Lung Grounded-SAM (LuGSAM): A Novel Framework for Integrating Text prompts to Segment Anything Model (SAM) for Segmentation Tasks of ICU Chest X-Rays

Dhanush Babu Ramesh [1], Rishika Iytha Sridhar [2], Pulakesh Upadhyaya [2], and Rishikesan Kamaleswaran [2]

[1]Georgia Institute of Technology
[2]Affiliation not available

October 31, 2023

Abstract

Chest radiography is a commonly utilized imaging technique for acquiring Chest X-Ray (CXR) images due to its cost-effectiveness and its role in diagnosing lung?related disorders. Nevertheless, interpreting CXR images can be challenging, and the process of separating the lung field from CXR images can be a valuable tool for assessing and diagnosing lung diseases. While various segmentation methods exist, this study primarily focuses on META's latest Segment Anything Model (SAM). SAM is an Artificial Intelligence (AI) model designed to segment objects within an image. This research aims to harness SAM's capabilities for segmenting CXR images. Additionally, we explore the potential of another novel model called Grounding DINO. Grounding DINO is a zero-shot object detection model that utilizes a Swin (Shifted Windows) transformer for extracting image features and BERT (Bidirectional Encoder Representations from Transformers) for extracting textual information. It is primarily employed to detect objects in an image based on a provided text prompt, creating bounding boxes around the objects when certain text and box thresholds are met. These bounding boxes are then used as prompts for SAM to generate segmentation masks. The proposed framework has been assessed on CXRs obtained from patients at Emory Hospital in Atlanta, Georgia, USA and further evaluated using NIH clinical center's CXR image dataset.

Lung Grounded-SAM (LuGSAM): A Novel Framework for Integrating Text prompts to Segment Anything Model (SAM) for Segmentation Tasks of ICU Chest X-Rays

Dhanush Babu Ramesh, Rishika Iytha Sridhar, Pulakesh Upadhyaya and Rishikesan Kamaleswaran

***Abstract*— Chest radiography is a commonly utilized imaging technique for acquiring Chest X-Ray (CXR) images due to its cost-effectiveness and its role in diagnosing lung-related disorders. Nevertheless, interpreting CXR images can be challenging, and the process of separating the lung field from CXR images can be a valuable tool for assessing and diagnosing lung diseases. While various segmentation methods exist, this study primarily focuses on META's latest Segment Anything Model (SAM). SAM is an Artificial Intelligence (AI) model designed to segment objects within an image. This research aims to harness SAM's capabilities for segmenting CXR images. Additionally, we explore the potential of another novel model called Grounding DINO. Grounding DINO is a zero-shot object detection model that utilizes a Swin (Shifted Windows) transformer for extracting image features and BERT (Bidirectional Encoder Representations from Transformers) for extracting textual information. It is primarily employed to detect objects in an image based on a provided text prompt, creating bounding boxes around the objects when certain text and box thresholds are met. These bounding boxes are then used as prompts for SAM to generate segmentation masks. The proposed framework has been assessed on CXRs obtained from patients at Emory Hospital in Atlanta, Georgia, USA and further evaluated using NIH clinical center's CXR image dataset.**

***Index Terms*— Chest X-rays, Grounding DINO, Object Detection, Segment Anything Model, Lung Segmentation.**

I. Introduction

The development of foundational models has allowed for rapid advancements in medicine such as in medical imaging, drug discovery, and personalized medicine. While a number of machine learning (ML) and Deep Learning (DL) methods have been contributed for segmentation tasks, a general challenge is that such models are often trained for specific tasks. Training a model involves gathering a substantial amount of data and creating numerous ground-truth masks, leading to increase in number of annotation for the annotators [1]. An ambitious and unmet solution is to have an all-in-one, one-for-many models which can solve segmentation problems with no re-training or fine-tuning which can be refereed to as zero-shot learning models. In this direction, the Segment Anything Model (SAM) was introduced by Meta Research as a novel and state-of-the art foundational model for segmentation tasks [2] . This benchmark model was trained on >1 billion masks and >1 million natural images and has broad applicability to various image segmentation tasks [3]. In this study, we evaluate the performance of SAM using tailored anatomical text prompts on Chest X-ray (CXR) images. The objective of this work is to present the first comprehensive study in the assessment of SAM's abilities to perform segmentation on CXR images using text prompts.

Medical image segmentation plays a crucial role in identifying disorders and delineating structures of interest, making it a well-established technique [4]. X-ray imaging, specifically CXR, is a widely used imaging modality because of its simple procedure to obtain images from patients and in diagnosing lung-related disorders such as pneumonia, pulmonary edema and Acute Respiratory Distress Syndrome (ARDS) [5], [6]. In early stages of Coronavirus Disease 2019 (COVID-2019), the preliminary step was to obtain a Computed Tomography (CT) or CXR from the patient to visualize the spread of infection in the lung nodules [7] . However, due to substantial costs, as well as the concerns pertaining to potential airborne contamination and limited resource availability, the adoption of CT scans was often hindered in clinical settings [8] . Consequently, CXR emerged as a more commonly observed diagnostic modality. This preference for CXRs was further motivated by their advantages, including lower ionizing radiation exposure and portability [9].

Despite their advantages CXRs face criticism for their low diagnostic sensitivity, necessitating precise, time-consuming interpretation by a radiologist for compensation [10]. They lack the fine details about pulmonary nodules, interstitial

The authors were supported by the National Institutes of Health under Award Numbers R01GM139967 and UL1TR002378

Dhanush Babu Ramesh, MS in Biomedical Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA (e-mail: dbabu6@gatech.edu)

Rishika Iytha Sridhar, MS in Biomedical Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA (e-mail: rsridhar40@gatech.edu)

Pulakesh Upadhyaya, Postdoctoral Fellow, Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, Georgia, USA (e-mail: pulakesh.upadhyaya@emory.edu).

