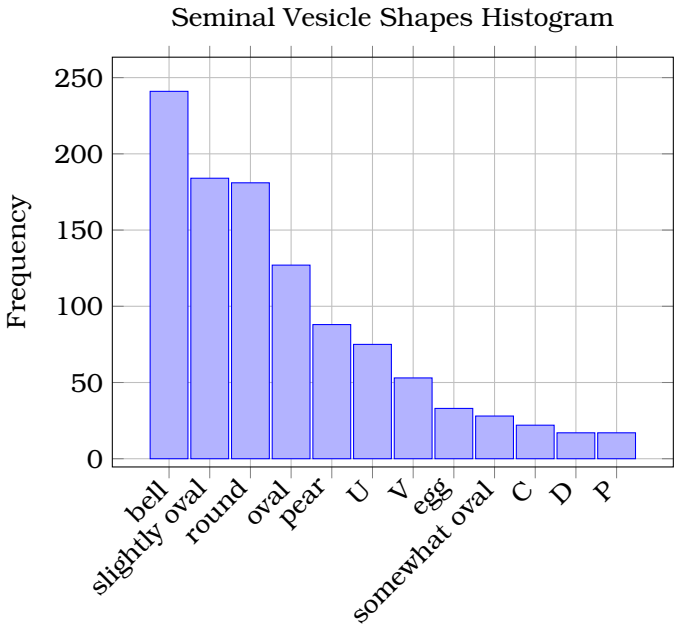


Figure 4.22. *Histogram of Seminal Vesicle Shapes*

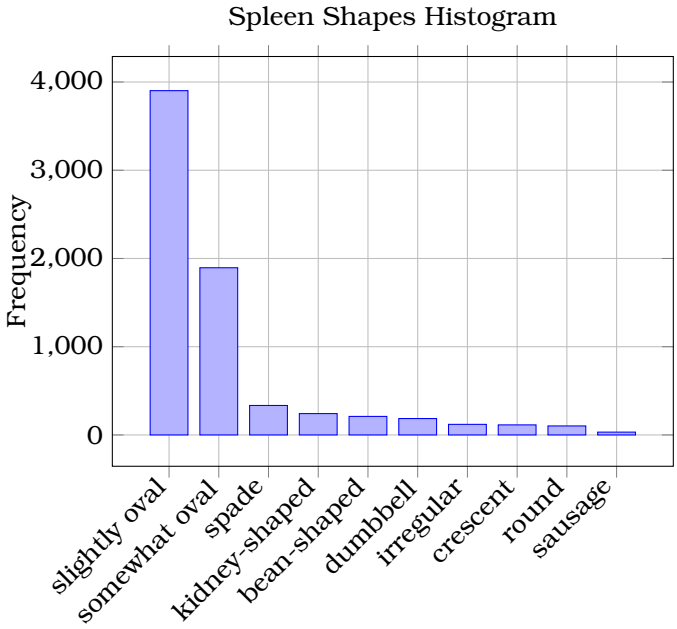


Figure 4.23. *Histogram of Spleen Shapes*

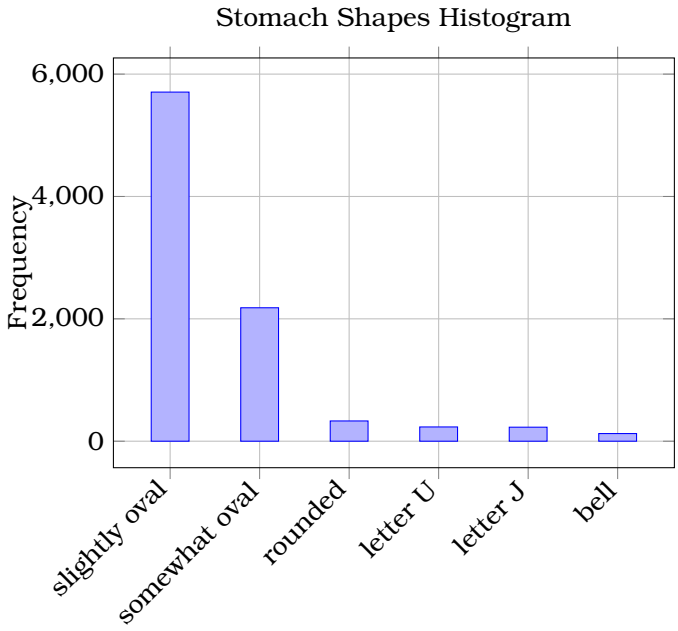


Figure 4.24. *Histogram of Stomach Shapes*

4.4 Visualization Tool

During data preparation, a tool (Figure 4.25) was developed to display abdominal Synthetic MRI or CT slices of patients and allow easy switching between them using a slider. This tool displays the actual image on the left and a label map on the right. Users can interact with the tool using the slider or by clicking on any region within the image or on the label map to reveal the associated mask label. Additionally, the tool provides the pixel value and the coordinates of the cursor on the image, improving the interpretability of the data.

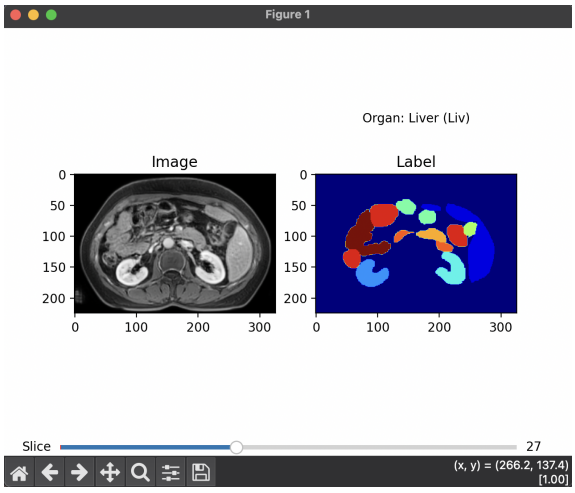


Figure 4.25. *Visualization tool for the patient's abdominal images.*

4.5 Data Augmentation

In the medical field, CT and MRI scans can appear in multiple orientations and in some cases, may even be flipped. To address this issue, we extended the dataset by incorporating rotated and flipped versions of the existing images. This augmentation enhances the model's ability to generalize across different orientations, improving its robustness when processing normally oriented scans. In addition, we applied resizing and random cropping to ensure that the model learns from objects of different scales and improve its robustness to object occlusion and localization errors.

As a result, 35 patients were randomly selected and their slices were appended to the dataset in altered forms:

- 5 patient's slices were rotated 90 degrees
- 5 patient's slices were rotated 180 degrees
- 5 patient's slices were rotated 270 degrees
- 5 patient's slices were flipped
- 5 patient's slices were flipped and rotated 90 degrees
- 5 patient's slices were flipped and rotated 180 degrees
- 5 patient's slices were flipped and rotated 270 degrees

4.6 Project Phases

4.6.1 Generation of textual descriptions for the masks

The first stage of the experiments involved prompting LLaVA-Med to create relative captions for the images. These captions should describe in detail the organs depicted, forming image-text pairs. The aim was to ultimately extract the knowledge of the fine-tuned LVLM in the biomedical field and distill it to groundingDINO. This step is essential for enabling the latter model to comprehend medical terminology and accurately produce a bounding box around the relevant organs.

To achieve this, key features were identified, namely shape, position, brightness and texture. Therefore, these constitute the primary attributes that LLaVA-Med will be asked to produce. However, texture was dropped as a feature. As for the relative position and brightness, deterministic algorithms were used instead to generate the textual descriptions, due to the inaccuracies of the LVLM. The overall process is illustrated in Figure 4.26 and the reasons for selecting this structure are presented in section 5.1.

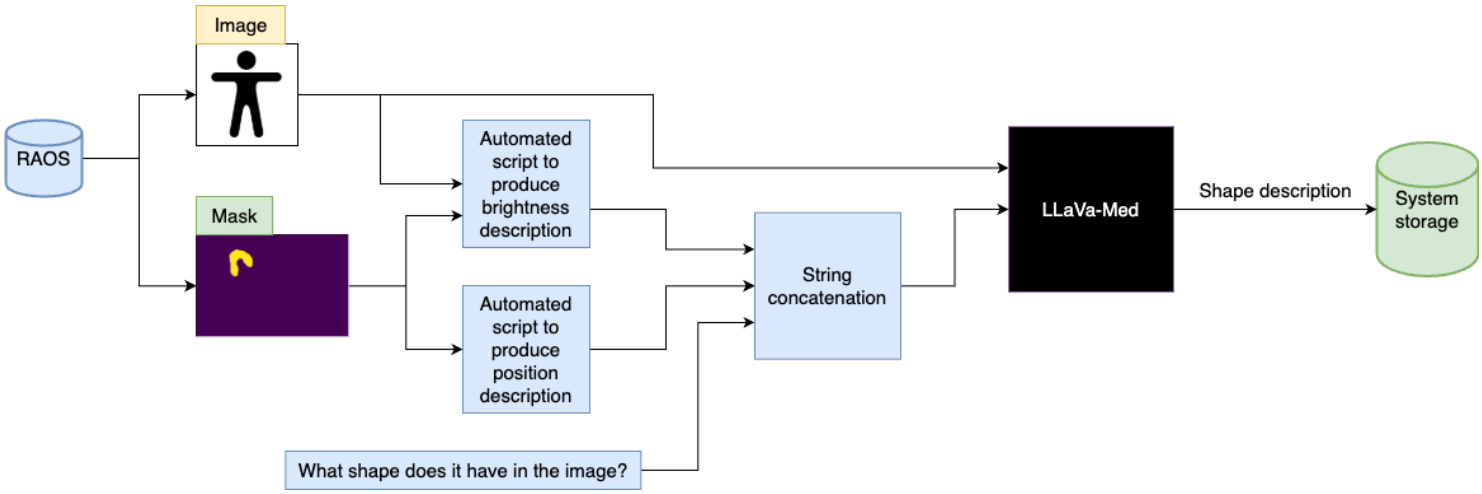
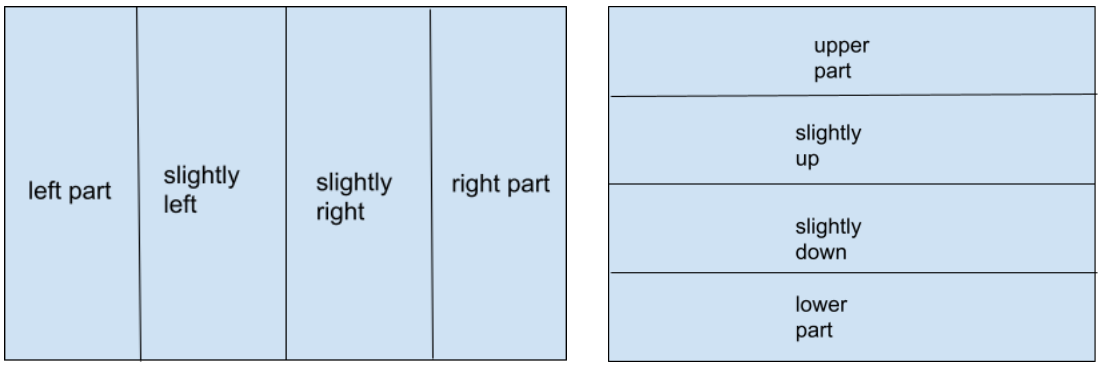


Figure 4.26. *Extraction of textual descriptions for the shape of each mask*

Relative position

The locations provided by LLaVA-Med were inaccurate, making them unsuitable for fine-tuning the GroundingDINO, as the erroneous information could be learned and result in inaccuracies during inference. To address this issue, the retrieval of organ positions was automated using Python scripts. For each organ mask, whether it represented a non-contiguous part of an organ or the whole organ, the center of mass was calculated. Subsequently, a sentence was generated based on the region of the image where the center of mass was located. The image divisions, are shown in Figure 4.27.



