Sim4Seg: Boosting Multimodal Multi-disease Medical Diagnosis Segmentation with Region-Aware Vision-Language Similarity Masks

Lingran Song, Yucheng Zhou, Jianbing Shen*

SKL-IOTSC, CIS, University of Macau
Lingran.Song@connect.um.edu.mo, jianbingshen@um.edu.mo

Abstract

Despite significant progress in pixel-level medical image analysis, existing medical image segmentation models rarely explore medical segmentation and diagnosis tasks jointly. However, it is crucial for patients that models can provide explainable diagnoses along with medical segmentation results. In this paper, we introduce a medical vision-language task named Medical Diagnosis Segmentation (MDS), which aims to understand clinical queries for medical images and generate the corresponding segmentation masks as well as diagnostic results. To facilitate this task, we first present the **M**ultimodal **M**ulti-disease **M**edical **D**iagnosis **S**egmentation (**M3DS**) dataset, containing diverse multimodal multi-disease medical images paired with their corresponding segmentation masks and diagnosis chain-of-thought, created via an automated diagnosis chain-of-thought generation pipeline. Moreover, we propose **Sim4Seg**, a novel framework that improves the performance of diagnosis segmentation by taking advantage of the **R**egion-Aware **V**ision-**L**anguage **S**imilarity to **M**ask (**RVLS2M**) module. To improve overall performance, we investigate a test-time scaling strategy for MDS tasks. Experimental results demonstrate that our method outperforms the baselines in both segmentation and diagnosis.

Code — https://github.com/SLR567/Sim4Seg
Dataset — https://github.com/SLR567/M3DS

1 Introduction

As an important clinical application, medical image segmentation aims to identify tissues, lesions, and organs in various medical images (Ramesh et al. 2021; Wang et al. 2022; Zhou et al. 2018). Although existing specialist models (Ronneberger, Fischer, and Brox 2015; Saha, Hosseinzadeh, and Huisman 2021) achieve impressive performance in specific tasks, they lack the capacity to directly provide diagnosis with explanations, which is an essential capability for real-world clinical workflows.

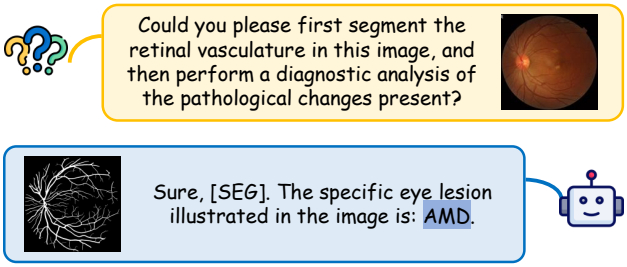

Figure 1: Medical Diagnosis Segmentation (MDS) task requires model to understand medical images and queries, then generate corresponding segmentation masks along with diagnoses.

As an extension of referring expression segmentation (Hu, Rohrbach, and Darrell 2016), reasoning segmentation has recently been proposed to generate fine-grained masks for objects referenced in text output. This paradigm acquires further exploration in medical image segmentation. In the general domain, recent studies (Lai et al. 2024; Zhou, Song, and Shen 2025b) have advanced fine-grained reasoning tasks for Large Vision-Language Models (LVLMs) using text prompts. These works effectively integrate visual encoders of LVLMs (Liu et al. 2023) with downstream task decoders (Kirillov et al. 2023), enhancing their visual grounding capabilities. Specifically, GSVA (Xia et al. 2024) addresses the distribution gap between multiple-target and empty-target scenarios. Separately, GLaMM (Rasheed et al. 2024) and PixelLM (Ren et al. 2024) enhance model versatility in language and vision modalities, respectively, enabling multi-granularity reasoning. READ (Qian, Yin, and Dou 2025) leverages points as prompts to enhance segmentation performance by improving fine-grained text-image correspondence. Models that combine reasoning and segmentation abilities hold significant potential in clinical applications. However, existing medical LVLMs focus primarily on segmentation capabilities (Tong et al. 2025) or leverage LVLMs for text-guided localization descriptions (Guo et al. 2024; Zhou, Song, and Shen 2025a). Developing a unified model that offers medical segmentation and explainable diagnosis capability remains an open challenge.