Rishikesan Kamaleswaran, Associate Professor, Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, and Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA (e-mail: rkamaleswaran@emory.edu).

patterns, and cavitations among others required for precise anatomical analysis which is crucial in many applications [11]. This has led to the emergence of Computer-Aided Diagnosis (CAD) in chest radiography, with the goal of improving diagnostic accuracy [12].

Significant advancements have been made in CAD research related to lung segmentation. Segmenting infected nodules of the lung can provide comprehensive information about the type, location, and characteristic of the disease, which helps clinicians to better understand the progression of diseases, aiding in treatment planning [13]. This has led to the development of ML and DL based segmentation pipelines that focus on specific diseases with improved accuracy. However, as methodologies mature, these ML algorithms need to be rendered robust to support transition from bench to bedside [14]. In such cases, foundational models like SAM prove to be an all-in-one automated solution that can provide accurate segmentation masks with just a click on the Region of Interest (RoI) without having to train on domain specific data [15].

In our previous work, we used SAM to perform segmentation on CXR images, and found that SAM performed considerably well on segmentation tasks in terms of stability score, when prompted with points or bounding boxes. Stability score is a metric that calculates the differences in segmentation output of multiple perturbed images, given an input image. The different types of segmentation like automatic, prompt-based, interface-based, and demographic based segmentation were explored and performed [16]. However, the text-prompt version of SAM has not been implemented yet and this work primarily focuses on providing customized anatomical text-prompts to SAM for segmentation of CXRs. Recent studies are unclear about the applicability of SAM model to medical image segmentation [17]. Hence, in this work, we not only use SAM for the interpretation of CXR images, but also take advantage of anatomical text-prompts to explore SAM's untapped potential in segmentation. Furthermore, text prompts can guide AI models like SAM to focus on the desired regions of interest which can result in efficient localization of objects. Also, using text-based cues can improve the model's understanding about the anatomical patterns and structures, thereby eliminating ambiguities when making segmentation decisions by providing relevant clinical information.

In this experiment, text prompts were provided to the Grounding DINO model to identify the object, and the detected bounding boxes were provided as prompts to SAM which produced the corresponding segmentation masks. The text prompts pertaining to the two lobes of the lungs were compared across the corresponding stability scores using histogram distribution plots. It was found that crafting a relevant prompt to the task yielded precise bounding boxes. It was also found that giving text prompts which were most relevant to the task at hand yielded precise bounding boxes.

The contribution of this paper is summarized as follows:

- A novel approach has been introduced in which text prompts are used to perform bounding box detection on ICU CXR images.
- The developed framework has been integrated into the SAM system, allowing lung segmentation on CXR images. The potential benefits of this development could aid in disease detection, planning of treatments, and other medical uses are significant.
- Additionally, CXR image binarization process has been performed, leading to enhanced object detection
- Performance comparison of customized anatomical text prompts and SAM versions.

II. Related Works

A. Image Labelling tasks

Image labelling is regarded as a critical problem in computer vision. One of the traditional labelling strategies is to combine a Large Language Model (LLM) with a Language Vision Model (LVM) so that they complement one another. Yu et al. [2023] employed ChatGPT as input to an Artificial Intelligence Generated Context (AIGC) model to produce images. It was then given to an LVM to generate labels, which Grounding DINO then turned into visual prompts. SAM was then used to segment the image based on the prompts provided [18].

B. SAM in medical imaging

SAM has been evaluated on a variety of medical imaging datasets. For instance, in a study by He et al [2023] SAM was evaluated on 12 available medical image segmentation datasets. The Dice overlap between the algorithm-segmented and ground-truth masks was used to determine accuracy. Furthermore, SAM was compared to five algorithms developed for medical image segmentation tasks. SAM's accuracy was assessed by segmentation ability score and Dice overlap in U-Net, image dimension, target region size, image modality, and contrast. It was found that SAM's accuracy was least affected by these factors [2]. In another approach Mazurowski et al comprehensively tested SAM's segmentation abilities across 19 medical imaging datasets by simulation with point and box prompts. SAM's performance varies depending on the dataset and task, excelling with clear prompts for well-defined objects but under-performing with ambiguity. Point prompts produce less effective outcomes than box prompts, and SAM performs better in single-point prompts. Although iterative use of multiple-point prompts marginally improves the performance of SAM, other techniques eventually outperforms SAM's point based segmentation [19]. However, despite the growing number of published papers on SAM's application in medical image segmentation, there is a noticeable gap in literature concerning the utilization of anatomical context text prompts specifically for segmenting CXR images. This particular approach, which incorporates anatomical context information, remains unexplored in relation to SAM's segmentation tasks.

III. Methodology

A. SAM Architecture

To summarize the architecture of SAM, it involves an image encoder, a prompt encoder which accepts sparse and dense

