PRS-Med: Position Reasoning Segmentation in Medical Imaging

Quoc-Huy Trinh[1,2], Minh-Van Nguyen[3], Jun Zeng[4], Debesh Jha[5,†], Ulas Bagci[2,†]
[1]Aalto University [2]Northwestern University [3]Technical University of Denmark
[4]Chongqing University of Posts and Telecommunications [5]University of South Dakota
[†]Co-advising https://huyquoctrinh.github.io/prsmed/

Abstract

Prompt-based medical image segmentation has rapidly emerged, yet existing methods rely on explicit prompts like bounding boxes and struggle to reason about the spatial relationships essential for clinical diagnosis. While general-domain models attempt complex coordinate regression, these approaches often lack the structured reliability required for medical applications. In this work, we introduce PRS-Med, a unified framework that adopts an elegant, clinical-first approach to position reasoning segmentation. By utilizing a medical vision-language model integrated with a segmentation decoder, PRS-Med mimics the structured "search patterns" used by radiologists to identify pathologies within specific anatomical zones. To support this robust reasoning, we present the Medical Position Reasoning Segmentation (PosMed) dataset, comprising 116,000 expert-validated, spatially grounded question-answer pairs across six imaging modalities. Unlike previous brittle attempts at spatial reasoning, PosMed leverages a scalable, deterministic pipeline validated by board-certified radiologists to ensure clinical accuracy. Extensive experiments demonstrate that our zone-based reasoning not only improves segmentation accuracy (mean Dice improvements up to +31.2%) but also provides a high-confidence interpretability layer that outperforms state-of-the-art complex reasoning models. By prioritizing functional reliability over unnecessary technical complexity, PRS-Med offers a practical and scalable baseline for the next generation of intelligent medical assistants.

1. Introduction

In the medical field in general and oncology in particular, doctors typically make diagnoses by examining potential tumor locations and types to evaluate tissue conditions. This makes position reasoning and segmentation visualization crucial for supporting early and accurate diagnoses. As medical assistant agents become more common, models like LLaVA-Med [24], Med-MoE [21], HuatuoGPT [7], and MedVLM-R1 [33] have been developed and shown potential in the diagnostic. However, they still face a challenge in position identification when doctors often need to identify unknown tumor locations to make a decision for their diagnosis. Additionally, the position information can help doctors recognize the growing tumors, which leads to more effective diagnosis and treatment. This technology can also help clinics create automated screening systems, reducing manual costs.

In the natural image domain, several works such as LISA [23], LLM-SEG [44], and SegLLM [47] have addressed the challenge of reasoning for segmentation, achieving notable success in enhancing object reasoning, identifying object positions through segmentation, and providing simple reasoning about objects. However, these Multimodal-LLMs are not well-trained on medical imaging, making their application to this field difficult. This is due to the complex nature of medical content and the difficulty of boundary learning in medical segmentation, which out-of-domain models struggle with. For position reasoning segmentation, a VLM's vision model needs to be well-trained on medical images to effectively distinguish and localize tumors and anatomies for reasoning.

We present **PRS-Med**, a framework for **P**osition **R**easoning **S**egmentation in medical imaging. This is a unified method that uses a Multimodal-LLM to perform position-reasoning segmentation from simple questions or commands. Our model outputs both a textual description and a segmentation mask that highlights the tumor location. PRS-Med acts as an intelligent assistant, answering a doctor's questions and visually indicating the position of tumors or anatomical structures in an interpretable way. In addition, general-domain models strive for pixel-perfect coordinate regression; PRS-Med prioritizes Clinical Action Zones, ensuring that the spatial reasoning mimics the structured language of board-certified radiologists. Our contributions are as follows:

- To address the lack of datasets for position reasoning in medical imaging, we create and release the Medical Position Reasoning Segmentation (PosMed) dataset

pipeline. This pipeline can build a coarse position reasoning dataset, designed to generate diverse, spatially grounded question-answer pairs in the medical context.
- We present PRS-Med, a position reasoning segmentation framework for Medical Imaging. It performs spatially-aware tumor segmentation using implicit natural language prompts.
- We are open-sourcing the dataset pipeline, model, and codebase to help the community develop spatially-aware multimodal LLMs in medical imaging.

2. Related Work

Medical Image Segmentation. Traditional Medical image segmentation has long relied on fully supervised CNN-based architectures like U-Net [36] and its variants, such as ResUNet++ [18], nnu-net [17], DoubleUNet [19], TransResUNet [41], Swin-UNet [5], and MEGANet [4]. More recently, several promptable segmentations have emerged as a response to the growing demand for interactive and context-aware medical AI. MedSAM [30], SAM-Med2D [12] adapts the Segment Anything Model (SAM) [22] to medical settings, supporting box- and point-prompted segmentation. However, MedSAM and SAM-Med2D still lack semantic understanding of positional cues within free-form text. In contrast, BiomedParse [51] directly uses text prompts to infer object shapes and positions, learning implicit position priors. Despite the promising results of previous works, they **lack the ability to segment with contextual reasoning**, nor does it support implicit or conversational prompts beyond class names prompt input. In this work, we propose the PRS-Med, which allows the MLLMs perform position reasoning with segmentation, enabling an interactive framework that responds to contextual questions and generates both position reasoning outputs and corresponding segmentation masks.

Multimodal-LLM in Medical Imaging. Multimodal large language models (MLLMs) have recently shown promising results in medical image reasoning tasks. Prominent methods such as Med-Flamingo [32], Med-MoE [21], GSCo [15], HuatuoGPT [7], and MedVLM-R1 [33] are built upon vision and Language models like LLaVA [26], Qwen2-VL [46], and Multimodal Llama [42] through the training technique via visual instruction tuning or reinforcement learning methods. Despite their promising results, these models struggle with medical image segmentation, which is a critical task for accurate disease diagnosis. Moreover, they also inherit spatial reasoning limitations from their Multimodal-LLMs, which are observed in SpatialVLM [6], Loc-VLM [35], and Spatialrgpt [9]. To address the spatial reasoning challenge, we propose PosMed, a position reasoning dataset to perform the localization understanding of the MLLMs in the medical domain. This represents an initial step toward spatial reasoning, and it can be extended for further grounding and spatial perception of the MLLMs in the medical domain.