(A) Horizontal division (B) Vertical division

Figure 4.27. *The image is divided into four equal parts along both the vertical and horizontal axes. Each part is assigned a descriptive label. The center of mass of each mask is used to determine its position, acquiring one descriptor for the horizontal axis and another for the vertical axis. For instance a generated description might be: "The Left Kidney is located slightly up and slightly left in the image and appears to be Very Light Gray."*

Brightness

An automated method for generating captions regarding the brightness of each organ was also implemented. Since the dataset described in section 4.1 includes pixel-level annotations, it allows for the computation of the mean brightness value for each segmented

entity. The brightness scale used in this process to classify the brightness of a mask is presented in Figure 4.28.

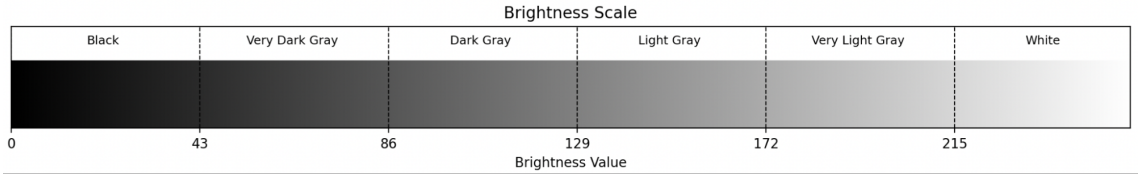
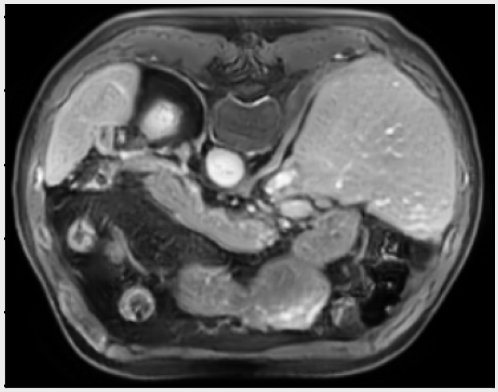


Figure 4.28. *Brightness scale utilized to categorize the brightness of each organ mask based on the mean pixel intensity value*

Shape

We ultimately decided to provide the LLaVA-Med with the information about the position and brightness to assist it in locating the organ within the image. The revised prompts for LLaVa-Med, incorporating both positional and brightness information, are illustrated in figure 4.29.



User: The Stomach is located on the lower part and slightly right in the image and appears to be Dark Gray. What shape does it have in the image?

LLaVA-Med: The stomach in the image has a round shape.

Figure 4.29. *Revised prompts, utilizing automated position and brightness descriptions, that should help the model identify the organ.*

4.6.2 Fine-tuning Grounding DINO

For this phase, the Open-GroundingDino [41] was utilized, a third-party implementation of GroundingDINO with fine-tuning capability. The diagram depicting the process is illustrated in Figure 4.30.

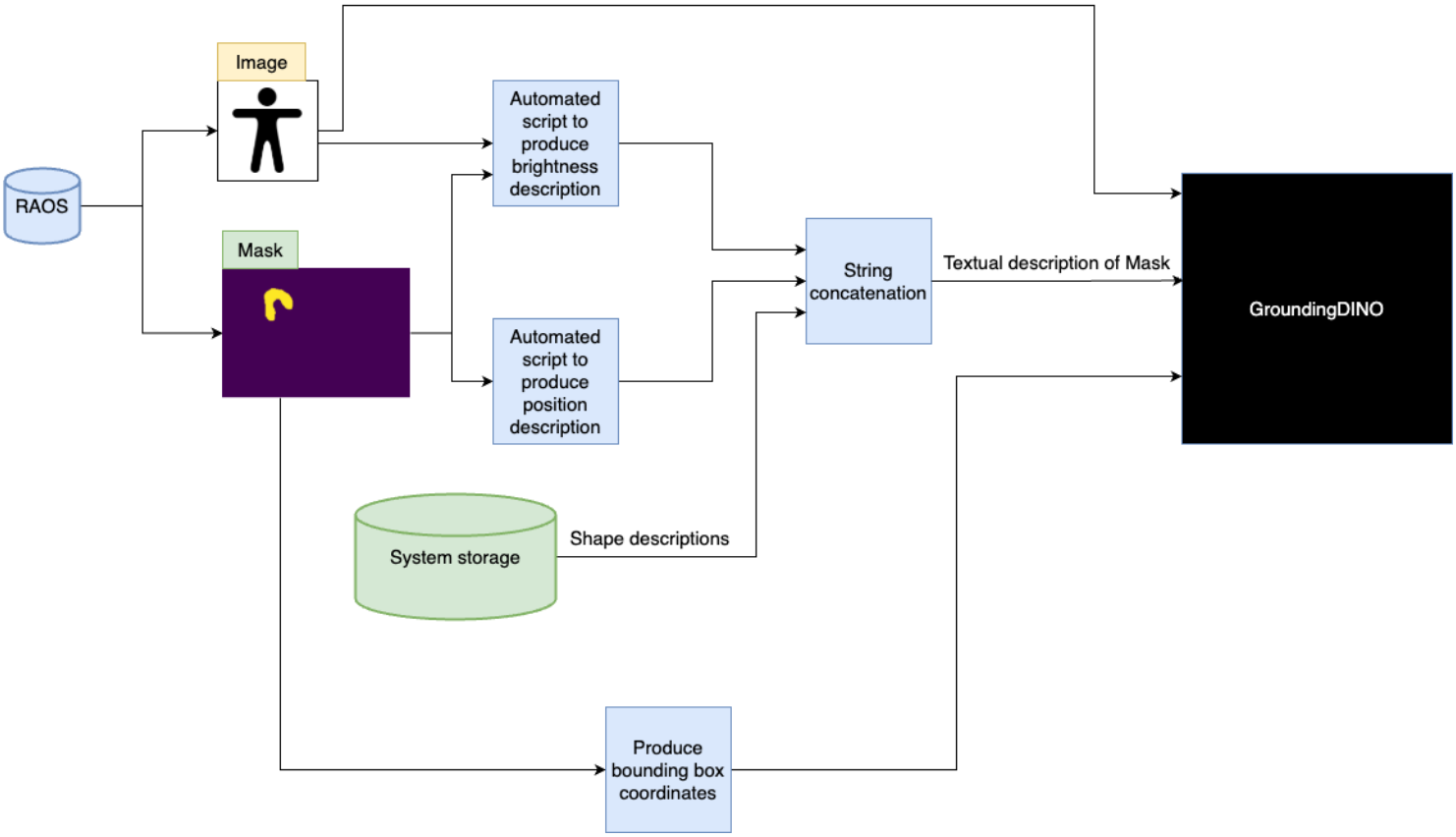
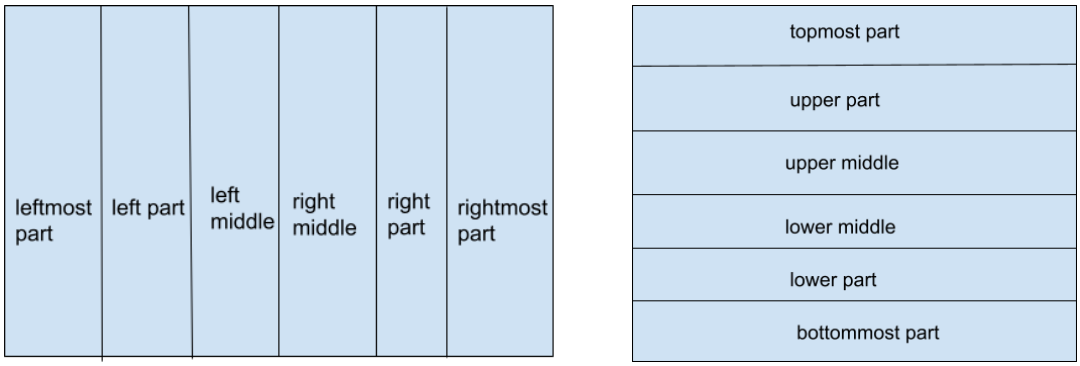


Figure 4.30. *Fine-tuning of GroundingDINO diagram*

The dataset used consisted of images from the RAOS dataset, as mentioned earlier, enriched with shape, position and brightness descriptions for each mask in every slice. The brightness description was calculated using the same methodology as in the prompts provided to Llava-Med. However, to enhance the accuracy of the position descriptions, the image was further subdivided and the number of possible combinations increased (Figure 4.31). For organ shape descriptions, the output of Llava-med was utilized and treated as ground truth.



(A) Horizontal division (B) Vertical division

Figure 4.31. *The image is divided into six equal parts along both the vertical and horizontal axis. The center of mass of each organ is used to determine its spacial description, acquiring one regarding the horizontal axis and another regarding the vertical axis. For instance, a generated description might state: "The Left Kidney is located on the topmost part and on the left part in the image and appears to be Very Light Gray."*

Before concatenating the descriptions obtained from LLaVA-Med (shape) and those from the deterministic algorithms (position and brightness), a filtering function was applied to the combined prompt for position and brightness. This step aimed to minimize redundancy by reducing the frequency of commonly used words, helping the model to focus on relevant information. In contrast, the shape prompt remained unfiltered for the primary dataset images but was filtered for the augmented, flipped or rotated, images. Without the full filtering regarding the augmented images, the model exhibited signs of collapse, failing to generate meaningful outputs. This issue may be attributed to the fact that in this case the same textual prompts would be combined with the original and the altered augmented versions of the same image, thus hindering the learning process.

An additional advantage of adding flipped and rotated images to the dataset is the improvement in precision with respect to the task of object detection. Without the inclusion of augmented images, the model failed to generate any results for object detection. However, with the augmented dataset, its performance improved significantly, demonstrating greater accuracy in detecting objects. Nevertheless, further training is required to optimize the model's performance on this task.

Hyperparameters

Grounding DINO was trained using Swin-B [17] (swin_B_384_22k) as the image backbone and BERT-base-uncased [18] as the text encoder. The model operates with 900 queries and supports a maximum text token length of 256. It features six encoder and six decoder layers, with a hidden dimension of 256 and a feedforward dimension of 2048. The transformer uses 8 attention heads with ReLU activation. Training involves batch size 4, a base learning rate of 0.0001 and weight decay of 0.0001. Data augmentation is applied with scales ranging from 480 to 800 and a maximum image size of 1333. The optimizer follows a learning rate drop at epochs 4 and 8. For loss computation, the Hungarian matching is used with classification, L1 and GIoU losses. The model is further

enhanced by text cross-attention and a fusion layer.

Environment

Our environment is equipped with an NVIDIA A10G GPU, running driver version 550.127.05 and CUDA 12.4. The training process took approximately 44 hours for the synthetic MRI images, while the LLaVA-Med required around 24 hours to produce the shape descriptions.

4.6.3 Evaluation of the fine-tuning

Evaluation metrics

The experiment involves prompting Grounding DINO with a textual description of the mask, which in turn generates a bounding box that serves as input to the SAM2 model. The SAM2 model then produces the final output, a segmentation mask of the target object. The performance of this pipeline is assessed using two evaluation metrics: Intersection over Union (IoU) and precision. For the comparison with other models in the next chapter we will use the Dice similarity coefficient and wall distance.

Intersection over Union (IoU)

As the name suggests, intersection over union (IoU) is defined as the ratio of the common pixel area between the predicted and the ground truth mask to the total area covered by both masks. A value of one indicates a perfect match, while lower values signify a deviation from the ground truth. In mathematical terms:

$$\text{IoU} = \frac{\text{Area of Intersection}}{\text{Area of Union}} = \frac{|A \cap B|}{|A \cup B|} \tag{4.1}$$

where A and B are the ground truth and predicted masks, respectively.