In this work, we propose a novel medical vision-language

*Corresponding author. This work was supported by the National Natural Science Foundation of China (No.624B2002) and the Jiangyin Hi-tech Industrial Development Zone under the Taihu Innovation Scheme (EF2025-00003 SKL-IOTSC).

task, Medical Diagnosis Segmentation (MDS) (illustrated in Figure 1), which requires a model to understand medical images with queries and generate both segmentation masks and corresponding diagnosis results. To support the MDS task and facilitate future research, we introduce the Multimodal Multi-disease Medical Diagnosis Segmentation (M3DS) dataset. M3DS contains 10 subsets across various modalities and disease types. Each sample contains an image, ground truth mask, query, diagnosis result, and a diagnosis chain-of-thought (CoT). We first select multi-modality, multi-disease medical image segmentation data from publicly available datasets. Subsequently, we construct CoT question-answering pairs for M3DS using our automated pipeline, which leverages the open-source medical LVLM HuatuoGPT-Vision (Chen et al. 2024). Unlike traditional medical image segmentation or medical VQA datasets (Cheng et al. 2025; Zhang et al. 2023), M3DS unifies segmentation and diagnosis reasoning, enhancing pixel-level explainability for medical VQA while providing reasoning-based interpretability for segmentation, serving as a valuable resource for future research.

Building upon existing reasoning segmentation methods (Lai et al. 2024; Lan et al. 2025; Li et al. 2023b), we introduce Sim4Seg. This model leverages Region-Aware Vision-Language Similarity Masks (RVLSMs) derived from text-image token embedding similarity of the last hidden state to enhance diagnosis segmentation performance. Fine-tuning Sim4Seg on M3DS dataset improves its medical diagnosis segmentation capability. Furthermore, we develop a test-time scaling strategy especially designed for MDS task to optimize overall performance.

In summary, the main contributions of this paper are as follows:

- We introduce M3DS dataset, a unified resource comprising segmentation masks and diagnosis CoT generated by our automated pipeline, which facilitates future research.
- We propose Sim4Seg, a novel framework incorporating Region-Aware Vision-Language Similarity to Mask (RVLS2M) module to enhance medical image segmentation and diagnosis capabilities.
- We design a test-time scaling strategy specifically for medical image diagnosis and segmentation, leading to higher performance.
- Extensive experiments demonstrate that Sim4Seg outperforms existing models on while exhibiting robust cross-dataset and cross-modality generalization capabilities.

2 M3DS Dataset

To enhance the segmentation and diagnosis capabilities of Sim4Seg, we constructed the M3DS dataset. Firstly, we collected medical image segmentation datasets with corresponding disease categories. Subsequently, we designed a multi-role CoT data generation pipeline for M3DS dataset.

2.1 Data Collection

We constructed M3DS by integrating ten distinct sub-datasets from diverse sources. Specifically, FracAtlas (Abedeen et al. 2023) contains 4,083 X-Ray images, including 717 fracture cases, each annotated with polygonal segmentation masks. Bone fracture (Roboflow100 2023) comprises 458 X-Ray images, with 437 containing fracture regions. Bone Fracture Detection (BFD) (Darabi 2024) includes 2,060 fracture X-Ray images with pixel-level segmentation masks at six anatomical sites. ISBI (Gutman et al. 2016) and ISIC (Codella et al. 2018) contain 1,279 and 2750 images of benign or malignant skin lesions with corresponding segmentation masks, respectively. Kvasir-SEG (Jha et al. 2020) provides 1,000 polyp images paired with segmentation masks. BUSI (Al-Dhabyani et al. 2020) comprises 780 breast ultrasound images with 437 benign, 210 malignant and 133 normal cases. TN3K (Gong et al. 2023) includes 3,493 thyroid nodule images with corresponding masks. ChestX-Det (Lian et al. 2021) contains 3,578 images sourced from NIH ChestX-14 (Wang et al. 2017), annotated by three radiologists across 13 abnormality categories. FIVES (Kai et al. 2022) features 800 fundus photographs with manually annotated masks. The modality of these data varied from X-Ray, dermoscopy, endoscopy, ultrasound, to fundus photography. Specific details regarding the segments used from each sub-dataset are summarized in Table 1.

