Leveraging Spatial Statistics for Domain Adaptation of Vision Language Models in Medical VQA

A thesis submitted in partial fulfillment of the requirements for the award of the degree of

Master of Technology

in

Computer Science

submitted by

Himanshu Raj

(Roll No. CS2416)

under the supervision of:

Prof. Utpal Garain

Computer Vision & Pattern Recognition Unit,

Indian Statistical Institute, Kolkata

Indian Statistical Institute

Kolkata-700108, India

June 2026

Declaration of Authorship

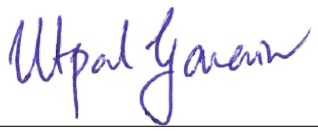
I, **Himanshu Raj**, declare that this Master's thesis titled, *"Leveraging Spatial Statistics for Domain Adaptation of Vision Language Models in Medical VQA"* is a result of my own research carried out under the guidance of Prof. Utpal Garain. This work forms a part of the broader doctoral research being pursued by Aditya Shankar Pal towards the fulfillment of the requirements for the Ph.D. degree at the Indian Statistical Institute, Kolkata. To the best of my knowledge contains no materials previously published or written by any other person, or substantial proportions of material which has formed the basis of award of any other degree or diploma at ISI Kolkata or any other educational institution, except where due acknowledgement is made in this thesis. I authorise ISI Kolkata to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Date: 10th June 2026

Himanshu Raj
Roll No: CS2416
Mtech in Computer Science
Indian Statistical Institute, Kolkata

Certificate of the Supervisor

This is to certify that the dissertation titled **“ Leveraging Spatial Statistics for Domain Adaptation of Vision Language Models in Medical VQA”** submitted by **Himanshu** to Indian Statistical Institute, Kolkata, in partial fulfillment for the award of the degree of **Master of Technology in Computer Science** is a bonafide record of work carried out by him/her under our supervision and guidance. The dissertation has fulfilled all the requirements as per the regulations of this institute and, in our opinion, has reached the standard needed for submission.



10/June/2026

Prof. Utpal Garain
Computer Vision & Pattern Recognition Unit
Indian Statistical Institute, Kolkata

Acknowledgements

I am truly grateful to **Prof. Utpal Garain** for giving me the opportunity to work under his supervision during my master's program. I feel fortunate to have learned from his experience and guidance. He not only supported me throughout the journey but also gave me the freedom to explore and learn independently during my dissertation. His thoughtful feedback and encouraging nature played a key role in helping me grow academically and confidently carry out my research.

I would like to sincerely thank **Mr. Aditya Shankar Pal** Senior Research Fellow, (CVPR Unit) for his invaluable support and guidance throughout the course of my dissertation. I genuinely believe that this work would not have been possible without his help. He patiently assisted me in understanding the challenging concepts and guided me through many aspects of implementation and coding. What I deeply appreciate is that, while offering constant support, he also gave me the freedom to learn and explore on my own throughout the course of this work. Even during the stressful period of placements, he selflessly took the time to support me and share my workload. I am truly grateful for his generosity, kindness and constant encouragement, which played a significant role in helping me complete this dissertation.

Finally, I would like to express my heartfelt thanks to my family, friends and batchmates for their unwavering support and encouragement throughout this journey. Their patience, understanding and constant motivation have been a source of strength for me. This achievement would not have been possible without their presence and belief in me at every step.

Abstract

Recent advances in Vision–Language Models (VLMs) have demonstrated strong performance in Medical Visual Question Answering (Medical VQA) task. Although they perform very well within their domains, these models often experience issues with their generalization ability on unknown clinical distribution data because of different imaging technologies and patient groups used in various medical facilities. Generalization problems faced by these models make their practical application in the field of VLM-based medical VQA systems rather difficult. To overcome this limitation we proposed our method named Spatial Semantics Aware Domain Adaptation (SSADA), which is an integrated framework that combines both finetuning and prompt-based in-context learning for domain adaptation. Our proposed approach, SSADA, includes the following three important components: (i) Mask-Aware Finetuning (MAFt) to make localization aware finetuning, (ii) Anatomy Aware Instance Normalization (AAIN) for handling intensity or distribition shift, and (iii) Weighted Multi-Modal Example Retrieval (WMMER) for semantically consistent example selection during inference. We evaluate the proposed framework on three publicly available Medical VQA benchmarks, SLAKE, VQA-Med 2019, and OmniMedVQA–RadImageNet, under cross-domain settings and compare it against standard finetuning techniques. Experimental results demonstrate the effectiveness of SSADA in improving cross-domain generalization.

Keywords: Medical Visual Question Answering, Vision-Language Models, Domain Adaptation, Spatial Statistics, In-Context Learning, Cross-Domain Generalization.

Publications

1. Aditya Shankar Pal, Himanshu Raj, Joy Mahapatra, Utpal Garain, "Leveraging Spatial Statistics for Domain Adaptation of Vision Language Models in Medical VQA" (Submitted at *British Machine Vision Conference, 2026*).

Contents

List of Figures

List of Tables

List of Abbreviations

AAIN Anatomy-Aware Instance Normalization.

CT Computed Tomography.

DHN-NCE Decoupled Hard Negative Noise Contrastive Estimation.

GQA Grouped-Query Attention.

ICL In-Context Learning.

IN Instance Normalization.

InfoNCE Information Noise-Contrastive Estimation.

LLM Large Language Model.

LoRA Low-Rank Adaptation.

M-ROPE Multimodal Rotary Position Embeddings.

MAFt Mask-Aware Finetuning.

MedSAM Medical Segment Anything Model.

MedVQA Medical Visual Question Answering.

MLP Multi-Layer Perceptron.

MRI Magnetic Resonance Imaging.

ROPE Rotary Position Embeddings.

SAM Segment Anything Model.

SigLIP Sigmoid Language-Image Pretraining.

SSADA Spatial Semantics Aware Domain Adaptation.

ViT Vision Transformer.

VLM Vision-Language Model.

VQA Visual Question Answering.

WMMER Weighted Multi-Modal Example Retrieval.

Chapter 1

Introduction

1.1 Introduction

Recent advances in large-scale Vision–Language Models (VLMs), such as Qwen2-VL [2], Gemma [3], LLaMA [4], and SmolVLM [5], have demonstrated strong multimodal performance in Medical Visual Question Answering (Medical VQA) [6, 7], a multimodal medical reasoning task, largely driven by large-scale pretraining and efficient finetuning techniques. Through effective cross-modal fusion of textual and visual representations, these models consistently outperform earlier task-specific architectures that process modalities independently. Despite their strong in-domain results, domain adaptation that is, domain generalization to unseen clinical distributions remains insufficiently explored in Medical VQA [8]. Addressing domain adaptation is particularly critical in this context because medical images contain complex and fine-grained spatial semantics that differ substantially from conventional natural images [9, 10]. Large differences exist across hospitals because of imaging methods.Scanning protocols often vary between different healthcare institutions [11].Patient populations also differ from one hospital to another.These differences create significant distribution shifts in medical data.As a result, domain adaptation becomes a major challenge and is essential for reliable Medical VQA system deployment.

Medical Visual Question Answering (MedVQA) systems that can be relied upon have the capability to assist in decreasing radiologists' work and offer quick second opinions. Nonetheless, clinical settings are characterized by great levels of diversity, and it is possible for a model to be trained using only one kind of modality or scan. For example, during training, the model could encounter mostly axial CT scans or examples concentrate on a certain part of the body. However, when testing if examples of sagittal MRI scans, or examples of new body part or anatomical regions come the performance of the model would decrease. Diagnosis in a medical setting largely relies on the identification of

small pathological characteristics because they normally occur at specific areas of the body. Therefore, Medical VQA models should be tested thoroughly using various kinds of questions. These tests should confirm whether the model is capable of detecting not only the imaging modality or anatomical plane but also the location of the organ. If performance drops across these categories, cross-domain robustness is affected. As a result, the clinical usefulness of the VQA system becomes limited.

To address this domain adaptation in VLMs, two common techniques are widely used: finetuning and prompt-based adaptation [12, 13]. Nevertheless, both approaches shows inherent limitations in their underlying mechanisms. Finetuning techniques often overlook explicit spatial (or localization-aware) semantics [14], such as region-of-interest masking, leaving the integration of spatially grounded clinical knowledge underexplored. Moreover, finetuning may encourage shortcut learning—capturing superficial or dataset-specific correlations rather than making use of anatomically meaningful spatial information which reduces robustness under distribution shift [14]. In contrast, prompt-based domain adaptation methods face two fundamental challenges: (i) insufficient modeling of cross-domain alignment [6], particularly in capturing both fine-grained visual details and global multimodal semantics during prompt construction, and (ii) difficulty in selecting representative and domain-relevant in-context examples for stable and reliable adaptation [15]. Therefore, naively applying these techniques within VLMs for Medical VQA is unlikely to effectively address the domain adaptation challenges inherent in medical imaging scenarios.

In this thesis, we aim to reduce the existing gaps in domain adaptation for Medical VQA using VLMs by jointly integrating finetuning and prompt-based in-context learning within a unified framework for domain adaptation, rather than focusing exclusively on either strategy as in prior works.

1.2 Contributions of the thesis

In this thesis, We propose the Spatial Semantics-Aware Domain Adaptation (SSADA) framework, which consists of three key modules:

(i) **Mask-Aware Finetuning** (MAFt),

(ii) **Anatomy-Aware Instance Normalization** (AAIN), and

(iii) **Weighted Multi-Modal Example Retrieval** (WMMER).

MAFt addresses the lack of explicit localization modeling because it uses region-of-interest masking during finetuning. This helps the model learn spatial information more effectively as the learned features align closely with anatomical structures.

AAIN reduces variations between different imaging domains because it replaces standard normalization methods with anatomical region-based instance normalization. This reduces differences across domains. Therefore , also help in improving robustness to intensity and contrast changes in medical images.

WMMER improves in-context learning through better example selection as it chooses representative and domain-relevant examples. The selection is based on both text and image similarity. This allows the model to use semantic and visual information together. After adding WMMER it makes our SSDA framework few-shot domain adaptation approach for MedVQA.