Precision

Precision is defined as the ratio of true positive values to the sum of true positive and false positive values. It quantifies the proportion of the predicted mask that correctly overlaps with the ground truth mask. A precision value of one indicates that all predicted pixels belong to the ground truth mask, whereas lower values suggest that some pixels that do not belong in the mask have been classified as mask pixels. Mathematically, precision is expressed as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{4.2}$$

Dice Similarity Coefficient

The dice similarity coefficient is expressed mathematically as follows:

$$\mathrm{DSC}(A,B) = \frac{2|A \cap B|}{|A| + |B|}. \tag{4.3}$$

where A and B are the ground truth and predicted masks, respectively. A score of 1 is the perfect match, while 0 indicates no common area.

Wall Distance

The Wall Distance expresses the average absolute difference between the boundaries of the predicted and the ground truth bounding boxes. In two dimensional bounding boxes the boundaries are four while in three dimensional bounding boxes the boundaries are six. We showcase the Wall Distance of a 2D bounding box:

$$\mathrm{WD} = \frac{1}{4}\sum_{i=1}^{4} |\hat{d}_i - d_i| \tag{4.4}$$

where d_i and $\hat{d}_i$ are the boundaries of the predicted and ground truth bounding boxes respectively, belonging to the same side.

Evaluation

The fine-tuned Grounding DINO model was evaluated on a dataset consisting of synthetic MR images of type "Delay" and CT scans from 67 patients with slices of their abdominal area. The testing pipeline, as illustrated in Figure 4.32, consists of Grounding DINO and a segmentation model, such as SAM2 or Med-SAM2. In this process, the image, along with a detailed description of each mask, are provided as input to Grounding DINO, which generates a bounding box. This pair (image, bounding box) is then used as the input for the segmentation model that processes it and produces a segmentation mask.

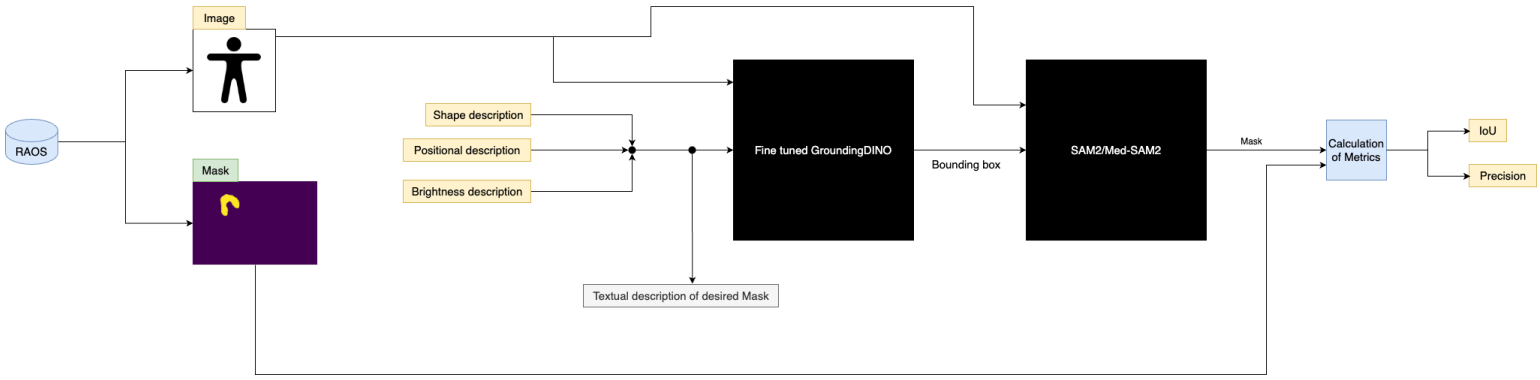


Figure 4.32. *Inference mode architecture*

Chapter 5

Results

5.1 Prompting LLaVA-Med

Initially, the ability of LLaVA-Med to detect organs within an image was evaluated. Its responses were not always accurate, as unannotated organs in our dataset were identified and occasionally, it failed to detect relevant ones. To improve the approach, we decided to adjust the prompts to explicitly mention the names of the relevant organs before asking other questions about them as depicted in figure 5.1.

In the next example illustrated in figure 5.2 the image is flipped and we explicitly mention the relevant organs before assigning a task to the model. The answer indicates that the model lacks robustness to this flipped version of the image, as it incorrectly returns the general anatomical position of the liver (upper right instead of lower left). Regarding kidney shape, LLaVA-Med correctly identifies them as "bean shaped". However, it also provides extraneous general information, such as the function of the kidneys and a comparison of their vertical positions in three dimensions, which does not contribute to groundingDINO's ability to generate accurate bounding boxes. This example suggests that the model prioritizes learned knowledge over direct image-based reasoning. Instead, the locations it returns are the anatomical positions of the organs in the human body and not the relative positions of the organs in the image.

In the subsequent experiment (Figure 5.3), we rotated the image 180 degrees to evaluate whether the model's descriptions of organ positions became more accurate than previously. Although the positional descriptions improved, it is hypothesized that the responses are not directly inferred from the image, but rather align with the model's learned anatomical knowledge. Furthermore, the model continues to avoid elaborating on the position and shape of some organs, failing to describe all the requested structures.

We further concentrated on asking about the quadrants to which each organ belongs (Figure 5.4), as the model demonstrated an understanding of this task and provided consistent responses without including general information. To enhance the model's performance, each task was segmented into distinct prompts, simplifying the process and task clarity.

Attempts to reason with the model

From the responses, it is evident that LLaVA-Med has learned the general shapes of the organs, although in many cases the answers reflect the three-dimensional shape rather than the actual shape observed in the image. For example, the model may describe the duodenum as "C-shaped" even when it appears ovoid in the given image.

One approach to address the problem of the 3D shape descriptions that it returns is to present this task as a multiple choice question for LLaVa-Med between the 2D descriptions of the organ and let the model deduce the correct shape. However, this strategy proved ineffective, as demonstrated in Figure 5.5.

Regarding shape and texture, another reasoning-based prompting strategy was also utilized to enhance the quality and reliability of the responses. This method involved providing detailed descriptions of the characteristics of each shape and texture while guiding the model through a step-by-step reasoning process. By analyzing the available information and identifying the relevant features, the model was expected to provide more precise predictions (Figures 5.5, 5.7). However, this proved ineffective as the model's reasoning capabilities are limited.

Texture descriptions

Regarding texture, a review of the training set of the LLava-Med was conducted to identify frequently occurring terms associated with texture. Based on this analysis, the model was asked whether an organ exhibited a homogenous or heterogenous texture, leveraging its familiarity with these terms (Figure 5.6). However, the model demonstrated limited reliability and struggled to recognize homogenous and heterogenous textures. This issue may be attributed to the fact that there are numerous synthetic MRI types, each displaying organs with varying colors and textures. Consequently, the model's responses were inaccurate, likely due to its lack of exposure to the specific synthetic MRI type. Given these challenges, **texture was ultimately discarded** as a descriptive feature for each organ.

5.2 Evaluation on Synthetic MRI of type "Delayed"

The model was fine-tuned for eleven epochs and this updated version of Grounding DINO demonstrated enhanced capabilities, surpassing the original foundation model in the task of segmenting medical images (Table 5.10). However, the performance differs across the organs, as some are more clearly visible than others due to higher contrast at their borders and distinct texture. A significant challenge is that certain organs are not easily distinguishable to the untrained eye, necessitating expert knowledge for accurate segmentation.

During the testing phase, a series of evaluations were performed using prompts with varying levels of detail, including or excluding position, shape and brightness descriptions of the masks. Initially, the model was provided with only positional information, followed by a combination of positional and shape data and finally, with position, shape and

brightness details for each organ. As anticipated, the model's performance improved as more descriptive and inclusive input was provided (Table 5.11).

Another evaluation was conducted to assess the impact of bigger sentences on the model's performance. In the previous tests, the model was prompted with only one full sentence regarding the shape and keywords for the position and brightness. However, in this phase, two and three full-sentence descriptions were given. The results indicate that this excess of information distracts the model, leading to slightly inferior results. While shorter prompts provide more focused input, full-sentence prompts necessitate that the model filter out irrelevant details, which may result in the model overlooking essential information (Tables 5.5 and 5.6).

Fine-tuned Grounding DINO (Epoch 11) + SAM2	IoU	Precision
Liver	0.7491	0.7996
Spleen	0.7869	0.8399
Left Kidney	0.8344	0.8770
Right Kidney	0.8153	0.8582
Stomach	0.6430	0.7549
Gallbladder	0.4746	0.5584
Esophagus	0.5132	0.6663
Pancreas	0.3246	0.3831
Duodenum	0.3641	0.4523
Colon	0.4988	0.6183
Intestine	0.3063	0.3541
Left Adrenal	0.2499	0.3186
Right Adrenal	0.3187	0.4171
Segmentation Overall Average	0.5230	0.6010

Table 5.1. *Performance results*

Fine-tuned Grounding DINO (Epoch 11) + Med-SAM2	IoU	Precision
Liver	0.7281	0.8194
Spleen	0.7815	0.8763
Left Kidney	0.7989	0.9235
Right Kidney	0.7915	0.9036
Stomach	0.6312	0.7738
Gallbladder	0.4637	0.5387
Esophagus	0.5432	0.6648
Pancreas	0.3055	0.3916
Duodenum	0.3642	0.4635
Colon	0.4967	0.6139
Intestine	0.2965	0.378
Left Adrenal	0.233	0.2896
Right Adrenal	0.3334	0.4088
Segmentation Overall Average	0.5133	0.6157

Table 5.2. *Performance results*

Fine-tuned Grounding DINO (Epoch 11) + SAM2 (Position only)	IoU	Precision
Liver	0.7557	0.8051
Spleen	0.7677	0.8160
Left Kidney	0.8205	0.8608
Right Kidney	0.7996	0.8388
Stomach	0.5759	0.6674
Gallbladder	0.3927	0.4663
Esophagus	0.4758	0.6228
Pancreas	0.2641	0.3090
Duodenum	0.2509	0.3022
Colon	0.2906	0.3613
Intestine	0.1744	0.1980
Left Adrenal	0.1670	0.2223
Right Adrenal	0.2199	0.2991
Segmentation Overall Average	0.4232	0.4785

Table 5.3. *Performance results*

Fine-tuned Grounding DINO (Epoch 11) + SAM2 (Position and shape)	IoU	Precision
Liver	0.7509	0.8010
Spleen	0.7830	0.8353
Left Kidney	0.8295	0.8705
Right Kidney	0.8106	0.8517
Stomach	0.6419	0.7539
Gallbladder	0.439	0.5173
Esophagus	0.4916	0.6321
Pancreas	0.3175	0.3718
Duodenum	0.3254	0.3987
Colon	0.4431	0.5517
Intestine	0.2859	0.3333
Left Adrenal	0.2103	0.2794
Right Adrenal	0.2916	0.3869
Segmentation Overall Average	0.4993	0.5726

Table 5.4. *Performance results*

Fine-tuned Grounding DINO (Epoch 11) + SAM2 (Unfiltered prompts natural language shortened version with 2 sentences)	IoU	Precision
Liver	0.7461	0.7957
Spleen	0.7432	0.7956
Left Kidney	0.8304	0.8712
Right Kidney	0.8138	0.8553
Stomach	0.6350	0.7652
Gallbladder	0.4557	0.5473
Esophagus	0.4915	0.6363
Pancreas	0.3088	0.3659
Duodenum	0.2570	0.3207
Colon	0.4616	0.5756
Intestine	0.237	0.2786
Left Adrenal	0.2043	0.2657
Right Adrenal	0.2828	0.3710
Segmentation Overall Average	0.4855	0.5597