Dataset	Modality					Size		
	X-Ray	DS	End	US	FP	Train	Val	Test
FracAtlas	✓					574	82	61
bone fracture	✓					311	83	43
BFD	✓					1804	173	83
ISBI		✓				899	277	101
ISIC		✓				2000	150	600
Kvasir-SEG			✓			801	99	100
BUSI				✓		532	55	60
TN3K				✓		2001	878	614
ChestX-Det	✓					2478	388	101
FIVES					✓	600	99	101
Total	4	2	1	2	1	12000	2284	1864

Table 1: Overview of M3DS dataset. In particular, DS stands for Dermoscopy, End refers to Endoscopy, US denotes Ultrasound, FP represents Fundus Photography, and BFD signifies Bone Fracture Detection.

2.2 Chain-of-Thought Reasoning Data Generation

To construct the M3DS dataset, we designed a multi-role CoT diagnosis data generation pipeline specifically for MDS tasks, leveraging HuatuoGPT-Vision (Chen et al. 2024) model as medical assistant and critical assistant. As illustrated in Figure 3, we first assemble the collected medical images, the corresponding questions, and the diagnosis results into a structured prompt. To guide the model toward a step-by-step understanding of the image and diagnosis reasoning, our prompt instructs the medical assistant to begin by identifying the image modality, then progressively analyze the medical image and finally derive the diagnosis. This prompt is then fed into the medical assistant to generate the

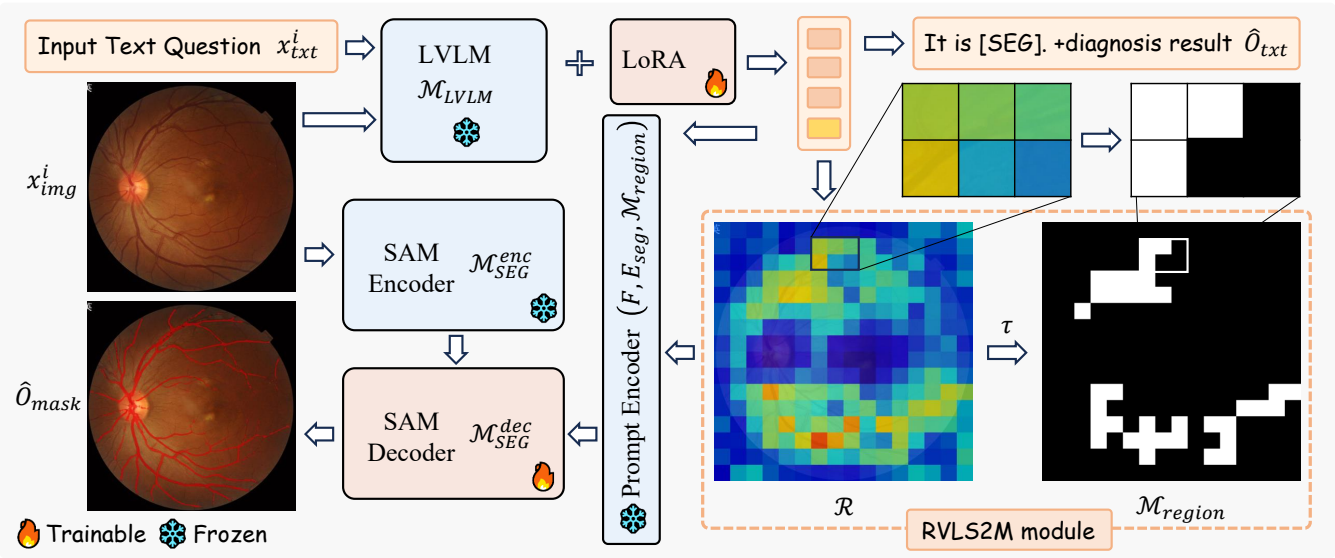

Figure 2: Model Architecture of Sim4Seg. LVLM generates the corresponding output according to the given medical image and query. The output hidden states are fed into the RVLS2M module, from which $\mathbf{M}_{region}$ is obtained to prompt the SAM (Kirillov et al. 2023) model.

