UNIVERSITY OF CALIFORNIA
Los Angeles

Applying Medical Language Models to Medical Image Analysis

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Computer Science

by

Danfeng Guo

2024

ABSTRACT OF THE DISSERTATION

Applying Medical Language Models to Medical Image Analysis

by

Danfeng Guo
Doctor of Philosophy in Computer Science
University of California, Los Angeles, 2024
Professor Demetri Terzopoulos, Chair

Medical image analysis powered by deep learning computer vision models has achieved significant advancements in the past decade. Deep learning models have demonstrated remarkable capabilities in a wide range of tasks, including medical image classification, detection, and segmentation. However, the limited availability of annotations has become a persistent challenge. Annotating medical images requires specialized professional knowledge, making it a costly process. This dissertation aims to relieve the reliance on medical image annotations by leveraging medical reports directly, which are usually associated with corresponding medical images and readily available. This thesis delves into the application of vision-language models, including large vision-language models, for enhancing medical image analysis. Existing vision-language models are modified and applied for three critical tasks: disease diagnosis, disease segmentation and medical report generation. In particular, the main contributions include: (1) proposing two prompting strategies to improve the accuracy of disease diagnosis through visual question answering in large vision language models; (2) introducing a disease segmentation model using medical reports as weak supervision; (3) evaluating medical large vision-language models in terms of the hallucination in generated reports across multiple complex diseases and applying existing techniques to mitigate the diagnostic errors in generated reports.

The dissertation of Danfeng Guo is approved.

Kadambi Achuta

Kai-Wei Chang

Yingnian Wu

Demetri Terzopoulos, Committee Chair

University of California, Los Angeles

2024

To my wife, my parents, my friends,

and everyone who offered a helping hand in the tough years.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my academic advisor, Professor Demetri Terzopoulos, sincerely for his invaluable mentorship, steadfast support, and patience, as well as for all the opportunities that he made possible for me during my PhD journey. In 2019, he graciously offered me this PhD opportunity which paved the way for my transition from computer vision to pioneering AI domains. This opportunity enriched my learning and broadened my vision on the future of AIs. During my PhD study, he granted me great autonomy which allowed me to explore the research topics of my interest. Without his unwavering support, I would not have had the chance to explore my interest in natural language processing, VL models, LLMs, and generative AIs, and the achievements documented in this thesis would not have been possible.

I would also like to extend my heartfelt thanks to my mentors, Sanchit Agarwal and Professor Mohit Bansal. During my three summers at Amazon, they provided me with lots of valuable guidance on the proper methodologies for conducting scientific research. They taught me the art of identifying and refining a research topic, navigating experimental problems and honing my scientific writing. Their suggestions largely shaped the way I do research today. They also encouraged me to learn new research areas, allowing me to explore unfamiliar topics in each internship. Beyond research, they also offered great help on my confusion related to my career. Without their help, I would never have attained my present achievements.

I am thankful for the companionship of all my friends during the past five years, especially Dingtong Yang. It was uplifting to speak with him during my times of despair. He advised me to evaluate my progress by days, even when my goals seemed out of reach. His attentive listening and encouraging words helped me overcome the challenges in my life. The times we spent together, dining and traveling across the USA were among the highlights of my PhD journey, creating memories I will treasure forever.

I will always remember Professor Fabien Scalzo, who extended an olive branch to me in 2017 when I just transferred to the area of AI and struggled to find a position. During

the time working in his lab, he provided valuable guidance. I am also deeply grateful for his recommendation, which set me on my PhD journey.

I also would like to express my gratitude to my psychological counsellors, Dr. Elizabeth Hernandez and Dr. Silvia Liu. Their kindness and attentive listening made a significant difference. They assisted me in stress management, and taught me several valuable techniques that can continue to benefit me in the future.

Lastly, I thank my family. Thank you my parents for their support of my education all the way to my PhD degree. Thank you to my wife for her companionship during the hard times; I did not feel alone while she was with me and we went through all the difficulties together.

VITA

2010–2015	B.S. in Electrical Engineering and B.B.A. in Management The Hong Kong Polytechnic University Hong Kong, China
2015–2017	M.S. in Electrical Engineering University of California, Los Angeles Los Angeles, California
2017–2018	Research Assistant Neurovascular Imaging Research Core University of California, Los Angeles Los Angeles, California
2018–2019	Algorithm Engineer Keya Medical Beijing, China
2021–2023	Applied Scientist Intern Amazon.com Services LLC Sunnyvale, California

PUBLICATIONS

Danfeng Guo, Demetri Terzopoulos. Weakly supervised zero-shot medical image segmentation using image-text contrastive learning, *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, Athens, Greece, 2024.

Danfeng Guo, Arpit Gupta, Sanchit Agarwal, Jiun-Yu Kao, Tagyoung Chung, and Mohit Bansal. Prompting vision-language models for aspect-controlled generation of referring expressions. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Mexico City, Mexico, 2024

Danfeng Guo, Arpit Gupta, Sanchit Agarwal, Jiun-Yu Kao, Shuyang Gao, Arijit Biswas, Chien-Wei Lin, Tagyoung Chung, and Mohit Bansal. GRAVL-BERT: Graphical Visual-Linguistic Representations for Multimodal Coreference Resolution, *the 29th International Conference on Computational Linguistics*, Gyeongju, Republic of Korea, 2022, pp. 285–297.

Danfeng Guo and Demetri Terzopoulos. A Transformer-Based Network for Anisotropic 3D Medical Image Segmentation, *2020 25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy, 2021, pp. 8857–8861.

Danfeng Guo, Haihua Wei, Pengfei Zhao, Yue Pan, Hao-Yu Yang, Xin Wang, Junjie Bai, Kunlin Cao, Qi Song, Jun Xia, Feng Gao, and Youbing Yin. Simultaneous Classification and Segmentation of Intracranial Hemorrhage Using a Fully Convolutional Neural Network, *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, Iowa City, IA, USA, 2020, pp. 118–121.

Hai Ye, Feng Gao, Youbing Yin, Danfeng Guo, Pengfei Zhao, Yi Lu, Xin Wang, Junjie Bai, Kunlin Cao, Qi Song, Heye Zhang, Wei Chen, Xuejun Guo, and Jun Xia. Precise diagnosis of intracranial hemorrhage and subtypes using a three-dimensional joint convolutional and recurrent neural network. In *European Radiology*, Apr 2019, pp. 6191–6201.

Yannan Yu, Danfeng Guo, Min Lou, David Liebeskind, and Fabien Scalzo. Prediction of hemorrhagic transformation severity in acute stroke from source perfusion MRI. In *IEEE Transactions on Biomedical Engineering*, Sep 2018, pp. 2058–2065.

CHAPTER 1

Introduction

Powered by deep learning, computer vision has developed rapidly in the past decade. Deep learning models have achieved state-of-the-art performance in many computer vision tasks such as image classification, segmentation and object detection. Inspired by these successes in mainstream computer vision, there has been a surge in applying these models to specialized fields, such as medical image analysis. Medical image analysis usually refers to identifying or predicting the abnormalities in medical scans (CT, X-ray, MRI, etc). It includes tasks such as detection, classification, segmentation, risk prediction, and report generation. Many deep learning computer vision models have been modified, applied to medical images, and have achieved excellent performance (Imran et al., 2020; Guo and Terzopoulos, 2021).

The development of deep learning models for medical image analysis is hindered by the limited availability of annotations. Most of these models are fully-supervised and rely heavily on either image-level or pixel-level annotations for training. Unfortunately, annotating medical images is an expensive and labor-intensive process because it requires a substantial level of knowledge/expertise plus tedious labor. Among all tasks, medical image segmentation requires the greatest effort because it demands annotations down to the pixel level. This relative scarcity of annotated data prevents supervised models from growing larger and performing better. This can also undermine the generalizability of models and succumb to overfitting, which is fatal given that medical AIs are ultimately intended to be deployed across hospitals and clinics for diagnostic purposes.

Numerous attempts have been made to address this problem. They mainly fall into three categories:

- Data augmentation: The simplest strategy is to augment the input images by applying transformations to them, such as flipping, rotation, noising, etc. Besides image manipulation, image generation models such as GANs (Goodfellow et al., 2014) have been applied to generate synthetic medical images (Skandarani et al., 2021), as have diffusion models (Kazerouni et al., 2023).

- Few-shot/unsupervised learning: Few-shot learning models need only a small quantity of annotations to train. They usually train a supervised model using limited labels, then use the trained model to generate pseudo labels for the unlabeled data, and finally both the real and pseudo labeled data are used to train the final models (Jiao et al., 2023). Unsupervised learning uses only medical images, without any labels. It usually involves clustering pixels using crafted pixel features (Abdal et al., 2021; Hamilton et al., 2022).

- Weakly-supervised learning: Instead of using the exact pixel-level annotations for the segmentation ROIs, one can use coarse annotations such as rough contours (Liu et al., 2022), boxes (Tian et al., 2021), polygons (Wang et al., 2020), points (Bearman et al., 2015), or image-level tags (Selvaraju et al., 2017; Fan et al., 2020). These coarse annotations require less effort to create. In the medical area, most weakly-supervised models are for medical image segmentation.

Additionally, models trained using unsupervised and weakly-supervised learning can also be used to generate new labels. The generated labels are combined with real labels for training, which further improves model performance (Xu et al., 2019).

In recent years, the unification of computer vision and natural language processing has become a leading trend of AI research. Sophisticated Vision-Language (VL) models are able to perform VL tasks such as Image Captioning (IC) and Visual Question Answering (VQA) (Lu et al., 2022; Wang et al., 2022c). They fuse visual features and text features, and use cross-attention mechanisms to select the related visual and text features. Furthermore, the recently popular pretrained Large Language Models (LLMs) (Radford et al., 2019; OpenAI, 2022; OpenAI et al., 2023; Wu et al., 2023a; Chiang et al., 2023) have expanded

to multimodal domains. Large Vision-Language Models (LVLMs) can perform various tasks such as captioning, summarizing and question answering (Liu et al., 2023c).

The success of VL models, including LVLMs, provide insights for medical image analysis research, particularly in how reliance on human produced annotations can be mitigated—by leveraging medical reports. Authored by clinicians, medical reports contain the key findings and diagnoses corresponding to medical images. Medical reports are widely available clinically because a medical image is usually associated with a medical report. They can be fetched effortlessly, and the text can be cleaned automatically. The textual features in medical reports can automatically be linked to visual features in medical images and used to train models to perform medical diagnosis tasks. This can significantly increase the quantity of training data, thus reducing the need for human annotation.

Technically, VL models can perform medical image classification tasks through contrastive learning or VQA. The contrastive learning method classifies an image by selecting the category word whose features are the most similar to the input image features. VQA returns "Yes" or "No" answers to questions regarding the presence of a specific disease in an image. VL models can also generate captions (reports) for medical images that also contain the classification (diagnosis) of diseases. These approaches could effectively serve as an alternative to traditional medical image classification or detection tasks. Moreover, medical reports can provide weak supervision for training medical image segmentation models (Xu et al., 2022).

1.1 Thesis Scope and Contributions

This thesis focuses on the usage of medical reports and medical VL models on three medical image analysis tasks: medical image classification, medical image segmentation, and medical report generation. For medical image classification, we focus on the diagnostic VQA accuracy of LVLMs and improve performance through prompting. For medical image segmentation, we propose a weakly-supervised learning framework to train the

medical image model only with medical reports. For medical report generation, we examine the diagnostic accuracy of the reports generated by medical LVLM and apply existing methods to mitigate its hallucination.

1.1.1 Medical Image Classification with Vision-Language Models

One prevalent approach for leveraging language models in the image classification task is through contrastive learning. VL models that are pretrained using the contrastive learning strategy can perform zero-shot classification (Radford et al., 2021). Contrastive learning enables models to align the visual and textual features. To classify an image, the model can assess the image alongside category-specific words and compare the similarity scores to make a classification. However, models trained using contrastive learning easily fail to identify minority classes (rare diseases), because the infrequent occurrence of these minority classes in the training corpus makes it difficult for the models to learn their features.

In recent years, research on LLMs has yielded astonishing achievements. Language models with billions of parameters have demonstrated excellent capabilities in a wide range of application scenarios (OpenAI, 2022; OpenAI et al., 2023; Chiang et al., 2023). The success of LLMs quickly extended to the VL domain. The visual features can be integrated into LLMs through training an adapter that projects visual features into those that can be interpreted by LLMs (Li et al., 2023b; Zhang et al., 2023b; Liu et al., 2023c). Medical image classification can be approached through medical VQA of LVLMs. The users pose questions regarding the presence of an object and the LVLMs respond based on their understanding of the images. VQA has become a basic skill of LVLMs and VQA accuracy serves as a test metric for most models (Li et al., 2023b; Zhang et al., 2023b; Zhu et al., 2023; Liu et al., 2023c). LVLMs have already been pretrained on medical datasets (Li et al., 2023a; Liu et al., 2023d; Singhal et al., 2023) and the models have been tested by medical VQA tasks (Lau et al., 2018; He et al., 2020). However, in existing datasets, a large portion of the questions involve simple questions such as "what is the

modality of this image" or "what is the organ/tissue in this image". Medical LVLMs have yet to be thoroughly evaluated on VQA accuracy across complex diseases. Additionally, general VQA models are usually tested by the commonly known accuracy, which is the percentage of correctly answered questions. However, this is not suitable for medical VQA models, because they usually suffer from the data imbalance problem, which involves many minority diseases. Medical image classification metrics, such as the Precision, Recall, and F1, are more suitable for the evaluation of medical VQA models.

Moreover, for medical LVLMs, the problem of data imbalance is more severe, exacerbated by the fact that many diseases are minority categories in medical datasets and the models are trained on large-scale data. Models may easily fail to learn the features of less common diseases. Addressing data bias typically involves strategies like including more data with better quality. However, given the scarcity of medical data, significantly enlarging the dataset may not be feasible. Traditional models tend to re-sample the data such that the positive and negative cases are relatively balanced. However, this method poses challenges when the data involves multiple categories of disease. In addition, re-sampling may not align well with the training needs of LLMs, which generally requires a large amount of data to train. All these problems highlight the need of a cost-effective approach to navigate the problem of minority categories in datasets.

There are several strategies to enhance the question answering of LLMs/LVLMs. Examples include chain-of-thought prompting (Zheng et al., 2023), self-consistency (Wang et al., 2023), and retrival-based augmentation (Caffagni et al., 2024). All these methods involve fine-tuning the models, which is expensive. Training-free methods to improve the VQA accuracy are desirable.

The crux of our study resides in the VQA of medical LVLM. An existing medical LVLM, LLaVA-Med (Li et al., 2023a), is tested for chest X-ray VQA across 5 categories of diseases. The results show that the model has low accuracy especially on minority diseases. To enhance the VQA accuracy, we propose two prompting strategies. The first involves enriching prompts with detailed descriptions of the queried disease. The descriptions include how the queried disease is defined and how it appears in images. The

second involves introducing an auxiliary weak-learner model as another agent. We train a small image classifier and fine-tune it to identify negative images accurately. Then, the negative predictions of this classifier are appended to the prompt as a reference for the LVLM. In summary, our contribution includes

1. We test LLaVA-Med in terms of diagnostic accuracy across 5 categories of diseases and show that it suffers from severe hallucination.

2. We improve the VQA accuracy by prompting the model with detailed descriptions of diseases.

3. We introduce a low-cost weak learner model as a reference for LLaVA-Med, and this effectively reduces the false positive (FP) answers.

We run our tests on the MIMIC-CXR-JPG (Goldberger et al., 2000) and Chexpert (Irvin et al., 2019) datasets. The results show that our prompt strategies improve the F1 score significantly on most disease categories (highest +0.27). We also show that our weak-learner-prompting strategy is applicable to the general domain. It reduces the false negative predictions of general domain LVLMs and improves the Recall by around 7% on POPE metrics (Li et al., 2023e).

Lastly, we train a traditional VL model on medical images and reports using contrastive learning, then test its performance on the medical image classification task. The pretrained VL model also serves as the backbone of our medical image segmentation model (introduced in Section 4.1).

1.1.2 Medical Image Segmentation with Language Models

As stated previously, annotation for medical image segmentation tasks is the most expensive among all medical image annotations because it demands pixel-level labels. An approach to addressing this problem is weakly supervised learning. It uses only coarse annotations: rough contours (Liu et al., 2022), boxes (Tian et al., 2021), points (Bearman et al., 2015), or image-level tags (Selvaraju et al., 2017; Fan et al., 2020). Semi-supervised

segmentation models use some portion of labeled data combined with unlabeled data (Lai et al., 2021). While these approaches mitigate annotation cost, medical image segmentation models have yet to be liberated from their annotation needs.

For common object segmentation, the prevalent approach to leveraging text as weak supervision is to train the VL models by contrastive learning that aligns the visual features with the textual features. Then, the visual encoder scans through the image to identify the regions/patches whose features are highly similar to those of the text descriptions (e.g., "*a black cat*"). Finally, the selected regions/patches may be processed to obtain refined ROI boundaries. Works of this type include Li et al. (2022a), Strudel et al. (2022), Mukhoti et al. (2022), Yi et al. (2023), Ren et al. (2023), and Xu et al. (2023). A mask generator can also be trained such that the masked visual features are closely matched with text features (Liang et al., 2023; Lai, 2024). The aforecited efforts cast light on how language models can be leveraged in the medical image segmentation task. To train a medical image segmentation model, one can link regional features to the corresponding statements in reports (e.g., "*tumor found in lower left lung*").

However, there are difficulties in applying these methods to medical image segmentation; one being that the identification of most diseases requires not only regional but also global information. For example, the diagnosis of Cardiomegaly requires identifying the heart and assessing its size relative to the chest. This cannot be done by merely comparing patch features. Moreover, existing training schemes fail to deal with normal, negative cases. For a normal image, no segmentation should be generated, so the VL models have no local features with which to align. Hence, current text-supervised image segmentation strategies, which are often image-patch-based, should be revisited and modified before they can successfully serve the purposes of medical image segmentation.

To overcome the aforementioned difficulties, we propose a novel strategy for text-supervised medical image segmentation. We utilize a medical language model pretrained with VL contrastive learning. The model encodes medical reports such that the encoded features are aligned with their corresponding medical image features. In the training stage, we use a positive and a negative prompt to guide the training. The visual encoder

learns to extract image features related to the positive prompt (e.g., "*tumor is seen.*"). The visual decoder learns to generate a filtering mask that is applied to the original image such that no salient features can be extracted, and the new image is aligned with the negative prompt (e.g., "*no tumor.*"). The training is weakly-supervised, merely with medical reports and images, without any segmentation annotations. We test our model on the Chexlocalize (Saporta et al., 2022) and SIIM-ACR (Zawacki et al., 2019) datasets. For the zero-shot segmentation of atelectasis, cardiomegaly, edema, pleural effusion, and pneumothorax, our zero-shot model outperforms other weakly/semi-supervised full-shot models by a significant margin.

1.1.3 Medical Report Generation

Medical report generation is in the realm of IC. Most image captioning models can be fine-tuned to serve in medical report generation. However, compared with IC, medical report generation has stricter requirements. Whereas IC aims to capture only the salient features of an image, generated medical reports should encompass all abnormal findings. Additionally, the diagnoses contained in the medical reports must be precise, given that they will serve the purposes of patient care. On the contrary, the primary concern of IC is often the human-like quality of the generated captions, and they are measured by similarity scores such as BLEU (Papineni et al., 2002) and CIDEr (Vedantam et al., 2015). Recent advancements in IC models have enabled the generation of highly human-like medical reports that achieve impressive similarity scores (Wang et al., 2022e; Li et al., 2022c). However, high similarity scores do not usually lead to high diagnostic accuracy. For example, changing the word 'left' to 'right' in generated texts may not cause a large drop in similarity scores, but it can result in erroneous facts.

It has been found that medical report generation models can easily gain high similarity scores while the diagnosis accuracy is poor (Liu et al., 2019a; Boag et al., 2020; Miura et al., 2021). This brings up the hallucination problem. Hallucination of image-to-text generation refers to the situation when the generated text includes content not shown

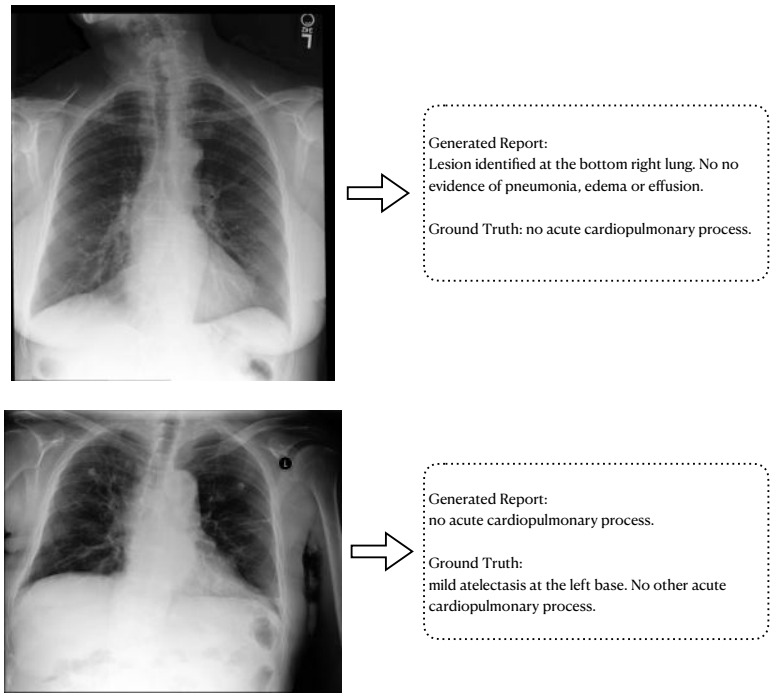


Figure 1.1: Hallucinations in medical image generation.

in the input image or that is contradictory to the image. In medical report generation, hallucination usually refers to the report containing an incorrect diagnosis; e.g., mistakes on the types of diseases and their locations. Figure 1.1 shows an illustrative example, where the top instance is an erroneous generation that reports a lesion whereas the patient is actually normal, and the bottom instance is an erroneous generation that reports normal whereas the image shows Atelectasis. Given their applications to medical diagnosis, hallucination becomes a fatal flaw in medical report generation models, since any wrong diagnosis can cause significant loss to both patients and clinicians.

In recent years, LVLMs have developed fast, demonstrating the ability to generate responses to questions that involve images (Liu et al., 2023c; Zhang et al., 2023b; Li et al., 2023b). They can be applied to medical image analysis. Albeit the impressive performance of LVLMs on various tasks, they still suffer from hallucination problem (Huang et al., 2024). While there have been efforts to fine-tune LVLMs for medical-specific tasks (Liu et al., 2023d; Singhal et al., 2023; Li et al., 2023a), the effectiveness of these models in producing accurate medical reports remains underexplored. In addition, to mitigate

hallucinations, strategies like multi-task learning (Wang et al., 2022b), prompting (Cheng et al., 2023) and contrastive decoding (Li et al., 2023d) have been proposed. It is still unknown whether these strategies can help reduce hallucinations of multimodal medical LLMs.