We evaluate the effectiveness of the SSADA framework on three widely recognized benchmark datasets: SLAKE [16], VQA-Med-2019 [17], and OmniMedVQA–RadImageNet [18], specifically under cross-domain evaluation settings.

To ensure the validity of the cross-domain evaluations, we conduct extensive data preprocessing, including calculating unique dataset statistics - such as specific mean and standard deviation values across different combinations of anatomical locations and modality to tackle this issue of domain shifts present across the Training and Testing sets.

1.3 Organisation of thesis

This thesis is organized into five chapters. An overview of each chapter is presented below:

Chapter 1: Introduction This chapter presents a summary of the research problem, explains the rationale for conducting the study, and emphasizes the main goals and significant contributions of the work.

Chapter 2: Foundation Knowledge This Chapter provides the prerequisite theoretical background necessary to understand the proposed framework Spatial Semantics-Aware Domain Adaptation (SSADA). It reviews the evolution of Medical Visual Question Answering, the architecture of MedCLIP model and its underlying logic, and the mechanics of image segmentation networks.

Chapter 3: Methodology This chapter details the proposed Spatial Semantics-Aware Domain Adaptation (SSADA) framework. It provides a detailed, mathematical and structural explanation of the three core modules: Mask-Aware Finetuning (MAFt), Anatomy-Aware Instance Normalization (AAIN), and Weighted Multi-Modal Example Retrieval (WMMER). The chapter explain how these modules are integrated to address the limitations of standalone finetuning and prompt-based adaptation.

Chapter 4: Experiments and Results This chapter presents the empirical evaluation of the SSADA framework. This chapter outlines the dataset preparation, the statistical

benchmarking of the SLAKE, VQA-Med-2019, and OmniMedVQA-RadImageNet datasets, and the specific implementation details of the VQA systems. It provides a rigorous quantitative analysis and compraison of the model's performance finetuned on different techniques across various categories (such as modality, organ, and plane accuracy) of the testing dataset. This chapter also contains qualitative visual examples and comprehensive ablation studies to validate the contribution of each individual module .

Chapter 5: Conclusion and Future Work This chapter summarizes the primary contributions and findings of the research. This work analyzes the importance of domain adaptation in medical imaging from the clinical perspective and provides ideas for future investigation into how the framework can be scaled up to handle entirely novel diagnostic modalities, along with enhanced optimization of the search process.

Chapter 2

Background

This chapter provides the theoretical background necessary to understand the methodologies proposed in this thesis. In this chapter We are going to discuss the fundamental concepts of Medical Visual Question Answering (MedVQA), SAM , MedSAM and prompt-based segmentation paradigms using MedCLIP-SAM. After that we are going to see cosine similarity definition which is used as a foundational metric for measuring similarity in embedding representation, followed by an analysis of domain shift, instance normalization, and the mechanics of in-context learning. Collectively, these principles provide the background for the adaptation framework presented in this work.

2.1 Medical Visual Question Answering (MedVQA)

Medical Visual Question Answering (MedVQA) is a multimodal task that need an intelligent system to answer medical queries based on medical images, such as radiology scans or pathology slides. The standard MedVQA pipeline involves extracting visual features from the medical scan and extracting semantic features from the text query. After extracting the visual and semantic features these both features are then projected into a shared multimodal latent space. Let I represent the input medical image and Q represent the text question. The objective is to learn a mapping function f_θ that maximizes the probability of the correct answer A:

$$\hat{A} = \arg\max_{A \in \mathcal{A}} P(A|I, Q; \theta) \tag{2.1}$$

where $\mathcal{A}$ is the answer space and θ represents the model parameters. A major bottleneck in MedVQA is the lack of large-scale annotated datasets, which make model prone to overfitting on source domains.

2.2 SAM and MedSAM

One of the most critical aspects in analyzing medical images is spatial localization since it allows one to identify the areas where a disease is situated. Therefore, these areas usually need to be segmented so that a proper examination can be conducted. SAM [19] proposes a unique way to approach visual segmentation by presenting a model that has a flexible architecture and is capable of performing zero-shot segmentation on various images.

There are three main components of SAM – image encoder responsible for extracting features of the input image, prompt encoder that takes into account the user-defined sparse data, and mask decoder that combines information about the image and prompt into the final segmentation mask.

SAM proves to perform well in natural images but not as good when used in medical images since medical images have particular contrast patterns and boundaries as well as their grayscale intensities differ from natural images'. To overcome these issues, medical extensions like MedSAM [20] was developed. It becomes possible due to fine-tuning the mask decoder on medical images' datasets. The fine-tuned model shows a better match to the correct anatomical boundaries.

Despite the progress achieved, it is still necessary to solve the problem of using geometrical prompts for segmentation masks predictions. They involve manual selection of key points or drawing of a bounding box. To eliminate this dependency, it is essential to integrate segmentation models with language models. For instance, MedCLIP-SAM framework addresses this crucial problem by proposing zero-shot mask prediction without prompt inputs (see Section 2.3).

2.3 Promptable Segmentation: MedCLIP-SAM

MedCLIP-SAM [21] addresses the above problem in the context of dense prediction problems, such as generating masks using the combination of CLIP and Segment Anything Model (SAM). Similar to conventional approaches, it uses a dual-encoder architecture, consisting of vision and text encoders that map input data into a common high-dimensional multimodal embedding space, yielding normalized vectors of image and text, named $\mathbf{I}_{p,i}$ and $\mathbf{T}_{p,i}$.

Traditional contrastive learning approaches strictly follow a one-to-one correspondence principle using the InfoNCE loss function, regarding all other pairs as negatives for a particular pair of vectors in a batch. But when dealing with medical datasets, we frequently encounter an issue of a large number of false negatives, meaning that the different pictures

have the same description. Furthermore, the negative-positive-coupling (NPC) effect inherent to standard InfoNCE can lead to sub-optimal learning efficiency. To resolve this, MedCLIP-SAM fine-tunes a domain-specific foundation model (BiomedCLIP) by employing a Decoupled Hard Negative Noise Contrastive Estimation (DHN-NCE) loss.

Here, we use the DHN-NCE loss function, which decouples the optimization and removes the positive part from the denominator while implementing dynamic scaling based on hard negative sampling. For a batch of size B, the image-to-text ($\mathcal{L}_{v\to t}$) and text-to-image ($\mathcal{L}_{t\to v}$) losses are formulated as:

$$\mathcal{L}_{v\to t} = -\sum_{i=1}^{B} \frac{\mathbf{I}_{p,i}\mathbf{T}_{p,i}^{\top}}{\tau} + \sum_{i=1}^{B} \log\left(\sum_{j\neq i} e^{\mathbf{I}_{p,i}\mathbf{T}_{p,j}^{\top}/\tau} \mathcal{W}_{v\to t}^{\mathbf{I}_{p,i}\mathbf{T}_{p,j}}\right) \tag{2.2}$$

$$\mathcal{L}_{t\to v} = -\sum_{i=1}^{B} \frac{\mathbf{T}_{p,i}\mathbf{I}_{p,i}^{\top}}{\tau} + \sum_{i=1}^{B} \log\left(\sum_{j\neq i} e^{\mathbf{T}_{p,i}\mathbf{I}_{p,j}^{\top}/\tau} \mathcal{W}_{t\to v}^{\mathbf{T}_{p,i}\mathbf{I}_{p,j}}\right) \tag{2.3}$$

where τ is the temperature parameter. The variables $\mathcal{W}_{v\to t}$ and $\mathcal{W}_{t\to v}$ represent hardness weighting functions that apply exponential scaling (parameterized by β_1 and β_2) to penalize highly similar but non-matching pairs. The overall cross-modal objective is the sum of both components:

$$\mathcal{L}_{\text{DHN-NCE}} = \mathcal{L}_{v\to t} + \mathcal{L}_{t\to v} \tag{2.4}$$

For automated mask creation, MedCLIP-SAM utilizes the robust representations learned via DHN-NCE to generate spatial prompts directly from clinical text. Given a text prompt, the framework extracts pixel-wise saliency maps (using gradient-based activation methods like gScoreCAM or Multi-modal Information Bottleneck mechanisms). These saliency maps are then converted into explicit visual bounding boxes. Finally, these derived bounding boxes are passed into SAM, which outputs precise, zero-shot or weakly supervised segmentation masks of the target pathology or anatomy, entirely bypassing the need for manual prompt engineering by clinicians.

2.4 Cosine Similarity in Multimodal Embedding Spaces

For the Weighted Multi-Modal Example Retrieval (WMMER), the similarity measure used to compute the intra-modal similarity for example retrieval is cosine similarity [22]. The target input is compared to the candidates' exemplars in the support set by computing their embedding vector similarity. However, the comparison is made within the same modality to enable intra-modal comparisons.

Given two embedding vectors $\mathbf{x}_a$ and $\mathbf{x}_b$ from the same modality (either visual or textual)

in an n-dimensional space, the cosine similarity is simplified to their dot product:

$$\text{Sim}(\mathbf{x}_a, \mathbf{x}_b) = \mathbf{x}_a \cdot \mathbf{x}_b = \sum_{i=1}^{n} x_{a,i} x_{b,i} \tag{2.5}$$

This metric is bounded within the interval $[-1, 1]$, where 1 denotes maximum similarity.

The average similarity score for each candidate in the WMMER algorithm can be calculated using the weighted convex combination of intra-modal cosine similarities. The advantage of such dual modality is that WMMER will select the examples which not only correspond to the target scan but also relate to the clinical question asked, thus providing relevant information to the Vision-Language Model (VLM).

2.5 Domain Shift and Instance Normalization

Domain shift might have an impact on data obtained by medical imaging because of changes in imaging modalities like CT scans, MRIs, and ultrasounds, change in vendors of the imaging equipment used, and acquisition processes. This will cause variations in terms of noise, contrast, and histograms of intensity distributions.