CoT of diagnosis reasoning. To ensure the quality and reliability of the generated result, a review step is incorporated into the pipeline, where the critical assistant evaluates the output. If the generated CoT is rejected by the critical assistant within the maximum allowed review rounds, the failure feedback is sent back to the first step to trigger a new output. To further guarantee the effectiveness of the generation, we added a human-assisted review phase after critical assistant evaluation. Using this multi-role diagnosis CoT generation pipeline, we train the Sim4Seg model on our constructed M3DS dataset.

3 Methodology

In this section, we present our model architecture in Section 3.1, which contains our proposed Region-Aware Vision-Language Similarity to Mask (RVLS2M) module. Then, we introduce training objectives in Section 3.2, and a test-time scaling strategy designed for MDS task in Section 3.3.

3.1 Sim4Seg Model Architecture

Let $X_{img} \in \mathbb{R}^{n\times h\times w\times c}$ be the set of input images, with $\mathrm{x}^i_{img} \in X_{img}$ representing a single image. Here n, h, w, and c denote the image count, height, width, and channel count, respectively. The corresponding set of text inputs, X_{txt} representing text queries, contains elements $\mathrm{x}^i_{txt} \in X_{txt}$. The primary task of MDS task is to predict text output $\hat{O}_{txt}$ and its corresponding segmentation mask $\hat{O}_{mask}$ as

$$\Theta_{MLE} = \arg\max_{\Theta} \mathcal{M}_\theta \left(\hat{O}_{txt}, \hat{O}_{mask} \mid X_{img}, X_{txt}; \Theta\right),$$

where the medical diagnosis segmentation model is denoted by $\mathcal{M}_\theta$, and Θ represents its parameters. As shown in Figure 2, this framework comprises an LVLM $\mathcal{M}_{LVLM}$ and a segmentation backbone $\mathcal{M}_{SEG}$, formally defined as $\mathcal{M}_\theta = \mathcal{M}_{LVLM} \oplus \mathcal{M}_{SEG}$. Here, $\oplus$ indicates a cascading operation between modules. The output segmentation mask $\hat{O}_{mask} \in \{0,1\}^{h\times w}$ is a binary matrix. To enable $\mathcal{M}_\theta$ to generate mask embeddings, LISA (Lai et al. 2024) expands the text vocabulary of $\mathcal{M}_{LVLM}$ by introducing a special token. The input image x^i_{img} is divided into uniform patches and processed by CLIP (Radford et al. 2021) encoder. During training, this special token is incorporated into x^i_{txt}, and both x^i_{img} and x^i_{txt} are put into $\mathcal{M}_{LVLM}$, producing language response $\hat{O}_{txt}$ as

$$\hat{O}_{txt} = \mathcal{M}_{LVLM}\left(\mathrm{x}^i_{img}, \mathrm{x}^i_{txt}\right). \tag{1}$$

Before predicting the binary segmentation mask, $\mathcal{M}_\theta$ generates a response $\hat{O}_{txt}$ containing the special token representing the target object. Following LISA (Lai et al. 2024), the last hidden layer embedding $\tilde{\mathbf{E}}_{seg}$ associated with this special token is extracted from $\mathcal{M}_{LVLM}$. This is subsequently projected through a multilayer projection layer ϕ to obtain the refined feature $\mathbf{E}_{seg}$ as

$$\mathbf{E}_{seg} = \phi\left(\tilde{\mathbf{E}}_{seg}\right). \tag{2}$$

Meanwhile, the visual backbone $\mathcal{M}^{enc}_{SEG}$ extracts visual features $\mathbf{F}$ from the input image x^i_{img}, formulated as