In this thesis, we generate medical reports using both traditional VL models and LVLMs. We then evaluate the generated reports in terms of the accuracy of generation, which includes the diagnosis accuracy of 5 critical medical findings. Furthermore, we explore existing strategies to improve the faithfulness of the reports generated by LVLMs, including instructional prompting and contrastive decoding. Our experiments show that both approaches gain a certain level o f improvement while, in general, medical LVLMs still suffer from poor report generation performance.

1.2 Overview

The remainder of this dissertation is structured as follows:

Chapter 2 reviews the development of medical image classification, segmentation and medical report generation. For classification, it introduces how VL models can be applied to it, and discusses the development of LVLMs as well as the hallucination problem in the LVLM VQA. For segmentation, it briefly introduces existing strategies for weakly-supervised segmentation and existing text-supervised segmentation models in general domain. For medical report generation, it introduces the development of IC models in general domain and their applications to medical report generation. It also talks about the IC of LVLMs and the hallucination problem of generated reports.

Chapter 3 introduces our two prompting strategies to improve the VQA accuracy of LVLMs. The corresponding experiment results are reported after the methodologies.

Chapter 4 introduces the proposed medical image segmentation model using medical reports as supervision, followed by the experiment results.

Chapter 5 introduces the fine-tuning technique for LVLM and the two applied methods

to control the hallucination (instructional prompting and contrastive decoding). Then it lists the corresponding experiment results.

Chapter 6 draws conclusions from our research, discusses its limitations, and proposes avenues for future work.

CHAPTER 2

Related Work

2.1 Medical Image Classification

2.1.1 Traditional Medical Image Classification with Deep Learning

Traditionally, deep learning-based image classification is performed using models built upon convolutional neural networks (CNN). Theses models can be fine-tuned for medical image classification (Ye et al., 2019; Guo et al., 2020; Kim et al., 2022). The training is usually performed in a fully-supervised manner. CNN-based models have achieved excellent performance on many medical image classification datasets (Pham et al., 2020; Seyyed-Kalantari et al., 2020). Since 2020, there have been efforts in applying transformers (Vaswani et al., 2017) to computer vision models. Transformer-based computer vision models (Dosovitskiy et al., 2021; Liu et al., 2021b) have also been applied to medical image classification tasks (Manzari et al., 2023; Almalik et al., 2022). Both CNNs and vision transformers can be modified and fine-tuned easily to excel in particular performance metrics that are crucial in medical applications. For example, Yuan et al. (2021) introduce a surrogate loss function to help maximize the AUC score. This gives classification models more flexibility when addressing the unique challenges found in real-world medical applications.

2.1.2 Medical Image Classification through Contrastive Learning

The first model to perform image classification through VL contrastive learning is ConVIRT (Zhang et al., 2022). It consists of a visual encoder and a text encoder. They are trained to

maximize the agreement between the features of the images and their corresponding reports. It can be easily transferred to image classification tasks by fine-tuning a classification layer after the visual encoder. Later, CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) were able to perform image classification tasks in a zero-shot/weakly-supervised manner using crafting text prompts for categories (e.g., *"an image of {object}"*). The encoded image features are matched with the text features of multiple prompts. An image is classified as the class of the prompt who has the highest similarity score. CLIP has also been adapted to medical image classification (Jang et al., 2022). Contrastive learning is also an effective pretraining strategy for VL models because it helps the model learn to align the high-level features from images and texts. To date, several medical VL models have been pretrained using contrastive learning; e.g., Wang et al. (2022d) and Wu et al. (2023b). They can be fine-tuned in a few-shot manner for various tasks including classification, detection, and segmentation.

2.1.3 VQA of Medical LVLMs

LVLMs are built upon LLMs. A pretrained visual encoder is used to extract the visual features and an adapter module is used to project the extracted features to the ones that can be understood by the LLM. Models of this type include Liu et al. (2023c), Zhu et al. (2023), and Zhang et al. (2023b). During training, the visual encoder and the LLM are usually fixed. VQA is an essential skill of LVLMs. Given an input image, the models should be able to answer questions regarding that image correctly. For medical LVLMs, given a medical scan, models such as LLaVA-Med (Li et al., 2023a) and Med-PALM (Liu et al., 2023d) are able to answer questions regarding the types of modalities, the scanned organs, and medical indicators such as opacity. They have demonstrated fine performance on medical VQA datasets such as VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021a), and Path-VQA (He et al., 2020). However, most medical questions in existing datasets are simple. Medical LVLMs have not been tested on a broader range of complex diseases.

2.1.4 Hallucination of LVLM VQA

The hallucination problem of LVLM usually refers to the model generating a response that is not consistent with the input image. For VQA, the models may make mistakes on the existence of an object, the location of an object, the attributes of an object, or the mutual relationship between objects in the generated answers. Li et al. (2023e) find that the frequent objects are easily hallucinated by LVLMs. The models tend to mention the existence of a frequent object even if it is not in the image. Qian et al. (2024) and Liu et al. (2023b) show that LVLMs sometimes presume the assumptions in questions are true and easily give wrong answers when they are asked about some objects that do not exist in the given image.

Hallucination can be incurred by bias in the training data, missing fine-grained visual features, and LLM decoding strategies (Liu et al., 2024). For data bias, the imbalanced distribution of data is an important aspect. When most of the answers to a question is "Yes" in training data, the model tends to answer "Yes" consistently. Missing fine-grained visual features is usually caused by the pretraining of the visual encoder. Most LVLMs use the visual encoder of CLIP trained through contrastive learning. The resulting visual encoder mainly focuses on salient features while ignoring the fine-grained features (Jain et al., 2023). For decoding strategies, most LVLMs choose the next word as the one having maximum conditional probability given previous texts and the input image. This criteria can lead to hallucination when the model overly relies on the knowledge learned in its training texts. There are also other causes such as model simplicity and insufficient attention (Liu et al., 2024).

Strategies to mitigate hallucination of LVLMs mainly fall into two categories, prompt engineering and model improvement. For prompt engineering, Liu et al. (2023b) leverage the visual instructions, constructed from the bounding box information in the input image to prompt the LLMs. Zheng et al. (2023) use chain of thought to prompt the models to perform step-by-step visual-language reasoning like humans, which at last leads to the correct answers. Wang et al. (2023) generate multiple chains of thought and use

the one with majority vote as the answer. Caffagni et al. (2024) prompt the model with explanations on the terms in questions. For model improvement, Sun et al. (2023a) improve the visual and text feature alignment through reinforcement learning to reduce hallucination. Leng et al. (2023a) propose a contrastive decoding strategy to reduce the models' reliance on pretrained knowledge. Favero et al. (2024) and Zhao et al. (2024) also focus on the inference stage and propose specialized decoding strategies to mitigate hallucination. Besides the two main strategies mentioned above, there are also other strategies to reduce hallucination. Zhou et al. (2024) design a post-processing model to detect the potential hallucinated objects and rephrase the generations. Sun et al. (2023b) adapt a reinforcement learning strategy that uses human evaluation on the hallucination level to improve the model.

Hallucination of LVLMs can be evaluated by two approaches. The first one is VQA. The ground truth information of the input images is leveraged to construct questions regarding the existence of objects in the images (e.g., Is there a black cat in the image?). There are also questions asking about the objects which do not exist in the images. The models are measured in terms of the percentage of correctly answered questions. Metrics of this type include POPE (Li et al., 2023e), CIEM (Hu et al., 2023), and NOPE (Lovenia et al., 2023). The other way is to use pre-designed prompts to let the models produce various generations, then evaluate them. Examples include CHAIR (Rohrbach et al., 2018), which counts the hallucinated objects in generated image captions, and MMHAL-BENCH (Sun et al., 2023a), which uses GPT-4 (OpenAI et al., 2023) to compare the generations with human answers and determine if there is hallucination.

2.2 Medical Image Segmentation

2.2.1 Fully-Supervised Medical Image Segmentation

Most medical image segmentation models are trained on fully-supervised manner with pixel-level annotations. The models are trained to perform classification for each pixel.

Most models are built upon U-Net (Ronneberger et al., 2015). Examples include Weng et al. (2019), Guo et al. (2020) and Huang et al. (2020). Later, the convolutional layers in computer vision models were partially replaced by transformers, and transformer-based models are commonly used in medical image segmentation tasks (Guo and Terzopoulos, 2021; Yan et al., 2022; Cao et al., 2023).

2.2.2 Weakly Supervised Image Segmentation

According to the types of annotations used for training, most existing works on weakly supervised segmentation can be categorized into the following groups: bounding boxes, marks, and image-level tags. An example of the first type is Tian et al. (2021). The model is trained such that the resulting masks have the same size as the annotated boxes, and closely-connected pixels are assigned the same labels. Wang et al. (2020) require polygon annotations; a patch-level classifier is trained using polygon annotations first and then the positive patches are aggregated to form the final segmentation mask. Models of the second type usually need a point (Bearman et al., 2015) or scribble (Lin et al., 2016) on the target area. The loss function is modified such that the outputs have high probability on the annotated pixels. For the third type, notable methods include Grad-CAM (Selvaraju et al., 2017), Grad-CAM++ (Chattopadhay et al., 2018), and Eigen-CAM (Muhammad and Yeasin, 2020). They first train a model for image classification. Then extract the segmentation map by either computing the gradient of classification scores with respect to the feature maps of the convolutional layers, or by computing the projection of those feature maps on their eigenvectors. Li et al. (2018) use Grad-CAM output to mask out the original image and train the model to classify the masked image as negative. This further improves the segmentation accuracy.

Weakly supervised segmentation in the medical domain remains a popular topic. Most approaches in recent years utilize image-level labels (Lerousseau et al., 2020; Chen et al., 2022; Qian et al., 2022; Giancardo et al., 2023; Li et al., 2023c), which take the least effort to acquire. Some approaches use bounding boxes (Mahani et al., 2022; Cai et al., 2022;

Du et al., 2023a), and others use patch-level labels (Dang et al., 2022).

2.2.3 Segmentation With Text Supervision

Following the success of CLIP-based models, researchers have shown that they can be also applied to segmentation because the text can serve as weak supervision. The key idea is based on cross-modal feature alignment and pixel grouping. The input image is usually divided into patches, and encoded by a patch-based encoder such as ViT (Dosovitskiy et al., 2021). The target class is represented as text prompts (e.g., *"A running bus."*) and encoded by the text encoder. The text feature is compared with each patch feature to construct a similarity matrix indicating the target locations. Patches with high similarity with the prompts are merged to form the segmentation masks for the corresponding classes. Approaches of this kind include Li et al. (2022a), Strudel et al. (2022), Mukhoti et al. (2022), Yi et al. (2023) and Ren et al. (2023). Among those, Yi et al. (2023) introduce a maximum response selection mechanism to let the model focus on the keywords and corresponding patches. The resulting similarity matrix is post-processed to have smoother and more accurate boundaries. Ren et al. (2023) align text features with the averaged features of images from multiple views to reduce the ambiguity of text supervision.

Instead of directly aligning text features with patch features, GroupViT (Xu et al., 2022) inserts "group tokens" in each layer of the ViT image encoder and uses them to merge the patch features. Xu et al. (2023) also insert learnable tokens. The visual features are first mapped to learnable tokens, then mapped to segmentation classes. The text feature is compared with those inserted token features and the ones with high similarity are selected to form the segmentation masks. Unlike the aforementioned approaches that directly generate masks from the region-text feature alignment, Yu et al. (2023), Cha et al. (2023), and Liang et al. (2022) add a visual decoder to generate masks and use them to crop the original image. Their models are trained to align the cropped image features and the prompt features. This helps the models learn local features.

2.3 Medical Report Generation

2.3.1 Medical Report Generation

Medical report generation can be performed by IC models, which traditionally usually consist of a visual encoder that extracts features from medical images, and a text decoder to generate reports. Examples of this type include Jing et al. (2018), Liu et al. (2019b), and Chen et al. (2020). Recent unified VL models (Lu et al., 2022; Wang et al., 2022c) can also be fine-tuned for medical report generation. These models also use a visual encoder (usually a transformer-based encoder) to extract the image features. The image features are then concatenated with the prompt embeddings and sent to a multilayer VL transformer. The cross-attention layers of the VL transformer select the related visual features for text generation. One can prompt the model to perform IC; e.g., "Describe the given image". A challenge of image-to-text generation is the alignment of visual and text features. Contrastive learning, introduced previously in Section 2.1.2, can address this problem. The visual encoder and text encoder can be pretrained to align the visual and text features. Li et al. (2022b) pretrain the model using an additional contrastive loss and this achieves better accuracy on language generation task. Recently, there have been several works applying IC models to medical report generation. Alfarghaly et al. (2021a) pretrain a medical image classification model to predict a group of medical tags and send the predicted tags to the language model for text generation. Li et al. (2022d) build a knowledge cluster from the training reports to guide the visual encoder to extract the image features. Li et al. (2023f) improve the cross-modal alignment while fusing visual and text features. Wu et al. (2023c) divide medical report generation into two steps: first they generate a one-sentence "impression" and then generate the multiple-sentence medical findings.

Since 2022, LLMs such as ChatGPT (OpenAI, 2022) and GPT4 (OpenAI et al., 2023) have achieved great success on content generation. Efforts have also been made to combine LLMs with visual encoders such that the models can generate content based upon visual inputs. The core strategy is to freeze the LLM and only train a small model that is

responsible for converting visual features into text features that can be understood by the LLM. The resulting LVLMs are able to perform common VL tasks such as IC and VQA. BLIP-2 (Li et al., 2023b) freezes the LLM and visual encoder and then it trains a Q-Former to align the visual features with text features. MiniGPT-4 (Zhu et al., 2023) freezes the LLM, visual ecndoer, and the Q-Former in BLIP-2 and only trains a single linear layer that is inserted after the Q-Former. LLaMA-Adapter (Zhang et al., 2023b) feeds the encoded visual features to a trainable adapter that is attached to multiple transformer layers in LLaMA (Touvron et al., 2023). LLaVA (Liu et al., 2023c) simplifies the structure and it only adds a trainable projecting matrix after the visual encoder to convert the encoded visual features. The converted features are concatenated with language instruction embeddings and fed to Vicuna (Chiang et al., 2023).

The popularity of large models has quickly spread to the medical area. Nori et al. (2023) tested GPT-4 on medical tasks and the results show that GPT-4 already exceeds the performance of previous models on benchmark datasets without being fine-tuned on medical tasks. Most medical LLMs focus on fine-tuning general LLMs on medical datasets. Two notable medical LLMs are Med-PALM (Liu et al., 2023d) and Med-PALM 2 (Singhal et al., 2023). They apply instruction prompt tuning on a small set of examples to make the model align with medical tasks. PMC-LLaMA (Wu et al., 2023a) fine-tunes LLaMA-7b (Touvron et al., 2023) on medical academic papers. HuatuoGPT (Zhang et al., 2023a) combines real-world data with distilled ChatGPT data for fine-tuning. For LVLMs, Visual Med-Alpaca (Han et al., 2023) trains a captioning model to convert medical images into text prompts, then sends them to LLaMA-7b. LLaVA-Med (Li et al., 2023a) fine-tunes LLaVA on medical datasets and achieves state-of-the-art performance on several medical tasks.

2.3.2 Evaluation of Medical Report Generation Models

General IC models are usually trained with cross-entropy loss and measured in terms of similarity scores such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie,

2005), and CIDEr (Vedantam et al., 2015). Although these methods and metrics prove effective on the captioning of general images, they are not suitable for medical image generation. Liu et al. (2019a) and Boag et al. (2020) have shown that, despite the fact that medical reports generated by many models can achieve high similarity scores, their diagnostic accuracy is poor. In Liu et al. (2019a), the model achieves 1.159 CIDEr score while having only 0.3 average Precision in terms of diagnosis accuracy. To compensate for the limitations of similarity scores, researchers use the clinical efficacy (Liu et al., 2019a) to evaluate the factual correctness of the generated reports. Clinical efficacy assesses the Precision, Recall, and F1 for each medical finding that may exist in the medical images. To measure the clinical efficacy, the generated reports are parsed and the all medical findings are categorized as positive, negative or unknown. They are compared with the ground truth class labels. This approach provides a more comprehensive and medically relevant assessment of a model's performance, focusing on the accuracy of the generated content rather than the linguistic similarity to ground truth reports.

2.3.3 Hallucination of Medical Report Generation Models

Hallucination, as previously introduced in Section 2.1.4, refers to the situations where the model generates content that violates the given facts. For medical report generation, the model may talk about medical findings that do not exist in the input image, or ignore medical findings that appear in the input image.

A common strategy to reduce hallucination is to introduce an additional classification task and perform multitask training. This helps models learn the fine-grained features of each disease. One can also adjust the weights of minority classes. Wang et al. (2022f) introduce a medical concept generation network trained to classify multiple medical concepts. The output of the concept generation network is used by the text decoder to generate the full report. Wang et al. (2022b) add classification tokens to the text decoder. The tokens are trained to classify specific diseases in the input image. Alfarghaly et al. (2021b) and Yang et al. (2023) pretrain their models for disease classification and

fine-tune them for report generation. To date, multitask learning has been mostly applied to traditional VL models rather than large models.

Prompt engineering is another effective strategy to mitigate hallucinations (Tonmoy et al., 2024). Studies show that LLMs can be prompted to generate more reliable answers. In Cheng et al. (2023), prompts are used to help GPT-3 balance demographic distribution and reduce social biases. FreshLLM (Vu et al., 2023) utilizes a search engine to inject up-to-date world knowledge into prompts to help the model answer questions that require information after the model training date. Lester et al. (2021) insert soft prompts which are learned during fine-tuning.

The decoding strategy of LLMs can also be improved to reduce hallucination. Contrastive decoding Li et al. (2023d) utilizes an expert model and an amateur model in the inference stage and the decoding objective is to select the words with the maximum difference between the likelihood of the two models. O'Brien and Lewis (2023) show that contrastive decoding improves the reasoning skills of LLMs. Leng et al. (2023b) develop a visual constrastive decoding strategy for LVLMs that mitigates object hallucination. There are also other decoding strategies. Critic-driven decoding (Lango and Dusek, 2023) combines the probability of LLM with that of a classifier that checks if the generation so far matches the given text. DoLa (Chuang et al., 2023) selects the next-token distribution by contrasting the differences in logits obtained from projecting the output of different layers to the vocabulary space.

CHAPTER 3

Medical Image Classification with Language Models

3.1 Methodology

In this thesis, we study medical image classification through both LVLM VQA and contrastive learning. This chapter focuses on applying LVLMs to medical VQA and introduces our two prompting strategies to improve the VQA performance. The details of using contrastive learning to perform medical image classification is discussed in Section A.1.

3.1.1 LVLMs

Figure 3.1 illustrates the structure of common LVLMs. They are based on a pretrained unimodal LLM such as Llama (Touvron et al., 2023) and Vicuna (Chiang et al., 2023). A visual encoder is applied to extract the image features, and the extracted features are projected to the text features space through an adapter. The projected visual features are concatenated with the text prompt embeddings and fed to the LLM. The adapter usually consists of several linear layers with non-linear activations. The visual encoder is a pretrained image encoder such as ViT (Dosovitskiy et al., 2021) or traditional CNNs. During training, the visual encoder and the LLM are usually frozen.

In our work, we choose the pretrained LLaVA-Med (Li et al., 2023a) as our model. LLaVA-Med is a medical LVLM built upon LLaVA (Liu et al., 2023c). The model structure resembles Figure 3.1. It uses pretrained Vicuna (Chiang et al., 2023) as the LLM and pretrained ViT encoder from CLIP (Radford et al., 2021) as the visual encoder. The adapter is simply a trainable projection matrix. Both the visual encoder and LLM

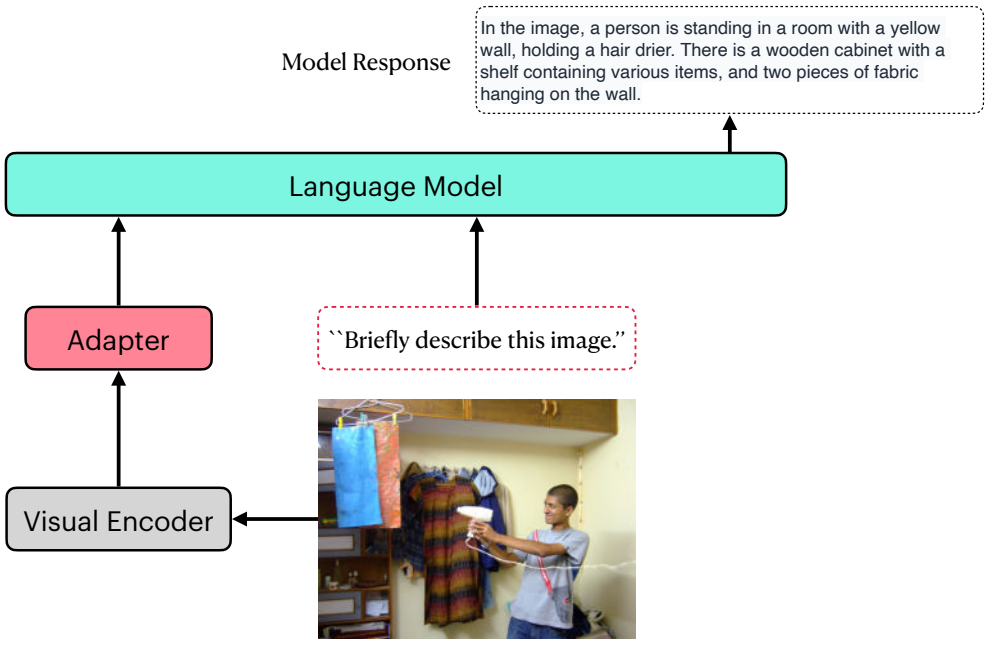


Figure 3.1: Structure of common LVLMs.

weights are frozen during training. LLaVA-Med fine-tunes LLaVA with two steps. Firstly, it fine-tunes LLaVA to generate medical reports from input medical images. Secondly, it uses GPT-4 to generate various questions from the ground truth reports and fine-tunes the model for question answering.

3.1.2 Medical Multimodall LLM VQA

Nowadays, most medical LLMs are trained for medical VQA. Medical image classification can be performed by asking questions related to various diseases, e.g., "Does this image have lung lesion?".

To reduce the model hallucination and improve the VQA accuracy, we propose two prompting strategies at the inference stage: providing the model with detailed explanations on the queried disease, and asking the model to refer to a weak learner.