Reasoning Image Segmentation. Recent advancements in reasoning segmentation have begun to integrate high-level reasoning, particularly through the use of the Multimodal-LLMs. Notable works in the domain include LISA [23], LLM-Seg [45], and SegLLM [47], which include a special [SEG] token used to compact segmentation-specific embeddings from the model. Although the initial success of previous works, they are still limited to the ability to perform single-object identification, and cannot perform long-form reasoning for segmentation. Moreover, recent MLLMs in the medical domain restrict the ability to diagnose, which is difficult to deal with, with more advanced reasoning abilities such as spatial reasoning and grounding ability. To address this limitation, the PosMed dataset is designed for dealing with the limitations in the position reasoning segmentation. In addition, PRS-Med is a unified model with the reasoning and segmentation ability for position-aware reasoning, along with the segmentation visualization. Our contributions aim to enhance the ability of the MLLMs in spatial reasoning, along with the segmentation visualization for the interpretation, which is important in the clinician application.

3. PosMed dataset

To address the challenge of the segmentation along with the position reasoning, we propose the PosMed dataset pipeline (presented in Figure 1), a two-stage dataset pipeline for position reasoning segmentation, which leverages the segmentation mask annotations to create the position annotation for the position reasoning, and by leveraging LLM collaborates with the doctor, we can semi-automate the dataset creation process. As a result, we release the PosMed dataset, a large-scale position reasoning segmentation.

Data source. We create the PosMed dataset by curating and extracting the spatial information from several data-sources, including BUSI [1] for breast; Brain MRI [10, 11] for brain tumors in the MRI; LungCT dataset [49] for the lesions in the CT modality, and Lung Xray [13, 34] for the lesion in the Xray modality; Kvasir-SEG [20], ClinicDB [3], CVC300 [43], ColonDB [40], and ETIS [38] for the polyp in the endoscopy images; and ISIC [14] for the skin lesion in the RGB image domain. In overall, there are total of 38731 images in our dataset. These data sources ensure the diversity and the variety in the medical image modalities, anatomy types, and tumor types in the medical domain.

Dataset Construction Pipeline. To begin, we leverage the GPT-4 model to generate 55 different question-and-answer templates (pairs) based on the mentioned question-answer pair. All of the question and answer pairs always include these two pieces of information <*tumor/anatomy name*>, and <*position*>. These templates are then validated by

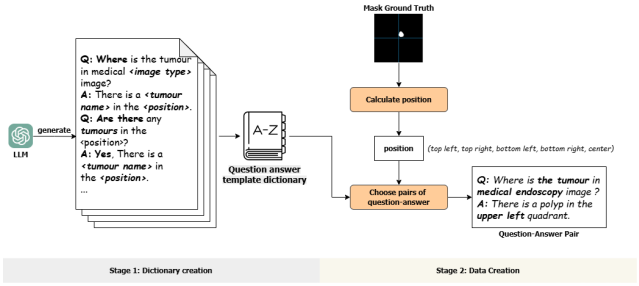

Figure 1. Visualization of two stages of the PosMed dataset pipeline.

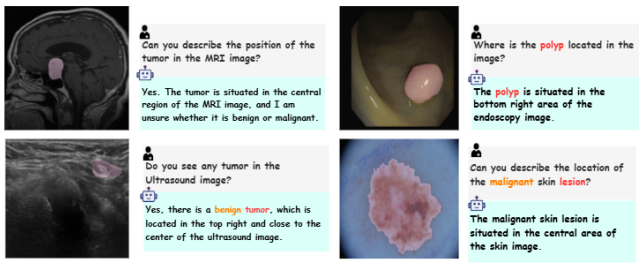

Figure 2. Visualization of the segmentation masks and question-answer pairs from the PosMed dataset.

three doctors to ensure the correctness in the medical context, and to ensure that when combined with the tumor name and positional information, the resulting sentences are coherent and contextually appropriate to provide the necessary information to the doctor.

Then, to extract positional information from the segmentation mask follow the Algorithm 1. Given a binary mask X_{mask}, we first derive the bounding box $\{x, y, w, h\}$ for the location of the tumor/anatomy. From this, we calculate the center point of the tumor as $x_{\text{center}} = \{x + \frac{w}{2}, y + \frac{h}{2}\}$. Next, we divide the image into anatomical action zone (AAZ) top left, top right, bottom left, and bottom right—as illustrated in Figure 1. The choice of a region extraction is a deliberate design decision to minimize spatial hallucination and maximize diagnostic reliability, which allows for the creation of massive, expert-validated datasets (116,000 pairs) without the noise inherent in manual free-form labeling. Based on the location of x_{center}, we determine which AAZ region the tumor lies in and assign it a corresponding label. In addition to handling cases where tumors are located near the image center, we also compute the distance between x_{center} and the geometric center of the image. If this distance falls below a predefined threshold τ pixel value, we label the tumor as being near the center.

After getting the position information, we integrate the extracted positional information along with the tumor/anatomy type from the dataset with the question-and-answer templates to generate the final dataset of spatially grounded tumor descriptions. Afterward, we employ the GPT model to polish the question and answer pairs to mitigate the coarse problems in linguistics. Finally, we filter out the repeated question and answer pairs in each image to get about 310,086 pairs. The example of question and answer results is illustrated in Figure 2.