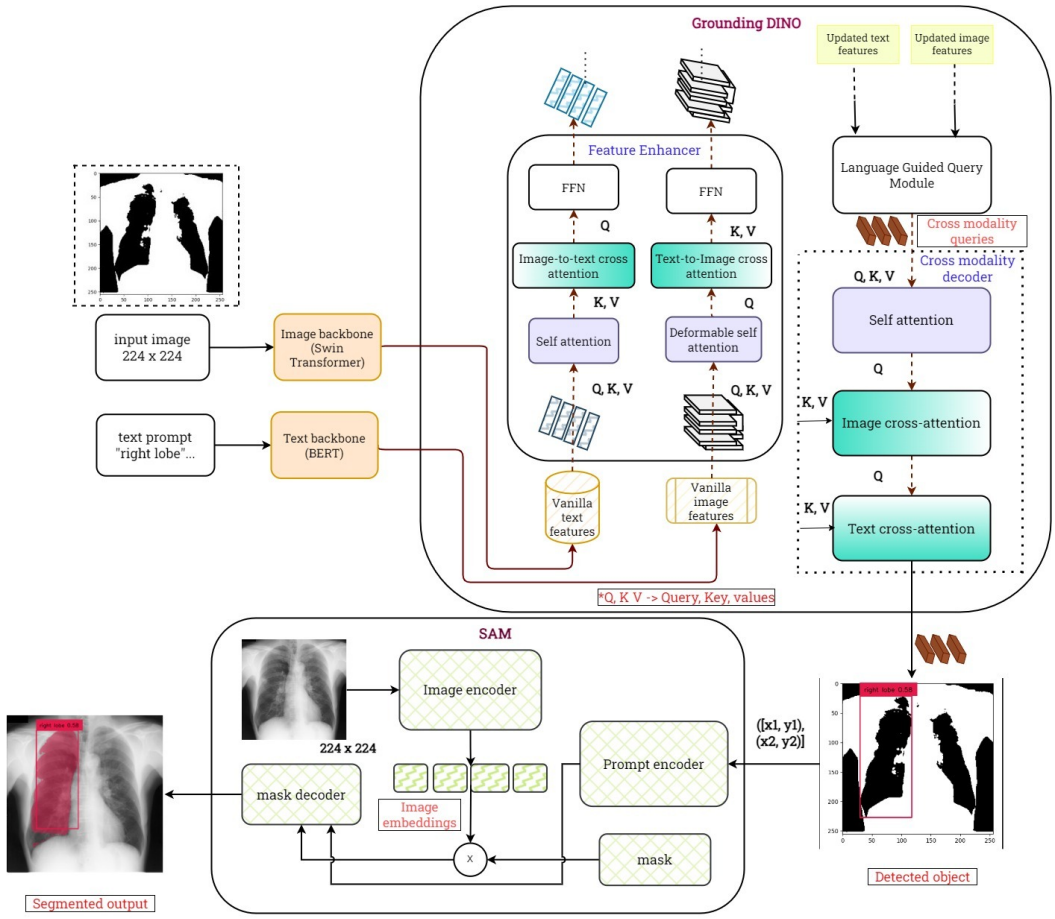


Fig. 1. Complete pipeline of Lung Grounded-SAM with the image and text backbone for extracting image embeddings and text tokens.

prompts, and a mask decoder. The image encoder component of SAM is a Masked Auto Encoder (MAE) and is designed to handle input upto a resolution of 3x1024x1024. The backbone of SAM is the Vision Transformer (Vit) which comes in three variants namely Vit-base (Vit-b), Vit-large (Vit-l), and Vit-huge (Vit-h) [3]. This backbone allows SAM to capture fine-grain details in the input.

The main difference in these variants is the number of parameters it was trained on [20]. Vit works by splitting an input image to a number of fixed-size non-overlapping patches which are projected linearly. The patches are referred to as tokens which are assigned to a class label for classification. Additionally, positional embedding is done to get information about the location of the patches in the image. These patches are then fed to a pure transformer which works on self-attention with which it can capture long-range dependencies and contextual information from the images [21].

Models like Vit have found their application in medical image segmentation obtained from Computed Tomography (CT), MRI's etc. [22]. It was found that adding attention to these networks eliminated irrelevant regions in an image while highlighting salient features about the task at hand [23]. Hence, attention mechanism plays a crucial role in foundational models like SAM. The model was trained using a combination of focal loss and dice loss as the loss functions. For optimization, the AdamW optimizer is employed with β_1 = 0.9 and β_2 = 0.999. Additionally, a linear learning rate warmup is applied for 250 iterations, followed by a step-wise learning rate decay schedule.

Furthermore, SAM is also associated with the zero-shot learning property which enables it to perform equally well on images that were not encountered during training. This eliminates time complexity and the need to re-train the model. Fine-tuning the model on custom datasets can improve accuracy for the specific task, enabling real-time deployment. Prompting these models with specific and relevant prompts can also improve model's performance.

B. Grounding DINO Architecture

Grounding DINO is a state-of-the-art zero-shot detector used in object detection. It accepts an (image, text) pair as input and outputs multiple object boxes. For example, if an input image has a scissor on a table; it locates the scissor and table and extracts the word "scissor" and "table" as labels.

It is built upon the DETR (Detection Transformers)-like model named DINO. This can identify and localize objects in an image when given a textual description. The model understands the language and visual content of the image and can associate the visual elements to a text or message. The model consists of a Swin transformer for extracting