Normalization methods are extremely useful in addressing the problem discussed above. To be more specific, Instance Normalization (IN) has proven to be very effective in style transfer and domain adaptation tasks. Different from Batch Normalization, where statistics are calculated based on mini-batches, IN calculates statistics for each channel of every single sample, separately, using μ and σ^2:

$$\text{IN}(x) = \gamma \left(\frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} \right) + \beta \tag{2.6}$$

where x is the input feature map, ϵ is a small constant for numerical stability, and γ, β are learnable affine parameters. By stripping the instance-specific style statistics (e.g., the scanner contrast) while preserving the core spatial content, IN reduces visual discrepancy. However, standard IN operates globally across the entire feature map. Adapting this principle to operate locally within specific anatomical boundaries forms the theoretical basis for mitigating complex cross-domain variations.

2.6 In-Context Learning in Vision-Language Models

In-Context Learning (ICL) [23] is an emerging capability of large-scale foundation models where the model learns to perform a task simply by conditioning on a few demonstration

examples provided within the input prompt, entirely bypassing gradient updates.

In a multimodal setting, an ICL prompt consists of a sequence of context examples $\{(I_1, Q_1, A_1), \ldots, (I_k, Q_k, A_k)\}$ followed by the target query $(I_{\text{target}}, Q_{\text{target}})$. The model's predictive performance is highly sensitive to the selection of these context examples. Randomly selecting examples often introduces noise, whereas retrieving examples that share both semantic similarity (question intent) and visual similarity (anatomical/pathological appearance) with the target query drastically reduces the inferential gap. Consequently, establishing optimized retrieval mechanisms based on high-dimensional similarity metrics is essential for maximizing the zero-shot and few-shot capabilities of MedVQA systems under distribution shift.

Chapter 3

Methodology

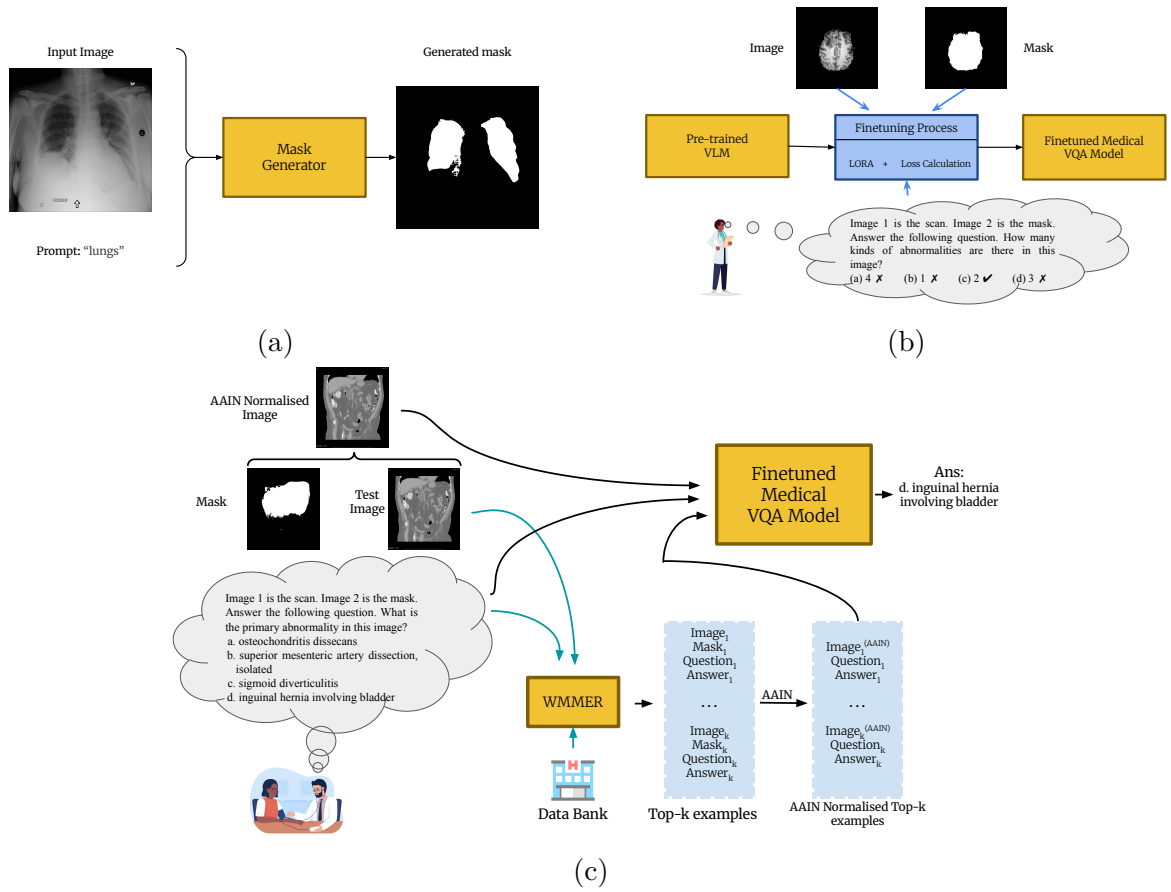


Figure 3.1: Overview of the proposed SSADA framework. (a) Illustration of the anatomical mask generation process. (b) Mask-Aware Finetuning (MAFt) on the source domain. (c) Illustration of the inference phase, where the target image is normalized using AAIN, followed by WMMER for few-shot example retrieval.

This section outlines the data preprocessing technique and presents the proposed Spatial Semantics Aware Domain Adaptation (SSADA) framework.

3.1 Data Preparation

Existing medical VQA benchmarks (SLAKE, VQA-Med 2019) are predominantly formulated as open-ended generation tasks, where model performance is often confounded by lexical variability, synonym ambiguity, and prompt-sensitive language generation. Consequently, models trained on one dataset often fail to generalize across datasets due to domain-specific answer distributions, reporting conventions, and modality-dependent linguistic priors [24]. Clinically equivalent responses may receive different evaluation scores despite conveying similar semantic meaning, thereby limiting the reliability and reproducibility of benchmarking [25, 26].

To address this limitation, we reformulate medical VQA as a clinically constrained multiple-choice reasoning task by constructing semantically plausible distractors from the same organ system and imaging modality as the target sample. For each data instance with image I_i, question Q_i, and true answer A_i, a set of candidate response options is generated. We use seprate answers from whole dataset as distractors based on the filter that we created. In order to ensure clinical relevance in options generated, all the answers in the database are first categorized based on distinct clinical categories (such as imaging modalities, anatomical locations, and types of pathology). The distractors for Q_i will be selected only from the same semantic class as A_i. If Q_i is asking about an anomaly, the distractors will be drawn from other questions related to anomalies for the same anatomical location. This organ and modality-based distractor generation approach ensures clinical plausibility, leading to significant inter-class confusability, and helps reduce shortcut learning due to surface language priors. Furthermore, the formulation guarantees that the medical VQA task does not become a mere problem of generating answers but remains a constrained problem of differential diagnosis than the generation of an open-ended answer would be.

The generation of high-quality options plays an essential role in making sure that there is proper balance and challenge posed by the questions If the right answer points to lung abnormality, like ‘lung opacity’, the other options come from similar anatomical abnormality, like ‘pleural effusion’ or ‘pneumothorax’ , not random options that aren’t even close, like fractures way up top in the skull.

During both training and inference, these options get appended after the main question in this set format: “Question: [the Original Question Here] Options: (A) [Option 1] (B) [Option 2] (C) [Option 3] (D) [Option 4]. Pick the right one, please.” Doing this strictly keeps everything standardized and let the system’s output stay standardized, which makes it easy to measure exact-match accuracy.

3.2 Spatial Semantics Aware Domain Adaptation Framework

The proposed framework for cross-domain adaptation comprises three main components, namely Mask Aware Finetuning (MAFt), Anatomy-Aware Instance Normalization (AAIN), and Weighted Multi-Modal Example Retrieval (WMMER).

3.2.1 Mask Aware Finetuning (MAFt)

We begin by generating the mask for the anatomical region. To obtain anatomical region-level annotations without requiring manual pixel-wise labeling, we employ MedCLIP-SAM [21], a MedCLIP in conjunction with the Segment Anything Model for automatic mask generation. MedCLIP is used to encode anatomical region prompt and identify semantically aligned regions within the medical image. These region cues are subsequently provided as prompts to SAM to generate consistent segmentation masks. For a given image and its associated question, we invoke a VLM to infer the anatomical region of interest. The image and a composite prompt – comprising the associated question with an anatomical region localization query – are passed to a VLM, which returns the anatomical region relevant to the image and question. The inferred anatomical region with original image is subsequently provided as input to MedCLIP-SAM to obtain the region mask. An example demonstration the process of mask generation using MedCLIP-SAM is illustrated in Figure 3.1a. The generated mask using the model is further utilised during the training step and the mask-based normalisation during inference.

During the supervised fine-tuning (SFt) phase (Figure 3.1b), the VLM is optimized exclusively on the source domain. We integrate the weakly supervised anatomical region mask obtained from the MedCLIP-SAM directly into the training loop. Prior to feature extraction, each training image I_i is paired with its corresponding binary mask M_i. We are passing both I_i and M_i as two seprate image in the finetuning and prompt which specify one as scan and another is segmentation mask of anatomical region of the scan. The model is finetuned end-to-end using a standard cross-entropy loss on the generated target answer. By utilising the masks during the finetuning, the VLM learns to weight the clinically relevant anatomical regions, preventing the model from overfitting to spurious background correlations or domain-specific scanner noise.

Consequently, the visual tokens generated by the VLM's vision transformer predominantly represent the features of the target organ, forcing the cross-attention layers of the model to ground the textual question directly to the isolated pathology rather than relying on background surrounding pixels.