Prompt with detail explanation Given the imbalanced training data, models might not be able to learn the features of the minority diseases. To compensate for the

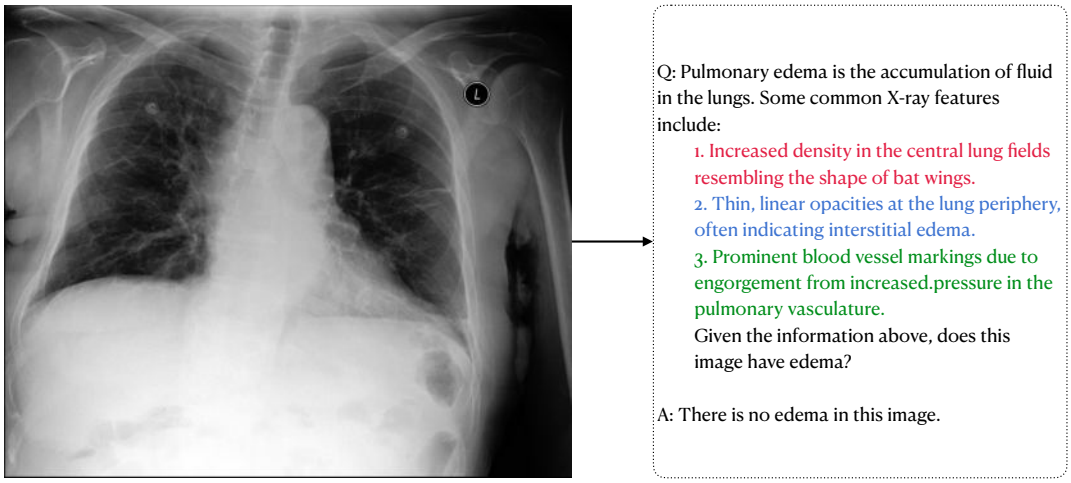


Figure 3.2: Example of prompting LVLM for medical VQA using disease explanations.

insufficient training, we provide a detailed explanation for the queried disease as prompt at the inference stage. The explanation briefly defines the asked disease and lists several key findings in medical images that may indicate its existence. An example is shown in Figure 3.2. In the beginning, the model is informed that Pulmonary Edema is defined as the accumulation of fluid in lungs. Then several chest X-ray findings that may suggest its existence are provided. The model can determine if the given image has Pulmonary Edema by linking the given key findings with the image features. The prompt templates for all diseases are listed in Appendix B.

Prompt through weak learner For traditional image classification models, data sampling is a commonly-used strategy to handle imbalanced datasets. Without re-sampling, models might consistently return negative predictions for a minority diseases. In contrast, models trained on sampled datasets often exhibit enhancements in terms of the Precision and Recall scores. However, this method may not be applicable for LVLMs for two-fold reasons. Firstly, it is difficult to balance a dataset containing many categories of diseases. Secondly, LVLMs usually demand much larger dataset and the fine-tuning is also expensive.

Although direct re-sampling may not be suitable for LVLMs, one can still let the LVLM benefit from training on sampled data by leveraging small models trained on sampled datasets to assist the LVLM. Our method resembles the multiagent LLM system such as Du et al. (2023b), where two LLMs debate with each other and hallucination can be corrected by referring to the other model's generation. Given that traditional image classifiers are smaller, it is feasible to train multiple small classifiers each of which is trained on sampled datasets of one disease. Those models can be further fine-tuned to optimize a single aspect, such as fewer false positives (FP) or fewer false negatives (FN). The classifiers are applied to the medical images and return preliminary predictions. These predictions are selectively included in the prompts as references for the LVLM. Hence, LVLMs can benefit indirectly from the nuanced understanding these specialized models can provide. This method is meaningful because clinicians usually need to balance the trade-off between overtreatment and undertreatment when making decisions. For example, they may prefer models having low FP rate if the cost of overtreatment is higher than that of undertreatment.

An example is shown in Figure 3.3, which asks about the presence of Edema. We first provide the model with the detailed description of it. Then, we aim to use the weak learner to suppress the FPs. The input image is sent to an Edema classifier, which has been fine-tuned on balanced dataset for high sensitivity and high true negative rate. If the prediction is negative, we craft the prompt "For this image, another agent thinks the probability of Edema is 0.1" and append it after the disease descriptions. The probability is manually chosen instead of the actual predicted probability because the decision threshold has been fine-tuned and no longer 0.5. We do not use zero probability because we do not want the model to overly trust the weak learner. Although in this example we only target on reducing FPs, our strategy can also be applied to reduce FNs, which can be simply done by fine-tuning the classifier for high true positive rate and applying the prompt for positive predictions.

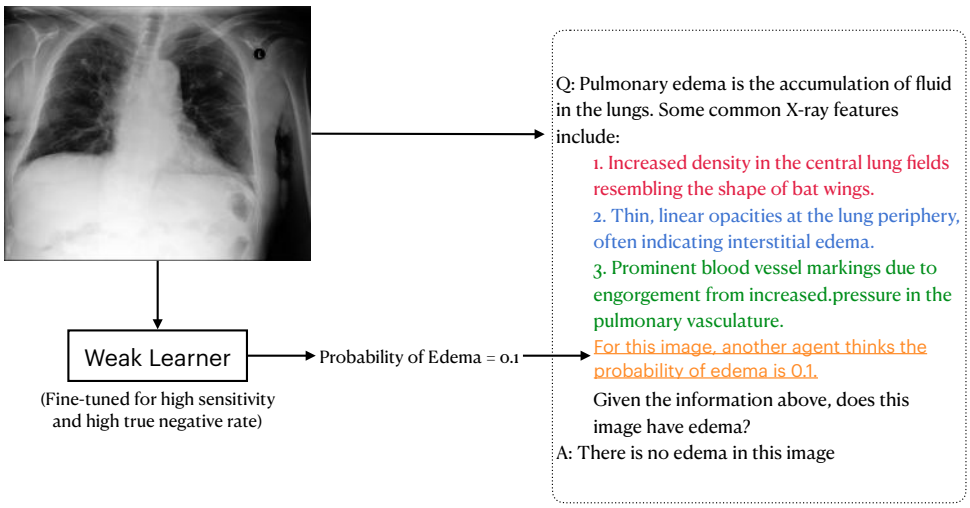


Figure 3.3: Example of prompting LVLM for medical VQA using both disease explanations and reference predictions.

3.2 Experiments and Results

3.2.1 Datasets

Test datasets The MIMIC-CXR-JPG (Goldberger et al., 2000) (MIMICCXR) and Chexpert (Irvin et al., 2019) test sets are used to evaluate the zero-shot performance. They include 5,159 and 668 images, respectively. Both datasets are chest X-rays.

MIMICCXR includes images and medical reports covering 13 diseases/findings (Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, Pleural Effusion, Pneumonia, Pneumothorax, Pleural Other, Support Devices). The raw reports are parsed and rough image-level tags are automatically generated by a rule-based approach (Irvin et al., 2019). Each label contains four values: 1 (positive), 0 (negative), -1 (uncertain) and missing. For simplicity, we treat both uncertain and missing as negative. MIMICCXR training set is also used to train the weak learner models. It contains 227,827 chest X-rays with reports.

Chexpert covers the same 13 categories in MIMICCXR. However, it does not include medical reports and only has the image-level labels. There is no overlap between

		Positive	Negative			Positive	Negative
MIMIC-CXR-JPG(5,159)	Atelectasis	1,034	4,125	**Chexpert**(668)	Atelectasis	178	490
	Cardiomegaly	1,258	3,901		Cardiomegaly	175	493
	Consolidation	326	4,833		Consolidation	35	633
	Edema	959	4,200		Edema	85	583
	Enlarged Cardio.	200	4,959		Enlarged Cardio.	298	370
	Fracture	167	4,992		Fracture	6	662
	Lung Lesion	202	4,957		Lung Lesion	14	654
	Lung Opacity	1,561	3,598		Lung Opacity	310	358
	Pleural Effusion	1,542	3,617		Pleural Effusion	120	548
	Pleural Other	119	5,040		Pleural Other	8	660
	Pneumonia	539	4,620		Pneumonia	14	654
	Pneumothorax	144	5,015		Pneumothorax	10	658
	Support Devices	1,457	3,702		Support Devices	315	353

Table 3.1: Splits of positive and negative cases for the 14 categories of MIMIC-CXR-JPG and Chexpert test sets. The 'Uncertain' is also seen as negative. "Enlarged Cardio." refers to Enlarged Cardiomediastinum.

MIMICCXR and Chexpert.

The split of categories (excluding normal) in MIMICCXR and Chexpert test sets is shown in Table 3.1. One can find that, almost all diseases are minor classes, whose positive data is much less than negative data.

We select the five diseases in the Chexpert Competition (Irvin et al., 2019) (Atelectasis, Cardiomegaly, Consolidation, Edema and Pleural Effusion) for test. More results on other medical findings are listed in the Appendix C.

LLaVA-Med pretraining dataset LLaVA-Med is trained on the PMC-15M (Zhang et al., 2024) dataset. PMC-15M contains image-text pairs of multiple modalities, e.g., CT, MRI, X-ray, etc. Note that PMC-15M does not overlap with MIMICCXR or Chexpert. In its first stage, 467,710 image-report pairs are selected for training. In the second stage, 56,708 question-answer pairs are created from the data of the first stage to fine-tuned the model. Table 3.2 shows the count of reports in the LLaVA-Med training data (second stage) which mention one of the five test diseases as positive. Compared with the total amount of data, all five categories are minorities.

	Positive Cases
Atelectasis	64
Cardiomegaly	31
Consolidation	335
Edema	1276
Pleural Effusion	260

Table 3.2: Counts of positive cases for the 5 test categories in LLaVA-Med training dataset.

Setting	Prompt
PLAIN	"Does this image have {target}?"
EXP	"{Descriptions} Given the information above, does this image have {target}?"
FINAL	"{Descriptions} For this image, another agent thinks the probability that it has {target} is {n} percent. Given the information above, does this image have {target}?"

Table 3.3: Prompt templates used in three settings. {target} is the disease asked in the questions. {Descriptions} contains the disease descriptions listed in Appendix B. {n} is the crafted predicted probability of the weak learner.

3.2.2 Implementation Details

As mentioned in Section 3.1, we use the pretrained LLaVA-Med, without any further fine-tuning. We convert the classification task into VQA by using the prompt template shown in Row 1 of Table 3.3. This is referred as the PLAIN setting of LLaVA-Med in our experiments. We first run pretrained LLaVA-Med on the PLAIN setting. Then, we add disease explanations (Row 2 of Table 3.3) to the PLAIN setting and this setting is referred as "EXP" setting. Following that, we further add the predictions of weak learners into the prompts, noted as the "FINAL" setting (Row 3 of Table 3.3).

Our weak learner is designed to suppress FP predictions, which will be justified in experiments. We use pretrained ResNet50 (He et al., 2016) as our model. For each disease, the training dataset is sampled such that the ratio of positive and negative cases is $2:1$. The model is trained for 10 epochs with $1e-4$ learning rate. The training process is monitored by the AUC score and the one with the highest validation AUC is kept.

Then, the decision threshold is fine-tuned to optimize the weighted sum of specificity and negative predictive value (NPV), which is

$$w_1 \frac{TN}{TN + FP} + w_2 \frac{TN}{TN + FN} \tag{3.1}$$

where w_1 and w_2 are preset as 0.2 and 0.8. The weak learners are applied to the medical images to obtain preliminary predictions for each disease. Then, only the negative predictions are selected to craft the additional prompts.

Lastly, the answers returned by LLaVA-Med are of various formats, e.g., "This image has Edema", "Edema is found", and "The fluid in the lung indicates Edema". An off-the-shelf Llama-7B (Touvron et al., 2023) is used to summarize the long answers into "Yes/No" such that the accuracy can be computed easily.

3.2.3 Results

Most existing medical image classification models report their results on MIMICCXR and Chexpert in terms of the AUC-ROC scores. However, this is not applicable to our context because the model generates a sequence of text instead of probabilities. One state-of-the-art work that uses F1 scores is Tiu et al. (2022), which reports the F1 scores for Atelectasis, Cardiomegaly, Consolidation, Edema and Pleural Effusion on Chexpert dataset. It also includes the F1 scores of radiologists. Their results can serve as a comparison benchmark for our experiments. Table 3.4 compares the radiologist scores, state-of-the-art scores, and LLaVA-Med VQA scores. It shows that, the VQA performance of LLaVA-Med (PLAIN setting) is unsatisfactory. The model is far from being deployed for real-world applications. The last columns shows the results after applying our two prompting strategies. Although still lower than radiologist level, a significant improvement can be observed, especially on Atelectasis, Cardiomegaly and Edema where the increase of F1 is around 0.17-0.21.

To demonstrate the efficacy of our prompting strategies, starting from the plain setting, the diseases explanations are provided first (the EXP setting). Then, based on the results,

	Radiologist	(Tiu et al., 2022)	PLAIN	FINAL
Atelectasis	69.2	64.6	26.5	41.3
Cardiomegaly	67.8	74.3	24.0	51.5
Consolidation	38.5	33.3	11.7	12.2
Edema	58.3	60.2	25.4	42.6
Pleural Effusion	73.7	70.4	35.5	46.8

Table 3.4: Comparison of F1 scores (represented as %) on 5 diseases in Chexpert dataset. Radiologist means the score of radiologist diagnosis. PLAIN is the PLAIN setting of LLaVA-Med VQA. FINAL is the best result achieved by applying both of our prompting strategies.

weak learners are tailor made to improve the performance on specific aspects (the FINAL setting).

3.2.3.1 Adding Disease Descriptions

Table 3.5 shows the Precision, Recall and F1 scores of plain setting and the exp setting on MIMICCXR and Chexpert test set. On MIMICCXR, after adding diseases explanations, the F1 scores of Atelectasis, Cardiomegaly, Edema, and Pleural Effusion show an increase, while the Consolidation F1 only has trivial increase. On Chexpert, after adding diseases explanations, the F1 scores of Atelectasis, Cardiomegaly, Edema show an increase, while the other two have no increase.

From the Precision and Recall scores it can be noticed that, adding the explanations generally leads to a large increase on the Recall while only has minimal influence on the Precision. For minority diseases such as Consolidation whose F1 is dominated by the low Precision, improving the Recall would not have much effect on the F1.

3.2.3.2 Referring to Weak Learners

Starting from the first prompting strategy, we use our second strategy to further improve the accuracy. Based on the analysis of Table 3.5, the performance bottleneck is the Precision. Table 3.6 counts the true positive (TP), false positive (FP) and false negative (FN) predictions of LLaVA-Med prompted by disease explanations on Chexpert test set.

		MIMIC-CXR-JPG		Chexpert	
Diseases	Metrics	PLAIN	EXP	PLAIN	EXP
Atelectasis	Precision	19.5	20.0	30.5	26.5
	Recall	41.5	92.9	44.4	91.6
	F1	26.5	33.0	36.5	41.0
Cardiomegaly	Precision	25.8	24.6	27.1	26.0
	Recall	22.5	89.4	20.0	86.3
	F1	24.0	38.6	23.0	40.0
Consolidation	Precision	6.8	6.3	6.0	5.2
	Recall	42.3	98.5	40.0	97.1
	F1	11.7	11.9	10.4	9.8
Edema	Precision	19.6	18.5	11.7	13.7
	Recall	36.0	72.7	29.4	76.5
	F1	25.4	29.5	16.8	23.2
Pleural Effusion	Precision	30.4	30.0	22.3	17.9
	Recall	42.8	92.7	49.2	90.0
	F1	35.6	45.3	30.7	29.9

Table 3.5: LLaVA-Med VQA performance evaluated by Precision, Recall and F1 scores of 5 diseases on MIMICCXR and Chexpert test set.

It is observed that the large number of FP cases is the performance bottleneck. Hence, the weak learners should be designed to suppress the FP predictions.

Table 3.7 compares the performance on Chexpert before and after referring to the weak learner. It shows that the prediction accuracy (F1) can be largely increased by providing reference predictions into the prompts. The F1 scores of Cardiomegaly, Edema and Pleural Effusion increase by 0.115, 0.194 and 0.089, respectively. To further demonstrate

	TP	FP	FN
Atelectasis	163	453	15
Cardiomegaly	151	430	24
Consolidation	28	557	7
Edema	65	410	20
Pleural Effusion	108	495	12

Table 3.6: Count of TP/FP/FN cases for LLaVA-Med prompted with disease explanations (EXP setting).

Diseases	Metrics	EXP	FINAL
Atelectasis	Precision	26.5	28.8
	Recall	91.6	83.1
	F1	41.0	42.8
Cardiomegaly	Precision	26.0	38.1
	Recall	86.3	79.4
	F1	40.0	51.5
Consolidation	Precision	5.2	7.5
	Recall	97.1	34.3
	F1	9.8	12.2
Edema	Precision	13.7	36.8
	Recall	76.5	50.6
	F1	23.2	42.6
Pleural Effusion	Precision	17.9	25.0
	Recall	90.0	85.0
	F1	29.9	38.8

Table 3.7: Diagnosis accuracy on Chexpert dataset for LLaVA-Med EXP setting and LLaVA-Med using both disease explanations and references of trained weak learners (FINAL setting).

the efficacy of our prompting strategy, we count the FP predictions for FINAL setting and compare them with the EXP setting, which is shown in Table 3.8. The reduction of FP cases is notable, especially on Edema, where the FP is reduced by 322 (78.5%).

This strategy can also be extended to general domain LVLMs. We study the performance of LLaVA (Liu et al., 2023c) and MiniGPT-v2 (Zhu et al., 2023) on POPE (Li et al., 2023e) metrics. POPE evaluates the hallucination of LVLMs by asking questions about the presence of objects. The POPE scores of LLaVA and MiniGPT-v2 have high Precision

	EXP	FINAL
Atelectasis	453	365
Cardiomegaly	430	226
Consolidation	557	149
Edema	410	88
Pleural Effusion	495	304

Table 3.8: Count of FP cases on Chexpert for LLaVA-Med EXP setting and WL setting.

	POPE Adversarial			POPE Popular			POPE Random		
	Prec	Recall	F1	Prec	Recall	F1	Prec	Recall	F1
LLaVA	91	78.8	84.5	95.2	78.8	86.2	97.4	78.8	87.1
with ref	88.4	85.7	87.0	92.8	85.7	89	97.3	85.7	91.1
MiniGPT-v2	88.2	77.2	82.3	92.7	77.2	84.2	97.2	77.2	86.1
with ref	86.8	84.2	85.5	91.9	84.2	87.9	97.3	84.2	90.3

Table 3.9: Comparison of POPE scores for models with and without referring to the predictions of weak learners. "Prec" refers to the Precision score.

and low Recall. Hence, the weak learner can be used to reduce the FN predictions. We select an off-the-shelf Fast-RCNN (Girshick, 2015) as the weak learner and fine-tune the detection threshold of bounding box scores for high Recall. Then, the positive predictions of the weak learner are added to the prompts. Results in Table 3.9 show that the Recall scores across three POPE categories are increased by around 7% and the Precision scores decrease slightly. The overall F1 scores are improved.

CHAPTER 4

Medical Image Segmentation with Language Models

4.1 Methodology

4.1.1 Pretraining

In the pretraining stage, we use contrastive learning to help our model learn to connect the visual and text representations. Specifically, the model learns to link the text representations of disease with corresponding image features. The steps are shown in Figure 4.1. The input data contains N medical images and the corresponding N medical reports, which describe the key observations in the medical images and include the diagnosis of diseases. We use the ViT (Dosovitskiy et al., 2021) to encode the images into sequences of patch features. The representation of each image is the global maximum of its patch features. The medical reports are encoded by a multi-layer transformer text encoder. The encoded visual and text representations are projected into the same dimension D.

Let $F_v \in \mathbf{R}^{N \times D}$ be the projected visual representation and $F_t \in \mathbf{R}^{N \times D}$ be the projected text representation. The similarity matrix is defined as $S = F_v * F_t^T \in \mathbf{R}^{N \times N}$, and $S(i, j)$ describes the closeness of image i and report j. We use the infoNCE loss (van den Oord et al., 2019) as our training loss. For the images, the contrastive loss is

$$\mathcal{L}_{\text{im}} = -\mathbb{E}_i \left[\log \frac{\exp S(i, j_c)}{\sum_{j \in N} \exp S(i, j)} \right], \tag{4.1}$$

where j_c is the text matched with image i. The loss for texts $\mathcal{L}_{\text{txt}}$ is defined in the same way. The final loss is the average of $\mathcal{L}_{\text{im}}$ and $\mathcal{L}_{\text{txt}}$.

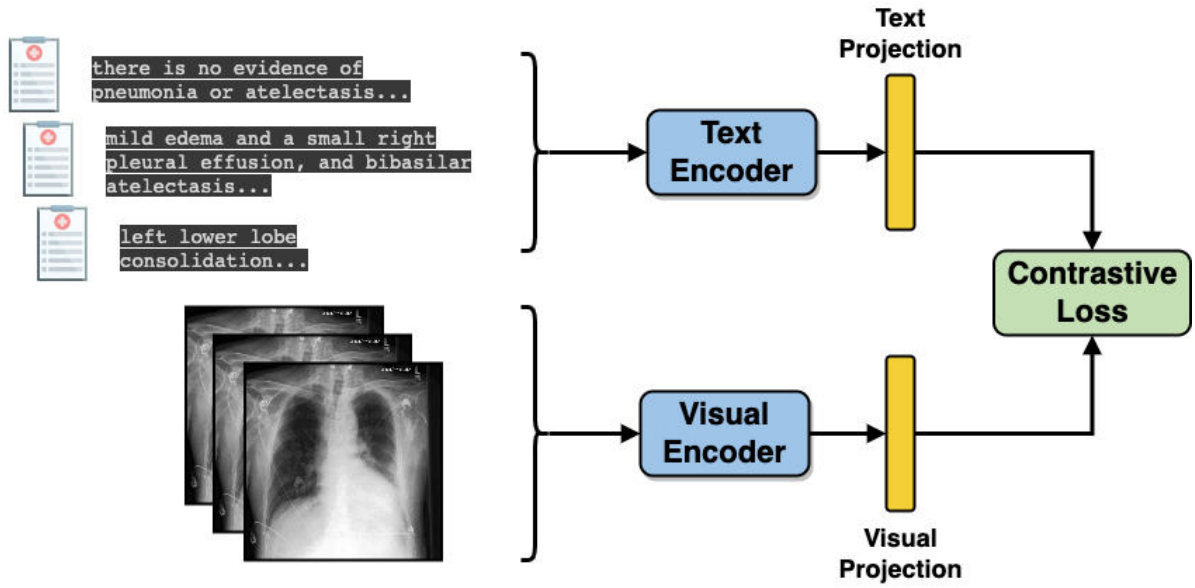


Figure 4.1: Pretraining of visual encoder and text encoder.