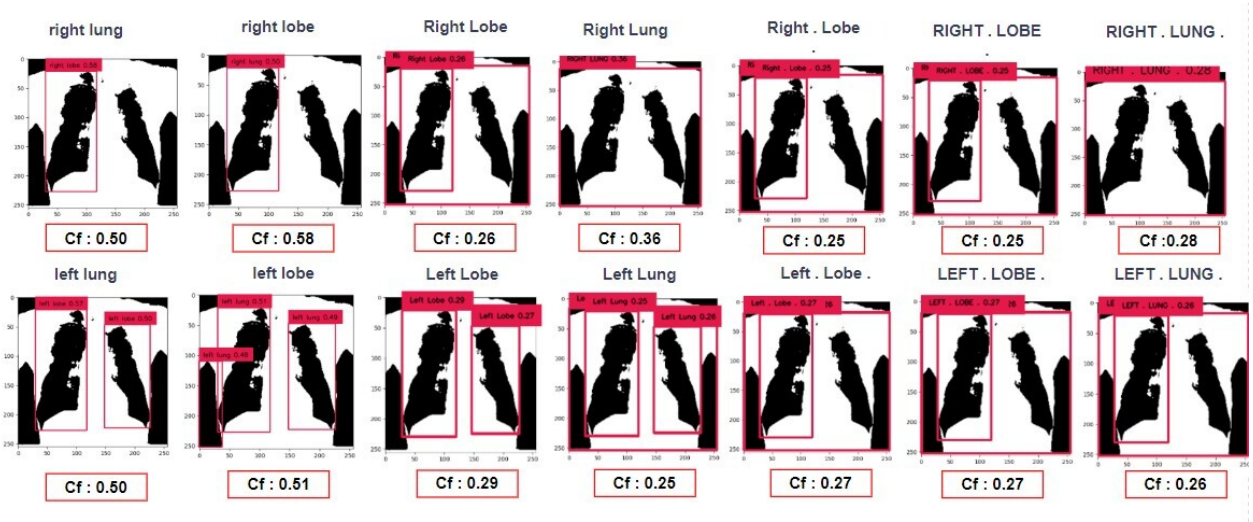


Fig. 2. Grounding DINO results when prompted with the anatomical text prompts. Maximum confidence scores and accurate object detection is associated with the prompts ”right lung”, ”right lobe”, ”left lung”, and ”left lobe”.

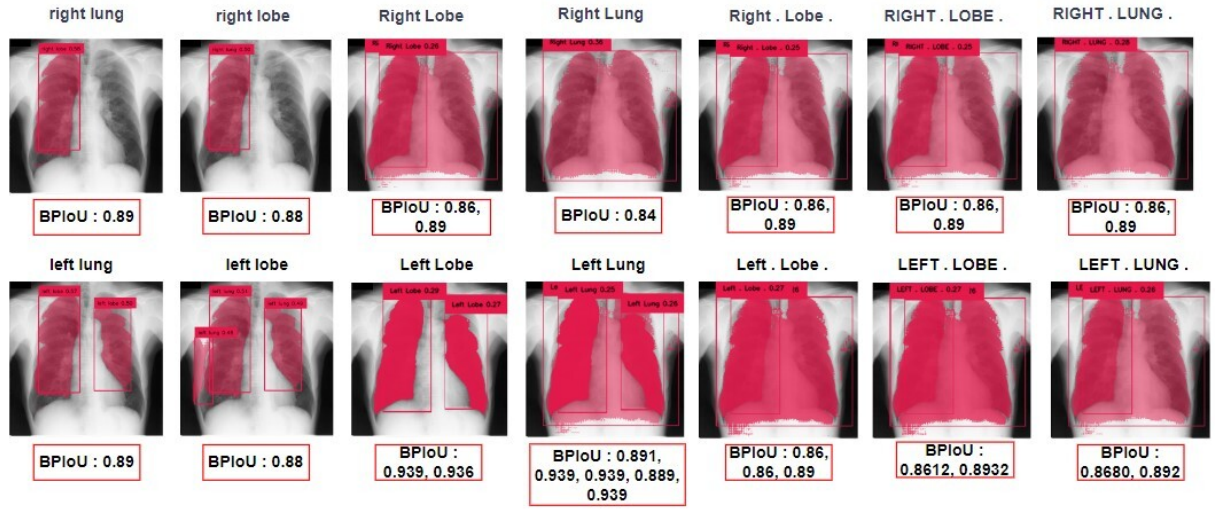


Fig. 3. Results of SAM for the corresponding anatomical prompts. It can be clearly seen that the prompts ”right lung”, ”right lobe”, ”left lung”, and ” Left Lobe” have achieved the best segmentation.

visual information from the images and BERT to extract textual descriptions. After the vanilla image and vanilla text features are extracted, it is fed to a feature enhancer block for cross-modality feature fusion, which utilizes deformable self-attention to enhance the features. Image-to-text cross-attention and text-to-image cross-attention are added for feature fusion. A language-guided query selection module is designed to select the features relevant to the input text. To merge the features of the text and image modality, a cross-modality decoder was created. Each cross-modality decoder layer feeds the results of each cross-modality query into a self-attention layer, an image cross-attention layer to combine image features, a text cross-attention layer to combine text features, and an Feed Forward Network (FFN) layer. Since text information must be injected into queries for improved modal alignment, each decoder layer contains an additional text cross-attention layer compared to the DINO decoder layer. The L1 loss and GIOU loss are used for bounding box regression. To add on, the model follows GLIP and uses contrastive loss between the predicted object and language tokens for classification. This model achieved an Average precision of 63.0 upon fine-tuning it with the COCO dataset [24].

C. Detection with Grounding DINO

The study was reviewed and approved by the Emory Institutional Review Board (IRB#STUDY00000302). The images obtained were resized to 224x224 for reducing computational complexity. The images were then converted to grayscale. Then, the image was fed to the Grounding DINO model for lung field detection. The box and text threshold were set to 0.25 and 0.25 respectively. To support multiple detections, irrelevant information from the images were suppressed by image binarization. This also helps in delineating the lung lobe structure from the background noises. Image binarization was done using OTSU thresholding with a threshold value set to 127 and the maximum value to

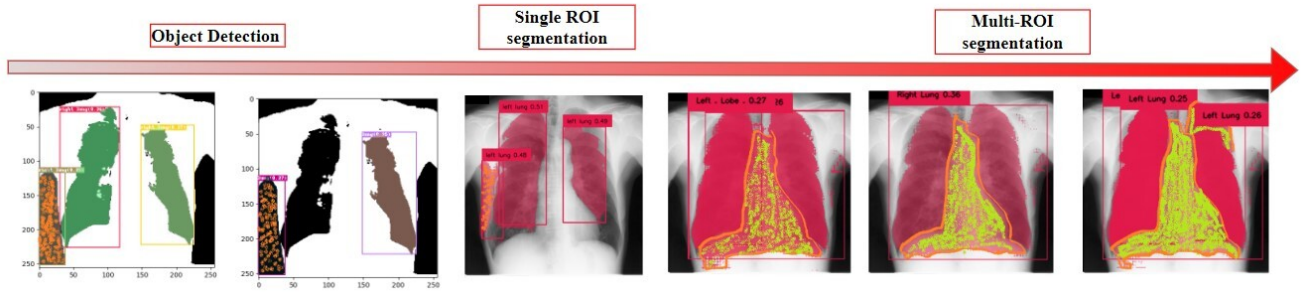


Fig. 4. The figure shows the ambiguous detections when prompted with lung anatomical prompts. It can be seen that the ambiguity begins in the DINO detection and passed to the successive stages. Similar was the case with single and multi-roi segmentations.