Expert-in-the-loop Validation. To ensure the **(1)** *medical accuracy* and **(2)** *clinical demand* of the PosMed dataset, we adopt a validation process supported by three board-certified radiologists; each sample is identified as 0 for unqualified and 1 for qualified. Each radiologist reviews the generated question–answer pairs independently, focusing on the *medical content* (correctness of anatomical or tumor descriptions), *positional reasoning* (relevant to the anatomy or tumor annotations in the image), and *clinical plausibility* (the question and answers are useful for radiologists). The final decision for each sample is obtained through major-

Algorithm 1 Extract tumor/anatomy AAZ regions from a binary segmentation mask with $\{TL, TR, BL, BR\}$ representing Top Left, Top Right, Bottom Left, Bottom Right.

Require: Image size (H, W); binary mask $X_{\text{mask}} \in \{0,1\}^{H \times W}$; center threshold $\tau > 0$
Ensure: $L \in \mathcal{L}$ (or `INVALID`), where $\mathcal{L} = \{\texttt{TL}, \texttt{TR}, \texttt{BL}, \texttt{BR}, \texttt{CENTER}\}$

$\Omega \leftarrow \{(u, v) \mid X_{\text{mask}}(u, v) = 1\}$ ▷ u: row (y), v: column (x)
if $\Omega = \emptyset$ **then**
  return `INVALID`
end if
$x_{\min} \leftarrow \min_{(u,v)\in\Omega} v$, $x_{\max} \leftarrow \max_{(u,v)\in\Omega} v$
$y_{\min} \leftarrow \min_{(u,v)\in\Omega} u$, $y_{\max} \leftarrow \max_{(u,v)\in\Omega} u$
$w \leftarrow x_{\max} - x_{\min}$, $h \leftarrow y_{\max} - y_{\min}$
$x_c \leftarrow x_{\min} + \frac{w}{2}$, $y_c \leftarrow y_{\min} + \frac{h}{2}$ ▷ tumor center from bbox
$x_I \leftarrow \frac{W}{2}$, $y_I \leftarrow \frac{H}{2}$ ▷ image center
$d \leftarrow \sqrt{(x_c - x_I)^2 + (y_c - y_I)^2}$
if $d \le \tau$ **then**
  return `CENTER`
end if
if $x_c < x_I \wedge y_c < y_I$ **then**
  $L \leftarrow \texttt{TL}$
else if $x_c \ge x_I \wedge y_c < y_I$ **then**
  $L \leftarrow \texttt{TR}$
else if $x_c < x_I \wedge y_c \ge y_I$ **then**
  $L \leftarrow \texttt{BL}$
else
  $L \leftarrow \texttt{BR}$
end if
return L

ity voting, where only samples approved by the majority of experts are retained in the dataset. In addition, qualitative feedback from radiologists is used to refine the question-and-answer templates, thereby correcting ambiguous expressions, enhancing clinical clarity, and reducing potential biases or inconsistencies. This expert-driven validation framework ensures that PosMed maintains high fidelity in both medical semantics and spatial reasoning. Overall, after the validation stage, we obtain 116000 filtered question and answer pairs from 38731 images. Our dataset pipeline allows for multi-modal scaling (6 modalities) that would be impossible with manual spatial annotations.

Dataset Statistic. Figure 3 summarizes the PosMed dataset distribution. Lung X-ray constitutes the largest portion (48.0%), followed by lung CT (33.0%) and Brain MRI (7.9%). The remaining samples come from polyp endoscopy, skin images, and breast ultrasound, which together form a smaller share but increase the dataset diversity across multiple anatomies and tumor types. In terms of sample composition, PosMed is dominated by anatomy-centric descriptions (84.3%), while tumor-centric samples account for 15.7%.

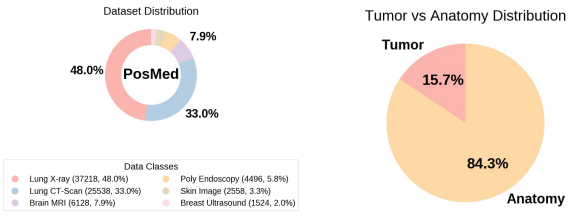

Figure 3. Statistical analysis of the PosMed. **Left:** Data-source distribution of six classes. **Right:**Tumor vs. anatomy composition.

4. PRS-Med

Overall Architecture. The primary goal of PRS-Med is to perform position reasoning segmentation, enabling the model to explain the location of tumors or anatomies in an image along with relevant medical information. Additionally, the segmentation head allows the model to perform tumor segmentation within the image using a single prompt. The overall architecture is illustrated in Figure 4.

This framework consists of three main modules. The first is the Vision-Language Model, we employ LLaVA-Med [24], as it is a well-trained Multimodal-LLM for the medical dataset. The second module is the Tiny SAM image encoder, employed from TinySAM [37], which is used to encode the input image. The third module is our proposed Prompt Mask Decoder, which includes our proposed fusion component that combines image features from the image encoder with the vision-language embeddings from the Medical Vision-Language Model to generate the final segmentation mask. In addition, we include a Language Model Head to perform the reasoning task.

During training, due to the challenges of fine-tuning the full LLaVA-Med model, we apply Low-Rank Adaptation (LoRA) [16] to enable the model to effectively learn position reasoning information from our prepared dataset.

Vision Backbone. The primary objective of the vision backbone is to extract pixel-level features from medical images to support conditional segmentation. For this purpose, we adopt the image encoder from TinySAM [37], which is based on the lightweight TinyViT architecture [48]. This design enables efficient image encoding while reducing computational resource requirements, and adapt to the medical domain without initializing weights from scratch.

Given a batch of b input images $X_{image} \in \mathbb{R}^{b\times 3\times W\times H}$, the images are processed through a tiny vision transformer model $\mathcal{F}vis$, consisting of approximately four transformer layers, to produce an image representation embedding $z_{image} \in \mathbb{R}^{b\times 256\times \frac{W}{16}\times \frac{H}{16}}$. The ablation for the choice of this TinyViT-based vision backbone is detailed in Section 5.