Table 5.5. *Example of input: "In the image, the Liver has a slightly oval shape. The Liver is located on the lower middle and on the right part of the image and appears to be Light Gray."*

Fine-tuned Grounding DINO (Epoch 11) + SAM2 (Unfiltered prompts natural language with 3 sentences)	IoU	Precision
Liver	0.7369	0.7863
Spleen	0.7814	0.8351
Left Kidney	0.8323	0.8739
Right Kidney	0.8146	0.8575
Stomach	0.6319	0.7619
Gallbladder	0.4647	0.5561
Esophagus	0.4981	0.6475
Pancreas	0.2968	0.3533
Duodenum	0.2962	0.3685
Colon	0.4357	0.5450
Intestine	0.2377	0.2818
Left Adrenal	0.2292	0.2974
Right Adrenal	0.3048	0.3981
Segmentation Overall Average	0.4842	0.5590

Table 5.6. *Example of input: "In the image, the Liver has a slightly oval shape. The Liver is located on the lower middle and on the right part of the image. The Liver in the image appears to be Light Gray."*

Fine-tuned Grounding DINO (Epoch 7) + SAM2	IoU	Precision
Liver	0.7548	0.8057
Spleen	0.7876	0.840
Left Kidney	0.8345	0.8775
Right Kidney	0.8196	0.8626
Stomach	0.6311	0.7402
Gallbladder	0.4402	0.5185
Esophagus	0.486	0.6211
Pancreas	0.3184	0.3753
Duodenum	0.3259	0.4006
Colon	0.5085	0.6311
Intestine	0.3052	0.3528
Left Adrenal	0.2262	0.2871
Right Adrenal	0.263	0.3391
Segmentation Overall Average	0.5188	0.5951

Table 5.7. *Performance results*

Original Grounding DINO + SAM2	IoU	Precision
Liver	0.1059	0.1078
Spleen	0.0289	0.0291
Left Kidney	0.11306	0.1171
Right Kidney	0.2639	0.2709
Stomach	0.0225	0.0250
Gallbladder	0.0085	0.0085
Esophagus	0.00161	0.0016
Pancreas	0.0177	0.0177
Duodenum	0.00262	0.0026
Colon	0.0230	0.0233
Intestine	0.02708	0.0271
Left Adrenal	0.00088	0.0009
Right Adrenal	0.0018	0.0018
Segmentation Overall Average	0.0487	0.0498

Table 5.8. *Performance results*

Original Grounding DINO + Med-SAM2	IoU	Precision
Liver	0.0612	0.0683
Spleen	0.021	0.0223
Left Kidney	0.1054	0.1207
Right Kidney	0.2504	0.2724
Stomach	0.0204	0.02414
Gallbladder	0.0042	0.0042
Esophagus	0.00012	0.00012
Pancreas	0.0085	0.0087
Duodenum	0.00055	0.00056
Colon	0.028	0.0298
Intestine	0.01681	0.01773
Left Adrenal	0.000409	0.000423
Right Adrenal	0.00123	0.00124
Segmentation Overall Average	0.0407	0.0445

Table 5.9. *Performance results*

Pipeline	IoU	Precision
Fine-tuned Grounding DINO (Epoch 11) + SAM2	0.5230	0.6010
Fine-tuned Grounding DINO (Epoch 11) + Med-SAM2	0.5133	0.6157
Original Grounding DINO + SAM2	0.0487	0.0498
Original Grounding DINO + Med-SAM2	0.0407	0.0445

Table 5.10. *Results based on model combination*

Prompt information	IoU	Precision
Position	0.4232	0.4785
Position + shape	0.4993	0.5726
Position + shape + brightness	0.5230	0.6010

Table 5.11. *Results based on prompt information for the combination of the Fine-tuned Grounding DINO (Epoch 11) with SAM2*

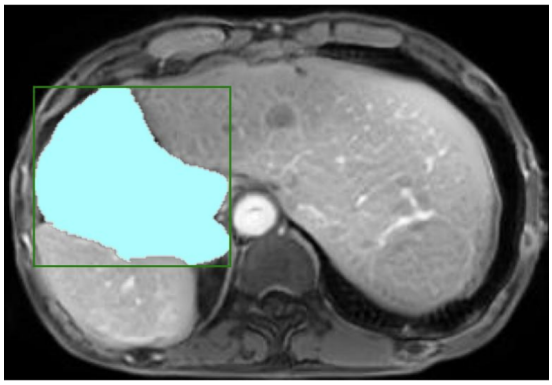


Figure 5.9. *Ground truth bounding box and mask [5]*

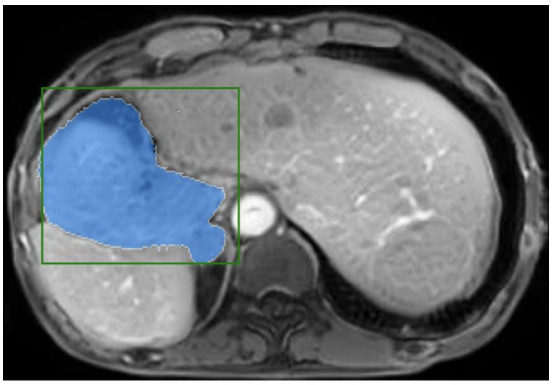


Figure 5.10. *Predicted bounding box and mask. IoU = 0.85 Precision = 0.997 [5]*

Figure 5.11. *Example of stomach grounded segmentation with input: "In the image, the Stomach has a slightly oval shape. Stomach on upper middle and on left part . Stomach Light Gray ."*

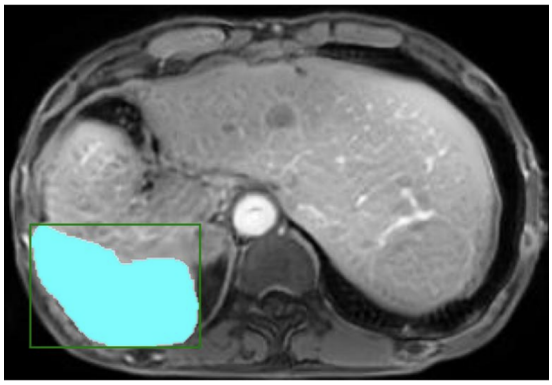


Figure 5.12. *Ground truth bounding box and mask [5]*

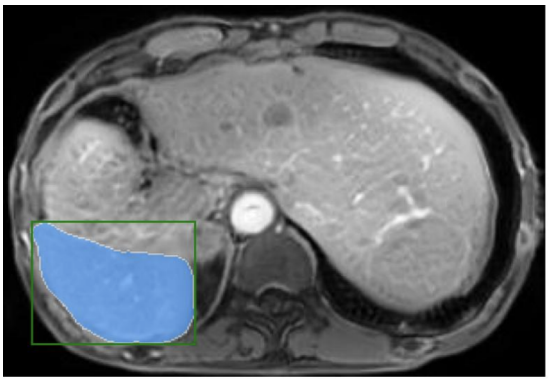


Figure 5.13. *Predicted bounding box and mask. IoU = 0.927 Precision = 0.969 [5]*

Figure 5.14. *Example of spleen grounded segmentation with input: "In the image, the Spleen has a shape that resembles a crescent. Spleen on lower part and on left part . Spleen Very Light Gray ."*

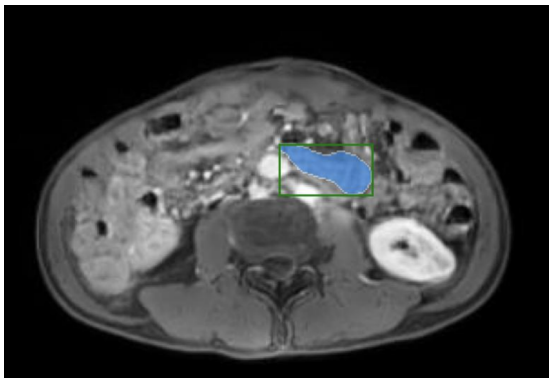


Figure 5.15. *Ground truth bounding box and mask [5]*

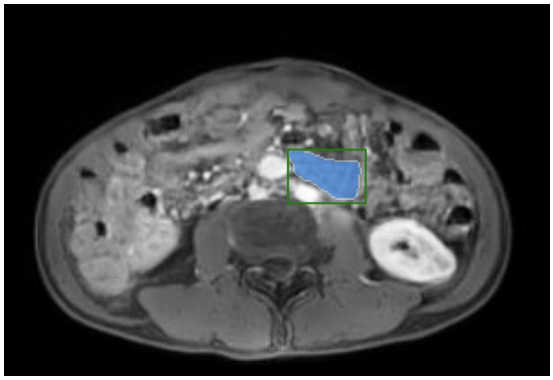


Figure 5.16. *Predicted bounding box and mask. IoU = 0.595 Precision = 0.784 [5]*

Figure 5.17. *Example of duodenum grounded segmentation with input: "In the image, the Duodenum has a C-shaped appearance. Duodenum on upper middle and on right middle . Duodenum Light Gray ."*

5.3 Evaluation on CT Scans

The developed pipeline was also evaluated on computed tomography scans. Following the same steps by first producing the textual descriptions, then training GroundingDINO we proceed on asking the model to predict the masks for the organs of the abdominal area of the same 67 patients as in the previous section.

Fine-tuned Grounding DINO (Epoch 11) + SAM2	IoU	Precision
Liver	0.7844	0.8200
Spleen	0.8311	0.8829
Left Kidney	0.8777	0.9288
Right Kidney	0.8689	0.9250
Stomach	0.7219	0.8462
Gallbladder	0.6155	0.7125
Esophagus	0.6553	0.8081
Pancreas	0.5152	0.5899
Duodenum	0.4490	0.5535
Colon	0.5822	0.7347
Intestine	0.3308	0.3770
Left Adrenal	0.5283	0.6468
Right Adrenal	0.4800	0.5863
Rectum	0.7242	0.8564
Bladder	0.8437	0.9023
Left Head of the Femur	0.7773	0.9074
Right Head of the Femur	0.7586	0.9000
Prostate	0.6632	0.7589
Seminal Vescicle	0.6205	0.7329
Segmentation Overall Average	0.5779	0.6702

Table 5.12. *Performance Results*

Fine-tuned Grounding DINO (Epoch 11) + Med-SAM2	IoU	Precision
Liver	0.7879	0.8395
Spleen	0.8435	0.9160
Left Kidney	0.8773	0.9448
Right Kidney	0.8563	0.9257
Stomach	0.7288	0.8368
Gallbladder	0.6108	0.6874
Esophagus	0.6800	0.7888
Pancreas	0.5056	0.5790
Duodenum	0.4601	0.5525
Colon	0.6030	0.6937
Intestine	0.3338	0.3717
Left Adrenal	0.4803	0.5725
Right Adrenal	0.4929	0.5709
Rectum	0.7418	0.8546
Bladder	0.8113	0.8952
Left Head of the Femur	0.6695	0.9011
Right Head of the Femur	0.6967	0.8859
Prostate	0.6445	0.7631
Seminal Vescicle	0.6246	0.7240
Segmentation Overall Average	0.5783	0.6588

Table 5.13. *Performance Results*

Fine-tuned Grounding DINO (Epoch 11) + SAM2 (Position only)	IoU	Precision
Liver	0.7358	0.7682
Spleen	0.8230	0.8729
Left Kidney	0.8698	0.9223
Right Kidney	0.8629	0.9194
Stomach	0.6222	0.7217
Gallbladder	0.5396	0.6216
Esophagus	0.6500	0.7945
Pancreas	0.4272	0.4870
Duodenum	0.2862	0.3516
Colon	0.3965	0.4958
Intestine	0.1449	0.1591
Left Adrenal	0.5025	0.6145
Right Adrenal	0.4445	0.5456
Rectum	0.7143	0.8545
Bladder	0.8160	0.8684
Left Head of the Femur	0.7210	0.8487
Right Head of the Femur	0.7292	0.8670
Prostate	0.5699	0.6492
Seminal Vescicle	0.5197	0.5937
Segmentation Overall Average	0.4592	0.5262