4.1.2 Model Design

We will now motivate our model design. The segmentation models with text supervision reviewed in the previous section either directly align regional representations with prompt representations (Li et al., 2022a; Strudel et al., 2022; Mukhoti et al., 2022) or generate segmentation masks to crop the original images and match the cropped images with prompts (Yu et al., 2023; Cha et al., 2023). We believe that these strategies are not suitable for most medical images, for several reasons. First, most medical images are grayscale and the abnormal regions usually appear as an area with high intensity. The ROIs of different diseases share a high similarity and one can hardly identify the correct type of disease simply by matching the text features and regional image features. Global image information is also needed. For example, the diagnosis of cardiomegaly requires the position of the ROI as well as its size relative to the chest. Figure 4.3 shows the ROIs of four diseases cropped from chest X-rays. They are all bright areas and highly similar to each other. Even pathologists may fail to identify the correct types merely by looking at those cropped regions. Moreover, existing training schemes fail to deal with normal cases. For a healthy scan, no segmentations are supposed to be generated and those models

have nothing to align with.

To address the two aforementioned problems, we perform image-text alignment instead of region-text alignment. On the one hand, the extracted image features contain global information and are better aligned with texts of different diseases. On the other hand, we can handle normal cases simply by aligning with texts such as *"a healthy scan"*.

As shown in Figure 4.2, our model consists of the visual and text encoder used in the pretraining stage, and a visual decoder. Given an input image that contains the target, we extract its visual representation using the same approach as pretraining. For the text, rather than using the medical report as the input, we create two prompts: "{*target*} *is seen.*" and "*no* {*target*}", where {*target*} is the name of segmentation category; e.g., pleural effusion. We denote the two prompts as P_{pos} and P_{neg}. We use the text encoder to obtain the representations of the two prompts.

Our model has two branches. In the first branch, the model is trained to align the representation of the input image and P_{pos}. We expect that the visual encoder can learn to extract features related to the {*target*}. In the second branch, we use a visual decoder to generate a filtering mask from the encoded image features. The mask is used to generate a healthy scan from a scan with the target disease. The new image is encoded by the visual encoder and its representation is aligned with the representation of P_{neg}. The image decoder is expected to generate masks which contain as much ROI as possible such that the filtered image is as close as possible to a healthy one, whose representation has high similarity with that of P_{neg}.

Note that during training, if the input image does not contain the target, the model will be directly trained to align the image representation with the representation of P_{neg}. Hence, image-level tags are needed. Instead of using human annotations, we use the Chexpert labeling toolbox (Irvin et al., 2019) to automatically parse the report and extract rough labels.

For the text encoder, we have two options: the multi-layer transformer of BLIP (Li et al., 2022b) which is trained on RefCOCOg (Mao et al., 2016) for image-text alignment,

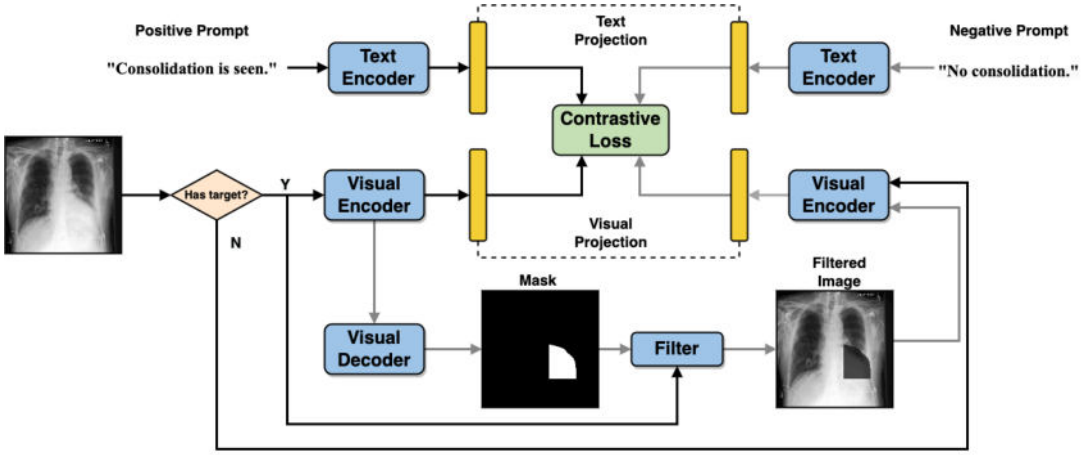


Figure 4.2: Model architecture for text-supervised segmentation.

and the Clinical BERT (Alsentzer et al., 2019) trained on MIMIC-III (Johnson et al., 2016) for medical report classification. For the visual decoder, we also consider two models, Segformer (Xie et al., 2021) and Segmenter (Strudel et al., 2021). Both models use ViT as visual encoders. The difference is that, for the decoder, Segformer uses linear layers and Segmenter uses transformer layers. We also consider that both the linear layers and transformer layers may not preserve spatial information among pixels well. Hence, we explore adding convolutional layers behind. We will compare all the above options in our Ablation Studies.

4.1.3 Image Filtering

In detail, given that for most diseases the intensity of the ROI is higher than for normal cases, we think the high-intensity ROIs can be filtered such that the generated image is close to normal ones. Then, the filtered images can be well aligned with the negative prompt. Given the original image $\mathcal{I}$ and mask $\mathcal{M}$, the filtered image is simply

$$\mathcal{I}_f = \mathcal{I}(1 - \mathcal{M}), \tag{4.2}$$

where $\mathcal{M}$ is the output of the image decoder and its values are between 0 and 1.

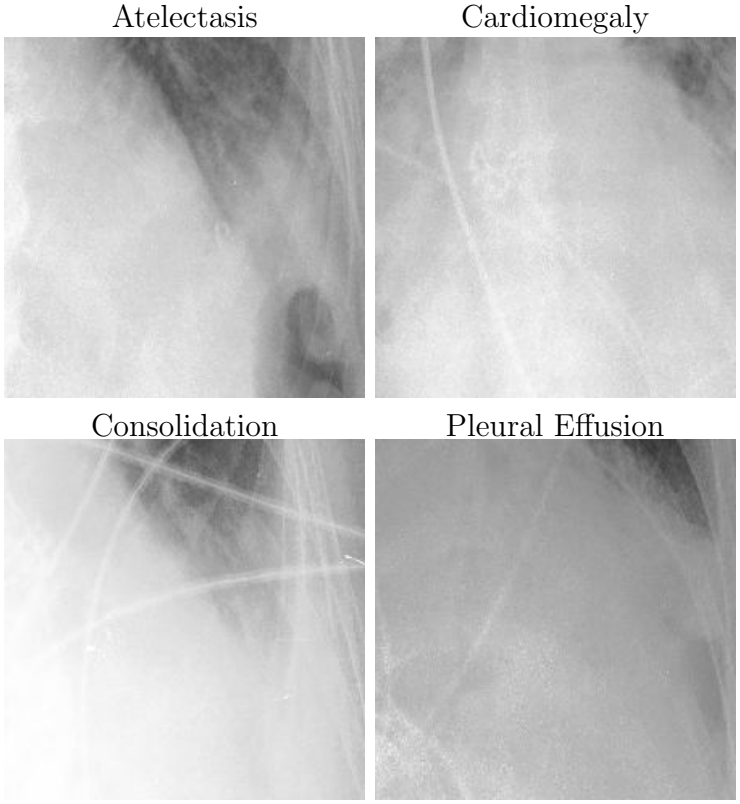


Figure 4.3: ROIs of Atelectasis, Cardiomegaly, Consolidation and Pleural Effusion, cropped from chest X-rays.

4.1.4 Loss Function

The loss function is still based on the infoNCE loss, but unlike pretraining, we remove $\mathcal{L}_{\text{text}}$:

$$\mathcal{L}_m = \begin{cases} \mathcal{L}_{\text{im}}(\mathcal{I}, P_{\text{pos}}) + \mathcal{L}_{\text{im}}(\mathcal{I}_f, P_{\text{neg}}) & \mathcal{I} \in \mathcal{I}_{\text{pos}} \\ \mathcal{L}_{\text{im}}(\mathcal{I}, P_{\text{neg}})) & \mathcal{I} \in \mathcal{I}_{\text{neg}}, \end{cases} \tag{4.3}$$

where I_{pos}, I_{neg} and I_f are images with positive labels, images with negative labels, and filtered images with positive labels. To avoid generating a large mask that erases almost all regions, we add another loss term $\mathcal{L}_{\text{area}} = mean(\mathcal{M})$ to restrict the size of the masks. The final loss is

$$\mathcal{L} = \mathcal{L}_m + \alpha \mathcal{L}_{\text{area}}, \tag{4.4}$$

where α is a hyperparameter to be tuned.

4.2 Experiments and Results

4.2.1 Datasets

The medical reports and images in MIMIC-CXR-JPG (Goldberger et al., 2000) are used for both pretraining and training. For validation and testing, we used the validation (187 images) and test (499 images) sets of the Chexlocalize X-ray image segmentation dataset (Saporta et al., 2022). There is no overlap between the MIMIC-CXR training set and Chexlocalize. We evaluated our model on the segmentation of 5 diseases: Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion. Furthermore, we also tested its zero-shot performance on the SIIM-ACR Pneumothorax Segmentation dataset (Zawacki et al., 2019) using the test split (1,737 images) from (Wang et al., 2022a; Wu et al., 2023b; Wan et al., 2023).

4.2.2 Implementation Details

In the pretraining stage, our visual encoder is the ViT in segformer-b1 (Xie et al., 2021). For the text encoder, we use the 12-layer transformer of BLIP (Li et al., 2022b). In the training stage, we fix the text encoder. For the visual decoder, we use the linear decoder of Segformer and add three convolutional layers after that. Our choices are justified by our Ablation Studies.

Parameter α in (Equation 4.4) is an important hyperparameter that controls the size of the generated masks. The value should be unique for each disease because different diseases usually have different sizes. Hence, we tune α for each of the 5 diseases separately. The α for Atelectasis, Consolidation, Edema, Pleural Effusion and Pneumothorax is 0.005 and for Cardiomegaly it is 0.001. Our above choices are justified by our ablation studies.

4.2.3 Results

For Chexlocalize, the results were evaluated in terms of mIOU with respect to human annotations. Our model was compared with the baseline model, a full-shot Grad-CAM

	Base	ChexSeg	GroupViT	Ours
Atelectasis	0.254	0.323	0.112	0.347
Cardiomegaly	0.452	0.461	0.341	0.488
Consolidation	0.408	0.110	0.231	0.332
Edema	0.362	0.257	0.401	0.447
Pleural Effusion	0.235	0.273	0.073	0.275

Table 4.1: Comparison of model segmentation performance measured by mIOU on Ateletasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion. (The performances reported by ChexSeg is only on 500 images annotated by their own radiologists.)

(Selvaraju et al., 2017) provided in the Chexlocalize publication, ChexSeg (Gadgil et al., 2021), and GroupViT (Xu et al., 2022), which we trained. The Grad-CAM (Selvaraju et al., 2017) is trained on the Chexlocalize training data with automatically extracted image-level tags. ChexSeg is a semi-supervised segmentation model. It is trained on Chexlocalize training data and uses pseudo segmentation labels generated by the baseline, combined with a portion of human-annotated segmentation labels. GroupViT is trained with images and medical reports from MIMIC-CXR-JPG, similar to our model, but the difference is that it generates masks by directly matching text prompts with encoded token features.

For SIIM-ACR, the performance was measured by the Dice score, and compared with other models that reported Dice scores.

Table 4.1 and Table 4.2 report our results on Chexlocalize and SIIM-ACR. For Chexlocalize, our model outperforms by a large margin on atelectasis, cardiomegaly, edema, and pleural effusion. Especially, the comparison with GroupViT also supports our position that directly aligning patch features with text prompts may not be suitable for medical images. For SIIM-ACR, our model also outperforms other semi-supervised models. In Row 5, we also post the result of the SIIM-ACR Challenge winner. There is still a large gap between our zero-shot performance and the top runner. Figure 4.4 and Figure 4.5 show segmentation examples compared with human-annotated labels for all 6 disease categories. The segmentations are shown as the heatmap of the filtering mask. In Figure 4.6, we also show the heatmap for normal patients.

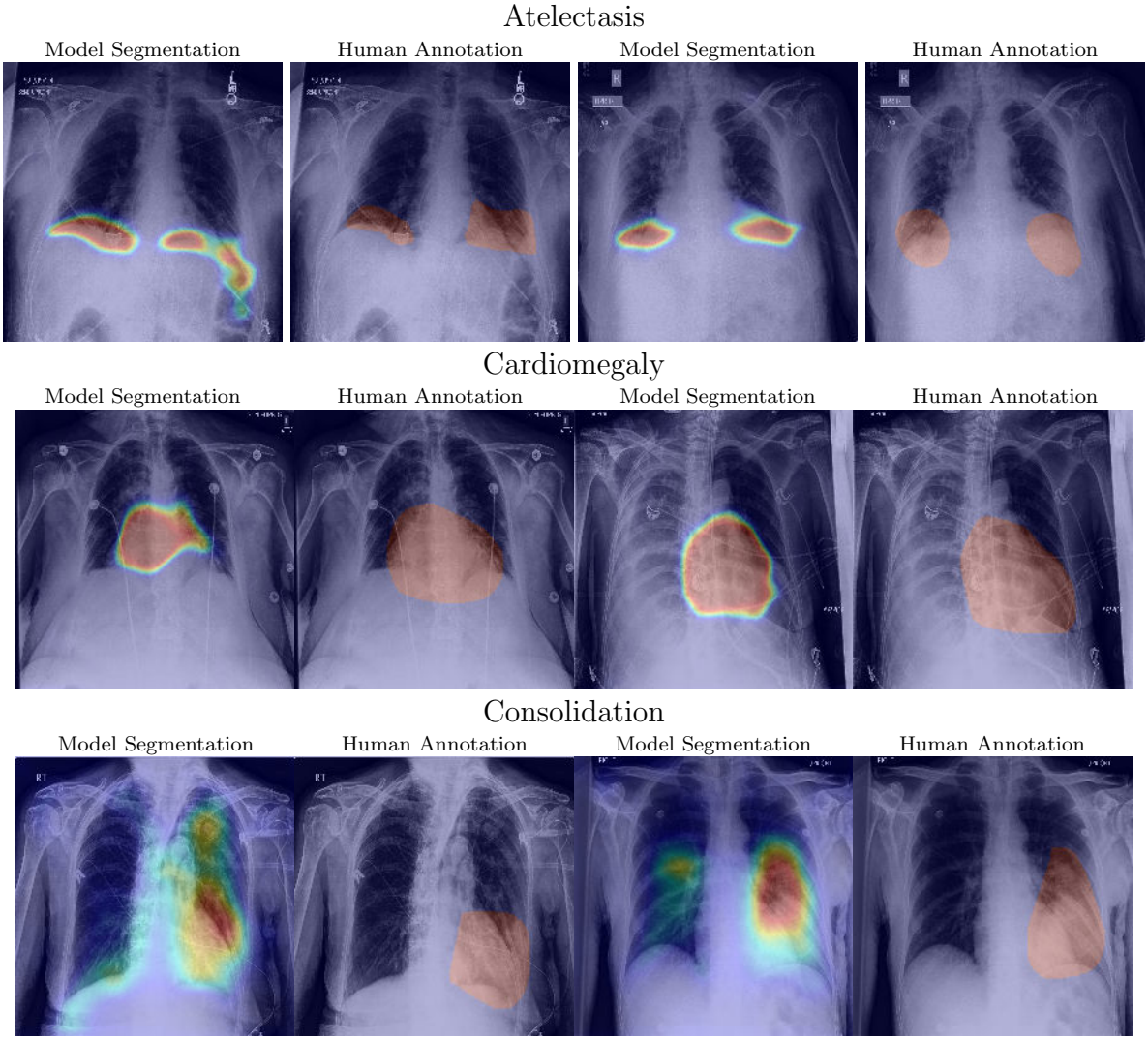


Figure 4.4: Visual comparison of segmentation heatmaps and human annotations for Atelectasis, Cardiomegaly and Consolidation.

	Dice	Remark
MGCA (Wang et al., 2022a)	59.3	10% data
MedKLIP (Wu et al., 2023b)	60.8	10% data
MedUniC (Wan et al., 2023)	62.2	10% data
IMITATE (Liu et al., 2023a)	61.7	10% data
Winnder (Anuar, 2019)	86.79	fully-supervised
Ours	68.64	zero-shot

Table 4.2: Dice on SIIM-ACR pneumothorax segmentation.

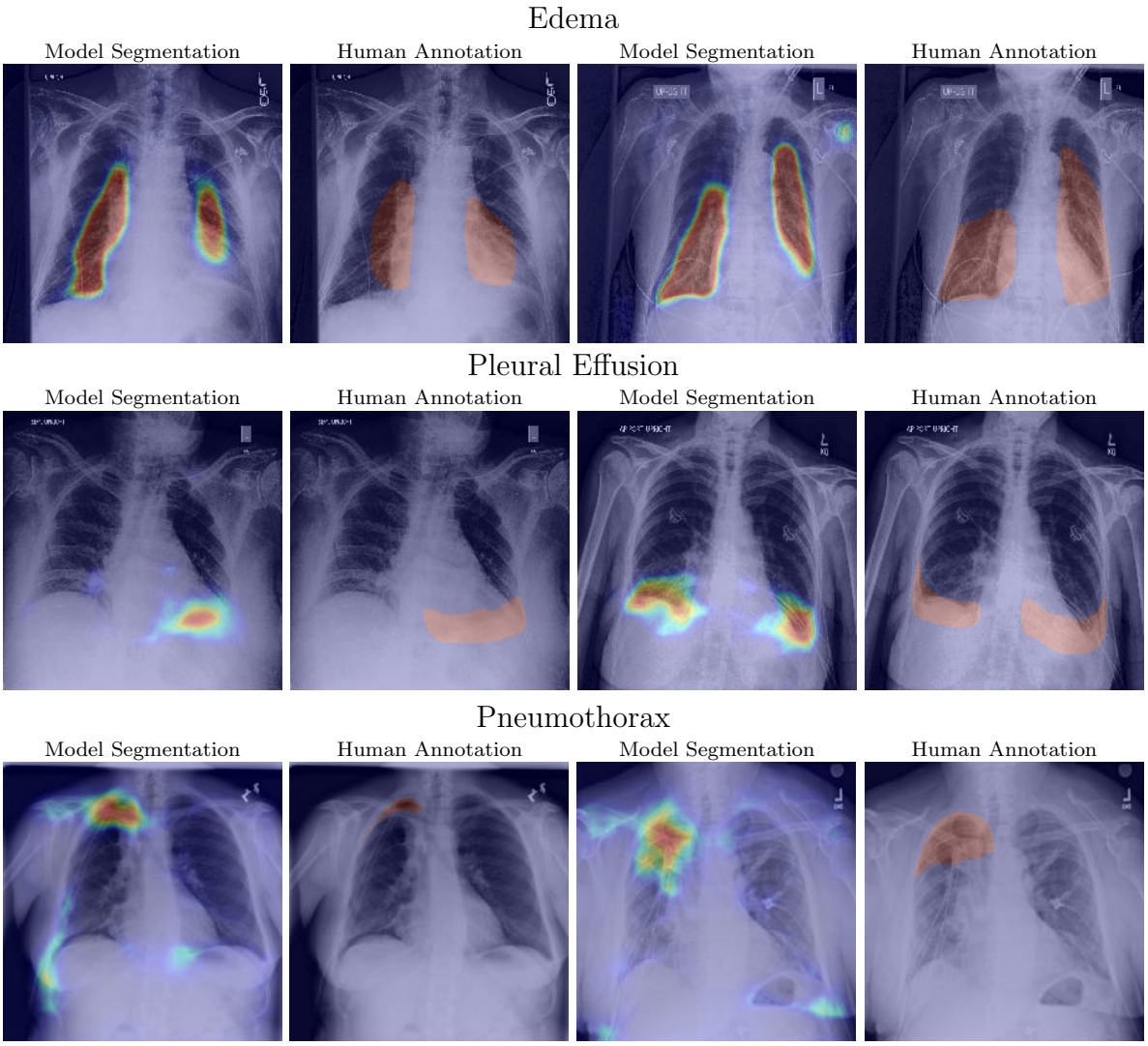


Figure 4.5: Visual comparison of segmentation heatmaps and human annotations for Edema, Pleural Effusion and Pneumothorax.

4.2.4 Ablation Studies

Choices of encoders/decoders: Table 4.3 compares the effects of different text encoders and visual decoders. The results are measured in terms of the mIOU for Pleural Effusion segmentation. As mentioned previously, for the text encoder, we consider Clinical BERT (Clin BERT) trained on medical reports for text classification and the text encoder of BLIP trained on RefCOCOg (BLIP BERT) for VL feature alignment. For the visual decoder, we consider the decoder of Segformer, which consists of linear layers, and the transformer decoder of Segmenter. We also modify the linear visual decoder by appending

Text Encoder	Visual Decoder	mIOU
Clin BERT	Linear	0.242
BLIP BERT	Linear	0.220
BLIP BERT	Transformer	0.185
Clin BERT	Linear + Conv	0.247
BLIP BERT	Linear + Conv	0.275

Table 4.3: mIOU on Pleural Effusion using different types of encoders and decoders. "Linear" refers to the linear decoder of Segformer and "Transformer" refers to the transformer decoder of Segmenter.

convolutional layers. Clin BERT (Row 1) yields a better score than BLIP BERT (Row 2). Row 2 and Row 3 show that the linear decoder of Segformer is better than the transformer decoder of Segmenter. A possible explanation could be that the transformer-based decoder has much larger size and easily overfits the data. Row 4 and Row 5 show that appending convolutional layers behind linear layers improves model performance. The best performer is BLIP BERT with the linear+convolutional decoder and it is used as our preferred architecture.

Cropping: We modified our model in the same way as (Yu et al., 2023; Cha et al., 2023). Instead of using the mask to filter the original image, we use it to crop the image and train our model to align local image features with Q_{pos}. For $\mathcal{I}_{\text{pos}}$, the new alignment loss is the combination of global and local infoNCE loss:

$$\mathcal{L}_m = \begin{cases} \mathcal{L}_{\text{im}}(\mathcal{I}, P_{\text{pos}}) + \mathcal{L}_{\text{im}}(\mathcal{I}_c, P_{\text{pos}}) & \mathcal{I} \in \mathcal{I}_{\text{pos}} \\ \mathcal{L}_{\text{im}}(\mathcal{I}, P_{\text{neg}})) & \mathcal{I} \in \mathcal{I}_{\text{neg}}, \end{cases} \tag{4.5}$$

where $\mathcal{I}_c$ is the cropped image. Lastly, as in (Equation 4.4), we include $\alpha\mathcal{L}_{\text{area}}$. The segmentation mIOUs for the five categories are shown in the right column of Table 4.4. The cropping approach underperforms by a large gap, which further supports our claim that the existing strategies are not suitable for medical images. The model fails to segment Atelectasis, Consolidation, and Pleural Effusion. This is also aligned with the observations in Figure 4.3 that one can hardly identify them simply using the cropped regions.