$$\mathbf{F} = \mathcal{M}^{enc}_{SEG}\left(\mathrm{x}^i_{img}\right). \tag{3}$$

Building on the capability of SAM (Kirillov et al. 2023), which supports various types of prompts, we investigate region-aware masks as prompts via region-aware vision-language similarity. Formally, let $\mathbf{E} = \{\mathbf{E}_1, \ldots, \mathbf{E}_n \mid \mathbf{E}_i \in \mathbb{R}^d\}$ denote the hidden states of $\mathcal{M}_{LVLM}$, where n is the token count, and d denotes the embedding dimension. $\mathbf{E}_{seg}$

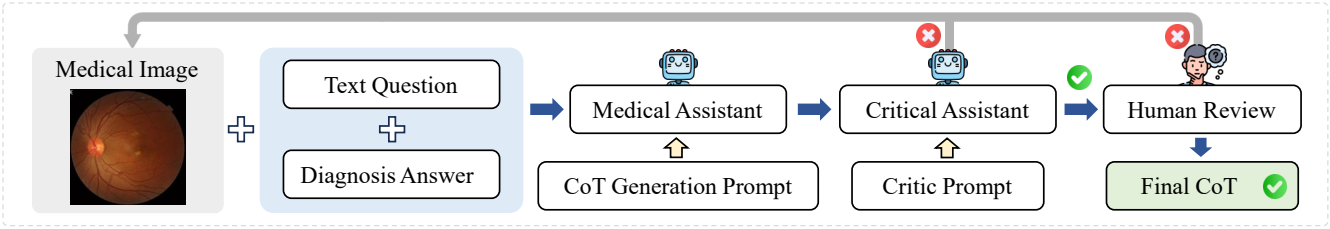

Figure 3: CoT generation pipeline for M3DS dataset construction.

in Equation 2 satisfies $\mathbf{E}_{seg} \in \mathbf{E}$. During training, visual features and text instruction x_{txt}^{i} are jointly used as input to LVLM. Consequently, it has $\mathbf{E}_{img} \subseteq \mathbf{E}$. The vision-language similarity score is defined as

$$\mathrm{Sim} = \mathbf{E}_{img} \cdot (\mathbf{E}_{seg})^{\mathrm{T}}, \tag{4}$$

where $\mathrm{Sim} \in \mathbb{R}^{n}$ measures the similarity between each image token and the special token generated by LVLM. Along with $\mathbf{E}_{seg}$ in Equation 2 and Sim defined in Equation 4, we propose the Region-Aware Vision-Language Similarity to Mask (RVLS2M) module detailed in Algorithm 1 for improving segmentation capabilities. After calculating the similarity score Sim, we normalize the similarity scores via softmax to enhance the separability of regions as follows

$$\mathrm{Sim}_{\mathrm{norm}} = \mathrm{softmax}(\mathrm{Sim}) = \left\{ \frac{\exp(s_i)}{\sum_{j=1}^{n} \exp(s_j)} \right\}_{i=1}^{n}, \tag{5}$$

where s_i denotes the i-th element in Sim. Next, we construct a region-aware vision-language similarity map. The normalized scores $\mathrm{Sim}_{\mathrm{norm}}$ are reshaped into a 2D map $M \in \mathbb{R}^{h' \times w'}$, with dimensions $h' = \lfloor \sqrt{n} \rfloor$ and $w' = \lceil n/h' \rceil$, such that

$$M_{u,v} = \mathrm{Sim}_{\mathrm{norm}}[u \cdot w' + v], \tag{6}$$

where $\forall u \in [0, h'-1], v \in [0, w'-1]$. This map is then divided into non-overlapping $g \times g$ grids. Within each grid cell, we compute region similarity $\mathcal{R}_{k,l}$ by average pooling

$$\mathcal{R}_{k,l} = \frac{1}{b^2} \sum_{i=bk}^{b(k+1)-1} \sum_{j=bl}^{b(l+1)-1} M_{i,j}, \tag{7}$$