Multimodal-LLM. Most current Multimodal-LLM backbones applied to the medical domain—such as Flamingo [2], LLaVA [26], Qwen-VL [46], and InternVL [8]—demonstrate strong reasoning capabilities. However, they cannot generally generate masks for visual recognition tasks and struggle to comprehend positional information, such as the position of objects within medical images. In this work, we leverage the LLaVA-Med [24], which is trained on the extensive medical image dataset as the Language Model backbone. To allow the LLaVA-Med segmentation ability along with the reasoning, we leverage the LLaVA-Med's last hidden state embedding $z_{emb} \in \mathbb{R}^{b\times l\times 4096}$ (where l is the token length) as the conditioning input for the segmentation head. In detail, we generate the z_{emb}, and the reasoning output $z_{txt} \in \mathbb{R}^{b\times l}$ from the input image $X_{image} \in \mathbb{R}^{b\times 3\times w\times h}$ and input text $X_{txt} \in \mathbb{R}^{b\times l\times d}$ (where d is the vocabulary size), we define F_{mllm} as a parametric function. The autoregressive process of the model is described in Equation 1 and Equation 2.

$$z_{emb} = \mathcal{F}_{mllm}(X_{image}, X_{txt}), \tag{1}$$

$$z_{txt} = \prod_{i=1}^{l} p_\theta(z^i_{txt}|X_{image}, X^{i-1}_{txt}) \tag{2}$$

where θ is the trainable parameter. In our case, θ is from the parameter of the parametric function F_{mllm}.

During training, due to the high computational cost associated with fully supervised fine-tuning of the LLaVA-Med model, we employ the LoRA method [16] as an adapter. This approach enables the model to learn reasoning from our position-aware medical reasoning dataset while

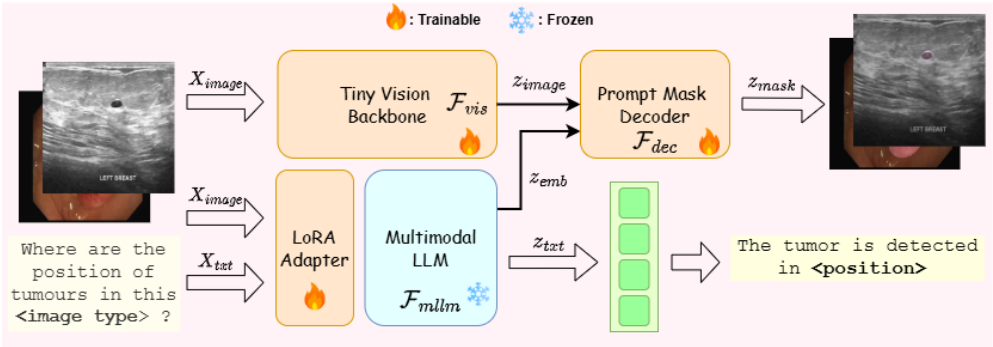

Figure 4. The architecture of PRS-Med comprises three primary components: (1) the Tiny Vision Backbone, (2) the Prompt Mask Decoder, and (3) the Multimodal-LLM. The framework accepts two input modalities: an image and a text-based prompt (e.g., a question). The image is processed through a vision encoder, while the prompt is embedded via a LoRA-adapted Multimodal-LLM. The fused representations are used to produce two outputs: a segmentation mask for the tumor regions, and a textual description specifying the tumor's location.

adapting to produce informative embeddings for the mask decoder, thereby mitigating catastrophic forgetting during multi-task MLLM training. The reason and LoRA hyperparameter choices are ablated in Section 6.3. By training the MLLM on clear spatial zones from PosMed, the model learns a strong spatial prior, which helps the Prompt Mask Decoder focus its attention, allowing it to capture small lesions that most grounded models miss.

Prompt Mask Decoder. The goal of this module is to predict the segmented mask from two inputs, including medical images representation feature z_{image} and the embedded image-text prompt z_{emb} from the Multimodal-LLM. This decoder module includes two parts: the fusion module and the mask prediction module. This design allows dynamic alignment between image regions and positional phrases, making better alignment between spatial features and medical vocabulary.

Fusion Module: Given the image representation from the vision encoder, denoted as $z_{\text{image}} \in \mathbb{R}^{b\times 256\times 16\times 16}$, and the conditioning input from the Multimodal-LLM, denoted as $z_{\text{emb}} \in \mathbb{R}^{b\times l\times 4096}$, the overall fusion process is formalized in Equation 3 and Equation 4.

where d_k is the scaling value, $MHA(.)$ is the Multi-head Attention layers, and $\sigma(.)$ is the softmax function.

First, the image representation z_{image} is reshaped to a new form $z_{\text{image}} \in \mathbb{R}^{b\times(16\times 16)\times 256}$ to enable interaction with the embedding $z_{\text{emb}} \in \mathbb{R}^{b\times l\times 4096}$ from the Multimodal-LLM. As shown in Equation 3, two projection layers, $\mathcal{F}^{\text{proj}}_{\theta_1}$ and $\mathcal{F}^{\text{proj}}_{\theta_2}$, are applied to project both features into a shared latent space of dimension 256. This alignment allows effective fusion through a cross-attention mechanism, which integrates the image features with the Multimodal-LLM's embeddings. The choice of cross-attention is motivated by the dynamic length of the z_{emb} sequences, making it a more flexible and suitable alternative to simple addition or concatenation. Following the fusion, a self-attention layer is employed to model the internal dependencies within the target sequence. The resulting fused representation, $z_{\text{fused}} \in \mathbb{R}^{b\times(16\times 16)\times d}$, is then reshaped to $\mathbb{R}^{b\times 256\times 16\times 16}$. Finally, as described in Equation 4, a skip connection is introduced to preserve gradient flow and mitigate the vanishing gradient problem during training.

Mask Prediction Module: The input z_{fused} is passed through a stack of transposed 2D convolutional layers, each followed by Batch Normalization and ReLU activation. This series of operations progressively upsamples z_{fused} to produce the final segmentation output $z_{\text{mask}} \in \mathbb{R}^{b\times 1\times 1024\times 1024}$.