Table 5.14. *Performance Results*

Fine-tuned Grounding DINO (Epoch 11) + SAM2 (Position and shape)	IoU	Precision
Liver	0.7661	0.7994
Spleen	0.8295	0.8814
Left Kidney	0.8775	0.9286
Right Kidney	0.8687	0.9248
Stomach	0.7137	0.8386
Gallbladder	0.6166	0.7148
Esophagus	0.6566	0.8055
Pancreas	0.4937	0.5646
Duodenum	0.4272	0.5254
Colon	0.5584	0.7064
Intestine	0.2892	0.3274
Left Adrenal	0.5267	0.6476
Right Adrenal	0.4739	0.5780
Rectum	0.7237	0.8530
Bladder	0.8417	0.9003
Left Head of the Femur	0.7164	0.8457
Right Head of the Femur	0.7176	0.8570
Prostate	0.6508	0.7441
Seminal Vescicle	0.5970	0.7023
Segmentation Overall Average	0.5540	0.6422

Table 5.15. *Performance Results*

Fine-tuned Grounding DINO (Epoch 11) + SAM2 (Unfiltered prompts natural language shortened version with 2 sentences)	IoU	Precision
Liver	0.7272	0.7585
Spleen	0.8280	0.8787
Left Kidney	0.8785	0.9300
Right Kidney	0.8687	0.9246
Stomach	0.6909	0.8130
Gallbladder	0.6049	0.7077
Esophagus	0.6565	0.8047
Pancreas	0.4506	0.5212
Duodenum	0.3806	0.4719
Colon	0.4993	0.6360
Intestine	0.2564	0.2956
Left Adrenal	0.5239	0.6512
Right Adrenal	0.4765	0.5859
Rectum	0.7250	0.8639
Bladder	0.8159	0.8725
Left Head of the Femur	0.7102	0.8360
Right Head of the Femur	0.7117	0.8514
Prostate	0.6481	0.7375
Seminal Vescicle	0.5236	0.6220
Segmentation Overall Average	0.5231	0.6088

Table 5.16. *Example of input: "In the image, the Liver has a slightly oval shape. The Liver is located on the lower middle and on the right part of the image and appears to be Light Gray."*

Fine-tuned Grounding DINO (Epoch 11) + SAM2 (Unfiltered prompts natural language with 3 sentences)	IoU	Precision
Liver	0.7754	0.8111
Spleen	0.8279	0.8789
Left Kidney	0.8781	0.9297
Right Kidney	0.8683	0.9242
Stomach	0.7117	0.8364
Gallbladder	0.6090	0.7156
Esophagus	0.6504	0.8054
Pancreas	0.4918	0.5695
Duodenum	0.3982	0.4941
Colon	0.5427	0.6858
Intestine	0.2702	0.3092
Left Adrenal	0.5287	0.6541
Right Adrenal	0.4851	0.5996
Rectum	0.7223	0.8650
Bladder	0.8304	0.8876
Left Head of the Femur	0.7159	0.8393
Right Head of the Femur	0.7368	0.8768
Prostate	0.6615	0.7587
Seminal Vescicle	0.5489	0.6511
Segmentation Overall Average	0.5443	0.6326

Table 5.17. *Example of input: "In the image, the Liver has a slightly oval shape. The Liver is located on the lower middle and on the right part of the image. The Liver in the image appears to be Light Gray."*

Fine-tuned Grounding DINO (Epoch 7) + SAM2	IoU	Precision
Liver	0.7859	0.8223
Spleen	0.8323	0.8849
Left Kidney	0.8764	0.9285
Right Kidney	0.8675	0.9251
Stomach	0.7187	0.8392
Gallbladder	0.6019	0.7010
Esophagus	0.6523	0.8040
Pancreas	0.5080	0.5837
Duodenum	0.4260	0.5174
Colon	0.5810	0.7358
Intestine	0.3264	0.3731
Left Adrenal	0.5305	0.6542
Right Adrenal	0.4763	0.5861
Rectum	0.7111	0.8459
Bladder	0.8463	0.9062
Left Head of the Femur	0.7717	0.9033
Right Head of the Femur	0.7581	0.8995
Prostate	0.6520	0.7521
Seminal Vescicle	0.6196	0.7277
Segmentation Overall Average	0.5741	0.6668

Table 5.18. *Performance Results*

Original Grounding DINO + SAM2	IoU	Precision
Liver	0.0906	0.0910
Spleen	0.0385	0.0385
Left Kidney	0.0433	0.0441
Right Kidney	0.0425	0.0430
Stomach	0.0321	0.0431
Gallbladder	0.0066	0.0066
Esophagus	0.0018	0.0018
Pancreas	0.0184	0.0184
Duodenum	0.0038	0.0044
Colon	0.0113	0.0131
Intestine	0.0259	0.0260
Left Adrenal	0.0014	0.0015
Right Adrenal	0.0015	0.0015
Rectum	0.0130	0.0130
Bladder	0.0717	0.0718
Left Head of the Femur	0.0299	0.0318
Right Head of the Femur	0.1072	0.1155
Prostate	0.0201	0.0201
Seminal Vescicle	0.0051	0.0051
Segmentation Overall Average	0.0293	0.0306

Table 5.19. *Performance Results*

Original Grounding DINO + Med-SAM2	IoU	Precision
Liver	0.1060	0.1098
Spleen	0.0422	0.0422
Left Kidney	0.0427	0.0442
Right Kidney	0.0326	0.0333
Stomach	0.0346	0.0433
Gallbladder	0.0077	0.0077
Esophagus	0.0011	0.0011
Pancreas	0.0205	0.0206
Duodenum	0.0038	0.0043
Colon	0.0102	0.0117
Intestine	0.0266	0.0268
Left Adrenal	0.0012	0.0012
Right Adrenal	0.0015	0.0015
Rectum	0.0123	0.0124
Bladder	0.0776	0.0777
Left Head of the Femur	0.0253	0.0295
Right Head of the Femur	0.0883	0.1053
Prostate	0.0215	0.0215
Seminal Vescicle	0.0062	0.0062
Segmentation Overall Average	0.02949	0.03134

Table 5.20. *Performance Results*

Pipeline	IoU	Precision
Fine-tuned Grounding DINO (Epoch 11) + SAM2	0.5779	0.6702
Fine-tuned Grounding DINO (Epoch 11) + Med-SAM2	0.5783	0.6588
Original Grounding DINO + SAM2	0.0293	0.0306
Original Grounding DINO + Med-SAM2	0.02949	0.03134

Table 5.21. *Results based on model combination*

Prompt information	IoU	Precision
Position	0.4592	0.5262
Position + shape	0.5540	0.6422
Position + shape + brightness	0.5779	0.6702

Table 5.22. *Results based on prompt information for the combination of the Fine-tuned Grounding DINO (Epoch 11) with SAM2*

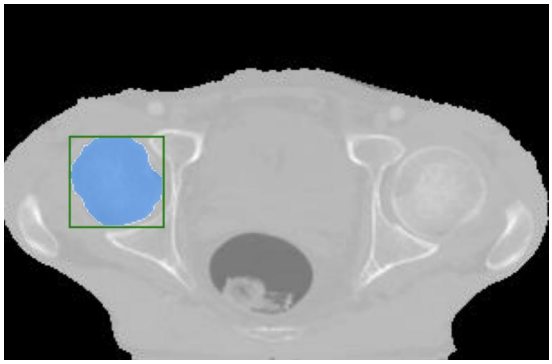


Figure 5.18. *Ground truth bounding box and mask [5]*

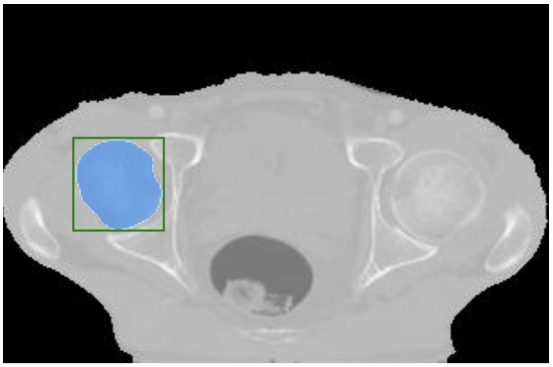


Figure 5.19. *Predicted bounding box and mask. IoU = 0.91 Precision = 0.98 [5]*

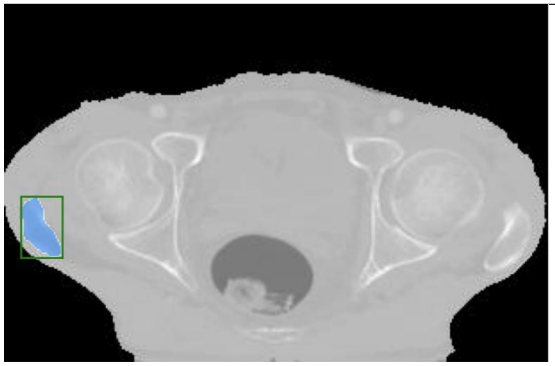


Figure 5.20. *Ground truth bounding box and mask [5]*

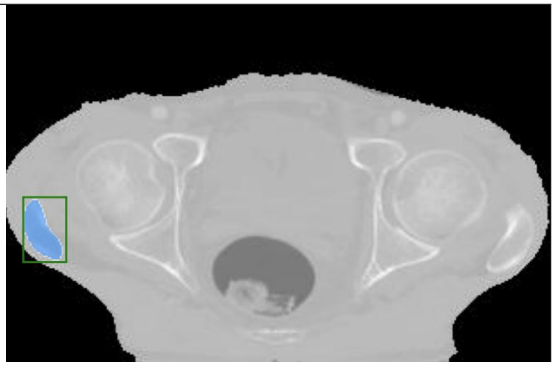


Figure 5.21. *Predicted bounding box and mask. IoU = 0.89 Precision = 0.95 [5]*

Figure 5.22. *Example of left head of the femur grounded segmentation with input: "In the image, the Left Head of the Femur has a rounded shape. Part Left Head Femur on lower middle and on **left part** Left Head Femur Very Light Gray ." for the first pair and "In the image, the Left Head of the Femur has a rounded shape. Part Left Head Femur on lower middle and on **leftmost part** Left Head Femur Very Light Gray ." for the second pair*

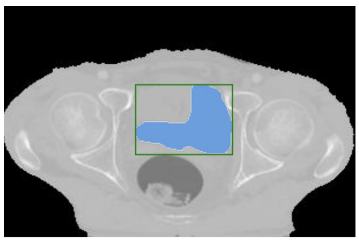


Figure 5.23. *Ground truth bounding box and mask [5]*

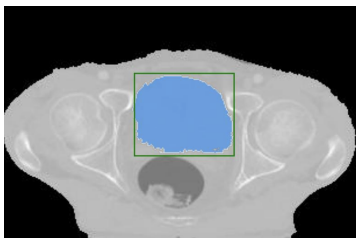


Figure 5.24. *Predicted bounding box and mask. IoU = 0.57 Precision = 0.58 [5]*

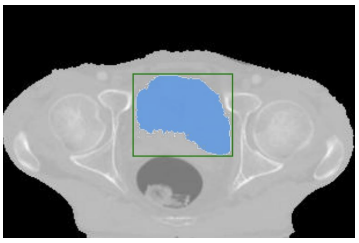


Figure 5.25. *Predicted bounding box and mask. IoU = 0.54 Precision = 0.46 [5]*

Figure 5.26. *Example of bladder grounded segmentation with SAM2 on the left and MedSAM2 on the right*