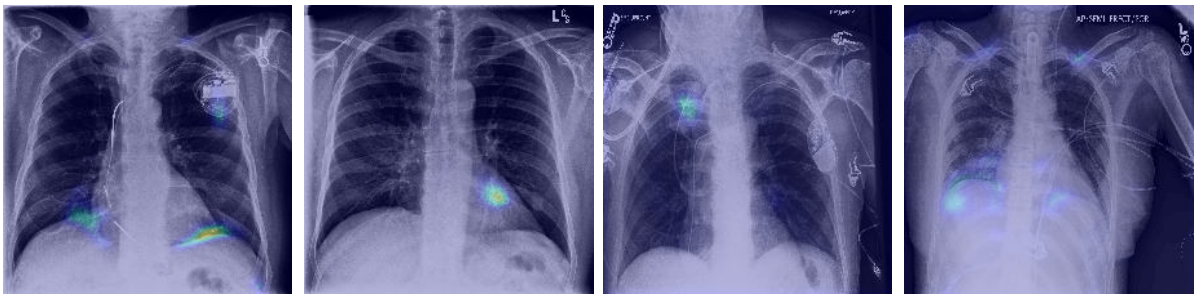


Figure 4.6: Segmentation heatmap for normal patients.

	Filtering	Cropping
Atelectasis	0.347	0.007
Cardiomegaly	0.488	0.138
Consolidation	0.332	0.094
Edema	0.447	0.160
Pleural Effusion	0.275	0.065

Table 4.4: Comparison of using image filtering and cropping.

The influence of α: In Figure 4.7 we plot the mIOU over α values from 1.0 to 0.0001 for all five Chexlocalize classes. When α is large, our model is heavily penalized for generating large masks and the mIOUs are lower. The accuracy first increases as α decreases, then starts to decrease after a certain α. Although the performance is largely affected by α, the scores are generally high for α between 0.0001 and 0.005, and some of the lowest scores still exceed those of other models.

Prompt engineering: Finally, we explored the effect of different prompts. We created a new pair of prompts, "$\{target\}$" for the positive prompt and "no $\{target\}$" for the negative. This pair is also used by Jang et al. (2022) for chest X-ray classification. We name our prompt as P_{default} and the new one as P_{new}. The performance comparison is shown in Table 4.5. Our P_{default} yields better performance.

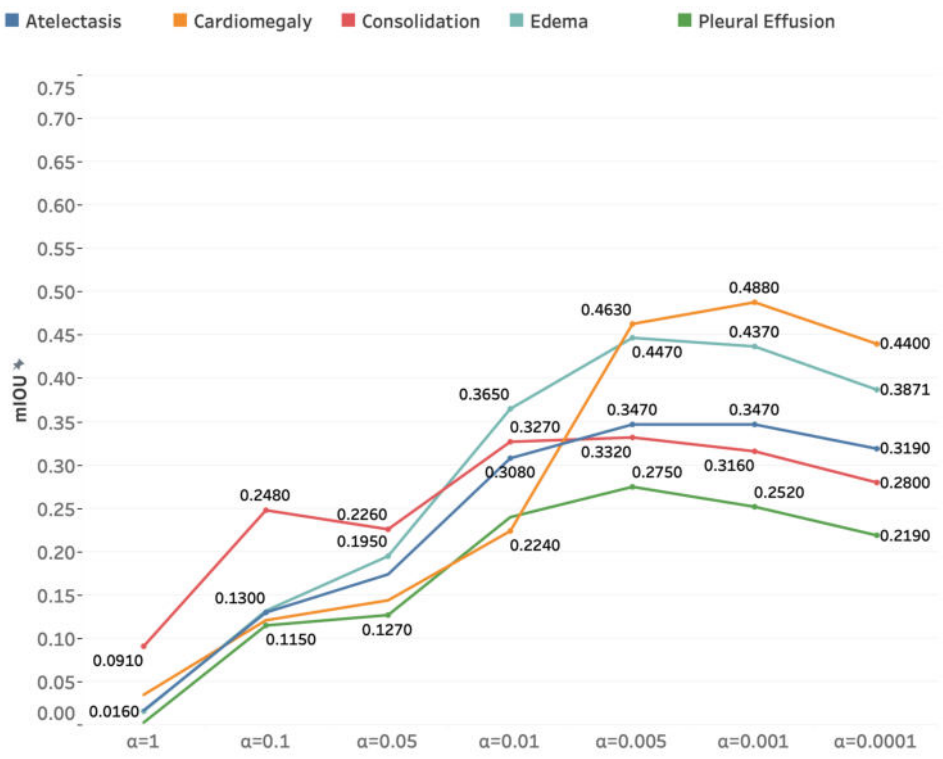


Figure 4.7: mIOU of five categories under different α values.

	P_{default}	P_{new}
Atelectasis	0.347	0.343
Cardiomegaly	0.488	0.373
Consolidation	0.332	0.286
Edema	0.447	0.423
Pleural Effusion	0.275	0.241

Table 4.5: mIOU for P_{default} and P_{new}

CHAPTER 5

Medical Report Generation from Medical Images

5.1 Methodology

5.1.1 Parameter-Efficient Fine-Tuning (PEFT) LLMs on Medical Datasets

Given the huge number of parameters, it is expensive and time-consuming to fine-tune LLMs. In addition, fine-tuning may also cause the model to lose the abilities learned from the prior extensive training. Hence, instead of fine-tuning the whole model, we use a PEFT method called Low-Rank Adaptation of Large Language Models (LoRA) (Hu et al., 2022). LoRA is based on the hypothesis that, the fine-tuned weights $W_{new} \in \mathbb{R}^{d \times k}$ are the original pretrained weights added by a low-rank sparse matrix $W_{new} = W_0 + \Delta W$. The low-rank matrix ΔW is further decomposed as the product of two smaller matrices $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times k}$, where r is the rank. During training, the pretrained weights W_0 are fixed. Only A and B are trainable. This method yields a large reduction in memory and storage usage. LoRA is suitable for our situation because we are fine-tuning a medical LVLM that is previously trained on medical images. Given that the training image domains only have small differences, it is reasonable to assume that W_{new} and W_0 are close and the assumption of LoRA holds.

5.1.2 LLM Prompt Engineering

Prompt Engineering is a popular strategy to align LLMs with specific tasks. In our work, we provide the LLM with instructional prompts and fine-tune it for report generation. The prompts serve as a step-by-step guide that tells the model how to write a medical

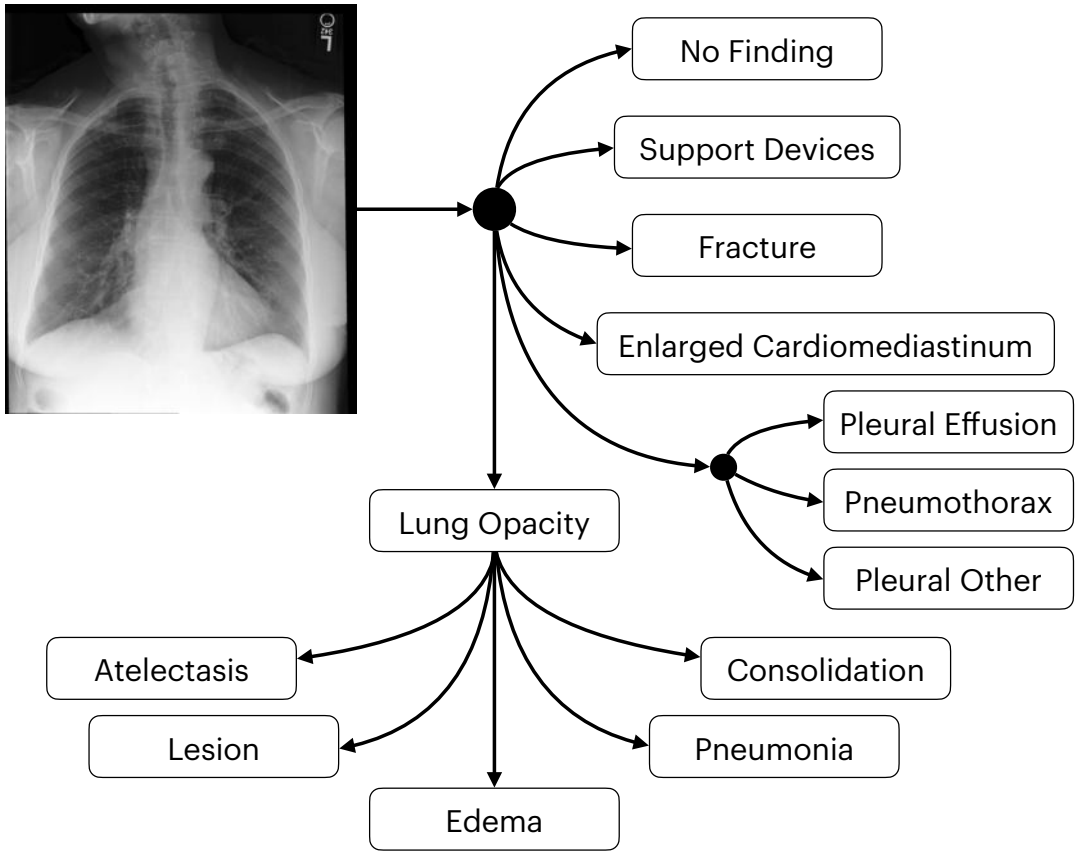


Figure 5.1: The flow graph of describing a chest X-ray.

report. A template is shown below.

> *Instructions: You are a helpful radiology assistant. Firstly describe what lines, tubes, and devices are present and each of their locations. Describe if fracture is present. Describe if there is any pleural abnormality. If there is, check if it is pneumothorax, pleural effusion, or other pleural disease. Describe if there is enlarged cardiomediastinum. If it is, check if there is cardiomegaly. Then, describe if lung opacity is present; if present, check whether atelectasis, consolidation, edema, lung lesion or pneumonia exists.*

The template follows the decision making process in (Irvin et al., 2019), shown in Figure 5.1.

5.1.3 Visual Contrastive Decoding

Visual contrastive decoding is designed to mitigate the over-reliance of VL models on the statistical bias and language priors. It is applied to the LLM decoding stage. Traditionally, the model simply selects the next word w_n with the maximum probability $p(w_n|w_{prev}, v)$, where w_{prev} is the text sequence before w_n and v is the input image. Visual contrastive decoding introduces a second probability $p(w_n|w_{prev}, v')$, where v' is obtained by distorting the input image with heavy noise. Then, the contrastive distribution is defined as

$$p_{vcd}(w_n|w_{\text{prev}}, v, v') = \text{softmax}[(1+\alpha)\,\text{logit}(w_n|w_{\text{prev}}, v) - \alpha\,\text{logit}(w_n|w_{\text{prev}}, v')]. \quad (5.1)$$

The contrastive distribution is designed to penalize the entire outputs affected by v'. This may potentially result in unfluent or unreasonable generations. Hence, p_{vcd} is only used for words with high probabilities:

$$p_{\text{vcd}}(w_n|w_{\text{prev}}, v) \geq \beta \max[p_{\text{vcd}}(w_n|w_{\text{prev}}, v)]. \quad (5.2)$$

In other words, in the output vocabulary, their contrastive probabilities are set as 0.

5.2 Experiments and Results

5.2.1 Datasets

We use the MIMICCXR (Goldberger et al., 2000) and Chexpert (Irvin et al., 2019) datasets. The MIMICCXR training set is used to fine-tuned the model. The test sets of MIMICCXR and Chexpert are used for testing. The model is evaluated across the same five diseases in Chapter 3 (Atelectasis, Cardiomegaly, Consolidation, Edema and Pleural Effusion). Table 5.1 counts the positive cases for the five diseases in the training set. Consolidation and Edema are minority categories. Even for Atelectasis, Cardiomegaly and Pleural Effusion, their positive cases are much fewer compared with the normal cases.

	Count
Atelectasis	65,047 (17.2%)
Cardiomegaly	64,346 (17.1%)
Consolidation	14,675 (3.9%)
Edema	36,564 (9.7%)
Pleural Effusion	76,957 (20.4%)

Table 5.1: Count of positive cases in the MIMIC-CXR-JPG training set. The percentages of the entire training data are bracketed.

	Atelectasis	Cardiomegaly	Consolidation	Edema	Pleural Effusion	Micro
(Chen et al., 2021)	31.1	40.4	9.4	33.8	36.6	34.6
(Miura et al., 2021)	53.0	46.6	5.5	58.5	73.1	56.7
(Yang et al., 2023)	-	-	-	-	-	35.2
(Wang et al., 2022b)	-	-	-	-	-	48.8
(Nicolson et al., 2023)	34.2	54.9	9.9	30.3	54.8	44.2
BLIP	30.8	38.2	4.0	34.7	53.5	39.0
Ours	31.8	42.5	0.6	18.4	51.0	38.6

Table 5.2: F1 score comparison for state-of-the-art medical report generation models on 5 categories of diseases on MIMIC.

5.2.2 Implementation Details

In our experiments, we use LLaVA-Med (Li et al., 2023a). The zero-shot performance on the MIMICCXR (Goldberger et al., 2000) and Chexpert (Irvin et al., 2019) datasets is first evaluated. Then, we fine-tune the model and apply both instructional prompts and visual contrastive decoding.

To evaluate the diagnosis accuracy of the generated reports, we use the tools in Irvin et al. (2019) to automatically parse and label the reports on the key medical finding categories. We mainly compare the F1 scores across the 5 categories used by the classification task in Section 3.2 (Atelectasis, Cardiomegaly, Consolidation, Edema and Pleural Effusion).

		Vanila	Fine-tuned on MIMIC	Inst	VCD
Atelectasis	Precision	20.8	32.1	32.0	31.8
	Recall	14.3	33.0	31.3	32.6
	F1	20.0	32.6	31.7	31.8
Cardiomegaly	Precision	26.0	30.0	29.7	29.6
	Recall	16.3	69.8	74.6	75.4
	F1	20.0	42.0	42.5	42.5
Consolidation	Precision	7.0	0	0	11.1
	Recall	4.6	0	0	0.3
	F1	5.5	0	0	0.6
Edema	Precision	10.7	48.3	47.6	40.7
	Recall	0.8	4.4	4.1	11.9
	F1	1.5	8	7.5	18.4
Pleural Effusion	Precision	28.2	58.6	57.5	56.2
	Recall	19.6	51.3	47.2	46.8
	F1	23.1	54.7	51.8	51.0

Table 5.3: Diagnosis accuracy of generated reports measured by the Precision, Recall and F1 scores of 5 diseases on MIMIC-CXR-JPG test set.

5.2.3 Results

Table 5.2 compares the best diagnostic accuracy we achieve using LLaVA-Med and that of other medical report generation models (Row 1-5). All the models are traditional VL models. We also compare the same pretrained BLIP model (Row 6) in Section 3.1. It is first trained in a multi-task scheme that minimizes image-text contrastive loss and language generation loss. Then the pretrained model is fine-tuned solely on the report generation task, with larger image size (384×384).

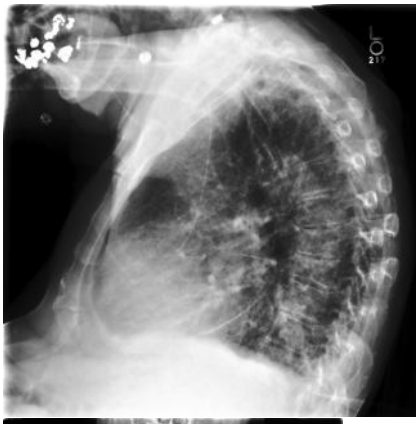
The results show that, compared with traditional medical report generation models, LLaVA-Med does not show an advantage. Figure 5.2 and Figure 5.3 show examples of correctly and incorrectly generated reports. The comparison between the ground truth reports and the generated reports clearly shows that similarity scores may not be suitable for medical report generation. In the examples, a single disease can be described through various valid expressions; e.g., “cardiomegaly” and “enlargement of the cardiac sihouette”.

		Vanila	Fine-tuned on MIMIC	Inst	VCD
Atelectasis	Precision	23.1	56.8	50.9	53.9
	Recall	10.1	25.8	31.5	35.4
	F1	14.1	35.5	38.9	42.7
Cardiomegaly	Precision	24.4	42.3	40.8	34.9
	Recall	16.0	54.9	64.6	62.3
	F1	19.3	47.8	50	44.8
Consolidation	Precision	0	0	0	0
	Recall	0	0	0	0
	F1	0	0	0	0
Edema	Precision	13.3	60.0	72	54.2
	Recall	2.35	3.53	21.2	15.3
	F1	4	6.67	32.7	23.9
Pleural Effusion	Precision	16.2	52.6	47.5	50.4
	Recall	17.5	50.8	46.7	47.5
	F1	16.8	51.7	47.1	48.9

Table 5.4: Diagnosis accuracy of generated reports measured by the Precision, Recall and F1 scores of 5 diseases on Chexpert test set.

Conversely, an expression similar to the ground truth might still be considered invalid.

We have conducted ablation studies on our two methods. We first evaluated the original LLaVA-Med (vanilla). Next we fine-tuned it on MIMICCXR for report generation with and without the instructional prompt. Lastly, we applied visual contrastive decoding to the LLaVA-Med fine-tuned with the instructional prompt. The results are reported in Table 5.3 and Table 5.4. Vanilla LLaVA-Med has the poorest diagnostic accuracy. The accuracy can be largely improved by fine-tuning it on MIMICCXR. The use of instructional prompting improves the performance on Chexpert while not on MIMIC. Visual contrastive decoding shows improvement only on a few categories. All settings fail in the diagnosis of Consolidation because of the imbalanced training data.



Ground Truth Report:
there may be some minimal residual elevation of pulmonary venous pressure and **small pleural effusion** with **bibasilar atelectasis**. central **catheter** remains in place.

Generated Report:
the patient has been **extubated** and the nasogastric tube was removed. the lung volumes have slightly decreased. the patient shows signs of mild fluid overload and a **small right pleural effusion**. **atelectasis at both lung bases**, left more than right. no pneumonia.

Ground Truth Report:
no evidence of pneumonia. **borderline size of the cardiac silhouette** without pulmonary edema. no pleural effusions. no inhomogeneous bone structure. mild tortuosity of the thoracic aorta.

Generated Report:
there is again **enlargement of the cardiac silhouette** without vascular congestion, pleural effusion, or acute focal pneumonia.

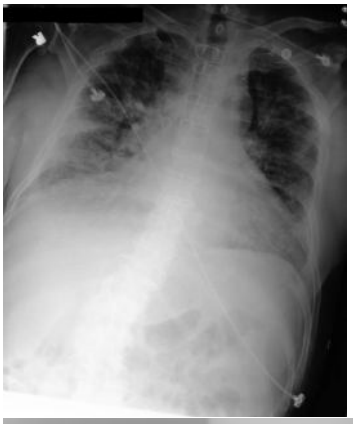
Ground Truth Report:
lung volumes remain low, **accentuating the cardiac silhouette** and bronchovascular structures. with this limitation in mind, cardiomediastinal contours are stable in appearance. persistent elevation of left hemidiaphragm with adjacent **atelectasis at the left lower lobe**. **right retrocardiac atelectasis** is also similar to the prior study.

Generated Report:
the lung volumes have decreased. there is a newly appeared **retrocardiac atelectasis**. moderate **cardiomegaly** persists. no overt pulmonary edema. no larger pleural effusions. no pneumonia.

Ground Truth Report:
no evidence of acute intrathoracic process.', 'no evidence of the et tube.

Generated Report:
there is little change and no evidence of acute cardiopulmonary disease. no pneumonia, vascular congestion, or pleural effusion.

Figure 5.2: Examples of correctly generated reports.



Ground Truth Report:
mild pulmonary edema, **mild pulmonary edema** and **mild-to-moderate cardiomegaly** with some mediastinal widening is worsened.

Generated Report:
there is little overall change. again there is **enlargement of the cardiac silhouette** with diffuse bilateral pulmonary opacifications consistent with **pulmonary edema**. in the appropriate clinical setting, superimposed **pneumonia** would have to be considered. the right ij **catheter** is unchanged.

Ground Truth Report:
one of the right chest tubes appears to have been removed. no definite pneumothorax is appreciated. post-surgical changes persist in the right hemithorax and there is extensive subcutaneous gas along the right lateral chest wall.

Generated Report:
there is little overall change. again there is **enlargement of the cardiac silhouette** with some elevation of pulmonary venous pressure. no evidence of acute focal pneumonia.

Ground Truth Report:
comparison is made with prior study performed a day earlier. interstitial opacities in the right lung have minimally increased, likely due to **edema**. in the right upper hemithorax several air-fluid levels are more conspicuous than in prior studies .. right perihilar opacity is grossly unchanged, allowing the difference in position of the patient. the left lung is clear. the right chest **tube** remains in unchanged position. right chest wall subcutaneous emphysema has improved.
Generated Report:
there is a new parenchymal opacity in the right upper lobe. in addition, there is a newly appeared right **pleural effusion**. the changes are highly suggestive of **pneumonia**. no other parenchymal changes. normal size of the cardiac silhouette. no pulmonary edema.

Ground Truth Report:
no acute intrathoracic process.

Generated Report:
there is no relevant change. the lung volumes are low. **moderate cardiomegaly** with mild fluid overload but no overt pulmonary edema. no pleural effusions. no pneumonia.

Figure 5.3: Examples of incorrectly generated reports.

		VQA	RG			VQA	RG
Atelectasis	Precision	31.0	50.9	Cardiomegaly	Precision	27.1	40.8
	Recall	44.4	31.5		Recall	20.0	64.6
	F1	36.5	38.9		F1	23.0	50.0
Consolidation	Precision	6.0	0	Edema	Precision	11.7	72.0
	Recall	40.0	0		Recall	29.4	21.2
	F1	10.4	0		F1	16.8	32.7
Enlarged Cardiomediastinum	Precision	49.3	28.6	Fracture	Precision	0.9	0
	Recall	12.4	0.7		Recall	33.3	0
	F1	19.8	1.3		F1	1.8	0
Lung Lesion	Precision	2.0	0	Lung Opacity	Precision	50.0	81.3
	Recall	71.4	0		Recall	70.3	25.2
	F1	3.9	0		F1	58.5	38.4
Pleural Effusion	Precision	22.3	47.5	Pneumonia	Precision	2.4	6.3
	Recall	49.2	46.7		Recall	21.4	7.1
	F1	30.7	47.1		F1	4.4	6.7
Pneumothorax	Precision	0	0	Support Devices	Precision	48.6	78.5
	Recall	0	0		Recall	81.6	63.8
	F1	0	0		F1	60.9	70.4

Table 5.5: Comparison of diagnostic accuracy (Precision, Recall, F1) for LLaVA-Med VQA and report generation (RG) on Chexpert.