3.2.2 Anatomy-Aware Instance Normalization (AAIN)

The difficulties faced in obtaining cross-dataset generalization in medical data due to some reasons are: 1) The variations in pixel intensities caused by changes in protocol used, scanner used and software parameters. 2) The tedious and time-consuming process of manually annotating the images using the expertise of the doctors. [11]. Standard instance normalization operates at the image level, computing channel-wise mean and variance over all spatial locations of an image. However, in medical images, the diagnostically relevant regions often occupy only a subset of pixels. Consequently, the global normalization process may attenuate discriminative intensity variations within the organ of interest by coupling them with potentially high-variance background regions. To minimise this effect, we propose Anatomy-Aware Instance Normalization (AAIN), a region-conditioned normalization technique that leverages anatomical priors from the generated mask to restrict normalization statistics to clinically relevant regions. Let $I \in \mathbb{R}^{H\times W\times C}$ be the raw image and $M \in \{0,1\}^{H\times W}$ be the binary organ mask obtained from MedCLIP-SAM. We decompose the image statistics based on M into two partitions: the foreground partition focusing on the anatomical region of interest, and the background partition focusing on the complementary regions. For each channel c, the foreground statistics are computed strictly within the masked region as

$$\mu_{fg}^{(c)}(I,M) = \frac{1}{|M|}\sum_{i,j} I_{i,j,c}\cdot M_{i,j} \tag{3.1}$$

$$\sigma_{fg}^{(c)}(I,M) = \sqrt{\frac{1}{|M|}\sum_{i,j}(I_{i,j,c}-\mu_{fg}^{(c)})^2\cdot M_{i,j}+\epsilon} \tag{3.2}$$

where $|M| = \sum_{i,j} M_{i,j}$ denotes the number of foreground pixels and ϵ is a stability constant. The background statistics $(\mu_{bg}^{(c)}, \sigma_{bg}^{(c)})$ are computed analogously using the inverse mask $(1-M)$. To achieve cross-dataset generalization, we align the statistics of images from the unseen target domain (Domain B) toward the canonical distributions of the source training domain (Domain A). Consequently, anatomical structures are statistically remapped to resemble their counterparts in Domain A. To account for modality and structural variations, these global reference statistics are calculated based on the specific imaging modality and organ. Let $\mathcal{D}_A = \{(I_{A,k}, M_{A,k})\}_{k=1}^{N}$ denote the set of source training images and their corresponding masks that match the given imaging modality and organ. The global foreground mean is defined as

$$\bar{\mu}_{fg}^{(c)} = \frac{1}{N_{fg}}\sum_{k=1}^{N}\sum_{i,j}\widetilde{I}_{A,k,i,j,c}\cdot\widetilde{M}_{A,k,i,j} \tag{3.3}$$

and the global foreground standard deviation is

$$\bar{\sigma}_{fg}^{(c)} = \sqrt{\frac{1}{N_{fg}} \sum_{k=1}^{N} \sum_{i,j} \left(\widetilde{I}_{A,k,i,j,c} - \bar{\mu}_{fg}^{(c)} \right)^2 \cdot \widetilde{M}_{A,k,i,j}} \tag{3.4}$$

where N_{fg} denotes the total number of foreground pixels across the specific images in Domain A. The global background statistics $\bar{\mu}_{bg}^{(c)}$ and $\bar{\sigma}_{bg}^{(c)}$ are defined analogously using $(1 - \widetilde{M}_{A,k}(x))$. If the imaging type or organ metadata are unavailable, the statistics are replaced with the overall global mean and standard deviation computed across the entire source dataset.

During inference, an image from domain B (I_B) is normalised as

$$\hat{I}_{fg}^{(c)} = \bar{\sigma}_{fg}^{(c)} \left(\frac{I_{B,c} - \mu_{fg}^{(c)}(I_B, M_B)}{\sigma_{fg}^{(c)}(I_B, M_B)} \right) + \bar{\mu}_{fg}^{(c)} \tag{3.5}$$

$$\hat{I}_{bg}^{(c)} = \bar{\sigma}_{bg}^{(c)} \left(\frac{I_{B,c} - \mu_{bg}^{(c)}(I_B, M_B)}{\sigma_{bg}^{(c)}(I_B, M_B)} \right) + \bar{\mu}_{bg}^{(c)} \tag{3.6}$$

Finally, the rescaled partitions are recombined as

$$I_{norm} = M_B \odot \hat{I}_{fg} + (1 - M_B) \odot \hat{I}_{bg} \tag{3.7}$$

where $\odot$ denotes the element-wise multiplication. By restricting the estimation of normalization statistics to anatomically relevant regions, AAIN aims to reduce the intensity discrepancies between domains while preserving structural boundaries.

Furthermore, it is important to note that AAIN is applied strictly as a pre-processing transformation during the inference phase. By remapping the target domain (Domain B) image statistics prior to tokenization, the VLM perceives the target image as if it were sampled from the source domain (Domain A) distribution, effectively neutralizing the domain shift without requiring any internal updates to the target model's parameters.

3.2.3 Weighted Multi-Modal Example Retrieval (WMMER)

Recent advancements in VLMs demonstrate significant performance gains through few-shot in-context learning (ICL). To further minimize the domain gap, we utilise a limited access to an annotated support set from the target domain, denoted as $\mathcal{S}_{adapt} = \{(I_s, Q_s, A_s)\}_{s=1}^{S}$, where I_s, Q_s, and A_s represent the image, question, and ground-truth answer, respectively. To dynamically construct the most informative few-shot prompt for a given test instance, we introduce Weighted Multi-Modal Example Retrieval (WMMER). Instead of relying solely on visual or textual similarity, WMMER retrieves exemplars by jointly optimizing for

both anatomical (visual) and semantic (question) relevance. Given a test query comprising an image I_q and a question Q_q, we extract their L_2-normalized feature embeddings using frozen visual and textual encoders, yielding $v_q = \mathcal{E}_V(I_q)$ and $t_q = \mathcal{E}_T(Q_q)$. The visual and textual embeddings (v_s, t_s) are pre-computed for all exemplars in $\mathcal{S}_{adapt}$. The multi-modal retrieval score for the s-th exemplar is formulated as a convex combination of visual and textual cosine similarities

$$S_{WMMER}(q, s) = w\ COS\_SIM(v_q,\ v_s) + (1 - w)\ COS\_SIM(t_q,\ t_s) \tag{3.8}$$

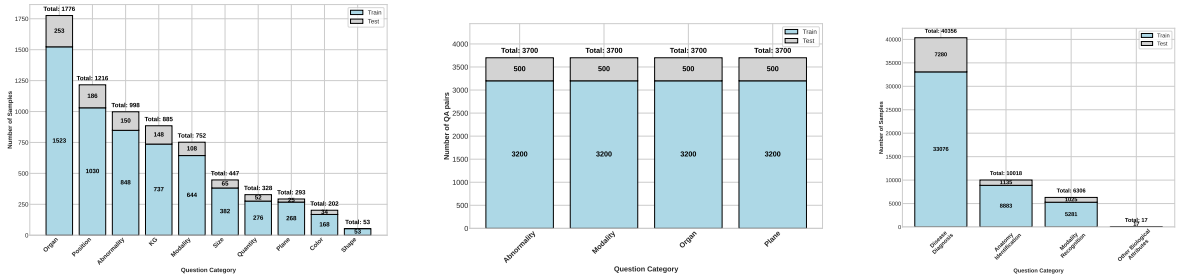
where $w \in [0, 1]$ is a tunable hyperparameter that dictates the relative importance of visual morphology versus query semantics. The top-K exemplars yielding the highest S_{WMMER} scores are retrieved and prepended to the test query, providing the model with explicitly aligned diagnostic context from the target domain.

The final multimodal input presented to the VLM follows an interleaved format: *"Context 1: [Image 1] [Question 1] [Answer 1] ... Context K: [Image K] [Question K] [Answer K]. Target: [Target Image] [Target Question]"*, guiding the model to utilize the provided structural reasoning to infer the target answer.

Chapter 4

Experiments and Results

4.1 Dataset Details



(a) SLAKE (b) VQA-Med 2019 (c) OmniMedVQA-RadImageNet

Figure 4.1: Distribution of question–answer (QA) pairs across different categories in the training and test partitions of the datasets.

We evaluate the proposed framework on three benchmark medical VQA datasets. The train–test distributions of question–answer (QA) pairs across different categories for the datasets used for evaluation are illustrated in Fig. 4.1.

- SLAKE [16]: It is a semantically labeled bilingual dataset containing radiology and clinical images with expert-annotated question–answer pairs across multiple categories (e.g., modality, organ, abnormality), for which we use only the English subset and official splits.
- VQA-Med 2019 [17]: This dataset consists of radiology images paired with visually answerable medical questions categorized into abnormality, modality, organ, and imaging plane identification, where we use the provided training split with an 80/20 train–validation division and evaluate on the validation set.

- OmniMedVQA–RadImageNet [18]: This dataset is derived from the RadImageNet subset of the OmniMedVQA benchmark, which is originally built upon the large-scale medical imaging collection introduced in [27]. The subset comprises 55,443 medical images and 56,697 question-answer pairs spanning three imaging modalities, namely CT, MRI, and ultrasound. In this study, we use the RadImageNet subset, which better aligns with the imaging modalities, anatomical regions, and question distributions in the SLAKE and VQA-Med 2019 datasets, thereby facilitating a more meaningful cross-domain evaluation. For experimental consistency, the dataset is partitioned into training, validation, and test sets using a stratified 70:10:20 split based on anatomical regions. The questions present in this dataset were already in Multiple choice format so we did not do any dataset preparation part mentioned in methodology chapter.

4.2 Vision-Language Baseline Architectures

To evaluate the robustness of the proposed Spatial Semantics-Aware Domain Adaptation (SSADA) framework, we benchmark its performance across five state-of-the-art vision-language foundation models spanning various parameter scales (from 2B to 12B parameters) and architectural paradigms. These models have used diverse strategies for cross-modal integration, ranging from cross-attention gated networks to natively unified multimodal transformers.

4.2.1 Gemma3-IT (12B)

Gemma3-IT (12B) [3] represents the advanced iteration of Google's open-weights LLM family, natively optimized for multimodal instruction following. Gemma3 represents an architecture that integrates vision-language capabilities via an inherently native, multi-modal transformer model, unlike previous vision-language systems that rely heavily on more loosely defined cross-modal projections. The model is able to map its inputs into a unified token space such that the autoregressive decoder layers can handle the image-text mixture at similar depths of computation.

In its 12B parameter configuration, the architecture manages a balance between compute efficiency and capacity using Grouped Query Attention (GQA) and multi-dimensional extensions of Rotary Position Embeddings (RoPE). The model is pre-trained primarily using large-scale corpora from visual data at high resolution and synthetic reasoning pathways, making Gemma3 highly tuned to fine-grained architecture and spatial features found in medical scans. Yet, it still is prone to domain-specific intensity variations before

fine-tuning.