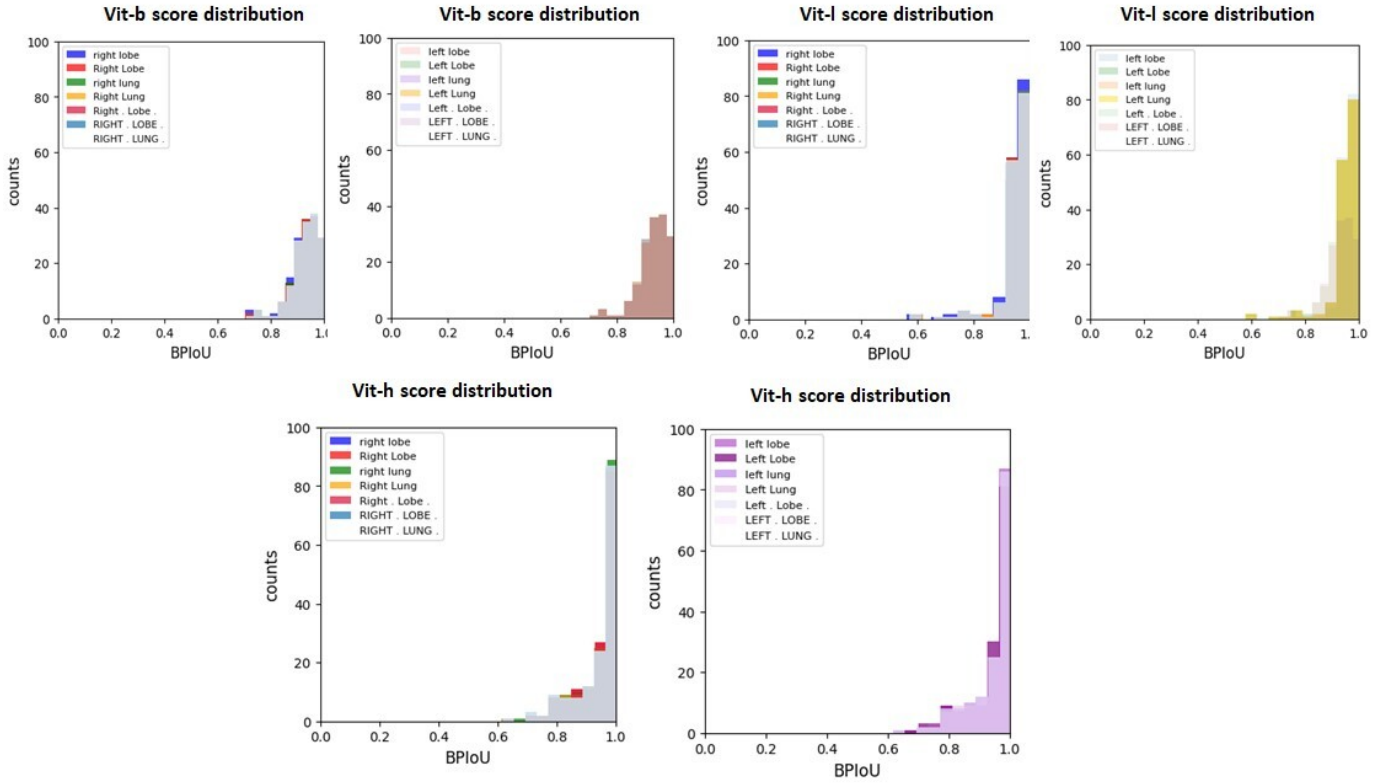


Fig. 5. Histogram plots of scores obtained from various versions of SAM for the anatomical text prompts. Most of the results are clustered on the right side of the histogram. However, we can observe that Vit-h and Vit-l model has the best distribution. Similar results are seen for the prompts ("right lobe", "Right Lobe"), ("Right . Lobe .", "RIGHT . LOBE".) irrespective of letter capitalization. The same is observed for prompts pertaining to the left lung.

196. The binarization B for a threshold T with respect to an image I is expressed as in (1).

$$\mathrm{B}(x_i, y_i) = \begin{cases} 1, & \text{if } \mathrm{I}(x_i, y_i) \geq \mathrm{T} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

D. Evaluation metrics

IoU is a generally used metric to assess quality of segmentation. IoU is measure of the overlap between the predicted and ground truth masks by dividing the intersection of the masks by the union of the masks. In this study, we used a metric called the Binarized Predicted Intersection over Union (BPIoU). The main idea was that if the binarization of CXR images yielded a higher IoU scores, then the segmentation accuracy was likely to be more, leading to precise segmentation masks. Conversely, lower stability scores resulted in poor segmentation masks, failing to produce the desired results. BPIoU (ψ) of an image $B(x_i, y_i)$ and a ground truth mask $G(x_i, y_i)$ is expressed as in (2).

$$\psi = \frac{|B(x_i, y_i) \cap G(x_i, y_i)|}{|B(x_i, y_i) \cup G(x_i, y_i)|} \tag{2}$$

IV. Results

A. Direct Object Recognition

Grounding-DINO incorporates language integration into closed-set detectors across various stages. Grounding DINO