5.3 Factual Mismatch Between VQA and Report Generation

Table 5.5 compares the zero-shot LLaVA-Med VQA accuracy and the zero-shot diagnostic accuracy of the reports generated by LLaVA-Med, on the Chexpert dataset. Both experiments are on the vanilla setting, without any fine-tuning or additional prompts. Notably there exists a significant inconsistency between the two tasks. On Atelectasis, Lung Opacity and Support Devices, the VQA results show a low Precision and high Recall, while the generated reports have high Precision and low Recall. On Consolidation, Fracture, and Lung Lesion, the extremely low Precision and high Recall may suggest a large number of FP predictions, while for report generation the model simply predicts most cases as negative.

One of the reasons leading to the mismatch could be insufficient training. Minority

classes such as Consolidation, Fracture, Lung Lesion, Pneumonia and Pneumothorax, are rarely mentioned in the training corpus; hence, the generated reports tend to omit them. When the model is asked about the presence of these diseases, it is unable to answer because it does not learn their features well.

CHAPTER 6

Conclusions, Discussion, and Future Work

This thesis explored the application of language models to medical image analysis. It focused on three tasks, medical image classification, medical image segmentation, and medical report generation from medical scans.

Medical Image Classification with Language Models: For medical image classification, we tested LLaVA-Med on VQA regarding the diagnosis of diseases and the results show that the model has unsatisfactory performance when asked questions regarding complex diseases. Two prompting strategies were proposed to improve the VQA accuracy: providing descriptions of diseases and providing the predictions of weak learners as references. The first one helped the model understand the minority diseases that it does not learn well at the training stage. Referring the predictions of weak learners can help improve the accuracy on specific aspects; e.g., suppressing FPs. More importantly, this strategy can be extended to the general domain. It is meaningful because our strategy can be applied not only to medical LVLMs but also to LVLMs in other specialized domains. However, the two strategies are not effective on diseases that have extremely scarce data; e.g., Consolidation, Fracture, Lung Lesion, Pneumonia, and Pneumothorax. For these categories, providing text descriptions might not suffice since the visual encoder does not learn the visual features either. Moreover, the data might not suffice to train the weak learners. An approach to handle these rare categories would be a promising direction for future research. Retrieval augmented generation (RAG) could be one potential solution. For a disease, one can provide not only the text description but also typical example images to help the model make decisions.

Medical Image Segmentation with Language Models: For medical image segmentation, we have devised an approach to medical image segmentation in a zero-shot learning manner with text supervision. Crafting two prompts P_{pos} and P_{neg}, we trained our model to generate a mask and used it to filter the original image. The model learns to align the original image with P_{pos} and the filtered image with P_{neg}. When trained on the MIMIC-CXR-JPG dataset and tested on the Chexpert and SIIM-ACR datasets, our zero-shot model outperformed weakly-supervised/semi-supervised full-shot learning models. A limitation of the proposed method is that it cannot be applied to low-intensity ROIs (e.g., air in organs). For these ROIs, the mask should be added instead of subtracted to make them normal. This can be addressed in future work. Moreover, it could be difficult to obtain a normal image simply by adding a mask to the original image directly. A synthetic image generation module can be considered. Another limitation is the minority classes; these appear rarely in the training data and the model may not be able to learn their features well. Hence, the model would be unable to distinguish P_{pos} and P_{neg} and the decoder would not be guided during training. Lastly, the choice of α has significant influence on the segmentation accuracy. In future work, instead of presetting, a learnable α may be desired.

Medical Report Generation: For medical report generation, we evaluated the accuracy of LLaVA-Med in terms of diagnostic accuracy. The results showed that it suffers from severe hallucination. To mitigate hallucination, we mainly applied two strategies, instructional prompting and contrastive decoding. Neither of the two strategies yielded significant improvement in accuracy. This indicates that the over-reliance on unimodal priors might not be the main cause of hallucination for medical LVLMs. The failure of the visual encoder may be the core problem. Most LVLMs use the ViT encoder, but it could be extremely hard for the ViT to learn the features of multiple diseases from multiple medical image modalities. Moreover, the imbalanced training data makes it more difficult to learn minority diseases. Lastly, the observed performance mismatch between the VQA and report generation tasks is meaningful. Mostly previous models were tested on either

VQA or captioning, while the consistency between them was rarely studied. In future work, the consistency between the two tasks could potentially serve as a new metric to evaluate the hallucination of LVLMs.

APPENDIX A

Medical Image Classification by Contrastive Learning

A.1 Classification with Pretrained VL Models

The method of image classification using language models follows ConVIRT and CLIP. The image encoder and text encoder are trained on the image captioning dataset using contrastive learning. The two encoders are trained to maximize the similarity between correct image-caption pairs while minimizing that between wrong image-caption pairs. This is usually achieved by computing the cosine similarity between all image and text features and applying cross-entropy loss or contrastive loss. At the inference stage, one matches the encoded image features with a sequence of class text features and selects the one with the highest score as the prediction.

A.1.1 Model Architecture

Our model is trained for two tasks: image-text contrastive learning and medical report generation, with the latter being an auxiliary task. The overall model architecture (Figure A.1) follows BLIP, which consists of three modules: a visual encoder, a text encoder, and a multimodal text decoder.

Visual/Text Encoder: The visual encoder is ViT encoder and the text encoder is a multi-layer transformer. They are unimodal encoders that extract visual features and text features respectively.

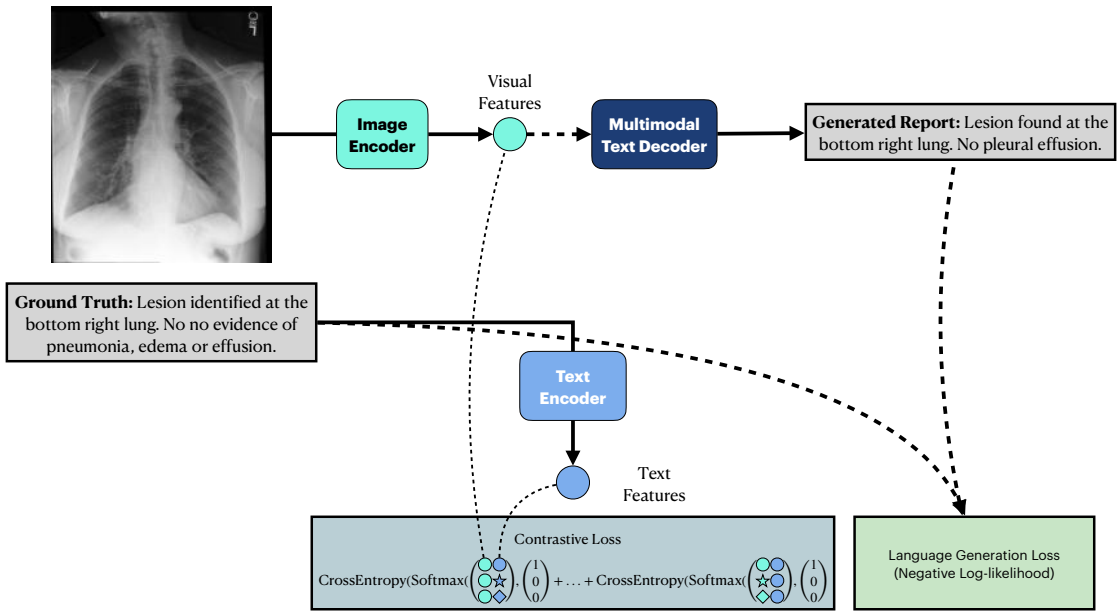


Figure A.1: Multitask training for contrastive learning models. The image encoder and text encoder are trained to minimize the contrastive loss and the language generation loss.

Multimodal Decoder: Similar to the multimodal encoder, it also shares parameters with the text encoder. The bi-directional self-attention is replaced by uni-directional self-attention. A special token is used to indicate the decoding mode.

A.1.2 Multitask Losses

Image-Text Contrastive Loss: This is the loss used by Li et al. (2021). It is the sum of softmax text-to-image and image-to-text similarity:

$$L_{t2i} = \frac{\exp S(I, T_j)}{\sum_{j \in N} \exp S(I, T_j)}, \quad L_{i2t} = \frac{\exp S(I_j, T)}{\sum_{j \in N} \exp S(I_j, T)}, \tag{A.1}$$

where S is the similarity computed using the image/text features encoded by the image/text encoder. We use cosine similarity for S.

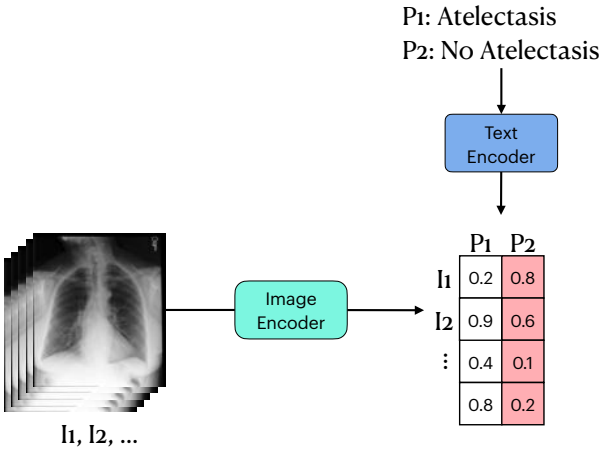


Figure A.2: Medical image classification using contrastive learning

Language Generation Loss: The visual encoder and the multimodal decoder also perform the report generation task. The loss function is the cross-entropy loss.

A.1.3 Inference

The inference steps are shown in Figure A.2. To perform disease classification, we create two prompts, "{*target*}" and "*no* {*target*}", where "target" is the disease word. They are encoded by the text encoder and matched with the image features. The one with higher similarity becomes the prediction result.

A.2 Datasets, Implementation Details, Experiments, and Results

We use the same datasets as the LLaVA-Med experiments: MIMICCXR and Chexpert. The model is trained on the MIMICCXR training set and tested on both the MIMICCXR and Chexpert test sets.

Table A.1 and Table A.2 report the classification F1 scores of the 5 disease categories on the MIMICCXR and Chexpert datasets. We compare the pure contrastive learning

	Plain	Plain + NLG
Atelectasis	17.44	41.5
Cardiomegaly	19.25	47.4
Consolidation	14.90	14.3
Edema	41.30	46.0
Pleural Effusion	61.10	64.0

Table A.1: Classification performance evaluated by F1 scores of 5 diseases on the MIMIC-CXR-JPG test set.

	Plain	Plain + NLG
Atelectasis	12.5	58.5
Cardiomegaly	13.8	61.8
Consolidation	17.6	18.5
Edema	42.7	49.7
Pleural Effusion	28.1	59.3

Table A.2: Classification performance evaluated by F1 scores of 5 diseases on the Chexpert test set.

setting and the multitask setting (contrastive learning and language generation). Adding the report generation task generally leads to a significant improvement.

Compared with the performance of LLaVA-Med in Table 3.4, BLIP outperforms LLaVA-Med on Cardiomegaly, Consolidation, Edema, and Pleural Effusion and has comparable performance on Atelectasis. Smaller models like BLIP still have advantages over large models.

APPENDIX B

Prompts with Medical Explanations

The explanations of medical findings used in Chapter 3 are listed below.

Atelectasis: Atelectasis refers to the partial or complete collapse of a lung or a section of lung. The features of atelectasis on an X-ray can vary depending on the cause and extent of the collapse. Some common X-ray features include: 1. The affected area may appear denser or whiter than normal lung tissue due to the collapse, leading to increased opacity on the X-ray. 2. The affected portion of the lung may appear smaller or compressed compared to the surrounding healthy lung tissue. 3. Atelectasis can cause a shift or displacement of nearby structures, such as the trachea or heart, toward the affected area. 4. In obstructive atelectasis (caused by a blockage in the airways), there might be signs of hyperinflation in the unaffected areas of the lung and a visible blockage or narrowing in the affected bronchus. 5. Linear or band-like opacities may be visible, often referred to as plate or band atelectasis, which can occur due to the collapse of small airways. Given the information above, does this image have Atelectasis?

Cardiomegaly: Cardiomegaly is enlargement of the heart. The definition is when the transverse diameter of the cardiac silhouette is greater than or equal to 50% of the transverse diameter of the chest (increased cardiothoracic ratio) on a posterior-anterior projection of a chest radiograph or a computed tomography. Given the information above, does this image have Cardiomegaly?

Consolidation: Consolidation on an X-ray refers to the filling of the lung's air spaces with fluid inflammatory exudate, or cellular material. Typical X-ray findings suggesting

consolidation include: 1. Areas of increased density in the lung tissue, appearing as an opaque or hazy patch on the X-ray. Given the information above, does this image have Consolidation?

Edema: Pulmonary edema is the accumulation of fluid in the lungs. Some common X-ray features include: 1. Increased density in the central lung fields resembling the shape of bat wings. 2. Thin, linear opacities at the lung periphery, often indicating interstitial edema. 3. Prominent blood vessel markings due to engorgement from increased pressure in the pulmonary vasculature. Given the information above, does this image have Edema?

Enlarged Cardiomediastinum: Enlarged cardiomediastinum refers to both an enlarged heart and widened mediastinum (the space in the middle of the chest containing the heart and other structures). Some common X-ray features include: 1. The width of the heart compared to the width of the chest appears larger than normal. The heart occupies more than 50% of the chest width on the X-ray. 2. The heart's outline appears larger and may extend beyond its usual boundaries, indicating cardiac enlargement. 3. The space between the lungs where the heart, major blood vessels, and other structures reside appears wider than normal. Given the information above, does this image have Enlarged Cardiomediastinum?

Fracture: X-ray findings that suggest the presence of a fracture typically include: 1. A line or gap in the normal bone structure. This could be a visible break, crack, or irregularity in the bone's smooth surface. 2. Bone segments may appear displaced or misaligned compared to their normal anatomical position. 3. Swelling or soft tissue changes around the site of the suspected fracture. 4. widening of the bone at the fracture site. Given the information above, does this image have Fracture?

Lung Lesion: Lung lesion could include tumors, nodules, or other abnormalities. Some common X-ray features include: 1. An abnormal area in the lung that appears denser or more opaque than the surrounding healthy lung tissue. 2. The lesion may have

well-defined or ill-defined margins. 3. Presence of calcifications within the lesion can sometimes be observed. 4. Any associated changes in the lung tissue surrounding the lesion, such as consolidation, collapse, or scarring.

Lung Opacity: The term "lung opacity" on a chest radiograph refers to areas in the normally dark-appearing lung that appear denser, hazy, or gray.

Pleural Effusion: Pleural effusion is the accumulation of fluid in between the parietal and visceral pleura. Some common X-ray features include: 1. Blunting of the costophrenic/cardiophrenic angle. 2. Fluid within the horizontal or oblique fissures. 3. Meniscus is seen. 4. Mediastinal shift occurs away from the effusion.

Pneumonia: Pneumonia is an infection that inflames the air sacs in one or both lungs. Some common X-ray features include: 1. Areas of increased density in the lung parenchyma, appearing as opacities or infiltrates. 2. Consolidation of an entire lobe or segment of the lung, presenting as a dense and homogeneous opacity with sharp margins. 3. Patchy opacities that may be multifocal and scattered throughout the lung fields.

Pneumothorax: Pneumothorax occurs when air accumulates in the pleural space. Some common X-ray features include: 1. A distinct dark area, often without lung markings, between the lung and chest wall on the affected side. 2. Partial or complete collapse of the lung on the affected side due to the presence of air, leading to reduced lung volume and a smaller appearance of the affected lung. 3. The edge of the collapsed lung may be shifted away from the chest wall, leading to a visible separation between the lung edge and the chest wall. 4. An increased angle between the chest wall and the diaphragm due to the absence of lung tissue in the pleural space. 5. In severe cases, a tension pneumothorax can cause displacement of the mediastinal structures (trachea, heart) toward the unaffected side.

Support Devices : Common support devices in chest X-ray include tubes, oxygen masks, sensor attachments, electrodes, catheters, probes.

APPENDIX C

Additional Experiments With the LLaVA-Med VQA Model

Table C.1 and Table C.2 report the results of LLaVA-Med PLAIN setting and EXP setting across another 7 key medical findings: Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, Pneumonia, Pneumothorax, and Support Devices on the MIMICCXR and Chexpert datasets. Providing diseases explanations generally yields better results, but this is not obvious. For Fracture, Lung Lesion, and Pneumothorax, where the Precision scores are extremely low, the EXP setting has almost no improvement. Another special category is Support Devices, which refers to any medical devices implanted into the chest. In the explanations, we provide only the names of several typical support devices. The model might have more difficulties understanding these specific names.

Diseases	Metrics	PLAIN	EXP
Enlarged Cardiomediastinum	Precision	4.3	3.9
	Recall	15	89
	F1	6.7	7.4
Fracture	Precision	3.7	2.7
	Recall	41.3	24.0
	F1	6.7	4.9
Lung Lesion	Precision	3.9	3.9
	Recall	77.2	100
	F1	7.4	7.5
Lung Opacity	Precision	31.4	30.4
	Recall	67.6	88.8
	F1	42.8	45.3
Pneumonia	Precision	11.4	10.5
	Recall	20.0	74.6
	F1	14.6	18.4
Pneumothorax	Precision	3.0	2.6
	Recall	16.7	78.5
	F1	5.1	5.1
Support Devices	Precision	29.2	27.7
	Recall	81.3	51.8
	F1	43	36.12

Table C.1: LLaVA-Med VQA performance of 7 diseases on the MIMICCXR-JPG test set

Diseases	Metrics	PLAIN	EXP
Enlarged Cardiomediastinum	Precision	49.3	44.1
	Recall	12.4	85.2
	F1	19.8	58.1
Fracture	Precision	0.9	0.6
	Recall	33.3	16.7
	F1	1.8	1.1
Lung Lesion	Precision	2.0	2.1
	Recall	71.4	100
	F1	3.9	4.1
Lung Opacity	Precision	50.0	47.2
	Recall	70.3	90.7
	F1	58.5	62.1
Pneumonia	Precision	2.4	1.7
	Recall	21.4	57.1
	F1	4.4	3.4
Pneumothorax	Precision	0	1.7
	Recall	0	90.0
	F1	0	3.3
Support Devices	Precision	48.6	42.6
	Recall	81.6	45.1
	F1	60.9	43.8

Table C.2: LLaVA-Med VQA performance of 7 diseases on Chexpert test set

APPENDIX D

Medical Image Segmentation with Transformers

D.1 Introduction

With the success of transformers (Vaswani et al., 2017) in language models, attempts have been made to apply them to computer vision models. Besides replacing 2D convolution layers with transformers (Dosovitskiy et al., 2021; Liu et al., 2021b), transformers can also be applied to address the anisotropic problem of 3D medical images.

Deep learning models trained on a dataset with some specific slice spacing may perform poorly on clinical images with a different slice spacing. The conventional approach to dealing with the variable slice spacing problem is to "re-slice" all images such that they have a common spacing and implement 3D convolutional neural networks (CNNs) that encode information across the slices (Imran et al., 2020). Anisotropic convolutional kernels and hybrid 2D/3D convolutions have also been employed. Ideally, however, a model should adapt to variable slice spacing, for example, by replacing 3D convolutions with recurrent networks to process information along the z-axis.

In our work, we propose a transformer-based approach to address the anisotropy problem. We modify the transformer and apply it to 3D medical images. Our model uses a self-attention mechanism to encode inter-slice information. It adapts to variable slice spacing, is computationally efficient, and consumes fewer resources compared to 3D convolutional and recurrent networks.

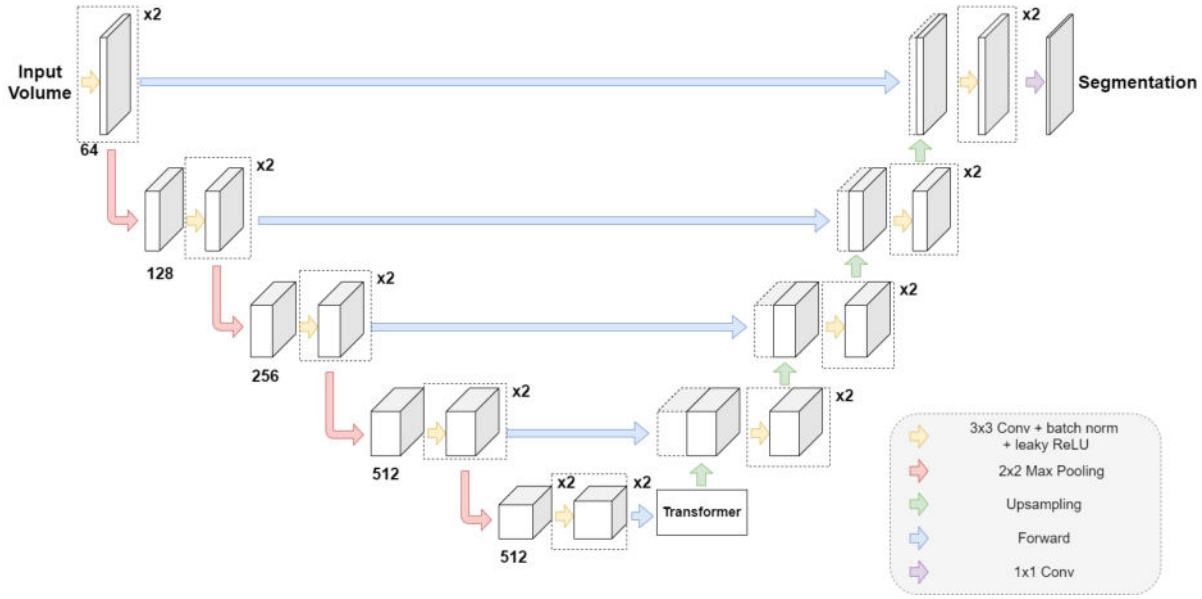


Figure D.1: Model architecture. Dashed rectangles denote structure repetition. The numbers below feature maps indicate the number of channels.

D.2 Methods

The core idea of our approach is to encode the information along the z axis by representing the feature maps of a slice as a weighted sum of these maps and the feature maps of its neighboring slices. Thus, all the feature maps of a sequence of slices are used to compute the weight distribution for each slice. Such a weight distribution reflects how neighboring slices are correlated with the target slice, and the weight decreases as the slice spacing grows.