4.2.2 Llama3.2-Vision-Instruct (11B)

Llama3.2-Vision-Instruct (11B) [28]is Meta's open multimodal foundation model designed to natively support image recognition, reasoning, and visual question answering. Structurally, the architecture integrates a pretrained Vision Transformer (ViT) image encoder with a Llama 3.2 text decoder. To bridge the structural gap between the visual and textual representations, the model infuses cross-attention layers directly into the transformer decoder stacks. These cross-attention gates selectively feed visual token representations into the language model layers during token generation.

A core advantage of the Llama3.2-Vision architecture is its native support for high-resolution image processing via a tile-based decomposition technique. Images which have been inputted and which are higher than regular resolutions are dynamically segmented into smaller tiles with equal aspect ratios, and these individual images will be independently encoded and finally integrated together with a downscaled global tile. Through such a technique, the algorithm would be able to capture detailed local abnormalities and contrasting edges like calcifications, thereby making it an excellent choice for analyzing medical images in distribution-shift settings.

4.2.3 Qwen2-VL Series (2B & 7B)

The Qwen2-VL [2] family is designed with important architectural features for handling multi-dimensional spatial grids. We test two different variants from this family for studying parameter scale scaling properties: Qwen2-VL-2B-Instruct and Qwen2-VL-7B-Instruct. These two models are equipped with the Naive Dynamic Resolution feature, which enables the vision transformer to handle images with any resolution and ratio without downsizing or altering geometrical scales.

Furthermore, the Qwen2-VL architecture introduces multimodal Rotary Position Embeddings (M-ROPE). Whereas conventional positional encodings treat image tokens in one-dimensional linear fashion, M-ROPE separates its positional indices into unique parts pertaining to the width, height, and time/text dimensions. Such structural representation provides an architectural advantage to the model, enabling it to understand exact absolute and relative positions and thus enabling it to perform well on tasks that require localized reasoning. The 2B version is more of an optimization of a basic model for resource-scarce clinical settings, while the 7B version boosts its hidden and attention layers to handle complex clinical diagnosis problems.

4.2.4 SmolVLM-Instruct (2B)

SmolVLM-Instruct (2B) [5] is a compact multimodal model developed by Hugging Face, specifically tailored for local deployment and parameter-efficient on-device operation. Structurally, SmolVLM addresses the significant compute footprint typically associated with vision encoders by coupling a lightweight language backend (SmolLM2-1.3B) with an ultra-efficient visual backbone, typically derived from the SigLIP (Sigmoid Language-Image Pretraining) architecture.

Visual tokens are projected into the text decoder's input embedding space via a streamlined, low-overhead multi-layer perceptron (MLP) adapter network. To maximize efficiency over extended context windows such as multi-turn dialogues or few-shot in-context learning prompts - SmolVLM utilizes an aggressive downsampling strategy that condenses spatial tokens without shedding high frequency semantic features.

Table 4.1: Architectural and structural profiles of the baseline Vision-Language Models used in the domain adaptation benchmarks.

Model Identifier	Parameter Scale	Attention Type	Vision Backbone Integration	Primary Spatial Alignment Paradigm
Gemma3-IT	12B	Grouped-Query	Unified Native Multimodal	Unified Embedding Space
Llama3.2-Vision-Instruct	11B	Grouped-Query	ViT via Interleaved Cross-Attention	Dynamic High-Resolution Tiling
Qwen2-VL-7B-Instruct	7B	Multi-Head	ViT via Dynamic Resolution MLP	Multimodal Rotary Embeddings (M-ROPE)
Qwen2-VL-2B-Instruct	2B	Multi-Head	ViT via Dynamic Resolution MLP	Multimodal Rotary Embeddings (M-ROPE)
SmolVLM-Instruct	2B	Grouped-Query	SigLIP via Low-Overhead MLP Projector	High-Density Token Downsampling

4.3 Implementation Details

We utilize a diverse set of VLMs, including Gemma3-IT (12B) [3], Llama3.2-Vision-Instruct (11B) [4], Qwen2-VL (2B & 7B) [2], and SmolVLM-Instruct (2B) [5], all obtained from the HuggingFace model hub. All models are finetuned and evaluated on a single NVIDIA A6000 GPU. We adopt LoRA (Low-Rank Adaptation) [29] with $r = 16$ and $\alpha = 32$, and apply gradient accumulation during training. Each model is trained for 2 epochs using the AdamW optimizer with a linear learning rate decay starting from 1×10^{-4}. All masks are generated using MedCLIP-SAM [21]. Visual and textual embeddings are computed using CLIP [30] and Sentence-BERT [31], respectively. The value of w in WMMER is fixed at 0.5 for all the experiments.

4.4 Results on Anatomical Region Label Prediction

A key step for the anatomical region mask generation is the anatomical region label. For this task, we employed the instruction-tuned MedGemma with 27B parameters

(MedGemma 27B-IT) [32]. To enable memory-efficient inference, the model was loaded using 4-bit NormalFloat (NF4) quantization. The model was evaluated on medical VQAs from the SLAKE, VQA-Med 2019, and OmniMedVQA–RadImageNet datasets. Ground-truth anatomical region labels were derived from the corresponding dataset metadata fields containing anatomical labels. Prior to inference, all unique anatomical regions present across the combined datasets were extracted to construct a candidate label space, which was subsequently incorporated into the prompting strategy to constrain the model outputs.

The model was prompted using a structured two-step reasoning framework designed to encourage both visual interpretation and anatomically grounded classification. The model was first instructed to describe the salient anatomical structures and imaging findings, followed by the prediction of the most appropriate anatomical region from the predefined candidate list. The exact prompt template used during inference is provided below:

Anatomical Region Label Prediction Prompt:

```
You are an expert radiologist analyzing a medical scan.

Step 1: Briefly describe the key anatomical structures
and findings visible in this image.

Step 2: Based on your visual analysis, identify the
specific body part or organ shown.

You MUST select exactly ONE of the following options:
[List of all possible anatomical locations].

You MUST use this exact format for your response:

Reasoning: [your analysis here]

Answer: [exact option here]
```

The final anatomical region prediction was extracted from the *Answer* field of the generated response. Using this framework, the model achieved anatomical region classification accuracies of 84% and 81% on the SLAKE and VQA-Med 2019 datasets, respectively, while achieving an accuracy of 80% on the OmniMedVQA–RadImageNet dataset.

4.5 Results on Generated Masks

To evaluate the consistency of the anatomical region segmentation masks generated by the MedCLIP-SAM model, we adopt the uncertainty-aware evaluation framework proposed by Sayed *et al.* [1]. Since the ground-truth annotation masks are unavailable in our setting, the evaluation is performed by comparing the segmentation masks obtained from the original images with the inverse-transformed masks generated from multiple augmented versions of the same images. As summarized in Table 4.2, the Dice similarity coefficient quantifies the spatial consistency between the original and augmentation-derived masks, while the uncertainty component measures voxel-wise disagreement across predictions from different augmentations. In our experiments, five augmentations are generated per image. Thus, this framework provides an unsupervised estimate of segmentation robustness and reliability for MedCLIP-SAM generated masks under image perturbations.

Table 4.2: The performance metrics corresponding to the anatomical region segmentation masks generated by MedCLIP-SAM across three datasets.

Dataset	Dice score	Uncertainty
SLAKE	0.6021	0.0793
VQA-Med 2019	0.5471	0.0793
OmniMedVQA–RadImageNet	0.5761	0.0830

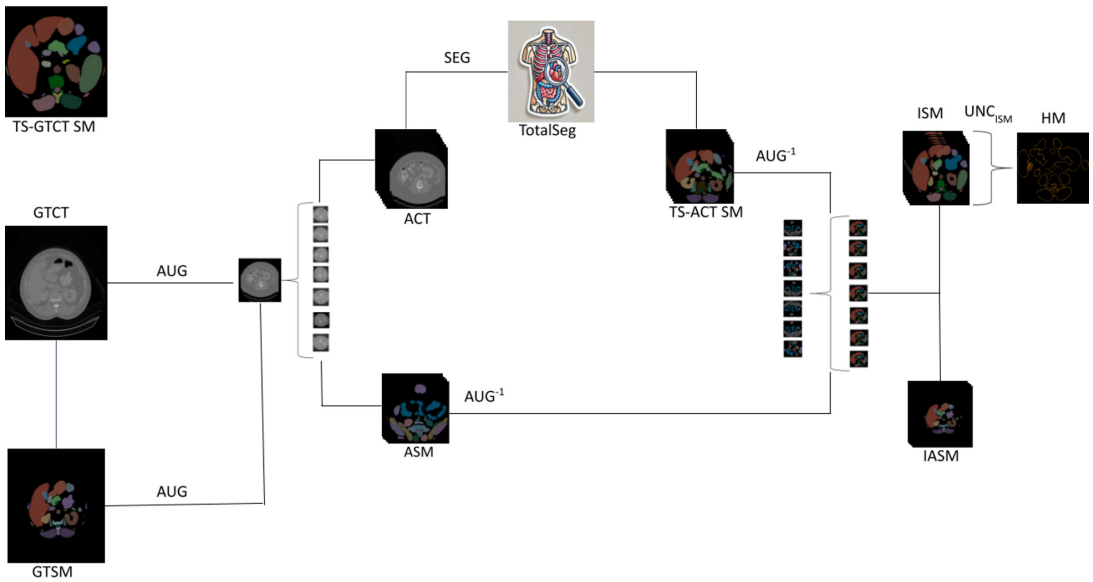


Figure 4.2: Overview of the uncertainty-aware evaluation framework. The original image and its augmented versions are processed by the segmentation model. The augmented predictions are then inverse-transformed back to the original coordinate space to compute the spatial consistency (Dice score) and voxel-wise uncertainty. Reproduced from Sayed *et al.* [1].