Objective Function. This model is post-trained by using the segmentation loss ($\mathcal{L}_{seg}$) and text generation loss $\mathcal{L}_{text}$. The overall objective function is depicted in Equation 5.

$$\mathcal{L} = \lambda_{seg}\mathcal{L}_{seg} + \lambda_{txt}\mathcal{L}_{text}. \tag{5}$$

where λ_{seg} and λ_{txt} show the importance of each loss in the overall framework. In our training setup, we set these values to 1 and 0.5, respectively.

Regarding $\mathcal{L}_{seg}$, we employ a combination of Binary Cross-Entropy and Dice loss [39], which is a common choice in image segmentation tasks. For $\mathcal{L}_{txt}$, we use the Categorical Cross-Entropy (CE) loss applied on the logit vectors of the tokens output. Let $\hat{y}_{mask}$ denote the ground truth mask and z_{mask} the predicted mask; similarly, let $\hat{y}_{txt}$ be the ground truth token index sequence and z_{txt} the predicted text logits. Equations 6 and 7 illustrate the formulations of the aforementioned loss functions $\mathcal{L}_{seg}$, and $\mathcal{L}_{txt}$.

$$\mathcal{L}_{seg} = \mathcal{L}_{BCE}(\hat{y}_{mask}, z_{mask}) + \mathcal{L}_{dice}(\hat{y}_{mask}, z_{mask}) \tag{6}$$

$$\mathcal{L}_{text} = \mathcal{L}_{CE}(\hat{y}_{txt}, z_{txt}). \tag{7}$$

By employing this objective function, PRS-Med can simultaneously learn position reasoning while also learning

$$z_{fused} = MHA(\sigma(\frac{\mathcal{F}_{\theta_1}^{proj}(z_{image})\mathcal{F}_{\theta_2}^{proj}(z_{emb})^T)}{\sqrt{d_k}})\mathcal{F}_{\theta_2}^{proj}(z_{emb})), \tag{3}$$

$$z_{fused} = z_{fused} + z_{image}. \tag{4}$$

to perform segmentation. Notably, during training, the decoder receives gradients not only from segmentation losses but also from textual reasoning losses, creating a feedback loop where segmentation informs reasoning and vice versa.

5. Experimental Setup

Dataset. We train on the *PosMed* dataset (consisting of six types of images: ultrasound, MRI, RGB image, CT Image, X-ray, and endoscopy), combining images from BUSI [1], BrainMRI [10, 11], ISIC [14], LungCT [49], LungXray [13, 49], Kvasir-SEG, and ClinicDB [3]. We follow the original train/test splits from each source dataset. To evaluate generalization on unseen distributions, we augment the endoscopy test set with CVC-300 [43], ETIS [38], and ColonDB [40], which are excluded from training.

Implementation Details. All experiments are conducted on two NVIDIA H100 80GB GPUs. We train for 20 epochs with a batch size of 8 per GPU using AdamW [29] with a learning rate of 1×10^{-4}. For LoRA, we set rank $r = 16$, $\alpha = 16$, and dropout $= 0.05$, with weights initialized from a uniform distribution. LoRA is applied to the query, key, value, and output projection matrices of all attention layers in LLaVA-Med. The TinySAM vision encoder is fully unfrozen during training.

Baselines. To compare our work with SOTA methods, following the reasoning segmentation task mentioned by LISA [23] and the diagnostic reasoning of medical MLLM [24], we conduct three benchmarks, including segmentation, position reasoning, and position understanding. For the Segmentation task, we compare our methods with the Foundation Segmentation model of medical imaging, such as SAM-Med 2D [52] (2024), and Biomedparse [51] (2024) (finetuned image encoder and decoder on our dataset), and the reasoning segmentation model, which is also finetuned on our dataset, is LISA [23] with two versions are 7B and 13 B. Regarding the SAM model in medical imaging, there is a challenge that most medical segmentation model is based on the box prompt. For this reason, we leverage the Grounding Dino [28] as the text understanding model to extract the boxes coordinates for the segmentation task. In the Position Reasoning benchmark, due to the lack of methods done reasoning segmentation, we reproduce the fine-tuning process on our dataset for the Multimodal-LLM for medical image, which includes LLaVA-Med [24] (2024), HuatuoGPT-Vision [7] (2024), Med-MoE [21], and MedVLM-R1 [33] (2025) to do the reasoning benchmark. In all of the comparisons, we do the fine-tuning of these methods on our PosMed dataset with the best-practice hyperparameter for each method for the fairest comparison.

Evaluation Metric. For evaluation, we use mDice and mIoU to benchmark segmentation performance, following standard practice in medical image segmentation. To assess the fluency and the accuracy of position reasoning, we report ROUGE score, and we use accuracy to measure the correctness of the model's reasoning answers.

6. Evaluation Results

6.1. Quantitative Results

Segmentation Task Results. To evaluate the overall performance of PRS-Med in the segmentation task, we compare our method with prior works as aforementioned. Table 1 presents results on radiology images of six different images and tissues, including Breast Ultrasound, Brain MRI, Lung CT-Scan, Lung X-ray, Polyp Endoscopy, and Skin Image.

As shown in Table 1, PRS-Med achieves competitive results with state-of-the-art. To the second-best method, the improvements (mDice, mIoU) are $(+3.4\%, +3.1\%)$ on Breast Ultrasound, $(+13.6\%, +13.2\%)$ on Brain MRI, $(+31.2\%, +41.5\%)$ on Lung CT-Scan, $(+0.1\%, +0.2\%)$ on Lung X-ray, $(+1.9\%, +1.7\%)$ on Polyp Endoscopy, and $(+0.8\%, +1.1\%)$ on Skin Images. These results highlight the generalization and robustness of PRS-Med across imaging modalities, anatomical structures, and tumors.

Position Reasoning Results. To assess the performance of PRS-Med, we evaluate the output of linguistic content with coherence and the accuracy of position reasoning with SOTA methods in Multimodal-LLM for medical images, which is shown in Table 2.