D.2.1 Task Formulation

We define our task as one of 3D semantic segmentation, because compared with tasks such as classification or detection, segmentation requires more spatial information since the regions of interest may have complex 3D structures and the anisotropy may have more influence on the result. Hence, given a 3D lung CT image $\mathcal{I}$ of size $N_x \times N_y \times N_z$, the task is to classify each pixel into one of two classes: cancer (1) and other (0). The model must learn the mapping from $\mathcal{I}$ to its label $\boldsymbol{\zeta} \in \{0,1\}^{N_x \times N_y \times N_z}$.

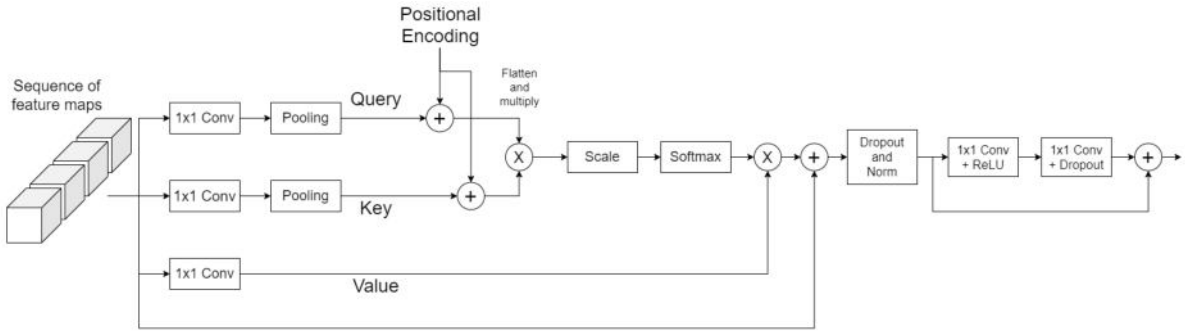


Figure D.2: Architecture of the transformer.

D.2.2 Network Structure

As shown in Figure D.1, our model uses the popular 2D U-Net (Ronneberger et al., 2015) as a backbone. It consists of a down-sampling stream, an up-sampling stream, and a transformer block. The feature maps of the down-sampling layers are forwarded to the corresponding up-sampling layers. At the bottom layer, the feature maps of slices are grouped as input sequences and passed to the transformer.

We use a transformer only at the bottom layer so that we can compare our approach fairly with a recurrent network structure, Convolutional LSTM (ConvLSTM) (Shi et al., 2015). If we were to use a transformer in each layer, then so should the ConvLSTM. However, this would greatly increase the size of the network and we would not have enough capacity to run experiments. We will discuss the possibility of applying the transformer at higher resolution levels in Section D.4.

Transformer Module: Figure D.2 reveals the architecture of our transformer module. The sequence of bottom layer feature maps is fed into three different local convolutional layers to become the queries Q, keys K, and values V. The d_k-dimensional queries and keys are used to compute a group of softmax weights and the d_v-dimensional values are multiplied by the computed weights. Thus, the new feature map of each slice is a weighted sum of its feature map and those of its neighboring slices. As in (Vaswani et al., 2017),

the attention function is

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{D.1}$$

Both queries and keys are feature maps and they are flattened. The product QK^T captures the correlations between the slices. The weighted sums of values are added back to the original values and normalized. This is passed to a residual module with two convolutional layers.For simplicity and a fair comparison in our study, we use only a single-layer transformer with one head.

Positional Encoding: Unlike recurrent network structures, the transformer cannot know the order (distance) of slices. To address this problem, they injected information about the sequence order. Following their work, we designed Positional Encoding (PE), which has the same dimensionality, d_k, as the queries and keys so that they can be merged. For slice j, where $0 \leq j < N_z$, it is an interleaving of sine and cosine functions:

$$\text{PE}(i, j) = \begin{cases} \sin(j/w^{\frac{i}{d_k}}), & i \text{ even}, \\ \cos(j/w^{\frac{i}{d_k}}), & i \text{ odd}, \end{cases} \tag{D.2}$$

where $0 \leq i < d_k$ and $w = 10^4$. The sinusoidal design allows generalizing to sequence larger than the ones in training set easily. Because the PE of slice $j+n$ can be represented as a linear function of the PE at slice j. Referring to Figure D.2, PE is added to the queries and keys after pooling. Vaswani *et al.* applied positional encoding to the queries, keys, and values. We add PE only to the queries and keys because the positional information is needed only to determine the relation of the slices in order to compute the attention weights. PE is no longer useful once we have the attention weights, so we do not pad it to the values.

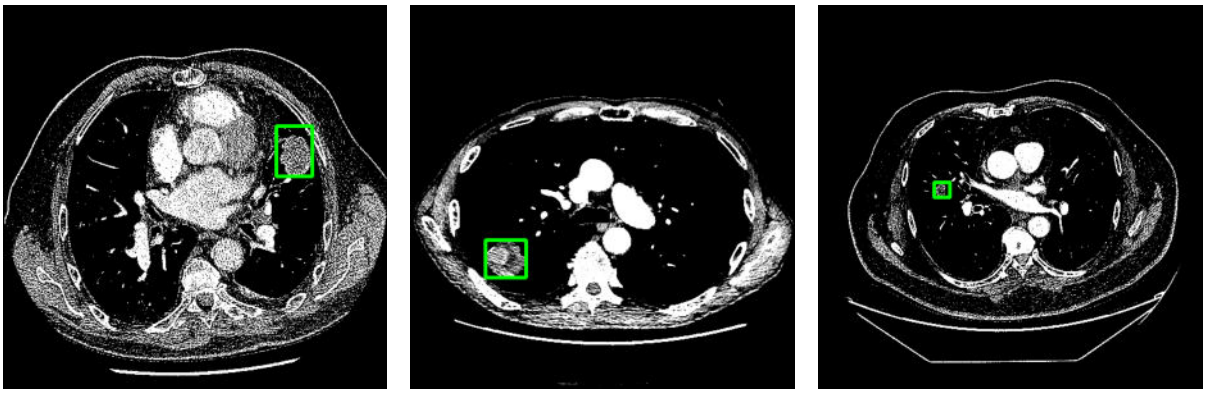


Figure D.3: Examples of lung CT slices. The cancer is highlighted by the green bounding box.

D.2.3 Loss Function

We have designed our loss function as a multi-class soft Dice loss

$$\mathcal{L} = 1 - \frac{2|L \cap S|}{|L| + |S|}, \tag{D.3}$$

where L denotes the ground truth label and S denotes the segmentation output.

D.3 Experiments and Results

D.3.1 Experimental Setup and Baseline Models

We conducted experiments using the lung cancer segmentation dataset published by the Medical Segmentation Decathlon (Simpson et al., 2019). It contains lung CT images as well as corresponding labels from 63 subjects (Figure D.3). All subjects have cancer. There are 20,707 slices in total, among which 2,027 are positive. The size of the images is 512×512 with the number of slices ranging from 112 to 636. The average voxel spacing is $0.771 \times 0.771 \times 1.245\,\text{mm}$. The spacing along the z axis is from 0.64 to 4.1 times that in the x, y plane. Given that both the number of positive slices and the size of the cancer lesion are small, instead of using all the negative slices, we sampled only those neighboring the positive slices. There are 1,890 negative slices in total. The image data

were partitioned into 50 subjects for training and validation and 13 for testing.

The intensity of input images were truncated within the range $[0, 200]$ Hounsfield Units in order to minimize the influence of non-ROIs such as air and bones. The truncated images were normalized to $[0, 1]$. The following image augmentation procedures were applied in the training stage: random shearing ($\pm 10\%$), zooming ($\pm 20\%$), rotation ($\pm 90°$), horizontal and vertical flipping, and shifting (± 20 pixels). The Adam optimizer was used and the learning rate was set to 10^{-5} with a decay factor of 0.9 for every 10 epochs.

We used the 3D U-Net and 2D U-Net with ConvLSTM (LSTMUNet) as comparison models. The structure of the LSTMUNet is similar to our transformer model, TSFMUNet, with the difference being that the transformer module is replaced by a one-layer, bi-directional ConvLSTM module. The number of layers, number of channels, and kernel size of the three models were kept as consistent as possible in order to make our comparisons more fair.

We also re-sliced the original training set such that the voxel spacings along the three dimensions are the same. Then we trained our three models on this isotropic dataset. For a meaningful comparison, we tested the models on the original anisotropic test set. The rationale is that by re-slicing the original data, we created a dataset that has a new distribution of voxel spacing different from the original dataset. Then, by testing the models on the original test data, we can compare their ability to adapt to variable spacing. A model is adaptable to variable spacing if it is trained on isotropic images and performs well on anisotropic images. The new dataset contains 2,314 positive slices and 2,142 negative slices. The average voxel spacing is $0.771 \times 0.771 \times 0.771\,\text{mm}$.

D.3.2 Results

We use the Dice score as the evaluation metric in all our experiments. The results of the models trained on the original data and re-sliced data are reported in Table D.1 and Table D.2, respectively. Table D.1 also reports the dice score of the 2D U-Net that serves as the backbone of LSTMUNet and of our TSFMUNet. In Table D.2, all the models were

Table D.1: Dice score comparison (original data).

Model	Dice Score
TSFMUNet	**0.8717**
LSTMUNet	0.8573
3D U-Net	0.7744
2D U-Net	0.7309

trained on isotropic images and tested on the dataset used for Table D.1, which reveals the drop of the Dice score compared with Table D.1. Figure D.4 shows 3D visualizations of the segmentation result of models trained on the original dataset. Figure D.5 shows those for the re-sliced dataset. The tables and figures confirm that our transformer-based model outperforms the baseline models on both datasets.

Comparing the results in the two tables, one sees a performance drop: 3D U-Net > LSTMUNet > TSFMUNet. Figure D.5 also reveals that the LSTMUNet and 3D UNet failed to segment the targets. This is expected because the 3D U-Net uses fixed kernels and therefore has the greatest dependency on voxel spacing. More specifically, the model trained on images with smaller inter-slice spacing will assume a tighter relationship between slices and the 3D kernels will have greater interaction between features from different slices. A large performance drop is to be expected when this model is tested on data with large spacing and less correlation between slices. Compared with the 3D U-Net, the LSTMUNet is less affected because it operates on z-axis information only at the bottom level. However, it too suffers from the same problem because the LSTM uses the same kernel to compute the next state from previous states. Our TSFMUNet achieves the least performance drop because it uses the self-attention mechanism and there are no kernels working along the z axis—inter-slice information is encoded by computing a weighted sum of neighboring slices based on their similarities.

Table D.2: Dice score comparison (re-sliced data).

Model	Dice Score	Performance Drop
TSFMUNet	**0.8674**	**0.0043**
LSTMUNet	0.8217	0.0356
3D U-Net	0.7261	0.0483

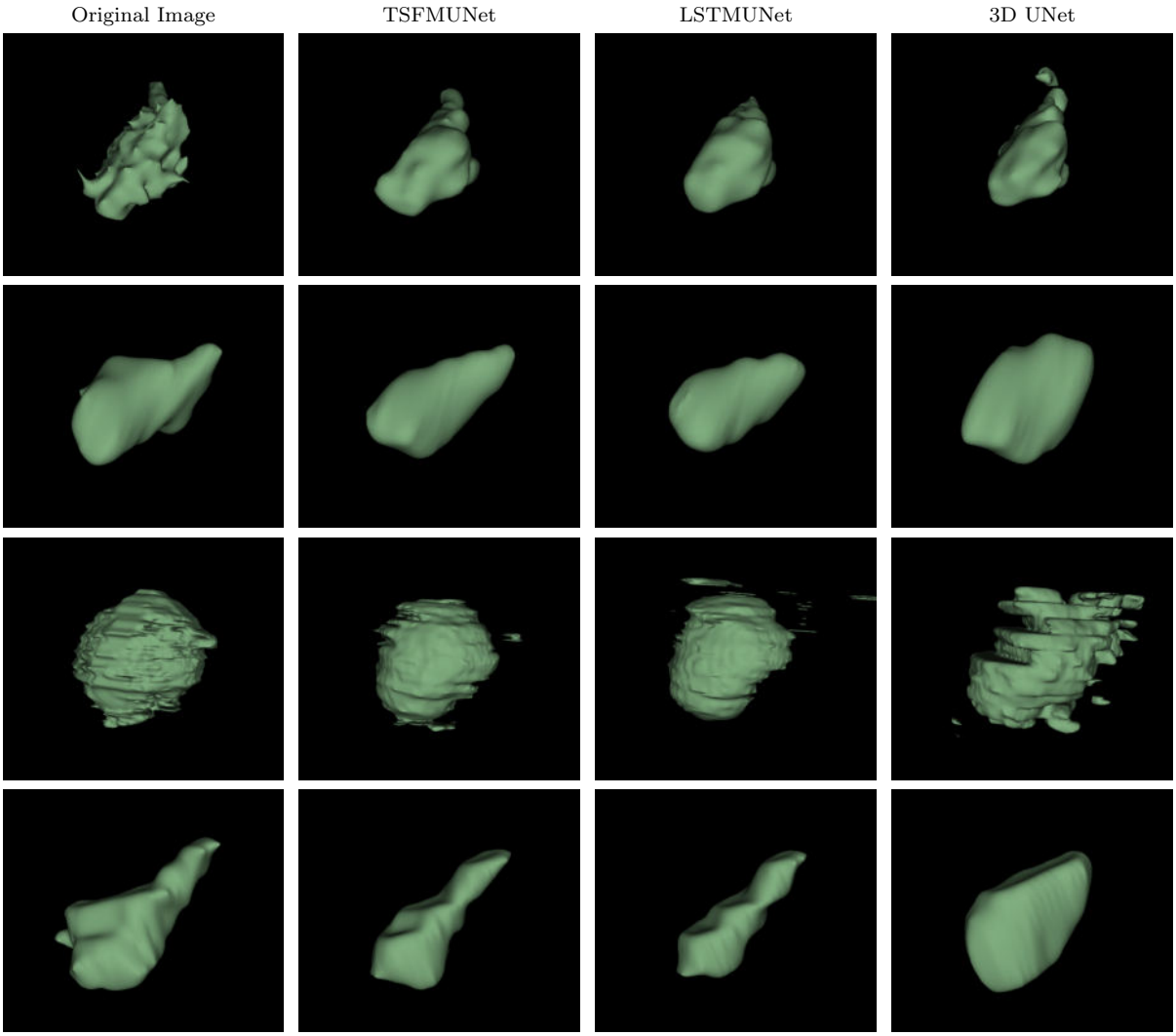


Figure D.4: Visualizations of the segmentation results on models trained on the original dataset. The left column shows the visualizations of ground-truth cancer lesion segmentations and the other columns show segmentation results by the 3 models.

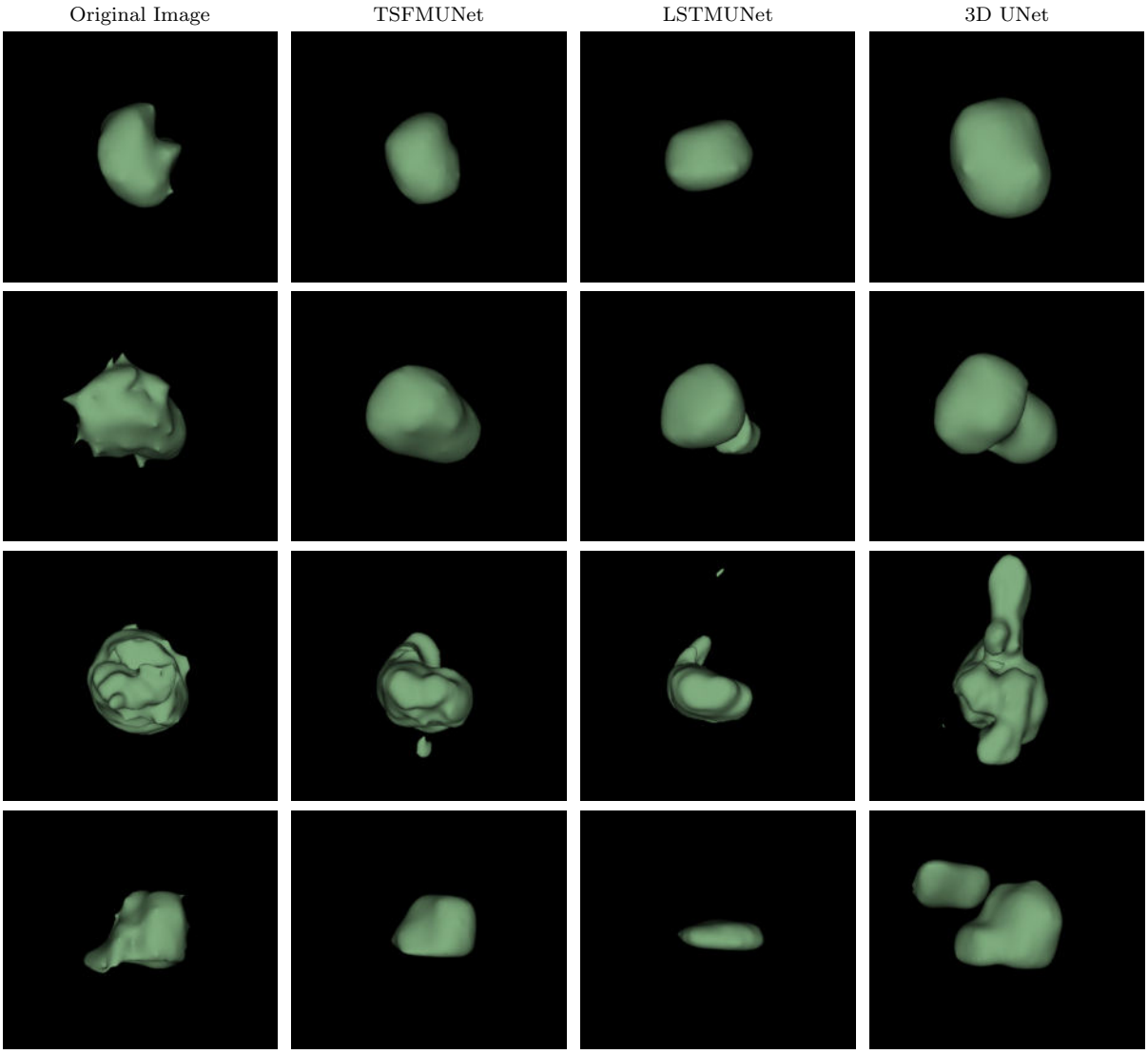


Figure D.5: Visualizations of segmentation results on models trained on re-sliced dataset. The plotted images come from the original non-resliced test dataset. The left column shows the visualizations of ground-truth cancer lesion segmentations and the other columns show segmentation results by the 3 models.

D.4 Conclusions and Discussion

We have proposed a transformer-based network to deal with the anisotropy problem in 3D medical image analysis. Experimental results with a lung cancer segmentation task reveal that our architecture outperforms baseline models. Our TSFMUNet model uses a self-attention mechanism to encode spatial information and it adapts to images with variable slice spacing. Moreover, unlike networks that have recurrent structures, our

model can be parallelized. Our model requires the least capacity among the tested models and is faster to train. To achieve the results in Table D.1, the 3D U-Net includes around 57 million parameters and the LSTMUNet has around 33 million parameters. However, our TSFMUNet model requires only about 21 million parameters.

In our study, for comparison purposes, we used only a one-layer, single-head transformer and incorporated it just on the bottom layer. Consequently, our model encodes inter-slice information only at the low-resolution level. It may capture only the general shape of the ROI but fail to align the detailed texture. In Figure D.4 and Figure D.5, the segmentations generated by our model are too smooth. Given that some detailed features will be lost in the downsampling process, we expect that the performance could be further improved by also using transformers on the other layers to encode texture information at the higher-resolution levels. In addition, recall that in our model the weight distribution of the self-attention mechanism is derived from the product QK^T of the queries and values. This may not be the best approach to capturing the correlation between slices because spatial information is lost when feature maps are flattened. Better representations of slice correlation will be developed in future work.