4.6 Results on Domain Adaptation

For each of the five VLMs, we evaluate four settings: (1) Zero-shot, where the pretrained model is directly tested on the target domain; (2) Supervised Finetuning (SFt) (No Mask), where the model is finetuned on the source domain without anatomical masks and evaluated on the target domain; (3) SFt (With Mask), where finetuning and evaluation incorporate anatomical region masks; and (4) SSADA (Full), where Mask-Aware Finetuning (MAFt) is performed on the source domain, followed by inference with AAIN and WMMER. For SLAKE and VQA-Med 2019, we report accuracy on the target-domain test set and present results only for the question types common to both datasets. Additional results are provided in the Supplementary Material. The best performance across the adaptation directions is achieved using top-1 example retrieval with WMMER.

Table 4.3: Accuracy of different fine-tuning strategies on VQA-Med 2019 and OmniMedVQA–RadImageNet datasets, evaluated across multiple VLMs finetuned on SLAKE.

Model	Method	VQA-Med 2019					OmniMedVQA–RadImageNet				
		Abnormality	Modality	Organ	Plane	Overall	Anatomy Identification	Disease Diagnosis	Modality Recognition	Other Biological Attributes	Overall
Gemma3-IT (12B)	Zero-shot	47.20	36.80	40.00	69.60	48.40	63.19	57.67	99.21	**35.29**	63.22
	SFt (No Mask)	63.20	81.80	81.40	**72.80**	74.80	70.71	57.94	**99.68**	0.00	64.78
	SFt (With Mask)	67.40	83.20	**82.20**	70.40	75.80	69.34	60.10	99.52	23.52	66.05
	SSADA	**70.20**	**86.70**	**82.20**	72.40	**77.87**	**73.48**	**63.65**	99.37	**35.29**	**69.32**
Llama3.2-VI (11B)	Zero-shot	28.40	40.20	45.20	84.20	49.45	71.98	26.63	73.37	41.17	37.16
	SFt (No Mask)	69.40	82.60	84.20	87.00	80.90	93.92	52.83	**99.90**	0.00	62.77
	SFt (With Mask)	73.40	87.80	**86.40**	85.00	83.15	93.03	55.09	**99.90**	41.17	64.81
	SSADA	**79.40**	**95.80**	**86.40**	**90.00**	**87.90**	**99.91**	**76.18**	99.32	**100.00**	**81.58**
Qwen2-VL (7B)	Zero-shot	49.20	63.80	80.80	**77.40**	67.80	36.36	39.81	98.24	29.41	45.73
	SFt (No Mask)	66.00	78.60	75.20	74.80	73.35	35.95	44.63	99.22	52.94	49.52
	SFt (With Mask)	68.80	81.00	**82.80**	76.00	77.15	44.58	47.75	99.71	70.59	53.04
	SSADA	**70.40**	**91.20**	81.20	72.80	**78.90**	**94.54**	**68.46**	**100**	**94.12**	**75.06**
Qwen2-VL (2B)	Zero-shot	50.60	73.40	77.80	53.20	63.75	51.81	62.11	99.52	**76.47**	64.43
	SFt (No Mask)	62.40	80.60	79.60	56.60	69.80	44.66	52.44	99.68	23.52	56.62
	SFt (With Mask)	**66.20**	82.20	**81.80**	58.60	72.15	50.92	57.14	**99.76**	23.52	60.73
	SSADA	64.20	**84.40**	79.00	**75.20**	**75.40**	**60.07**	**63.97**	92.55	41.17	**66.54**
SmolVLM (2B)	Zero-shot	47.40	53.40	74.60	**66.20**	60.25	62.82	53.26	88.68	29.41	58.20
	SFt (No Mask)	50.60	73.40	77.80	53.20	63.75	77.44	45.78	**99.22**	11.76	55.31
	SFt (With Mask)	54.60	72.60	78.60	56.60	65.60	73.74	52.10	98.24	5.88	59.50
	SSADA	**56.60**	**74.60**	**78.90**	57.60	**66.93**	**78.14**	**55.48**	97.17	**47.05**	**62.70**

The results for finetuning on the SLAKE dataset and testing on VQA-Med 2019 and OmniMedVQA–RadImageNet are given in Table 4.3. The "Overall" accuracy is computed as the weighted mean of the accuracies across all question types, and the best-performing results are highlighted in bold. Across both target datasets, the results demonstrate that SSADA achieves the highest overall accuracy across all five VLMs in the cross-domain setting, performing consistently better than or at par with other finetuning techniques across all the question categories.

The results for finetuning on the VQA-Med 2019 dataset and testing on the SLAKE

Table 4.4: Accuracy of different fine-tuning strategies on SLAKE and OmniMedVQA–RadImageNet datasets, evaluated across multiple VLMs finetuned on VQA-Med 2019.

Model	Method	SLAKE					OmniMedVQA–RadImageNet				
		Abnormality	Modality	Organ	Plane	Overall	Anatomy Identification	Disease Diagnosis	Modality Recognition	Other Biological Attributes	Overall
Gemma3-IT (12B)	Zero-shot	42.00	54.63	75.89	80.00	61.21	63.19	57.67	99.21	35.29	63.22
	SFt (No Mask)	48.00	**94.44**	79.45	68.00	74.24	74.02	60.19	99.68	0.00	66.96
	SFt (With Mask)	64.66	92.59	78.26	68.00	77.08	68.93	62.74	99.52	0.00	67.86
	SSADA	**86.20**	82.67	**91.67**	**80.24**	**84.00**	**83.72**	**65.84**	**99.76**	**76.47**	**72.75**
Llama3.2-VI (11B)	Zero-shot	32.67	66.67	52.57	32.00	49.27	71.98	26.63	73.37	41.17	37.16
	SFt (No Mask)	46.67	89.81	76.28	84.00	71.50	94.98	52.75	**99.90**	41.18	62.91
	SFt (With Mask)	62.67	88.89	78.26	88.00	74.34	94.89	55.38	99.80	17.66	64.87
	SSADA	**76.00**	**89.81**	**84.58**	**100.00**	**81.59**	**100.00**	**83.28**	99.71	**100.00**	**87.10**
Qwen2-VL (7B)	Zero-shot	48.67	66.67	71.54	84.00	66.00	36.36	39.81	98.24	29.41	45.73
	SFt (No Mask)	46.00	91.67	70.36	72.00	71.60	58.41	56.74	**99.32**	64.71	61.57
	SFt (With Mask)	66.00	89.91	70.36	76.00	74.44	56.22	59.43	**99.32**	23.53	63.30
	SSADA	**82.00**	**93.52**	**78.66**	**90.00**	**83.74**	**61.78**	**64.23**	97.37	**94.12**	**67.57**
Qwen2-VL (2B)	Zero-shot	64.67	82.41	78.66	32.00	68.66	51.81	**62.11**	**99.52**	**76.47**	64.43
	SFt (No Mask)	56.67	90.74	**80.63**	36.00	70.32	52.65	58.79	99.45	23.50	62.17
	SFt (With Mask)	58.00	**93.52**	76.68	72.00	72.38	50.96	60.44	99.37	23.53	63.05
	SSADA	**66.67**	**93.52**	79.05	**84.00**	**75.81**	**66.80**	61.81	98.26	23.53	**66.69**
SmolVLM (2B)	Zero-shot	43.33	68.52	61.26	28.00	59.05	62.82	53.26	88.68	29.41	58.20
	SFt (No Mask)	40.00	70.37	71.15	24.00	61.02	63.61	55.43	81.66	0.00	59.15
	SFt (With Mask)	47.33	86.11	72.73	28.00	64.64	64.23	56.60	90.93	70.59	61.15
	SSADA	**52.67**	**87.90**	**74.70**	**52.00**	**69.24**	**67.92**	**58.22**	**97.95**	**76.47**	**63.73**

Table 4.5: Accuracy of different fine-tuning strategies on SLAKE and VQA-Med 2019 datasets, evaluated across multiple VLMs finetuned on OmniMedVQA–RadImageNet.

Model	Method	SLAKE					VQA-Med 2019				
		Abnormality	Modality	Organ	Plane	Overall	Abnormality	Modality	Organ	Plane	Overall
Gemma3-IT (12B)	Zero-shot	42.00	54.63	75.89	80.00	61.21	47.20	36.80	40.00	**69.60**	48.40
	SFt (No Mask)	43.33	**80.55**	77.47	68.00	66.79	53.40	**87.80**	74.20	54.60	67.50
	SFt (With Mask)	60.66	76.85	80.63	72.00	72.57	62.20	87.00	72.60	54.00	69.05
	SSADA	**80.66**	78.70	**88.53**	**84.00**	**78.35**	**67.60**	86.80	**77.60**	60.00	**73.00**
Llama3.2-VI (11B)	Zero-shot	32.67	66.67	52.57	32.00	49.27	28.40	40.20	45.20	84.20	49.45
	SFt (No Mask)	38.67	77.80	71.93	68.00	63.36	57.40	77.00	78.20	65.60	69.66
	SFt (With Mask)	58.00	85.18	77.86	76.00	72.47	65.80	80.00	77.80	80.00	75.90
	SSADA	**71.33**	**92.38**	**85.37**	**88.00**	**80.90**	**71.00**	**83.40**	**82.00**	**84.40**	**80.45**
Qwen2-VL (7B)	Zero-shot	48.67	66.67	71.54	84.00	66.00	49.20	63.80	80.80	77.40	67.80
	SFt (No Mask)	42.00	81.48	63.64	68.00	63.08	53.00	76.80	81.00	77.20	72.00
	SFt (With Mask)	74.66	**90.74**	69.96	**84.00**	74.33	59.40	81.40	80.00	**78.80**	74.90
	SSADA	**86.67**	89.81	**74.70**	72.00	**79.92**	**64.60**	**85.20**	**84.60**	78.60	**78.25**
Qwen2-VL (2B)	Zero-shot	64.67	82.41	78.66	32.00	68.66	50.60	73.40	**77.80**	53.20	63.75
	SFt (No Mask)	42.00	75.00	79.44	68.00	59.45	59.20	72.60	71.20	63.40	66.60
	SFt (With Mask)	62.00	89.91	78.26	68.00	66.20	**62.20**	77.60	72.00	60.20	68.05
	SSADA	**66.00**	**90.74**	**80.63**	**76.00**	**69.63**	61.80	**81.20**	74.20	**69.40**	**71.65**
SmolVLM (2B)	Zero-shot	43.33	68.52	61.26	28.00	59.05	54.60	72.60	78.60	56.60	65.60
	SFt (No Mask)	40.67	38.89	52.57	**68.00**	49.66	45.40	69.80	75.40	50.20	60.20
	SFt (With Mask)	**63.33**	**75.00**	66.40	60.00	62.59	55.80	70.60	77.60	57.60	65.40
	SSADA	58.00	72.22	**71.94**	44.00	**64.54**	**58.00**	**75.40**	**78.80**	**62.20**	**68.60**

and OmniMedVQA–RadImageNet dataset are given in Table 4.4, while detailed results for additional categories are provided in the Table 4.6. The "Overall" accuracy is computed as the weighted mean of the accuracies across all question types, and the best-

performing results are highlighted in bold. For the SLAKE dataset, supervised finetuning with masking enhances modality prediction; however, improvements in abnormality, plane, and organ recognition remain limited compared to the proposed approach. For the OmniMedVQA–RadImageNet dataset, the proposed SSADA framework delivers substantial gains in anatomy identification, disease diagnosis and other biological attribute prediction, demonstrating improved robustness under domain shift.