As shown in Table 2, PRS-Med achieves the best ROUGE across all six datasets and obtains the highest accuracy on five out of six datasets. Compared with the strongest prior baseline (Med-MoE), the absolute gains (ROUGE, ACC) are: Breast Ultrasound $(+0.03, +11.29)$, Brain MRI $(+0.01, +5.00)$, Lung CT-Scan $(+0.02, -11.50)$, Lung X-ray $(+0.03, +10.36)$, Polyp $(+0.04, +16.79)$, and Skin Image $(+0.08, +4.23)$. Overall, PRS-Med consistently improves generation quality (ROUGE) while also strengthening task correctness (ACC), with the only exception on

Method	Breast Ultrasound		Brain MRI		Lung CT-Scan		Lung X-ray		Polyp Endoscopy		Skin Image	
	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑	mDice ↑	mIoU ↑
G-Dino + SAM-Med2D [31]	0.515	0.441	0.667	0.625	0.540	0.392	0.401	0.300	0.488	0.418	0.237	0.171
Biomedparse [51]	0.783	0.698	0.294	0.245	0.516	0.399	0.972	0.949	0.824	0.774	0.893	0.822
LISA-7B [23]	0.299	0.246	0.478	0.402	0.478	0.402	0.397	0.263	0.241	0.202	0.464	0.368
LISA-13B [23]	0.705	0.680	0.439	0.357	0.656	0.528	0.664	0.535	0.312	0.247	0.643	0.536
PRS-Med	**0.817**	**0.729**	**0.803**	**0.757**	**0.968**	**0.943**	**0.973**	**0.952**	**0.843**	**0.791**	**0.901**	**0.833**
vs previous works	**+0.034**	**+0.031**	**+0.136**	**+0.132**	**+0.312**	**+0.415**	**+0.001**	**+0.002**	**+0.019**	**+0.017**	**+0.008**	**+0.011**

Table 1. Quantitative results of PRS-Med across six medical image types. The highest score in each column is in **bold**; the second highest is underlined. PRS-Med shows its competitive results when it surpasses all of the previous works across six datasets.

Method	Breast Ultrasound		Brain MRI		Lung CT-Scan		Lung X-ray		Polyp		Skin Image	
	ROUGE ↑	ACC ↑	ROUGE ↑	ACC ↑	ROUGE ↑	ACC ↑	ROUGE ↑	ACC ↑	ROUGE ↑	ACC ↑	ROUGE ↑	ACC ↑
LlaVA-Med [24]	0.33	75.22	0.33	81.17	0.32	74.17	0.33	81.09	0.30	53.76	0.29	92.08
HuatuoGPT [7]	0.36	79.65	0.36	25.50	0.35	**88.17**	0.36	10.97	0.30	55.64	0.30	60.69
Med-MoE [21]	0.61	81.63	0.66	67.00	0.69	12.50	0.61	84.00	0.67	27.60	0.68	81.58
Med-VLMR1 [33]	0.28	50.44	0.28	12.50	0.27	72.22	0.28	3.13	0.25	32.58	0.24	13.98
PRS-Med	**0.64**	**92.92**	**0.67**	**86.17**	**0.71**	76.67	**0.64**	**94.36**	**0.71**	**72.43**	**0.76**	**96.31**
vs previous works	**+0.03**	**+11.29**	**+0.01**	**+5.00**	**+0.02**	**-11.50**	**+0.03**	**+10.36**	**+0.04**	**+16.79**	**+0.08**	**+4.23**

Table 2. Quantitative results of PRS-Med on the reasoning task across six medical image types. The highest score in each column is in **bold**, the second highest is underlined.

Lung CT-Scan where PRS-Med attains the best ROUGE but trails the accuracy of HuoGPT [7]. Notably, although PRS-Med and LlaVA-Med [24] share the same multimodal-LLM pretraining, PRS-Med yields substantially better results on position reasoning. We attribute this improvement to the jointly trained segmentation module, which injects explicit localization supervision during training and allows the LoRA adapters to better specialize for position-sensitive reasoning by updating their weights accordingly. These results also indicate that, during the learning of the segmentation, the MLLM does not forget its reasoning ability, which leads to the competitive performance with SOTA.

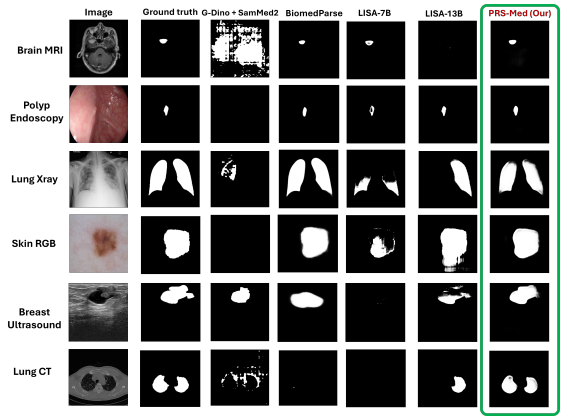

Figure 5. Comparison of PRS-Med with previous works. PRS-Med produces more accurate boundaries and captures small lesions missed by other methods.

6.2. Qualitative Analysis

In Figure 5, we present qualitative visualizations that highlight the improvements achieved by PRS-Med. The results clearly show that PRS-Med can capture small lesions and anatomies that previous baselines miss, consistently generating completed masks with the lowest loss. We attribute these improvements to the informative feature extraction of the lightweight vision encoder and the effectiveness of the fusion module. Overall, the results provide strong evidence for the promise of our approach. In addition, it shows that the discrete spatial grounding provided by PRS-Med acts as a 'visual prompt' that guides the decoder to identify small, high-stakes pathologies that generalist reasoning models overlook. However, through this visualization, we observe that the boundary problems are still the limitations of PRS-Med, and we are planning to improve in the future.