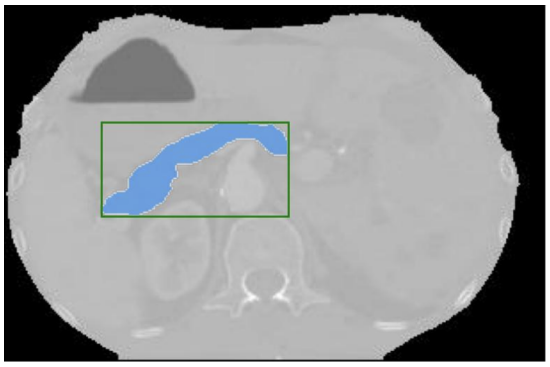


Figure 5.27. *Ground truth bounding box and mask [5]*

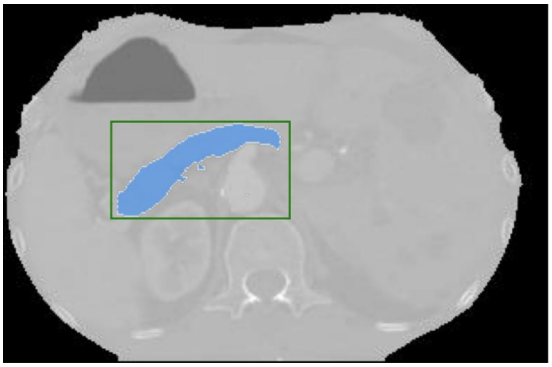


Figure 5.28. *Predicted bounding box and mask. IoU = 0.73 Precision = 0.89 [5]*

Figure 5.29. *Example of pancreas grounded segmentation with input: "In the image, the Pancreas has a slightly irregular shape. Pancreas on upper middle and on left middle . Pancreas Very Light Gray."*

5.4 Evaluation on out-of-distribution Synthetic MRI types

Up to this point, the model has been evaluated on datasets it was trained on. In this experiment, we assess the performance of the pipeline on out-of-distribution datasets. On the new Synthetic MRI images, the same patients are depicted with significant variations in organ brightness compared to the "Delayed" type. This means that the image depictions differ and that is also reflected on textual prompts regarding the brightness. Under these unseen circumstances we test whether the model understands the change and adapts to it.

Pre Artery type of Synthetic MRI	IoU	Precision
Liver	0.7166	0.7903
Spleen	0.6833	0.7694
Left Kidney	0.6740	0.7274
Right Kidney	0.5578	0.6058
Stomach	0.5818	0.6977
Gallbladder	0.3575	0.4267
Esophagus	0.4236	0.5426
Pancreas	0.1910	0.2356
Duodenum	0.2007	0.2658
Colon	0.4410	0.5885
Intestine	0.2908	0.3617
Left Adrenal	0.0197	0.0284
Right Adrenal	0.0996	0.1343
Segmentation Overall Average	0.4352	0.5235

Table 5.23. *Fine-tuned Grounding DINO (Epoch 11) + SAM2*

PV type of Synthetic MRI	IoU	Precision
Liver	0.7402	0.8076
Spleen	0.7547	0.8254
Left Kidney	0.8173	0.8739
Right Kidney	0.8172	0.8729
Stomach	0.6395	0.7526
Gallbladder	0.4103	0.4979
Esophagus	0.4859	0.6572
Pancreas	0.3341	0.4071
Duodenum	0.3439	0.4404
Colon	0.4703	0.6039
Intestine	0.2985	0.3569
Left Adrenal	0.2556	0.3367
Right Adrenal	0.2742	0.3706
Segmentation Overall Average	0.5071	0.5972

Table 5.24. *Fine-tuned Grounding DINO (Epoch 11) + SAM2*

Type of Synthetic MRI	IoU	Precision
Delay	0.5230	0.6010
Pre Artery	0.4352	0.5235
PV	0.5071	0.5972

Table 5.25. *Results across different types of Synthetic MRI*

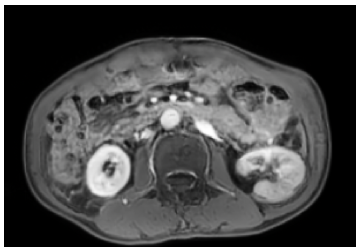


Figure 5.30. *Delay [5]*

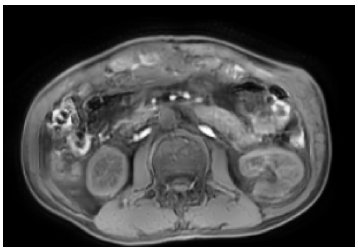


Figure 5.31. *Pre Artery [5]*

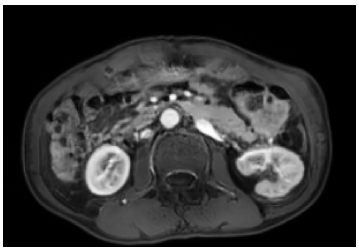
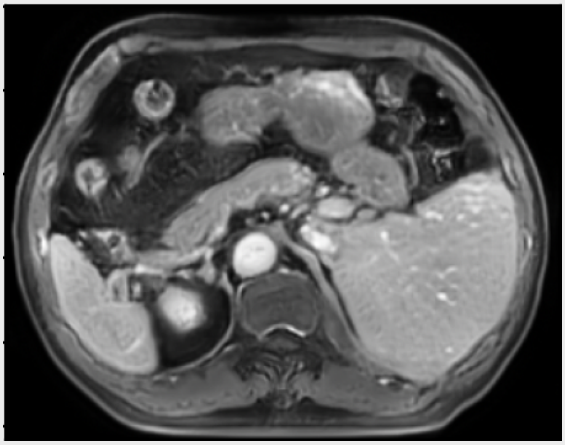


Figure 5.32. *PV [5]*

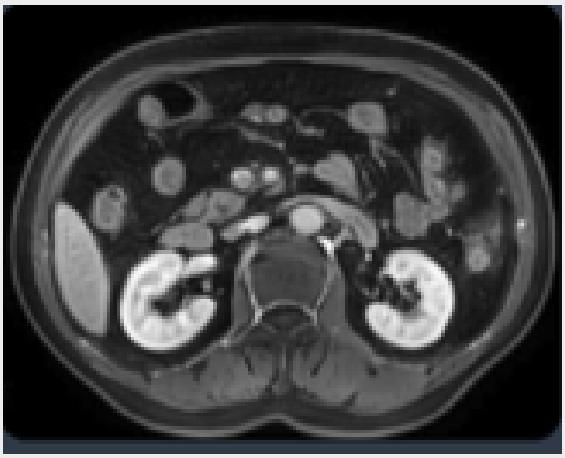
Figure 5.33. *Examples of different synthetic MRI types*



User: Describe the organs in this image in detail in regards to their position and shape.
LLaVA-Med: In this image, the organs are positioned in a way that is typical for a female. The liver is located on the right side and the spleen is on the left side. The liver appears to be enlarged, with a shape that resembles a "bunch of grapes." The spleen is also enlarged, with a shape that resembles a “sausage."

User: The visible organs in this image are the liver, spleen, duodenum, colon, stomach, pancreas, left and right adrenal. Describe them in detail in regards to their position and shape.
LLaVA-Med: In the image, the liver is located in the right upper quadrant and it appears to be enlarged. The spleen is situated in the left upper quadrant and it also appears to be enlarged. The duodenum is the first part of the small intestine and is located in the upper right quadrant. The colon is the large intestine and is situated in the lower right quadrant. The stomach is located in the upper left quadrant and the pancreas is in the upper left quadrant, just below the stomach. The left and right adrenal glands are small, triangular-shaped glands located on top of each kidney.

Figure 5.1. *LLaVA-Med misses some organs and does not return their shapes in the image. It may get confused when asked about multiple organs simultaneously*



User: In this image the organs are the two kidneys, the liver, the colon, the intestine, the pancreas and the duodenum. Describe them as they appear in the image.
LLaVA-Med: In the image, the two kidneys are visible as well as the liver, colon, intestine, pancreas and the duodenum. The kidneys are shown in their normal positions, with the liver located in the upper right part of the abdomen, the colon and intestine in the lower right part, the pancreas in the upper left part and the duodenum in the upper middle part of the abdomen. The organs appear to be in their usual anatomical positions and there are no visual abnormalities or pathologies mentioned in the context.

User: Describe the shape of the kidneys.
LLaVA-Med: The kidneys in the image are described as having a normal shape. They are bean-shaped organs with the right kidney being slightly higher than the left one. The shape of the kidneys is important for their proper functioning as it allows them to efficiently filter waste products and excess water from the blood.

User: Where is the liver here?
LLaVA-Med: In the image the liver is located in the upper right part of the abdomen.

User: Here the spinal cord is located in the lower part of the image in the middle. With that in mind, where is the liver?
LLaVA-Med: In the image the liver is located in the upper right part of the abdomen.

Figure 5.2. *The model focuses on the learned general anatomical knowledge rather than reasoning directly from the image*

User: The visible organs in this image are the liver, spleen, duodenum, colon, stomach, pancreas, left and right adrenal. Describe them in detail in regards to their position and shape.

LLaVA-Med: In the image, the liver is located in the right upper quadrant of the abdomen. It appears to be enlarged and has a nodular contour. The spleen is situated in the left upper quadrant and it also appears to be enlarged. The duodenum, which is the first part of the small intestine, is visible in the image. The colon is located in the lower right quadrant and it appears to be dilated. The stomach is situated in the upper left quadrant and it appears to be enlarged. The pancreas is visible in the image and it appears to be enlarged as well. The left and right adrenal glands are also visible in the image and they appear to be enlarged.

Figure 5.3. *Rotated version of the image to assess whether the positional descriptions improve.*

User: The visible organs in this image are the stomach, the liver, the spleen, the left kidney, the right kidney, the colon, the intestine, the right adrenal, the pancreas and the duodenum. Describe them in detail in regards to their position and say in which quadrant they belong or center of the image.

LLaVA-Med: In this image, the stomach is located in the left upper quadrant, the liver is in the right upper quadrant, the spleen is in the left upper quadrant, the left kidney is in the left lower quadrant, the right kidney is in the right lower quadrant, the colon is in the left lower quadrant, the intestine is in the right lower quadrant, the right adrenal is in the right lower quadrant, the pancreas is in the left lower quadrant and the duodenum is in the right upper quadrant.

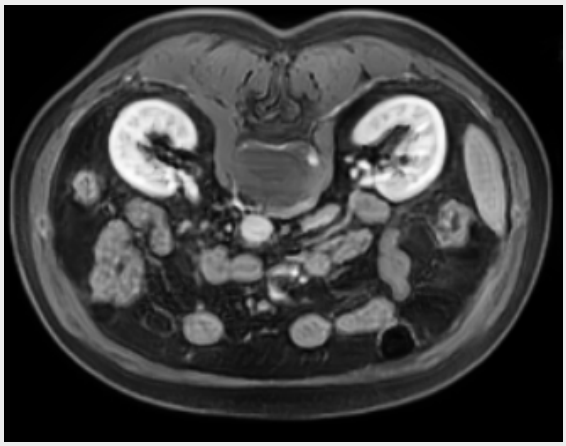
User: The visible organs in this image are the stomach, the liver, the spleen, the left kidney, the right kidney, the colon, the intestine, the right adrenal, the pancreas and the duodenum. Describe their shape in the image in detail.