with parameters $b = \lfloor \min(h', w')/g \rfloor$, $k, l \in [0, g-1]$. $\mathcal{R} \in \mathbb{R}^{g \times g}$ is the region-aware vision-language similarity matrix. Then, a binary segmentation mask $\mathbf{M}_{region}$ is generated by applying adaptive thresholding to $\mathcal{R}$, such that

$$\mathbf{M}_{region} = \mathbb{I}\left(\tau(\mathcal{R})\right), \tag{8}$$

where $\mathbb{I}(\cdot)$ is the indicator function. The specific selection of the τ-strategy will be detailed in Section 4. Finally, the binary segmentation mask $\mathbf{M}_{region}$ along with the special token embedding and visual features $\mathbf{F}$ from the input image x_{img}^{i} are fed into $\mathcal{M}_{SEG}^{dec}$, formulated as

$$\hat{O}_{mask} = \mathcal{M}_{SEG}^{dec}(\mathbf{F}, \mathbf{E}_{seg}, \mathbf{M}_{region}). \tag{9}$$

Algorithm 1: Region-Aware Vision-Language Similarity to Mask (RVLS2M)

Require: Image tokens $\mathbf{E}_{img} \in \mathbb{R}^{n \times d}$, special token embedding $\tilde{\mathbf{E}}_{seg} \in \mathbb{R}^{d}$, projection function ϕ, grid size g.
Ensure: Binary mask $\mathbf{M}_{region} \in \{0,1\}^{g \times g}$.

$\mathbf{E}_{seg} \leftarrow \phi(\tilde{\mathbf{E}}_{seg})$
$\mathrm{Sim} \leftarrow \mathbf{E}_{img} \cdot (\mathbf{E}_{seg})^{\mathrm{T}}$
$\mathrm{Sim}_{\mathrm{norm}} \leftarrow \mathrm{softmax}(\mathrm{Sim})$
$h' \leftarrow \lfloor \sqrt{n} \rfloor$, $w' \leftarrow \lceil n/h' \rceil$
$M \leftarrow \mathrm{reshape}(\mathrm{Sim}_{\mathrm{norm}}, [h', w'])$
$b \leftarrow \lfloor \min(h', w')/g \rfloor$
for $k \leftarrow 0$ **to** $g-1$ **do**
  for $l \leftarrow 0$ **to** $g-1$ **do**
    $\mathcal{R}_{k,l} \leftarrow \mathrm{mean}(M[bk : b(k+1), bl : b(l+1)])$
  end for
end for
$\mathbf{M}_{region} \leftarrow \mathbb{I}\left(\tau(\mathcal{R})\right)$
return $\mathbf{M}_{region}$

3.2 Training Objectives

By jointly optimizing the text generation loss $\mathcal{L}_{txt}$ in $\mathcal{M}_{LVLM}$ and the segmentation mask loss $\mathcal{L}_{mask}$ in $\mathcal{M}_{SEG}$, we enable $\mathcal{M}_{\theta}$ to perform MDS task. For $\mathcal{L}_{txt}$, we employ cross-entropy loss, and for $\mathcal{L}_{mask}$, we use a combination of binary cross-entropy (BCE) loss $\mathcal{L}_{bce}$ and DICE loss $\mathcal{L}_{dice}$. Finally, the overall loss objective $\mathcal{L}$ is the weighted average of $\mathcal{L}_{txt}$ and $\mathcal{L}_{mask}$, denoted by

$$\begin{aligned} \mathcal{L}_{mask} &= \lambda_{bce}\mathcal{L}_{bce} + \lambda_{dice}\mathcal{L}_{dice}, \\ \mathcal{L} &= \lambda_{txt}\mathcal{L}_{txt} + \lambda_{mask}\mathcal{L}_{mask}, \end{aligned} \tag{10}$$

where λ_{bce} and λ_{dice} are the weight of BCE loss $\mathcal{L}_{bce}$ and DICE loss $\mathcal{L}_{dice}$, respectively. The overall loss objective $\mathcal{L}$ is weighted by λ_{txt} for $\mathcal{L}_{txt}$ and λ_{mask} for $\mathcal{L}_{mask}$.