TABLE I
OBJECT DETECTION RESULTS

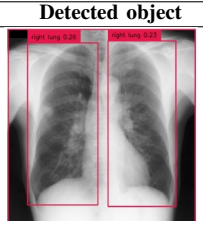
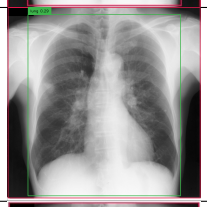
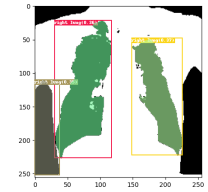
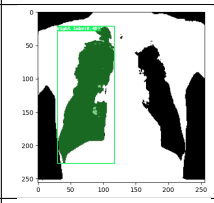
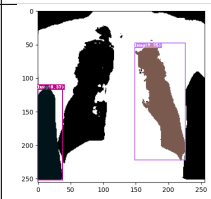
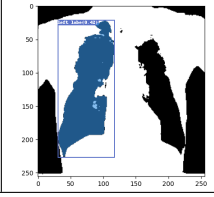
Prompt	Detected object	Comment
”right lung”		Detection is good since a higher confidence score is assigned to the right lobe.
”right lobe”		Detected both the lobes.
”left lung”		Detected both the lobes of the lung.
”left lobe”		Poor detection.

TABLE II
RESULTS OF GROUNDING DINO ON BINARY IMAGES.

Prompt	Detection	Comments
”right lung”		Identified the left lobe with a confidence score of 0.37.
”right lobe”		Correct identification of right lobe.
”left lung”		Correctly identified. But the unbound region is assigned a confidence score of 0.37.
”left lobe”		Has identified the right lobe.

accepts text prompts and provides a bounding box around the region of interest with a confidence score. This score ranges from 0 to 1. Usually, the object with the highest confidence score will correspond to the RoI. The box threshold and text threshold were specified according to the specific tasks. The results of the object detection when prompting the model with relevant text prompts are shown in Table 1. The prompt ”right lung” detected the right lobe of the lung with a score of 0.28 and also detected the left lobe with a confidence score of 0.23. However the algorithm works by selecting the bounding box rectangle with the maximum confidence score which is indicated in (3) where $[bi, bj] \in B$ represent bounding boxes in B. Furthermore, the prompts ”right lobe” and ”left lung” provided a localized bounding box with the two lobes of the lung. The prompt ”left lobe” gave poor detections which meant the model was not able to comprehend the prompt. These four anatomical prompts were experimented first to get an understanding about Grounding DINO.

$$(x_1, y_1) = \max\{(b_1, b_2), (b_3, b_4), \ldots, (b_n, b_m)\} \quad (3)$$

B. Performance on Binary Images

The performance of Grounding DINO on binary images was evaluated. The results show the ability of DINO to concentrate only on the lobes of the lungs, which are the primary regions of interest, contributes to a large portion of its accuracy in binary image detections. Furthermore, a text enhancement algorithm was incorporated to include the text ”all” with the textual prompts that was provided to the model. DINO effectively isolated the lung structures from extraneous features by taking advantage of the binary representation, resulting in robust detections.

As observed in Table 2, prompts like ”right lobe” accurately identified the RoI. The model was also provided with word-level prompts like ”right . lung .”, ”left . lung .”, resulting in accurate detections of RoI. However, the right lobe of the lung was identified with a confidence score of 0.41 when specified with the prompt ”left . lung .”.

C. Grounding DINO + SAM

The bounding box on the binary images is given as input to SAM. Only the bounding boxes associated with the maximum confidence score was selected and provided as an input to SAM. The resulting segmentation masks obtained from SAM was laid on the original image. This resulting model is the Grounded SAM which combines the zero-shot learning capabilities of both Grounding DINO and SAM. A variety of prompts related to detecting the left and right lobe of the lungs were tested and the segmentation results of the right and left lungs are presented in Figure 3 and Figure 4 respectively.

D. Prompt Ambiguity

Prompts can be ambiguous in the sense that it is unclear which object in the image is referred to by the prompt. When objects are nested within each other in an image, that is typical.

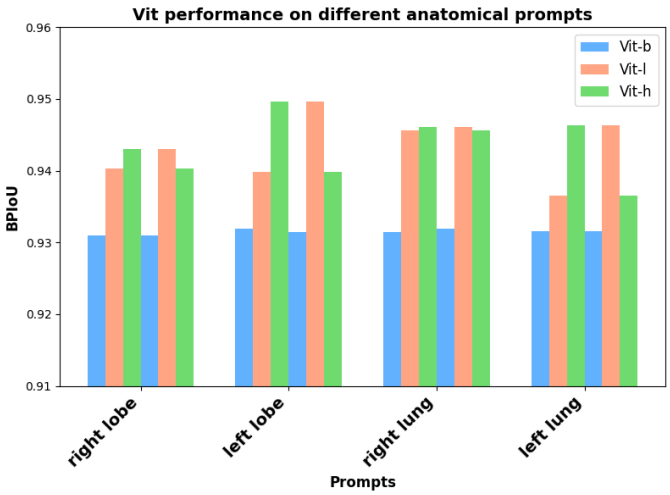


Fig. 6. A comparison of the performance of three different ViT models for CXR image segmentation on data obtained from Emory and NIH clinical center. The first three bars in each set correspond to the EMORY dataset, while the following three bars represent data from the NIH clinical center. The metrics used was the BPIoU. Variations in BPIoU was seen when prompted with the four prompts indicated on the x-axis. It can be seen that Vit-l followed by Vit-h has achieved better results compared to the three other models. The mean BPIoU score for Vit-l for the prompts "right lobe," "right lung," "left lobe," and "left lung" on data obtained from Emory was 0.943044, 0.949624, 0.946081, and 0.946362 respectively. Similarly, the scores obtained for Vit-l on the NIH data was found to be 0.943044, 0.949624, 0.946081, 0.946362 respectively. The results show that the pipeline generalizes well and also suggest that Vit-l can be integrated with Grounding DINO to provide robust segmentation.