LLaVA-Med: In this image, the stomach appears to be oval-shaped, the liver is oval-shaped, the spleen is oval-shaped, the left kidney is bean-shaped, the right kidney is bean-shaped, the colon is oval-shaped, the intestine is oval-shaped, the right adrenal is oval-shaped, the pancreas is oval-shaped and the duodenum is oval-shaped.

User: The visible organs in this image are the stomach, the liver, the spleen, the left kidney, the right kidney, the colon, the intestine, the right adrenal, the pancreas and the duodenum. Tell me about the brightness of each organ if it is white, gray or dark closer to black.

LLaVA-Med: In this image, the stomach, the liver, the spleen, the left kidney, the right kidney, the colon, the intestine, the right adrenal, the pancreas and the duodenum appear to be white. This means that they are well-visualized and easily distinguishable from the surrounding tissues.

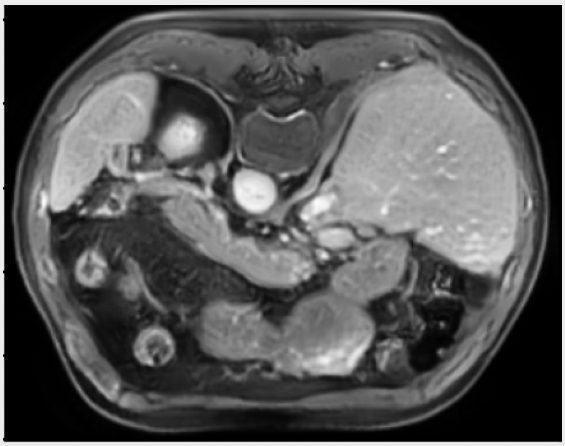
Figure 5.4. *The tasks were divided into separate propmts, requesting the position response to be in terms of quadrants. This was followed by inquiries about their shape and, finally brightness in a multiple-choice question form*



User: The Liver is located slightly up and on the right part of the image and appears to be Light Gray. Describe its shape in detail. If the liver appears to be elongated and small that means it has the shape of a sausage but if it is large it has its normal shape. What shape does it have here?
LLaVA-Med: In this image, the liver has a large shape, which is its normal shape.
User: Can you locate the liver in this image?
LLaVA-Med: Yes, the liver is located slightly up and on the right part of the image.
User: But it has the shape of a sausage because it is small and elongated.
LLaVA-Med: No, in this image, the liver has its normal shape, which is large.
User: The liver lies on the right of the right kidney. What shape does it have?
LLaVA-Med: The liver has a large shape, which is its normal shape.
User: No, it has the shape of a sausage.
LLaVA-Med: No, in this image, the liver has its normal large shape.
User: The liver here appears to cover a small area in regards to the area of the image, so it must be small and a small liver typically has the shape of a sausage. So what shape does it have here?
LLaVA-Med: In this image, the liver has the shape of a sausage because it is small.

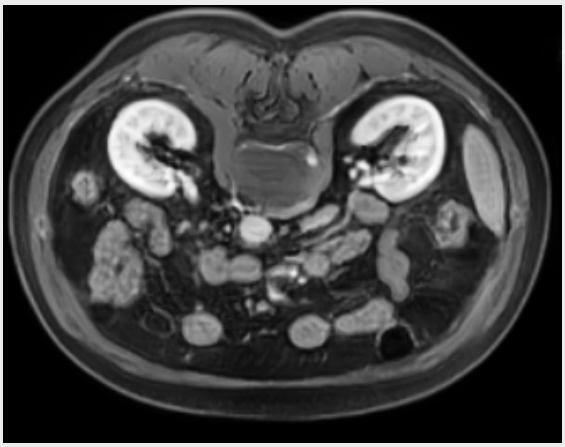
Another example for the same image
User: The Liver is located slightly up and on the right part of the image and appears to be Light Gray. If the liver appears to cover a small area in regards to the area of the image it must be small and a small liver typically has the shape of a sausage. Otherwise it has an irregular shape and is wedge-shaped. So what shape does it have here?
LLaVA-Med: In this image, the liver appears to be wedge-shaped.

Figure 5.5. *Experiments with multiple choice questions and reasoning based propmts*



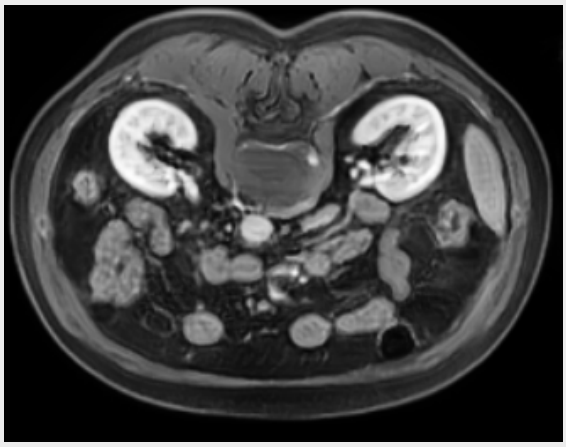
User: The Spleen is located slightly up and on the left part of the image. Does it have a homogenous or heterogenous texture? How did you come to the conclusion?
LLaVA-Med: The spleen appears to have a homogenous texture in the MRI image. This means that the spleen's appearance is uniform and consistent throughout the image.

Figure 5.6. *Example of prompt and response regarding texture*



User: The Liver is located slightly up and on the right part of the image and appears to be Light Gray. If the liver appears to cover a small area in regards to the area of the image it must be small and a small liver typically has the shape of a sausage. Otherwise it has an irregular shape and is wedge-shaped. So what shape does it have here?
LLaVA-Med: In this image, the liver appears to be wedge-shaped.

Figure 5.7. *Example of reasoning-based prompt for the shape of the liver*



User: The texture of the organ has to do with the inconsistencies that appear in the image. For example if there are arteries visible or other artifacts and noise then it is inconsistent. If there are different parts of the same organ with different brightness then it is inconsistent. The Liver is located slightly up and on the right part of the image and appears to be Light Gray. In this image what is the texture of the Liver?
LLaVA-Med: The texture of the Liver in this image is described as Light Gray.
User: I mean is it consistent or does it have irregularities in the pattern of the liver?
LLaVA-Med: The texture of the liver in this image is described as Light Gray, which suggests that it is consistent in appearance. However, it is important to note that the actual appearance of the liver in the image may vary depending on the imaging technique used and the specific details of the image.
User: The imaging technique is the delayed MRI. what is the texture of the liver?
LLaVA-Med: The texture of the liver in this delayed MRI image is described as Light Gray. This consistent appearance suggests that the liver tissue is uniform in this particular image. However, it is important to remember that the actual appearance of the liver in the image may vary depending on the imaging technique used and the specific details of the image.

Figure 5.8. *The model does not accurately interpret the concept of texture and instead mistakenly associates it with brightness*

Chapter 6

Discussion

6.1 Comparison with other Methods

In this section, we compare the performance of our pipeline with that of other models. Primarily, we compare it with MedLAM [31], a foundation model that localizes the requested organ and in combination with SAM [7] or MedSAM [24] segments it. We also present the performance of manual prompting [31] with SAM and MedSAM and lastly, the performance of nnU-Net [21], a fully supervised deep learning-based method.

It is important to note that prior work is tested on the WORD [9] dataset, which contains CT scans from the abdominal area, while our work is evaluated on the RAOS dataset. These two datasets share some patient CT scans. In particular, the annotations in WORD regarding common patients needed to be extended to match the 19 classes of RAOS instead of 16.

Table 6.1. *Dice scores (%) for region segmentation. Prior work is evaluated on the WORD dataset. Our framework is evaluated on RAOS, a continuation of WORD and thus, "-" refers to values missing from prior work.*

Regions	MedLAM		Man. Prompt		Fully Supervised nnU-Net		Our pipeline	
	SAM	MedSAM	SAM	MedSAM	5-shot	full	SAM2	MedSAM2
Liver	66.0	23.8	84.2	46.6	94.3	96.3	85.2	85.7
Spleen	61.7	36.3	85.3	65.0	90.9	95.7	89.4	90.4
Kidney L	82.1	70.7	92.1	84.1	83.4	94.7	93.0	93.0
Kidney R	88.3	77.3	92.9	86.4	86.0	95.2	92.4	91.7
Stomach	44.6	37.2	77.1	80.3	83.2	93.1	80.8	81.3
Gallbladder	13.1	10.2	72.7	68.8	54.7	77.0	70.7	70.5
Esophagus	36.6	27.8	67.0	63.1	72.0	81.5	77.2	79.0
Pancreas	29.7	21.4	64.4	46.9	72.0	84.7	61.6	60.9
Duodenum	26.0	21.1	54.1	51.0	53.7	77.3	53.9	54.9
Colon	25.6	26.6	41.8	44.1	74.7	86.1	67.4	68.7
Intestine	37.5	34.1	61.4	52.5	77.6	88.0	41.2	41.7
Adrenal	3.3	10.0	17.4	26.5	58.3	72.8	63.5	62.0
Left Adrenal	–	–	–	–	–	–	65.6	61.7
Right Adrenal	–	–	–	–	–	–	60.9	62.4
Prostate	–	–	–	–	–	–	75.3	74.3
Seminal Vesicle	–	–	–	–	–	–	72.9	73.2
Rectum	50.1	46.0	75.5	80.0	70.2	80.3	82.2	83.6
Bladder	65.3	59.1	83.0	82.9	84.0	91.8	89.9	87.9
Head of Femur L	81.7	71.5	90.5	80.3	43.1	31.2	84.9	77.3
Head of Femur R	80.1	74.3	89.1	83.2	47.8	40.98	83.3	79.1
Average	49.5	40.5	71.8	65.1	71.6	80.4	66.1	66.3

From Table 6.1 it is evident that our pipeline has a higher average dice score than MedLAM with SAM and MedSAM and is more successful in segmentation. It even surpasses in some classes the segmentation when the ground truth bounding boxes are provided to SAM or MedSAM. However, the fully supervised model nnU-Net with self adapting capabilities surpasses our results in most classes.

From Table 6.2 it is evident that our framework establishes better bounding boxes. We compare against **MedLAM**, **DetCo** [33], an unsupervised contrastive learning method for object detection, **Mask R-CNN** [13], a fully-supervised detection and segmentation approach and **MIU-VL** [30], a visual-language foundation model for medical image understanding. This model is not fine-tuned on the WORD dataset, which showcases that further training could significantly improve the results. Our performance is higher in most classes, except for the intestine.

Table 6.2. *Wall Distance (WD) results for each region and method on the RAOS dataset for our work and WORD for prior work. The best performance for each region is noted in bold*

Regions	**MedLAM**	**DetCo**	**Mask R-CNN**	**Our pipeline**	**MIU-VL**
Liver	10.5	73.8	7.1	**4.5**	80.4
Spleen	5.7	37.9	6.0	**1.7**	142.5
Kidney L	5.0	39.4	5.7	**0.9**	164.3
Kidney R	3.6	61.4	4.1	**1.0**	166.1
Stomach	17.7	49.3	8.8	**4.6**	102.8
Gallbladder	16.9	71.4	5.1	**2.5**	153.2
Esophagus	6.8	47.3	4.1	**1.5**	153.8
Pancreas	12.2	42.0	7.0	**5.1**	135.2
Duodenum	12.7	56.6	8.1	**5.9**	139.8
Colon	15.6	49.0	12.3	**9.8**	83.3
Intestine	15.4	50.7	**11.3**	25.3	106.0
Adrenal	7.6	52.3	6.2	**2.1**	160.2
Rectum	8.6	54.7	5.8	**2.2**	203.1
Bladder	8.1	65.1	3.7	**1.9**	167.7
Head of Femur L	5.3	55.8	3.3	**4.2**	175.1
Head of Femur R	5.9	52.1	4.1	**4.0**	163.4
Average	9.9	53.9	6.5	4.7	143.6

6.2 Conclusion

From the analysis of the sets of prompts and responses generated by LLaVA-Med, it can be concluded that there are limitations in the reasoning capability of the model. Despite the use of various prompting methods, such as multiple choice questions and rule based inference, the model seemed to lack an understanding of the relationship between the textual descriptions and the image features. Thus, it cannot be entrusted with producing the entirety of the textual descriptions as initially theorized.