6.3. Ablation Study

To assess the choice and the effectiveness of the module in our framework, we conduct several experiments to assess the performance and the limitations of each module. The experiments are conducted in the same training and testing dataset with the benchmark. Regarding the metrics, we calculate the average mDice and average mIoU for the segmentation results, and ROUGE and METEOR for reasoning results (to assess the coherence in reasoning results) on different modalities in our test dataset to have the best assessment of the robustness of each choice.

Vision Backbone	Param	mDice ↑	mIoU ↑
SAM-Med (Frozen)	21.52M	0.798	0.719
SAM-Med (Full)	292.60M	0.891	0.838
SAM-Med (LoRA)	47.84M	0.790	0.711
TinySAM (no pretrained)	31.49M	0.674	0.582
TinySAM (frozen)	21.73M	0.737	0.662
TinySAM (Full)	31.49M	**0.884**	**0.834**

Table 3. Comparison results of different vision encoder backbones.

MLLM	mDice ↑	mIoU ↑	ROUGE ↑	METEOR ↑
q,v	0.573	0.483	0.478	0.436
q, k, v	0.714	0.621	0.585	0.578
q,k,v,o	**0.879**	**0.827**	**0.654**	**0.599**

Table 4. Ablation study on LoRA target module (r=16) on Multimodal-LLM.

Design Choice Of Vision Encoder Backbone. As described in Section 4, we assess the choice of vision encoder with the results presented in Table 4. In our design, we consider two published pre-trained models, SAM-Med [50] and TinySAM [37], as our vision encoder. As demonstrated in Table 4, TinySAM gets higher performance than SAM-Med (LoRA) with similar trainable parameters in the general framework and gets lower results with the full supervised training version of SAM-Med. Given the trade-off between efficiency and model accuracy, TinySAM is chosen as our vision encoder.

Initialization Of TinySAM Image Encoder. We observe that the initialization of TinySAM significantly affects the overall results. For this reason, in Table 3, we assess the contribution of the TinySAM pretrained weight. Without the pretrained initialization, the overall results drop substantially, which shows the importance of the pretrained initialization to the overall framework.

Design Choice Of MLLM Backbone. To evaluate the choice of MLLM backbone for PRS-Med, we conducted experiments comparing three models: LLaVA-1.5 [25], LLaVA-1.6 [27], and LLaVA-Med [24] using the same 7B backbone and fine-tuned via the LoRA approach. The comparisons are on two tasks: segmentation and position reasoning (shown in Table 5). Results indicate that the performance of the LLaVA-Med baseline surpasses the remaining. This improvement can be attributed to LLaVA-Med's enhanced adaptation to the medical domain, which enables it to better handle tasks involving medical data.

LoRA Target Modules Choice. To assess the contribution of the target module from LoRA in the overall framework, we do the ablation to evaluate our choice of target module from LoRA, which is mentioned in Table 4. Our choice for the target modules (q,k,v,o) makes the overall framework

Metric	LLaVA-1.5 (LoRA)	LLaVA-1.6 (LoRA)	LLaVA-Med (LoRA)
Param	34.63M	31.49M	**31.49M**
Avg-mDice ↑	0.709	0.744	**0.879**
Avg-mIoU ↑	0.642	0.671	**0.827**
ROUGE ↑	0.414	0.508	**0.654**
METEOR ↑	0.385	0.432	**0.599**

Table 5. Ablation study for design choice of the Multimodal-LLM.

achieve significantly higher performances, which indicates that all of the projection weights allow LoRA to more effectively align cross-modal representations for both the reasoning and segmentation tasks.

7. Conclusion

In conclusion, we introduce **PosMed**, a large-scale dataset designed to benchmark and advance position-reasoning segmentation in medical imaging, together with **PRS-Med**, a strong baseline tailored to this challenge. Our experiments reveal two key gaps in current approaches: (i) text-referring segmentation methods remain brittle when the target is specified through spatial reasoning rather than direct textual cues, and (ii) existing medical MLLMs still struggle with position reasoning, a core capability for reliable perception in clinical settings. Importantly, results show that training and evaluating on PosMed, together with the PRS-Med framework, effectively addresses these limitations and yields competitive performance on position-reasoning segmentation. PRS-Med eschews the unnecessary complexity of pixel-coordinate regression in favor of a Categorical Spatial Grounding framework. This design choice ensures that the model's reasoning is not only technically sound but clinically actionable (simple), providing a direct bridge between visual evidence and the structured language of medical diagnosis. By releasing the PosMed data pipeline, model, and code, we hope to facilitate future work on spatially grounded multimodal reasoning and accelerate progress toward practical clinical applications.

Future work. We plan to extend PosMed and PRS-Med toward richer spatial reasoning beyond localization, anatomical distance, etc., instead of the coarse position only, enabling the model to identify relevant anatomy and tumors while also estimating clinically meaningful attributes such as lesion size and inter-structure distances.

Acknowledgement. U. Bagci is supported partially by NIH R01 CA268808 and R01 HL171376. D. Jha is supported in part by the U.S. Department of Education (P116Z240151 to the University of South Dakota and the South Dakota School of Mines & Technology). The views expressed are those of the author(s) and do not necessarily represent the official views of the U.S. Department of Education. We would like to thank to Gorkem Durak, and Halil Ertugrul Aktas to support in reviewing the quality of the dataset.