In such cases SAM offers multiple outputs to address this issue and assist in clarifying the prompts. This is a highly essential and practical element of SAM because, in the interactive segmentation setting, the user/clinician can be provided with numerous potential outputs from which they can choose the one that is closest to the object that they wanted. In our study, we display some of the results that SAM provided (Figure 5). Ambiguity originates in the DINO detection stage and subsequently propagates through the subsequent stages of the pipeline. This pattern of ambiguity propagation holds true for single and multi-ROI segmentations.

E. Comparison of SAM Versions

The SAM model employs the ViT-b, ViT-l, and ViT-h variations of the ViT architecture. These variations vary in the number of layers and hidden units used in the ViT backbone's self-attention and feed-forward layers. ViT-b, ViT-l, and ViT-h, in particular, have 91 million, 208 million, and 636 million parameters, respectively [25]. Figure 6 shows the histogram distribution of BPIoU scores for the three models. Based on these results we can observe that the Vit-h and Vit-l model has more number of data points clustered close to 1. The prompt "right lobe" achieves the highest BPIoU scores between the range of 0.95-1.0. Similarly for the prompt "left lobe" highest scores were seen. Vit-b showed good segmentation results with majority scores in the range 0.8-1.0 but the maximum number of data points that fall under the range is approximately 40 which is significantly less compared to Vit-l, and Vit-h.

V. Discussion

The findings of the study show that the lung region can be segmented effectively using SAM by specifying text prompts. Though SAM supports points, and bounding boxes as inputs, the ability of SAM to utilize textual information for segmenting RoI is undoubtedly a game-changer.

When CXR images were fed to Grounding DINO along with appropriate prompts, the object detection results ranged from good to poor. Using "right lung" as a prompt led to successful detections with higher confidence scores. However, prompts aimed at detecting the left lobe of the lung did not yield any detections. This can be due to a variety of reasons. Firstly, it could be attributed to the anatomical and structural differences in lungs among different patients. Secondly, there might be an overlap in tissue intensities since lung tissues have similar opacities.

Additionally, prompts relating to the two categories left and right lung were experimented and tested. It was found that prompts "right lung", and "right lobe" achieved the maximum BPIoU scores compared to other prompts when the box threshold and text threshold were set to 0.25 and 0.25 respectively. These prompts were able to detect the RoI with a confidence score of 0.58 and 0.50 respectively. Similar results were achieved for the prompts "left lung", and "left lobe". It should be noted that these prompts are sentence level and lack letter capitalization. Furthermore, the model was prompted with customized texts that had either the first letter of each word capitalized or the entire word capitalized like "Right Lobe", "Left Lobe", "RIGHT .LOBE .", and "LEFT . LOBE .". Both word and sentence-level prompts were included. It was found that all these prompts performed on the same level and produced similar results. An interesting observation was that, the prompt "Left Lung" yielded multiple detections and identified the left lobe of the lung with an IoU of 0.93 indicating a very good segmentation. As a result, multiple bounding boxes were obtained and all the bounding box coordinates were given as prompts to SAM, which in turn performed multiple detections. IoU scores of 0.86, 0.89 were achieved for prompts "Right . Lobe .", and "RIGHT . LOBE .". Prompts with similar results were eliminated from further study and only certain prompts were included in performance analysis of SAM as illustrated in Figure 7 [26].

VI. Conclusion

This study delves into CXR image segmentation, that is widely used for their cost-effectiveness and capacity to diagnose lung disorders. In this work, we have concentrated on investigating SAM, a cutting-edge segmentation model by META. SAM, an AI model with exceptional object segmentation capabilities, has proved its potential to revolutionize CXR image segmentation. Using SAM, we attempted to segment CXR images by specifying text prompts. Textual prompts were included by using a zero-shot object detector known as Grounding DINO. Grounding DINO is primarily used as an object detection tool based on a given text. In this work, we took advantage of Grounding DINO's ability to produce

bounding boxes around objects, thereby generating segmentation cues for SAM. The results of the research strongly indicate the potential integration of Grounding DINO into SAM, with a particular focus on the Vit-l version, which has shown improved segmentation. SAM's capacity to segment objects, along with the interpretability of Grounding DINO's bounding box cues, has shown to be a promising technique. With more research to address its shortcomings and enhance efficiency in medical image segmentation, SAM has the potential to become a revolutionary AI model in clinical diagnosis. We anticipate a future in which SAM, aided by improvements in AI and medical imaging, will play a crucial role in redefining lung disease identification, ultimately enhancing patient care and healthcare outcomes.

ACKNOWLEDGEMENT

Some results are based upon data provided by NIH Clinical Center :https://nihcc.app.box.com/v/ChestXray-NIHCC

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

REFERENCES

[1] L. Tang, H. Xiao, and B. Li, "Can SAM Segment Anything? When SAM Meets Camouflaged Object Detection," Apr. 2023, arXiv:2304.04709 [cs]. [Online]. Available: http://arxiv.org/abs/2304.04709

[2] S. He, R. Bao, J. Li, J. Stout, A. Bjornerud, P. E. Grant, and Y. Ou, "Computer-Vision Benchmark Segment-Anything Model (SAM) in Medical Images: Accuracy in 12 Datasets," May 2023, arXiv:2304.09324 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2304.09324

[3] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment Anything," Apr. 2023, arXiv:2304.02643 [cs]. [Online]. Available: http://arxiv.org/abs/2304.02643

[4] S. Yuheng and Y. Hao, "Image Segmentation Algorithms Overview," Jul. 2017, arXiv:1707.02051 [cs]. [Online]. Available: http://arxiv.org/abs/1707.02051

[5] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," Nov. 2020, arXiv:2001.05566 [cs]. [Online]. Available: http://arxiv.org/abs/2001.05566

[6] S. Hussain, I. Mubeen, N. Ullah, S. S. U. D. Shah, B. A. Khan, M. Zahoor, R. Ullah, F. A. Khan, and M. A. Sultan, "Modern Diagnostic Imaging Technique Applications and Risk Factors in the Medical Field: A Review," *BioMed Research International*, vol. 2022, p. 5164970, 2022.