At this point, it should be noted that LLaVA-Med as an LLMs is not deterministic, meaning it does not produce identical textual output for the same prompt, introducing a degree of unpredictability. In the case of the shape descriptions, this non-determinism does not constitute a significant issue, as the generated description generally aligns with the expected shape of the organ and in an open ended question other similar words can be selected. However, in the case of the organ's texture, when the model was prompted to classify the texture as either homogeneous or heterogeneous, its responses were not always consistent for the same image and prompt. This could also be detrimental for the position and brightness of the organs where the question has a multiple choice form and only one answer is correct.

As demonstrated in table 6.2 the localization that our pipeline offers is efficient and was the only task for which GroundingDINO was trained. The consecutive part of the framework, which generates the mask, achieves competitive results. However, there

are cases in which further fine-tuning is necessary for the segmentation model, as the bounding box is ideal.

When comparing the original GroundingDINO with its fine-tuned version, it is evident that the original does not comprehend the concepts presented and needs further training on the biomedical field as it detects the whole body. The segmentor MedSAM2 does not constitute a significant improvement of SAM2 as they achieve comparable results.

Regarding the textual prompts of the pipeline, they need to be as detailed as possible including information about the relative position in the image, brightness and shape of the mask to increase the performance of the pipeline. The model can also detect organs and segment them successfully when the prompts are different from the training set as was the case for the unfiltered version of the textual descriptions.

Finally, the performance of the model on unseen datasets was evaluated in which only the brightness of the organs changed, introducing new visual features. The result is that the model performed well when the organs remained clearly visible. Variations in organ intensity were also incorporated in the textual descriptions and the model understood this change in many cases. For images in which organs were blurred or exhibited significant variation in the pixel intensity, the model's performance declined. Overall, images with clear depictions, where the displayed organ's appearance deviated slightly from the training set, retained the best scores.

The proposed pipeline demonstrated adaptability in a closed setting, as demonstrated by the multiple evaluations on the datasets it was trained on. By closed setting it is inferred that images stem from the same distribution, for example have been generated by the same imaging technique and machine. As a result, this tool could be utilized as an assisting tool for medical professionals within a single clinical setting, trained on internal data.

6.3 Limitations

A key factor that limits the scores across all metrics is the inability to prompt neighboring non-contiguous areas of the same organ, as they cannot be exclusively prompted because the position description will be the same. For instance, as illustrated in Figure 6.1, different parts of the colon may not be distinguishable based on their position as they occupy the same general area.

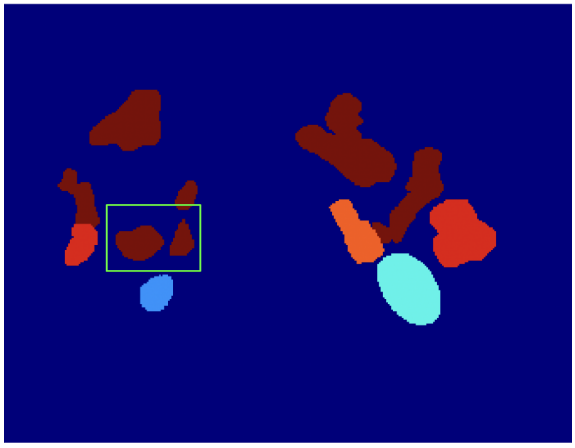


Figure 6.1. *Labelmap illustrating the challenge of distinguishing adjacent segments of the same organ based solely on position. The two parts inside the green bounding box will be prompted separately in the evaluation process as they are non-contiguous. However, if the same prompt is used for both, the model may generate identical outputs, affecting the final score.*

6.4 Future Work

In this work, the model was trained separately in two different settings (CT scans and synthetic MRI) and was evaluated in those particular closed settings. A possible future improvement would be to fine-tune the model on a large set of datasets spanning the whole medical imaging field covering multiple areas of the human body and containing images of different modalities (X-rays, multiple types of MRI, CT scans, Histology, Gross pathology). This would create a model with vast knowledge across the medical field that can be utilized in any setting as it is based on a large-scale dataset and will be generalizable.

Another direction for future development could be the continuous weight update of the LVLM model with the textual descriptions produced by the deterministic algorithms. This could enhance its reasoning capabilities and lead to more reliable answers.

Lastly, the application of the pipeline on 3D medical images could be explored by generating three dimensional bounding boxes and utilizing MedSAM2 to process and segment volumetric data. This could lead to improved results benefiting from the nature of the data, which is volumetric and the availability of all the slices in one prompt, leading to a more effective use of the memory mechanism of the segmentation model.

Bibliography

[1] Haotian Liu, Chunyuan Li, Qingyang Wu και Yong Jae Lee. *Visual Instruction Tuning*.

[2] Nikhila Ravi, Valentin Gabeur, Yuan Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao Yuan Wu, Ross Girshick, Piotr Dollár και Christoph Feichtenhofer. *SAM 2: Segment Anything in Images and Videos*.

[3] Jiayuan Zhu, Abdullah Hamdi, Yunli Qi, Yueming Jin και Junde Wu. *Medical SAM 2: Segment medical images as video via Segment Anything Model 2*. _eprint: 2408.00874.

[4] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu και others. *Grounding dino: Marrying dino with grounded pre-training for open-set object detection*.

[5] Xiangde Luo, Zihan Li, Shaoting Zhang, Wenjun Liao και Guotai Wang. *Rethinking Abdominal Organ Segmentation (RAOS) in the clinical scenario: A robustness evaluation benchmark with challenging cases*.

[6] Olaf Ronneberger, Philipp Fischer και Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*Nassir Navab, Joachim Hornegger, William M. Wells και Alejandro F. Frangi, επιμελητές, σελίδες 234–241. Springer International Publishing.

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan Yen Lo, Piotr Dollár και Ross Girshick. *Segment Anything*.

[8] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon και Jianfeng Gao. *Llava-med: Training a large language-and-vision assistant for biomedicine in one day*.

[9] Jianghong Xiao Jieneng Chen Tao Song Xiaofan Zhang Kang Li Dimitris N. Metaxas Guotai Wang Xiangde Luo, Wenjun Liao και Shaoting Zhang. *WORD: A large-scale dataset, benchmark and clinically applicable study for abdominal organ segmentation from CT image*. 82:102642. Publisher: Elsevier.

[10] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal

Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Eliz-

abeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk και Barret Zoph. *GPT-4 Technical Report*.

[11] Wei Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica και Eric P. Xing. *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality*.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit και Neil Houlsby. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*.

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár και Ross Girshick. *Masked Autoencoders Are Scalable Vision Learners*.

[14] Chaitanya Ryali, Yuan Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li και Christoph Feichtenhofer. *Hiera: A Hierarchical Vision Transformer without the Bells-and-Whistles*.

[15] Daniel Bolya, Chaitanya Ryali, Judy Hoffman και Christoph Feichtenhofer. *Window Attention is Bugged: How not to Interpolate Position Embeddings*.

[16] Junde Wu και Min Xu. *One-Prompt to Segment All Medical Images*. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, σελίδες 11302–11312.

[17] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin και Baining Guo. *Swin Transformer: Hierarchical Vision Transformer using Shifted Windows*.

[18] Jacob Devlin, Ming Wei Chang, Kenton Lee και Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*.

[19] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker και Daniel Rueckert. *Attention U-Net: Learning Where to Look for the Pancreas*.

[20] Xiaokun Liang, Na Li, Zhicheng Zhang, Jing Xiong, Shoujun Zhou και Yaoqin Xie. *Incorporating the hybrid deformable model for improving the performance of abdominal CT segmentation via multi-scale feature fusion network*. 73:102156.

[21] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen και Klaus H. Maier-Hein. *nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation*. 18(2):203–211. Publisher: Nature Publishing Group.

[22] Yucheng Tang, Dong Yang, Wenqi Li, Holger Roth, Bennett Landman, Daguang Xu, Vishwesh Nath και Ali Hatamizadeh. *Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis*.

[23] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger Roth και Daguang Xu. *UNETR: Transformers for 3D Medical Image Segmentation*.

[24] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You και Bo Wang. *Segment Anything in Medical Images*. 15(1):654.

[25] Ruining Deng, Can Cui, Quan Liu, Tianyuan Yao, Lucas W. Remedios, Shunxing Bao, Bennett A. Landman, Lee E. Wheless, Lori A. Coburn, Keith T. Wilson, Yaohong Wang, Shilin Zhao, Agnes B. Fogo, Haichun Yang, Yucheng Tang και Yuankai Huo. *Segment Anything Model (SAM) for Digital Pathology: Assess Zero-shot Segmentation on Whole Slide Imaging*.

[26] Maciej A. Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz και Yixin Zhang. *Segment Anything Model for Medical Image Analysis: an Experimental Study*. 89:102918.

[27] Junde Wu, Wei Ji, Yuanpei Liu, Huazhu Fu, Min Xu, Yanwu Xu και Yueming Jin. *Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation*.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger και Ilya Sutskever. *Learning Transferable Visual Models From Natural Language Supervision*.

[29] Jie Liu, Yixiao Zhang, Jie Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A. Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang και Zongwei Zhou. *CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection*. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, σελίδες 21095–21107.

[30] Ziyuan Qin, Huahui Yi, Qicheng Lao και Kang Li. *Medical Image Understanding with Pretrained Vision Language Models: A Comprehensive Study*.

[31] Wenhui Lei, Xu Wei, Xiaofan Zhang, Kang Li και Shaoting Zhang. *MedLSAM: Localize and Segment Anything Model for 3D CT Images*.

[32] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik και Kaiming He. *SlowFast Networks for Video Recognition*.

[33] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li και Ping Luo. *DetCo: Unsupervised Contrastive Learning for Object Detection*.

[34] Kaiming He, Georgia Gkioxari, Piotr Dollár και Ross Girshick. *Mask R-CNN*.

[35] Michael Baumgartner, Paul F. Jaeger, Fabian Isensee και Klaus H. Maier-Hein. *nnDetection: A Self-configuring Method for Medical Object Detection*. τόμος 12905, σελίδες 530–539.

[36] Ross Girshick. *Fast R-CNN*.

[37] Shaoqing Ren, Kaiming He, Ross Girshick και Jian Sun. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*.

[38] Paul F. Jaeger, Simon A. A. Kohl, Sebastian Bickelhaupt, Fabian Isensee, Tristan Anselm Kuder, Heinz Peter Schlemmer και Klaus H. Maier-Hein. *Retina U-Net: Embarrassingly Simple Exploitation of Segmentation Supervision for Medical Object Detection*.

[39] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang και Lei Zhang. *Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks*. _eprint: 2401.14159.

[40] *GPT-4o mini: advancing cost-efficient intelligence | OpenAI*.

[41] Wei Li Zuwei Long. *Open Grounding Dino:The third party implementation of the paper Grounding DINO*.

List of Abbreviations

AI	Artificial Intelligence
CNN	Convolutional Neural Network
CT	Computed Tomography
CV	Computer Vision
IoU	Intersection over Union
LLaVA	Large Language and Vision Assistant
LLM	Large Language Model
LM	Language Model
LMM	Large Multimodal Model
LVLM	Large Vision-Language Model
MAE	Masked Autoencoder
MRI	Magnetic Resonance Imaging
NLP	Natural Language Processing
SAM	Segment Anything Model
VLM	Vision-Language Model
VQA	Visual Question Answering