References

[1] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020. 2, 6

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 4

[3] Jorge Bernal, F. Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *CMIG*, pages 99–111, 2015. 2, 6

[4] Nhat-Tan Bui, Dinh-Hieu Hoang, Quang-Thuc Nguyen, Minh-Triet Tran, and Ngan Le. Meganet: Multi-scale edge-guided attention network for weak boundary polyp segmentation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 7985–7994, 2024. 2

[5] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218, 2022. 2

[6] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 2

[7] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, et al. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*, 2024. 1, 2, 6, 7

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 4

[9] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatial-rgpt: Grounded spatial reasoning in vision language models. *arXiv preprint arXiv:2406.01584*, 2024. 2

[10] Jun Cheng, Wei Huang, Shuangliang Cao, Ru Yang, Wei Yang, Zhaoqiang Yun, Zhijian Wang, and Qianjin Feng. Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PloS one*, 10(10): e0140381, 2015. 2, 6

[11] Jun Cheng, Wei Yang, Meiyan Huang, Wei Huang, Jun Jiang, Yujia Zhou, Ru Yang, Jie Zhao, Yanqiu Feng, Qianjin Feng, et al. Retrieval of brain tumors by adaptive spatial pooling and fisher vector representation. *PloS one*, 11(6): e0157112, 2016. 2, 6

[12] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiangand Hui Sun, Junjun He, Shaoting Zhang, Min Zhu, and Yu Qiao. Sam-med2d, 2023. 2

[13] Muhammad EH Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, et al. Can ai help in screening viral and covid-19 pneumonia? *Ieee Access*, 8:132665–132676, 2020. 2, 6

[14] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172, 2018. 2, 6

[15] Sunan He, Yuxiang Nie, Hongmei Wang, Shu Yang, Yihui Wang, Zhiyuan Cai, Zhixuan Chen, Yingxue Xu, Luyang Luo, Huiling Xiang, et al. Gsco: Towards generalizable ai in medicine via generalist-specialist collaboration. *arXiv preprint arXiv:2404.15127*, 2024. 2

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 4

[17] Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018. 2

[18] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Dag Johansen, Thomas De Lange, Pål Halvorsen, and Håvard D Johansen. Resunet++: An advanced architecture for medical image segmentation. In *Proceedings of the 2019 IEEE International Symposium on Multimedia (ISM)*, pages 225–2255, 2019. 2

[19] Debesh Jha, Michael A Riegler, Dag Johansen, Pål Halvorsen, and Håvard D Johansen. Doubleu-net: A deep convolutional neural network for medical image segmentation. In *2020 IEEE 33rd International symposium on computer-based medical systems (CBMS)*, pages 558–564, 2020. 2

[20] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-SEG: A Segmented Polyp Dataset. In *Multimedia Modeling*, 2020. 2

[21] Songtao Jiang, Tuo Zheng, Yan Zhang, Yeying Jin, Li Yuan, and Zuozhu Liu. Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. *arXiv preprint arXiv:2404.10237*, 2024. 1, 2, 6, 7

[22] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer White-

head, 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. 2

[23] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 1, 2, 6, 7

[24] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023. 1, 4, 6, 7, 8

[25] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 8

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2, 4

[27] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 8

[28] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55, 2024. 6

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[30] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15:654, 2024. 2

[31] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 7

[32] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367, 2023. 2

[33] Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*, 2025. 1, 2, 6, 7

[34] Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M Zughaier, Muhammad Salman Khan, et al. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in biology and medicine*, 132: 104319, 2021. 2

[35] Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S Ryoo, and Tsung-Yu Lin. Learning to localize objects improves spatial reasoning in visual-llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12977–12987, 2024. 2

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241, 2015. 2

[37] Han Shu, Wenshuo Li, Yehui Tang, Yiman Zhang, Yihao Chen, Houqiang Li, Yunhe Wang, and Xinghao Chen. Tinysam: Pushing the envelope for efficient segment anything model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 20470–20478, 2025. 4, 8

[38] Juan S. Silva, Aymeric Histace, Olivier Romain, Xavier Dray, and Bertrand Granado. Towards embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. *IJCARS*, pages 283–293, 2014. 2, 6

[39] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3*, pages 240–248, 2017. 5

[40] Nima Tajbakhsh, Suryakanth R. Gurudu, and Jianming Liang. Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information. *TMI*, pages 630–644, 2016. 2, 6

[41] Nikhil Kumar Tomar, Annie Shergill, Brandon Rieders, Ulas Bagci, and Debesh Jha. Transresu-net: Transformer based resu-net for real-time colonoscopy polyp segmentation. *arXiv preprint arXiv:2206.08985*, 2022. 2

[42] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[43] David Vázquez, Jorge Bernal, Francisco Javier Sánchez, Glòria Fernández-Esparrach, Antonio M. López, Adriana Romero, Michal Drozdzal, and Aaron C. Courville. A Benchmark for Endoluminal Scene Segmentation of Colonoscopy Images. *Journal of Healthcare Engineering*, 2017. 2, 6

[44] Junchi Wang and Lei Ke. Llm-seg: Bridging image segmentation and large language model reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1765–1774, 2024. 1

[45] Junchi Wang and Lei Ke. Llm-seg: Bridging image segmentation and large language model reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1765–1774, 2024. 2

[46] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 4

[47] XuDong Wang, Shaolun Zhang, Shufan Li, Konstantinos Kallidromitis, Kehan Li, Yusuke Kato, Kazuki Kozuka, and

Trevor Darrell. Segllm: Multi-round reasoning segmentation. *arXiv preprint arXiv:2410.18923*, 2024. 1, 2

[48] Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. Tinyvit: Fast pretraining distillation for small vision transformers. In *European conference on computer vision*, pages 68–85, 2022. 4

[49] J Yang, G Sharp, H Veeraraghavan, W Van Elmpt, A Dekker, T Lustberg, and M Gooding. Data from lung ct segmentation challenge (lctsc)(version 3)[data set]. *The Cancer Imaging Archive*, 2017. 2, 6

[50] Jin Ye, Junlong Cheng, Jianpin Chen, Zhongying Deng, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. Sa-med2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks. *arXiv preprint arXiv:2311.11969*, 2023. 8

[51] Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, Christine Moung-Wen, et al. Biomedparse: a biomedical foundation model for image parsing of everything everywhere all at once. *arXiv preprint arXiv:2405.12971*, 2024. 2, 6, 7

[52] Jiayuan Zhu, Abdullah Hamdi, Yunli Qi, Yueming Jin, and Junde Wu. Medical sam 2: Segment medical images as video via segment anything model 2. *arXiv preprint arXiv:2408.00874*, 2024. 6