The results for finetuning on the OmniMedVQA–RadImageNet dataset and testing on the VQA-Med 2019 and SLAKE dataset are given in Table 4.5, while detailed results for additional categories are provided in the Table 4.7. The "Overall" accuracy is computed as the weighted mean of the accuracies across all question types, and the best-performing results are highlighted in bold. For both datasets, the proposed SSADA framework yields the largest gains in abnormality prediction, suggesting that anatomically guided supervision enables the models to focus on diagnostically relevant regions. In addition to this, enhancements are constantly seen for both big and small VLMs, implying that the method introduced is highly efficient and generalizable. These results show that the use of anatomically informed adaptation leads to much greater robustness and generalization ability for visual question answering in medicine.

The annotations used for the experiments on domain adaptation were derived straight from the data set metadata, and included anatomical region tags such as lungs, brain, etc. For practical application of the models in the clinic, the region tag would either be annotated by medical professionals or automatically classified using the VLM, which proved to yield accuracies greater than 80% as demonstrated in Section 4.4.

However, a more in-depth analysis of the findings shows that the size of the model is correlated with the choice of adaptation technique. Traditionally, expanding the number of parameters (from 2B to 12B) is used as a default means of addressing domain shifts. At the same time, the results show that smaller models (Qwen2-VL and SmolVLM, both 2B) can compensate for the lack of parameters when paired with a complete version of the proposed solution and reach competitive performance compared to their much larger counterparts (11B and 12B). The data suggest that rather than adding more parameters, explicit spatial grounding and normalization can be used as a more efficient tool to ensure medical robustness.

Also, the discrepancy in accuracy depending on a question category helps to understand which mechanisms lie behind this process. The modality detection task is relatively simple to address with conventional finetuning since the required visual patterns are global (high contrast of CT images against the grain texture of ultrasounds). On the contrary, questions about abnormalities or particular slices require localized reasoning. The fact that there is a significant boost in accuracy in these categories supports the

Table 4.6: Accuracy of the remaining question categories on various finetuning techniques across multiple VLMs, where model is trained on VQA-Med 2019 and evaluated on SLAKE dataset.

Model	Method	Color	KG	Position	Quantity	Size	Overall
Gemma3-IT (12B)	Zero-shot	64.71	51.35	62.90	36.53	87.69	61.21
	SFt (No Mask)	**73.53**	79.73	73.66	71.15	89.23	74.24
	SFt (With Mask)	50.00	83.11	76.88	65.38	89.23	77.08
	SSADA	70.59	**83.11**	**83.87**	**73.08**	**90.77**	**84.00**
Llama3.2-VI (11B)	Zero-shot	35.29	35.14	55.91	25.00	**92.32**	49.27
	SFt (No Mask)	26.47	72.97	73.73	65.38	90.77	71.50
	SFt (With Mask)	32.35	68.92	74.19	73.08	92.31	74.34
	SSADA	**82.35**	**81.08**	**75.81**	**75.00**	84.62	**81.59**
Qwen2-VL (7B)	Zero-shot	47.08	70.95	63.44	57.69	90.77	66.00
	SFt (No Mask)	55.88	83.78	70.43	63.46	**92.31**	71.60
	SFt (With Mask)	50.00	81.08	70.97	73.08	**92.31**	74.44
	SSADA	**85.29**	**87.16**	**81.72**	**75.00**	90.77	**83.74**
Qwen2-VL (2B)	Zero-shot	38.24	68.92	66.13	19.23	92.31	68.66
	SFt (No Mask)	47.06	72.30	69.35	19.23	92.31	70.32
	SFt (With Mask)	41.18	77.70	73.11	**23.07**	**95.38**	72.38
	SSADA	**70.59**	**83.78**	**76.34**	15.38	83.08	**75.81**
SmolVLM (2B)	Zero-shot	26.47	68.24	60.75	50.00	81.54	59.05
	SFt (No Mask)	32.35	**72.30**	63.44	17.32	**86.15**	61.02
	SFt (With Mask)	32.35	66.89	**68.82**	21.15	**86.15**	64.64
	SSADA	**38.23**	71.62	66.67	**61.50**	**86.15**	**69.24**

initial assumption: without being constrained explicitly, language models are likely to employ shortcut learning; in this case, the mask technique is designed precisely to make the model ground its answers spatially.

4.7 Ablation Study

To assess the individual contributions of the proposed components, we perform an ablation study by isolating AAIN and WMMER within the SSADA framework. We train three representative VLMs — Gemma3-IT (12B), Llama3.2-Vision-Instruct (11B), and Qwen2-VL (7B)—using MAFt finetuning technique on the source dataset. During inference on the target dataset, we independently evaluate three variants: (i) AAIN, where only anatomical region-based normalisation is applied, (ii) WMMER_{s}, where only retrieval-based inference

Table 4.7: Accuracy of the remaining question categories on various finetuning techniques across multiple VLMs, where model is trained on OmniMedVQA–RadImageNet and evaluated on SLAKE dataset.

Model	Method	Color	KG	Position	Quantity	Size	Overall
Gemma3-IT (12B)	Zero-shot	64.71	51.35	62.90	36.53	**87.69**	61.21
	SFt (No Mask)	61.76	72.29	63.97	61.53	58.46	66.79
	SFt (With Mask)	**76.47**	**84.45**	66.66	59.61	60	72.57
	SSADA	55.88	74.32	**75.26**	**69.23**	67.69	**78.35**
Llama3.2-VI (11B)	Zero-shot	35.29	35.14	55.91	25.00	**92.32**	49.27
	SFt (No Mask)	20.58	61.48	67.40	69.23	69.23	63.36
	SFt (With Mask)	29.41	64.86	79.02	82.69	75.38	72.47
	SSADA	**76.47**	**78.37**	**82.79**	**88.46**	64.61	**80.90**
Qwen2-VL (7B)	Zero-shot	47.08	70.95	63.44	57.69	90.77	66.00
	SFt (No Mask)	11.76	67.57	65.59	63.46	86.15	63.08
	SFt (With Mask)	58.82	75.67	68.27	61.54	**92.31**	74.33
	SSADA	**76.47**	**80.41**	**74.73**	**73.08**	**92.31**	**79.92**
Qwen2-VL (2B)	Zero-shot	38.24	**68.92**	**66.13**	19.23	**92.31**	68.66
	SFt (No Mask)	23.52	50.68	54.79	63.46	75.38	59.45
	SFt (With Mask)	35.29	54.72	50	59.61	83.07	66.20
	SSADA	**52.94**	54.72	52.15	**71.15**	84.61	**69.63**
SmolVLM (2B)	Zero-shot	26.47	68.24	60.75	50.00	**81.54**	59.05
	SFt (No Mask)	14.71	61.49	46.24	**57.69**	64.62	49.66
	SFt (With Mask)	29.41	64.86	57.53	28.85	80	62.59
	SSADA	**38.24**	**70.27**	**63.98**	23.08	**81.54**	**64.54**

is performed by taking top-s examples, and (iii) the combined AAIN+WMMER$_s$ framework. The quantitative results for the cross-domain directions (*SLAKE* $\rightarrow$ *VQA-Med 2019* and *VQA-Med 2019* $\rightarrow$ *SLAKE*) are presented in Table 4.8 and Table 4.9, enabling a direct comparison of the relative impact of the individual components on domain adaptation performance. The detailed ablation results, highlighting the contribution of individual components (AAIN, WMMER$_s$ and AAIN+WMMER$_s$) on different question types, are provided in tables in this section.

Analyzing the individual impacts of different modules in the framework within the ablation tables shows relationship between normalization and retrieval. When evaluated independently, AAIN consistently elevates the baseline accuracy by neutralizing lower-level intensity mismatches, effectively presenting a standardized target image to the VLM. Conversely, WMMER introduces high-level semantic reasoning by retrieving historically relevant clinical cases.

Table 4.8: The question category-wise and overall accuracy of the SSADA finetuning technique using only AAIN and $WMMER_s$ across multiple VLMs, where models are trained on SLAKE and evaluated on VQA-Med 2019 dataset. $WMMER_s$ denotes that the top-s examples are retrieved for inference.

Model	Method	Abnormality	Modality	Organ	Plane	Overall
Gemma3-IT (12B)	**AAIN**	68.90	84.80	**83.40**	71.80	77.22
	$WMMER_1$	67.50	85.80	82.40	72.60	77.08
	$WMMER_1$ +AAIN	**70.20**	**86.70**	82.20	72.40	**77.87**
	$WMMER_2$	66.20	86.10	81.00	**73.80**	76.75
	$WMMER_2$ +AAIN	65.40	86.60	83.20	**73.80**	77.25
Llama3.2-VI (11B)	**AAIN**	76.30	93.50	84.80	89.10	85.92
	$WMMER_1$	73.10	89.50	**89.30**	87.00	84.73
	$WMMER_1$ +AAIN	**79.40**	**95.80**	86.40	**90.00**	**87.90**
	$WMMER_2$	61.60	80.40	79.60	75.00	74.15
	$WMMER_2$ +AAIN	78.20	83.40	78.80	72.40	78.20
Qwen2-VL (7B)	**AAIN**	72.40	82.60	82.20	77.00	78.55
	$WMMER_1$	69.20	82.20	**83.40**	72.60	76.85
	$WMMER_1$ +AAIN	**70.40**	**91.20**	81.20	72.80	**78.90**
	$WMMER_2$	65.20	84.00	79.80	80.20	77.30
	$WMMER_2$ +AAIN	64.80	84.60	81.20	**80.60**	77.80

Table 4.9: The question category-wise and overall accuracy of the SSADA finetuning technique using only AAIN and $WMMER_s$ across multiple VLMs, where models are trained on VQA-Med 2019 and evaluated on SLAKE dataset. $WMMER_s$ denotes that the top-s examples are retrieved for inference.