[7] R. Kundu, R. Das, Z. W. Geem, G.-T. Han, and R. Sarkar, "Pneumonia detection in chest X-ray images using an ensemble of deep learning models," *PLOS ONE*, vol. 16, no. 9, p. e0256630, Sep. 2021, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0256630

[8] A. Giełczyk, A. Marciniak, M. Tarczewska, and Z. Lutowski, "Pre-processing methods in chest X-ray image classification," *PLOS ONE*, vol. 17, no. 4, p. e0265949, Apr. 2022. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0265949

[9] A. Mittal, R. Hooda, and S. Sofat, "Lung field segmentation in chest radiographs: a historical review, current status, and expectations from deep learning," *IET Image Processing*, vol. 11, no. 11, pp. 937–952, Nov. 2017. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1049/iet-ipr.2016.0526

[10] N. Narayan Das, N. Kumar, M. Kaur, V. Kumar, and D. Singh, "Automated Deep Transfer Learning-Based Approach for Detection of COVID-19 Infection in Chest X-rays," *IRBM*, vol. 43, no. 2, pp. 114–119, Apr. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1959031820301172

[11] L. Delrue, R. Gosselin, B. Ilsen, A. Van Landeghem, J. De Mey, and P. Duyck, "Difficulties in the Interpretation of Chest Radiography," in *Comparative Interpretation of CT and Standard Radiography of the Chest*, E. E. Coche, B. Ghaye, J. De Mey, and P. Duyck, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 27–49, series Title: Medical Radiology. [Online]. Available: http://link.springer.com/10.1007/978-3-540-79942-9_2

[12] C.-C. Lee, E. C. So, L. Saidy, and M.-J. Wang, "Lung Field Segmentation in Chest X-ray Images Using Superpixel Resizing and Encoder-Decoder Segmentation Networks," *Bioengineering (Basel, Switzerland)*, vol. 9, no. 8, p. 351, Jul. 2022.

[13] W. Liu, J. Luo, Y. Yang, W. Wang, J. Deng, and L. Yu, "Automatic lung segmentation in chest X-ray images using improved U-Net," *Scientific Reports*, vol. 12, no. 1, p. 8649, May 2022. [Online]. Available: https://www.nature.com/articles/s41598-022-12743-y

[14] C. P. L and B. P, "A Study on Various Image Processing Techniques," Rochester, NY, May 2019. [Online]. Available: https://papers.ssrn.com/abstract=3388008

[15] J. Hofmanninger, F. Prayer, J. Pan, S. Röhrich, H. Prosch, and G. Langs, "Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem," *European Radiology Experimental*, vol. 4, no. 1, p. 50, Dec. 2020. [Online]. Available: https://eurradiolexp.springeropen.com/articles/10.1186/s41747-020-00173-2

[16] R. Iytha Sridhar and R. Kamaleswaran, "Lung Segment Anything Model (LuSAM): A Prompt-integrated Framework for Automated Lung Segmentation on ICU Chest X-Ray Images," May 2023. [Online]. Available: https://www.techrxiv.org/articles/preprint/Lung_Segment_Anything_Model_LuSAM_A_Prompt-integrated_Framework_for_Automated_Lung_Segmentation_on_ICU_Chest_X-Ray_Images/22788959/1

[17] Y. Zhang and R. Jiao, "Towards Segment Anything Model (SAM) for Medical Image Segmentation: A Survey," Aug. 2023, arXiv:2305.03678 [cs, eess]. [Online]. Available: http://arxiv.org/abs/2305.03678

[18] W. Sun, Z. Liu, Y. Zhang, Y. Zhong, and N. Barnes, "An Alternative to WSSS? An Empirical Study of the Segment Anything Model (SAM) on Weakly-Supervised Semantic Segmentation Problems," Jun. 2023, arXiv:2305.01586 [cs]. [Online]. Available: http://arxiv.org/abs/2305.01586

[19] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, "Segment anything model for medical image analysis: An experimental study," *Medical Image Analysis*, vol. 89, p. 102918, Oct. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841523001780

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," Jun. 2021, arXiv:2010.11929 [cs]. [Online]. Available: http://arxiv.org/abs/2010.11929

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," Aug. 2023, arXiv:1706.03762 [cs]. [Online]. Available: http://arxiv.org/abs/1706.03762

[22] K. He, C. Gan, Z. Li, I. Rekik, Z. Yin, W. Ji, Y. Gao, Q. Wang, J. Zhang, and D. Shen, "Transformers in Medical Image Analysis: A Review," Aug. 2022, arXiv:2202.12165 [cs]. [Online]. Available: http://arxiv.org/abs/2202.12165

[23] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning Where to Look for the Pancreas," May 2018, arXiv:1804.03999 [cs]. [Online]. Available: http://arxiv.org/abs/1804.03999

[24] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection," Mar. 2023, arXiv:2303.05499 [cs]. [Online]. Available: http://arxiv.org/abs/2303.05499

[25] M. Hu, Y. Li, and X. Yang, "SkinSAM: Empowering Skin Cancer Segmentation with Segment Anything Model."

[26] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 3462–3471, arXiv:1705.02315 [cs]. [Online]. Available: http://arxiv.org/abs/1705.02315