REFERENCES

Abdal, R., Zhu, P., Mitra, N. J., and Wonka, P. (2021). Labels4Free: Unsupervised segmentation using StyleGAN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13970–13979. 2

Alfarghaly, O., Khaled, R., Elkorany, A., Helal, M., and Fahmy, A. (2021a). Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, 24:100557. 18

Alfarghaly, O., Khaled, R., Elkorany, A., Helal, M., and Fahmy, A. (2021b). Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, 24:100557. 20

Almalik, F., Yaqub, M., and Nandakumar, K. (2022). Self-ensembling vision transformer (SEViT) forÂ robust medical image classification. In Wang, L., Dou, Q., Fletcher, P. T., Speidel, S., and Li, S., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 376–386, Cham. Springer Nature Switzerland. 12

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., and McDermott, M. (2019). Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics. 37

Anuar, A. (2019). SIIM–ACR Pneumothorax Segmentation. `https://github.com/sneddy/pneumothorax-segmentation`. 41

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics. 19

Bearman, A. L., Russakovsky, O., Ferrari, V., and Fei-Fei, L. (2015). What's the point: Semantic segmentation with point supervision. In *European Conference on Computer Vision*. 2, 6, 16

Boag, W., Hsu, T.-M. H., Mcdermott, M., Berner, G., Alesentzer, E., and Szolovits, P. (2020). Baselines for chest X-ray report generation. In Dalca, A. V., McDermott, M. B., Alsentzer, E., Finlayson, S. G., Oberst, M., Falck, F., and Beaulieu-Jones, B., editors, *Proceedings of the Machine Learning for Health NeurIPS Workshop*, volume 116 of *Proceedings of Machine Learning Research*, pages 126–140. PMLR. 8, 20

Caffagni, D., Cocchi, F., Moratelli, N., Sarto, S., Cornia, M., Baraldi, L., and Cucchiara, R. (2024). Wiki-LLaVA: Hierarchical retrieval-augmented generation for multimodal llms. 5, 15

Cai, W., Xie, L., Yang, W., Li, Y., Gao, Y., and Wang, T. (2022). DFTNet: Dual-path feature transfer network for weakly supervised medical image segmentation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 1–12. 16

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., and Wang, M. (2023). Swin-Unet: Unet-like pure transformer for medical image segmentation. In Karlinsky, L., Michaeli, T., and Nishino, K., editors, *Computer Vision – ECCV 2022 Workshops*, pages 205–218, Cham. Springer Nature Switzerland. 16

Cha, J., Mun, J., and Roh, B. (2023). Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 17, 35, 43

Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2018). Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847. 16

Chen, Z., Shen, Y., Song, Y., and Wan, X. (2021). Generating radiology reports via memory-driven transformer. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 49

Chen, Z., Song, Y., Chang, T.-H., and Wan, X. (2020). Generating radiology reports via memory-driven transformer. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, Online. Association for Computational Linguistics. 18

Chen, Z., Tian, Z., Zhu, J., Li, C., and Du, S. (2022). C-CAM: Causal CAM for weakly supervised semantic segmentation on medical image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11676–11685. 16

Cheng, S., Gan, Z., Yang, Z., Wang, S., Wang, J., Boyd-Graber, J., and Wang, L. (2023). Prompting GPT-3 to be reliable. In *International Conference on Learning Representations (ICLR 23)*. 10, 21

Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., and Xing, E. P. (2023). Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality. 2, 4, 19, 22

Chuang, Y.-S., Xie, Y., Luo, H., Kim, Y., Glass, J., and He, P. (2023). DoLa: Decoding by contrasting layers improves factuality in large language models. 21

Dang, V. N., Galati, F., Cortese, R., Di Giacomo, G., Marconetto, V., Mathur, P., Lekadir, K., Lorenzi, M., Prados, F., and Zuluaga, M. A. (2022). Vessel-CAPTCHA:

An efficient learning framework for vessel annotation and segmentation. *Medical Image Analysis*, 75:102263. 17

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. 12, 17, 22, 34, 70

Du, H., Dong, Q., Xu, Y., and Liao, J. (2023a). Weakly-supervised 3D medical image segmentation using geometric prior and contrastive similarity. *IEEE Transactions on Medical Imaging*, pages 1–1. 17

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. (2023b). Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*. 25

Fan, J., Zhang, Z., Tan, T., Song, C., and Xiao, J. (2020). CIAN: Cross-image affinity net for weakly supervised semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 2, 6

Favero, A., Zancato, L., Trager, M., Choudhary, S., Perera, P., Achille, A., Swaminathan, A., and Soatto, S. (2024). Multi-modal hallucination control by visual information grounding. 15

Gadgil, S. U., Endo, M., Wen, E., Ng, A. Y., and Rajpurkar, P. (2021). CheXseg: Combining expert annotations with DNN-generated saliency maps for X-ray segmentation. In Heinrich, M., Dou, Q., de Bruijne, M., Lellmann, J., Schläfer, A., and Ernst, F., editors, *Proceedings of the Fourth Conference on Medical Imaging with Deep Learning*, volume 143 of *Proceedings of Machine Learning Research*, pages 190–204. PMLR. 40

Giancardo, L., Niktabe, A., et al. (2023). Segmentation of acute stroke infarct core using image-level labels on CT-angiography. *NeuroImage: Clinical*, 37:103362. 16

Girshick, R. (2015). Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448. 33

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220. 6, 26, 39, 48, 49

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks. 2

Guo, D. and Terzopoulos, D. (2021). A transformer-based network for anisotropic 3d medical image segmentation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8857–8861. IEEE. 1, 16

Guo, D., Wei, H., Zhao, P., Pan, Y., Yang, H.-Y., Wang, X., Bai, J., Cao, K., Song, Q., Xia, J., Gao, F., and Yin, Y. (2020). Simultaneous classification and segmentation of intracranial hemorrhage using a fully convolutional neural network. *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 118–121. 12, 16

Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N., and Freeman, W. T. (2022). Unsupervised semantic segmentation by distilling feature correspondences. In *International Conference on Learning Representations*. 2

Han, T., Adams, L. C., Papaioannou, J.-M., Grundmann, P., Oberhauser, T., Löser, A., Truhn, D., and Bressem, K. K. (2023). MedAlpaca — an open-source collection of medical conversational ai models and training data. 19

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778. 28

He, X., Zhang, Y., Mou, L., Xing, E., and Xie, P. (2020). PathVQA: 30000+ questions for medical visual question answering. 4, 13

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*. 46

Hu, H., Zhang, J., Zhao, M., and Sun, Z. (2023). CIEM: Contrastive instruction evaluation method for better instruction tuning. 15

Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.-W., and Wu, J. (2020). UNet 3+: A full-scale connected unet for medical image segmentation. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1055–1059. 16

Huang, W., Liu, H., Guo, M., and Gong, N. Z. (2024). Visual hallucinations of multi-modal large language models. 9

Imran, A.-A.-Z., Hatamizadeh, A., Ananth, S. P., Ding, X., Tajbakhsh, N., and Terzopoulos, D. (2020). Fast and automatic segmentation of pulmonary lobes from chest CT using a progressive dense V-network. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 8(5):509–518. 1, 70

Irvin, J. A., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R. L., Shpanskaya, K. S., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C., Patel, B. N., Lungren, M. P., and Ng, A. (2019). CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI Conference on Artificial Intelligence*. 6, 26, 27, 36, 47, 48, 49

Jain, J., Yang, J., and Shi, H. (2023). VCoder: Versatile vision encoders for multimodal large language models. 14

Jang, J., Kyung, D., Kim, S., Lee, H., Bae, K., and Choi, E. (2022). Significantly improving zero-shot X-ray pathology classification via fine-tuning pre-trained image-text encoders. *ArXiv*, abs/2212.07050. 13, 44

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR. 13

Jiao, R., Zhang, Y., Ding, L., Xue, B., Zhang, J., Cai, R., and Jin, C. (2023). Learning with limited annotations: A survey on deep semi-supervised learning for medical image segmentation. *Computers in Biology and Medicine*, page 107840. 2

Jing, B., Xie, P., and Xing, E. (2018). On the automatic generation of medical imaging reports. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2577–2586, Melbourne, Australia. Association for Computational Linguistics. 18

Johnson, A., Pollard, T., Shen, L., Lehman, L.-w., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L., and Mark, R. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035. 37

Kazerouni, A., Aghdam, E. K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., and Merhof, D. (2023). Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88:102846. 2

Kim, H. E., Cosa-Linan, A., Santhanam, N., Jannesari, M., Maros, M. E., and Ganslandt, T. (2022). Transfer learning for medical image classification: A literature review. *BMC Medical Imaging*, 22(1):69. 12

Lai, X., Tian, Z., Jiang, L., Liu, S., Zhao, H., Wang, L., and Jia, J. (2021). Semi-supervised semantic segmentation with directional context-aware consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1205–1214. 7

Lai, Z. (2024). Exploring simple open-vocabulary semantic segmentation. *arXiv preprint arXiv:2401.12217*. 7

Lango, M. and Dusek, O. (2023). Critic-driven decoding for mitigating hallucinations in data-to-text generation. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2853–2862, Singapore. Association for Computational Linguistics. 21

Lau, J. J., Gayen, S., Ben Abacha, A., and Demner-Fushman, D. (2018). A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):1–10. 4, 13

Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., and Bing, L. (2023a). Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*. 15

Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., and Bing, L. (2023b). Mitigating object hallucinations in large vision-language models through visual contrastive decoding. 21

Lerousseau, M., Vakalopoulou, M., Classe, M., Adam, J., Battistella, E., Carré, A., Estienne, T., Henry, T., Deutsch, E., and Paragios, N. (2020). Weakly supervised multiple instance learning histopathological tumor segmentation. In Martel, A. L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M. A., Zhou, S. K., Racoceanu, D., and Joskowicz, L., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 470–479, Cham. Springer International Publishing. 16

Lester, B., Al-Rfou, R., and Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. 21

Li, B., Weinberger, K. Q., Belongie, S., Koltun, V., and Ranftl, R. (2022a). Language-driven semantic segmentation. In *International Conference on Learning Representations*. 7, 17, 35

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. (2023a). LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. 4, 5, 9, 13, 19, 22, 49

Li, J., Li, D., Savarese, S., and Hoi, S. (2023b). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*. 4, 9, 19

Li, J., Li, D., Xiong, C., and Hoi, S. (2022b). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR. 18, 36, 39

Li, J., Li, S., Hu, Y., and Tao, H. (2022c). A self-guided framework for radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 588–598. Springer. 8

Li, J., Li, S., Hu, Y., and Tao, H. (2022d). A self-guided framework for radiology report generation. In Wang, L., Dou, Q., Fletcher, P. T., Speidel, S., and Li, S., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 588–598, Cham. Springer Nature Switzerland. 18

Li, J., Selvaraju, R. R., Gotmare, A. D., Joty, S. R., Xiong, C., and Hoi, S. C. H. (2021). Align before Fuse: Vision and language representation learning with momentum distillation. In *Neural Information Processing Systems*. 60

Li, K., Qian, Z., et al. (2023c). Weakly supervised histopathology image segmentation with self-attention. *Medical Image Analysis*, 86:102791. 16

Li, K., Wu, Z., Peng, K.-C., Ernst, J., and Fu, Y. (2018). Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9215–9223. 16

Li, X. L., Holtzman, A., Fried, D., Liang, P., Eisner, J., Hashimoto, T., Zettlemoyer, L., and Lewis, M. (2023d). Contrastive decoding: Open-ended text generation as optimization. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics. 10, 21

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., and Wen, J.-R. (2023e). Evaluating object hallucination in large vision-language models. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics. 6, 14, 15, 32

Li, Y., Yang, B., Cheng, X., Zhu, Z., Li, H., and Zou, Y. (2023f). Unify, align and refine: Multi-level semantic alignment for radiology report generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2863–2874. 18

Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., and Marculescu, D. (2022). Open-vocabulary semantic segmentation with mask-adapted CLIP. *arXiv preprint arXiv:2210.04150*. 17

Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., and Marculescu, D. (2023). Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070. 7

Lin, D., Dai, J., Jia, J., He, K., and Sun, J. (2016). ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 16

Liu, B., Zhan, L.-M., Xu, L., Ma, L., Yang, Y., and Wu, X.-M. (2021a). Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. 13

Liu, C. et al. (2023a). IMITATE: Clinical prior guided hierarchical vision-language pre-training. 41

Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., and Wang, L. (2023b). Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*. 14

Liu, G., Hsu, T.-M. H., McDermott, M., Boag, W., Weng, W.-H., Szolovits, P., and Ghassemi, M. (2019a). Clinically accurate chest X-ray report generation. In Doshi-Velez, F., Fackler, J., Jung, K., Kale, D., Ranganath, R., Wallace, B., and Wiens, J., editors, *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 249–269. PMLR. 8, 20

Liu, G., Hsu, T.-M. H., McDermott, M., Boag, W., Weng, W.-H., Szolovits, P., and Ghassemi, M. (2019b). Clinically accurate chest X-ray report generation. In Doshi-Velez, F., Fackler, J., Jung, K., Kale, D., Ranganath, R., Wallace, B., and Wiens, J., editors, *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 249–269. PMLR. 18

Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023c). Visual instruction tuning. In *NeurIPS*. 3, 4, 9, 13, 19, 22, 32

Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., and Peng, W. (2024). A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*. 14

Liu, S., Liu, K., Zhu, W., Shen, Y., and Fernandez-Granda, C. (2022). Adaptive early-learning correction for segmentation from noisy annotations. *CVPR 2022*. 2, 6

Liu, Y., Wang, M., Liu, X., Chang, W., Sun, M., and Li, P. (2023d). Large language models encode clinical knowledge. *Nature*, 603:589–593. 4, 9, 13, 19

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021b). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022. 12, 70

Lovenia, H., Dai, W., Cahyawijaya, S., Ji, Z., and Fung, P. (2023). Negative object presence evaluation (NOPE) to measure object hallucination in vision-language models. 15

Lu, J., Clark, C., Zellers, R., Mottaghi, R., and Kembhavi, A. (2022). Unified-IO: A unified model for vision, language, and multi-modal tasks. In *ICLR*. 2, 18

Mahani, G. K., Li, R., Evangelou, N., Sotiropolous, S., Morgan, P. S., French, A. P., and Chen, X. (2022). Bounding box based weakly supervised deep convolutional neural network for medical image segmentation using an uncertainty guided and spatially constrained loss. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. 16

Manzari, O. N., Ahmadabadi, H., Kashiani, H., Shokouhi, S. B., and Ayatollahi, A. (2023). MedViT: A robust vision transformer for generalized medical image classification. *Computers in Biology and Medicine*, 157:106791. 12

Mao, J., Huang, J., et al. (2016). Generation and comprehension of unambiguous object descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20. 36

Miura, Y., Zhang, Y., Tsai, E., Langlotz, C., and Jurafsky, D. (2021). Improving factual completeness and consistency of image-to-text radiology report generation. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304, Online. Association for Computational Linguistics. 8, 49

Muhammad, M. B. and Yeasin, M. (2020). Eigen-CAM: Class activation map using principal components. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. 16

Mukhoti, J., Lin, T.-Y., Poursaeed, O., Wang, R., Shah, A., Torr, P. H. S., and Lim, S.-N. (2022). Open vocabulary semantic segmentation with patch aligned contrastive learning. 7, 17, 35

Nicolson, A., Dowling, J., and Koopman, B. (2023). Improving chest X-ray report generation by leveraging warm starting. *Artificial Intelligence in Medicine*, 144:102633. 49

Nori, H., King, N., McKinney, S. M., Carignan, D., and Horvitz, E. (2023). Capabilities of GPT-4 on medical challenge problems. 19

O'Brien, S. and Lewis, M. (2023). Contrastive decoding improves reasoning in large language models. 21

OpenAI (2022). ChatGPT - blog. `https://openai.com/blog/chatgpt`. 2, 4, 18

OpenAI et al. (2023). GPT-4 technical report. 2, 4, 15, 18

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics. 8, 19

Pham, H. H., Le, T. T., Tran, D. Q., Ngo, D. T., and Nguyen, H. Q. (2020). Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels. 12

Qian, Y., Zhang, H., Yang, Y., and Gan, Z. (2024). How easy is it to fool your multimodal LLMs? an empirical analysis on deceptive prompts. 14

Qian, Z., Li, K., et al. (2022). Transformer based multiple instance learning for weakly supervised histopathology image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 160–170, Cham. Springer Nature Switzerland. 16

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR. 4, 13, 22

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9. 2

Ren, P., Li, C., Xu, H., Zhu, Y., Wang, G., Liu, J., Chang, X., and Liang, X. (2023). ViewCo: Discovering text-supervised segmentation masks via multi-view semantic consistency. In *The Eleventh International Conference on Learning Representations*. 7, 17

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., and Saenko, K. (2018). Object hallucination in image captioning. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics. 15

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing. 16, 72

Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S., Nguyen, C., Ngo, V. D., Jayne, Seekins, D., Blankenberg, F., Ng, A. Y., Lungren, M. P., and Rajpurkar, P. (2022). Benchmarking saliency methods for chest X-ray interpretation. *Nature Machine Intelligence*, 4:867 – 878. 8, 39

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626. 2, 6, 16, 40

Seyyed-Kalantari, L., Liu, G., McDermott, M. B. A., and Ghassemi, M. (2020). CheXclusion: Fairness gaps in deep chest X-ray classifiers. *Pacific Symposium on Biocomputing*, 26:232–243. 12

Shi, X., Chen, Z., Wang, H., Yeung, D., Wong, W., and Woo, W. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *CoRR*, abs/1506.04214. 72

Simpson, A. L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., van Ginneken, B., Kopp-Schneider, A., Landman, B. A., Litjens, G. J. S., Menze, B. H., Ronneberger, O., Summers, R. M., Bilic, P., Christ, P. F., Do, R. K. G., Gollub, M., Golia-Pernicka, J., Heckers, S., Jarnagin, W. R., McHugo, M., Napel, S., Vorontsov, E., Maier-Hein,

L., and Cardoso, M. J. (2019). A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *CoRR*, abs/1902.09063. 74

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P., Prakash, S., Green, B., Dominowska, E., y Arcas, B. A., Tomasev, N., Liu, Y., Wong, R., Semturs, C., Mahdavi, S. S., Barral, J., Webster, D., Corrado, G. S., Matias, Y., Azizi, S., Karthikesalingam, A., and Natarajan, V. (2023). Towards expert-level medical question answering with large language models. 4, 9, 19

Skandarani, Y., Jodoin, P.-M., and Lalande, A. (2021). GANs for medical image synthesis: An empirical study. 2

Strudel, R. et al. (2022). Weakly-supervised segmentation of referring expressions. 7, 17, 35

Strudel, R., Garcia, R., Laptev, I., and Schmid, C. (2021). Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7262–7272. 37

Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L.-Y., Wang, Y.-X., Yang, Y., et al. (2023a). Aligning large multimodal models with factually augmented RLHF. *arXiv preprint arXiv:2309.14525*. 15

Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L.-Y., Wang, Y.-X., Yang, Y., Keutzer, K., and Darrell, T. (2023b). Aligning large multimodal models with factually augmented RLHF. *arXiv:2309.14525*. 15

Tian, Z., Shen, C., Wang, X., and Chen, H. (2021). BoxInst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5443–5452. 2, 6, 16

Tiu, E., Talius, E., Patel, P., Langlotz, C. P., Ng, A. Y., and Rajpurkar, P. (2022). Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406. 29, 30

Tonmoy, S. M. T. I., Zaman, S. M. M., Jain, V., Rani, A., Rawte, V., Chadha, A., and Das, A. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. 21

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). LLaMA: Open and efficient foundation language models. 19, 22, 29

van den Oord, A., Li, Y., and Vinyals, O. (2019). Representation learning with contrastive predictive coding. 34

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30. 12, 70, 72

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575. 8, 20

Vu, T., Iyyer, M., Wang, X., Constant, N., Wei, J., Wei, J., Tar, C., Sung, Y.-H., Zhou, D., Le, Q., and Luong, T. (2023). FreshLLMs: Refreshing large language models with search engine augmentation. 21

Wan, Z. et al. (2023). Med-UniC: Unifying cross-lingual medical vision-language pre-training by diminishing bias. 39, 41

Wang, B., Min, S., Deng, X., Shen, J., Wu, Y., Zettlemoyer, L., and Sun, H. (2023). Towards understanding chain-of-thought prompting: An empirical study of what matters. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Toronto, Canada. Association for Computational Linguistics. 5, 14

Wang, F. et al. (2022a). Multi-granularity cross-modal alignment for generalized medical visual representation learning. In *NeurIPS*. 39, 41

Wang, L., Ning, M., Lu, D., Wei, D., Zheng, Y., and Chen, J. (2022b). An inclusive task-aware framework for radiology report generation. In Wang, L., Dou, Q., Fletcher, P. T., Speidel, S., and Li, S., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 568–577, Cham. Springer Nature Switzerland. 10, 20, 49

Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., and Yang, H. (2022c). Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR. 2, 18

Wang, X., Chen, H., Gan, C., Lin, H., Dou, Q., Tsougenis, E., Huang, Q., Cai, M., and Heng, P.-A. (2020). Weakly supervised deep learning for whole slide lung cancer image analysis. *IEEE Transactions on Cybernetics*, 50(9):3950–3962. 2, 16

Wang, Z. et al. (2022d). MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. 13

Wang, Z., Tang, M., Wang, L., Li, X., and Zhou, L. (2022e). A medical semantic-assisted transformer for radiographic report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 655–664. Springer. 8

Wang, Z., Tang, M., Wang, L., Li, X., and Zhou, L. (2022f). A medical semantic-assisted transformer for radiographic report generation. In Wang, L., Dou, Q., Fletcher,

P. T., Speidel, S., and Li, S., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 655–664, Cham. Springer Nature Switzerland. 20

Weng, Y., Zhou, T., Li, Y., and Qiu, X. (2019). NAS-Unet: Neural architecture search for medical image segmentation. *IEEE Access*, 7:44247–44257. 16

Wu, C., Lin, W., Zhang, X., Zhang, Y., Wang, Y., and Xie, W. (2023a). PMC-LLaMA: Towards building open-source language models for medicine. 2, 19

Wu, C., Zhang, X., Zhang, Y., Wang, Y., and Xie, W. (2023b). MedKLIP: Medical knowledge enhanced language-image pre-training. *ICCV*. 13, 39, 41

Wu, X., Li, J., Wang, J., and Qian, Q. (2023c). Multimodal contrastive learning for radiology report generation. *Journal of Ambient Intelligence and Humanized Computing*, 14(8):11185–11194. 18

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12077–12090. Curran Associates, Inc. 37, 39

Xu, G., Song, Z., Sun, Z., Ku, C., Yang, Z., Liu, C., Wang, S., Ma, J., and Xu, W. (2019). CAMEL: A weakly supervised learning framework for histopathology image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2

Xu, J., Hou, J., Zhang, Y., Feng, R., Wang, Y., Qiao, Y., and Xie, W. (2023). Learning open-vocabulary semantic segmentation models from natural language supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2935–2944. 7, 17

Xu, J., Mello, S. D., Liu, S., Byeon, W., Breuel, T., Kautz, J., and Wang, X. (2022). GroupViT: Semantic segmentation emerges from text supervision. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18113–18123, Los Alamitos, CA, USA. IEEE Computer Society. 3, 17, 40

Yan, X., Tang, H., Sun, S., Ma, H., Kong, D., and Xie, X. (2022). AFTer-UNet: Axial fusion transformer UNet for medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3971–3981. 16

Yang, S., Wu, X., Ge, S., Zheng, Z., Zhou, S. K., and Xiao, L. (2023). Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, 86:102798. 20, 49

Ye, H., Gao, F., Yin, Y., Guo, D., Zhao, P., Lu, Y., Wang, X., Bai, J., Cao, K., Song, Q., Zhang, H., Chen, W., Guo, X., and Xia, J. (2019). Precise diagnosis of intracranial

hemorrhage and subtypes using a three-dimensional joint convolutional and recurrent neural network. *European Radiology*, 29:6191 – 6201. 12

Yi, M., Cui, Q., Wu, H., Yang, C., Yoshie, O., and Lu, H. (2023). A simple framework for text-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7071–7080. 7, 17

Yu, S., Seo, P. H., and Son, J. (2023). Zero-shot referring image segmentation with global-local context features. 17, 35, 43

Yuan, Z., Yan, Y., Sonka, M., and Yang, T. (2021). Large-scale robust deep AUC maximization: A new surrogate loss and empirical studies on medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3040–3049. 12

Zawacki, A. et al. (2019). SIIM-ACR pneumothorax segmentation. `https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation`. 8, 39

Zhang, H., Chen, J., et al. (2023a). HuatuoGPT: Towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885, Singapore. Association for Computational Linguistics. 19

Zhang, R., Han, J., Liu, C., Gao, P., Zhou, A., Hu, X., Yan, S., Lu, P., Li, H., and Qiao, Y. (2023b). LLaMA-Adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*. 4, 9, 13, 19

Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., Wong, C., Tupini, A., Wang, Y., Mazzola, M., Shukla, S., Liden, L., Gao, J., Lungren, M. P., Naumann, T., Wang, S., and Poon, H. (2024). BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. 27

Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. P. (2022). Contrastive learning of medical visual representations from paired images and text. In Lipton, Z., Ranganath, R., Sendak, M., Sjoding, M., and Yeung, S., editors, *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pages 2–25. PMLR. 12

Zhao, L., Deng, Y., Zhang, W., and Gu, Q. (2024). Mitigating object hallucination in large vision-language models via classifier-free guidance. 15

Zheng, G., Yang, B., Tang, J., Zhou, H.-Y., and Yang, S. (2023). DDCoT: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 5168–5191. Curran Associates, Inc. 5, 14

Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., and Yao, H. (2024). Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations*. 15

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. (2023). MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*. 4, 13, 19, 32