Model	Method	Abn.	Color	KG	Mod.	Organ	Plane	Pos.	Qty.	Size	Overall
Gemma3-IT (12B)	**AAIN**	55.33	47.06	79.73	91.67	75.89	80.00	76.88	71.15	**92.31**	77.76
	$WMMER_1$	80.00	70.59	83.11	91.67	82.21	76.00	80.11	75.00	89.23	82.17
	$WMMER_1$ +AAIN	**86.20**	70.59	83.11	82.67	**91.67**	**80.24**	83.87	73.08	90.77	84.00
	$WMMER_2$	82.00	79.41	83.10	**97.22**	82.60	76.00	79.56	75.00	89.23	83.34
	$WMMER_2$ +AAIN	82.00	**88.24**	**85.14**	92.59	85.38	68.00	**86.02**	**76.92**	89.23	**85.21**
Llama3.2-VI (11B)	**AAIN**	62.00	32.35	71.62	88.89	77.47	88.00	76.88	**76.92**	89.23	74.93
	$WMMER_1$	73.33	82.35	**83.11**	88.89	79.45	**100.00**	76.88	65.38	80.00	79.53
	$WMMER_1$ +AAIN	**76.00**	82.35	81.08	89.81	84.58	**100.00**	75.81	75.00	84.62	**81.59**
	$WMMER_2$	64.66	76.47	80.45	**93.51**	**89.44**	96.00	75.80	67.30	**90.76**	78.64
	$WMMER_2$ +AAIN	70.00	**85.29**	81.08	92.59	80.24	**100.00**	**77.42**	65.38	89.23	80.12
Qwen2-VL (7B)	**AAIN**	**84.00**	70.59	75.68	92.59	73.52	76.00	79.57	67.31	89.23	79.14
	$WMMER_1$	80.00	50.00	81.08	**95.37**	69.57	76.00	79.03	**76.92**	**95.38**	78.75
	$WMMER_1$ +AAIN	82.00	85.29	**87.16**	93.52	78.66	90.00	81.72	75	90.77	**83.74**
	$WMMER_2$	79.33	**88.24**	85.81	**95.37**	78.26	**96.00**	81.72	73.08	87.69	83.06
	$WMMER_2$ +AAIN	80.00	82.35	85.81	**95.37**	**80.63**	92.00	**82.26**	71.15	86.15	83.35

Interestingly, the data reveals a nuanced behavior regarding the number of retrieved exemplars (s). While one might assume that providing more context ($WMMER_2$) would strictly improve reasoning, our results indicate that $WMMER_1$ often yields superior or highly competitive accuracy compared to $WMMER_2$.We observed that the combination of both the AAIN and WMMER model yields optimal results in terms of stability on all

tested datasets. Thus, it is confirmed that the proposed two models are not redundant but are indeed complementary since each tackles a certain level within the domain shift hierarchy, specifically, AAIN closes the gap in the pixel distribution level, whereas WMMER closes the gap in diagnostic reasoning.

Chapter 5

Conclusion and Future Work

This paper addresses the underexplored challenge of domain adaptation in VLM-based Medical VQA. We propose a Spatial Semantics Aware Domain Adaptation framework that facilitates effective domain adaptation by leveraging spatially grounded learning, reducing cross-domain discrepancies, and selecting relevant domain-specific examples. We evaluated SSADA on SLAKE, VQA-Med-2019 and OmniMedVQA-RadImageNet, where it consistently outperformed the standard baselines. Extensive ablation studies further validate the contribution of each module. Also, our different experiment evaluations shows that explicitly grounding models with anatomical masks (MAFt) and aligning target distributions locally (AAIN) allows small with 2B parameter models and models like 11B or 12B models every segment has performance improvement better than baseline which shows consistency with different architecture of VLM. Moreover, in our experiments using ablation, we discovered an interesting characteristic about retrieval-based adaptation where choosing a single highly relevant multi-modal exemplar ($WMMER_1$) generally proves to be more effective than utilizing multiple exemplars, as this reduces the effect of semantically contradictory noise introduced during learning.

Despite these innovations, there are still some limitations to the existing method. Firstly, as shown in Table 4.2, the effectiveness of Anatomy-Aware Instance Normalization heavily relies on the quality of the masks generated by MedCLIP-SAM. While our experiment regarding uncertainty shows acceptable robustness to image perturbation, extremely distorted images can adversely impact mask quality, hence affecting the final results of the models. Secondly, it should be noted that the WMMER block assumes the existence of a manually annotated and pre-selected small subset of images in the target domain. In terms of clinical application, the proposed framework can be helpful for the doctors because it can generate reliable answers verifiable and understandable by the clinicians. The fact that the answers are generated within the constrained-answer settings also reduces the risk of the generation of clinically unrelated answers, which can be crucial in healthcare.

For future research, exploring more sophisticated prompting mechanisms, including Chain-of-Thought reasoning, as well as advanced fine-tuning methods, would help to assess model robustness better. Another direction for future improvement could be expanding the framework to process three-dimensional medical imaging volumes as opposed to 2D slices, as many diseases require understanding the continuous three-dimensional context. Finally, purely unsupervised domain adaptation without relying on the existence of a target-domain support set can potentially eliminate many difficulties and make deployment easier.

Bibliography

[1] Mattin Sayed, Sari Saba-Sadiya, Benedikt Wichtlhuber, Julia Dietz, Matthias Neitzel, Leopold Keller, Gemma Roig, and Andreas M Bucher. Evaluating medical image segmentation models using augmentation. *Tomography*, 10(12):2128–2143, 2024.

[2] Peng Wang and et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[3] Gemma Team and et al. Gemma 3 technical report, 2025.

[4] Meta AI. Llama 3.2 model card, 2024.

[5] Andrés Marafioti and et al. SmolVLM: Redefining small and efficient multimodal models. In *Second Conference on Language Modeling*, 2025.

[6] Hong Xia, Yifan Zhang, Hui Jia, Yanping Chen, Jing Xu, and Shiyong Li. A medical visual question-answering model based on multi-scale feature fusion and question feature enhancement. *Multimedia Systems*, 31(4):271, 2025.

[7] Yuetian Du, Yucheng Wang, Luyuan Chen, Jinjian Zhang, Wei Zhou, Ming Kong, and Qiang Zhu. CARE: Confidence-aware REasoning for reliable medical-VQA, 2025.

[8] Xingxuan Zhang, Jiansheng Li, Wenjing Chu, Junjia Hai, Renzhe Xu, Yuqing Yang, Shikai Guan, Jiazheng Xu, Liping Jing, and Peng Cui. On the out-of-distribution generalization of large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10315–10326, June 2025.

[9] Hasan Farooq, Murtaza Taj, Mehwish Nasim, and Arif Mahmood. Localization Lens for Improving Medical Vision-Language Models . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15968. Springer Nature Switzerland, September 2025.

[10] Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu:

A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

[11] Jee Seok Yoon, Kwanseok Oh, Yooseung Shin, Maciej A Mazurowski, and Heung-Il Suk. Domain generalization for medical image analysis: A review. *Proceedings of the IEEE*, 112(10):1583–1609, 2024.

[12] Kevin Vogt-Lowell, Noah Lee, Theodoros Tsiligkaridis, and Marc Vaillant. Robust fine-tuning of vision-language models for domain generalization. In *IEEE High Performance Extreme Computing Conference (HPEC)*, pages 1–7. IEEE, 2023.

[13] Cairong Zhao, Yubin Wang, Xinyang Jiang, Yifei Shen, Kaitao Song, Dongsheng Li, and Duoqian Miao. Learning domain-invariant prompt for vision-language models. *IEEE Transactions on Image Processing*, 33:1348–1360, 2024.

[14] Jiawei Chen, Yue Jiang, Dingkang Yang, Mingcheng Li, Jinjie Wei, Ziyun Qian, and Lihua Zhang. Can LLMs' Tuning Methods Work in Medical Multimodal Domain? . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15005. Springer Nature Switzerland, October 2024.

[15] Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. Mmed-rag: Versatile multimodal rag system for medical vision-language models. In *International Conference on Learning Representations (ICLR)*, 2025.

[16] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

[17] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019, 2019.

[18] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183, 2024.

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[20] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024.

[21] Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. MedCLIP-SAM: Bridging Text and Image Towards Universal Medical Image Segmentation . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15012. Springer Nature Switzerland, October 2024.

[22] Peter D Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188, 2010.

[23] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[24] Qingyi Si, Fandong Meng, Mingyu Zheng, Zheng Lin, Yuanxin Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. Language prior is not the only shortcut: A benchmark for shortcut learning in vqa. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3698–3712, 2022.

[25] Omid Reza Abbasi, Ali Asghar Alesheikh, and Aynaz Lotfata. Semantic similarity is not enough: A novel nlp-based semantic similarity measure in geospatial context. *IScience*, 27(6), 2024.

[26] Shaochen Xu, Zihao Wu, Huaqin Zhao, Peng Shu, Zhengliang Liu, Wenxiong Liao, Sheng Li, Andrea Sikora, Tianming Liu, and Xiang Li. Reasoning before comparison: Llm-enhanced semantic similarity metrics for domain specialized text analysis. *arXiv preprint arXiv:2402.11398*, 2024.

[27] Xueyan Mei, Zelong Liu, Philip M Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E Link, Thomas Yang, et al. Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 4(5):e210315, 2022.

[28] Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *IEEE Transactions on Medical Imaging*, 2026.

[29] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[31] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[32] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.