3.3 Test-Time Scaling for Medical Diagnosis Segmentation Task

To enhance model performance during inference, we propose a test-time scaling strategy designed for MDS task that leverages multi-path reasoning from LVLMs. Given an input medical image x_{img} and its corresponding query x_{txt}, our method generates diverse outputs through a two-stage reasoning process. Firstly, LVLM $\mathcal{M}_{LVLM}$ generates m diverse diagnosis reasoning paths, represented by

$$\{O_{txt}^{i}\}_{i=1}^{m} = \mathcal{M}_{LVLM}(\mathrm{x}_{img}, \mathrm{x}_{txt}), \tag{11}$$

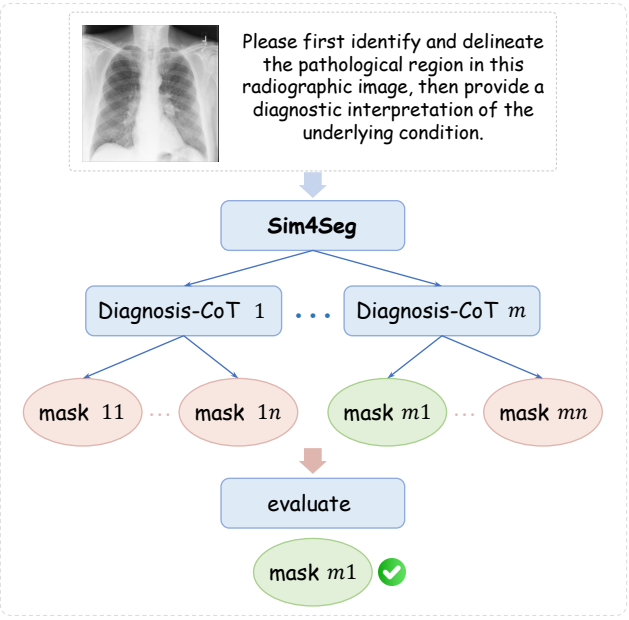

Figure 4: Test-Time Scaling strategy for MDS task.

where each $O^i_{txt} \in \mathbb{R}^d$ represents one of the m CoT reasoning output paths. For each path i, we extract the last hidden layer embedding $\tilde{\mathbf{E}}^i_{seg}$ corresponding to the special token in O^i_{txt}, and project it to obtain the refined feature $\mathbf{E}^i_{seg}$. Combined with image tokens $\mathbf{E}_{img}$ derived from the vision encoder, binary region masks $\mathbf{M}^i_{region}$ are generated through the RVLS2M module. Then, the segmentation model $\mathcal{M}_{SEG}$ produces $m \times n$ masks, formulated as

$$O^{i,j}_{mask} = \mathcal{M}_{SEG}\left(\mathrm{x}_{img}, \mathbf{E}^i_{seg}, g(\mathbf{M}^i_{region}, \theta_j)\right), \quad (12)$$

with $j = 1, \ldots, n$, and $g(\cdot)$ denoting a stochastic perturbation parameterized by $\theta_j \sim \mathcal{T}$ to ensure diversity in mask generation. The final segmentation mask is selected by maximizing an evaluation metric $\mathcal{Q}$ as follows

$$O^{final}_{mask} = \arg\max_{O^{i,j}_{mask}} \mathcal{Q}\left(O^{i,j}_{mask}, \hat{O}_{mask}\right), \quad (13)$$

where $\mathcal{Q}$ is the quality metric computed as the average of gIoU and cIoU. As illustrated in Figure 4, this strategy generates $m \times n$ candidate masks and selects one based on evaluation performance.

4 Experiments

4.1 Experimental Setting

Dataset and Metric. In our experiments, we trained Sim4Seg model on the training split (12,000 samples) of the proposed M3DS dataset and evaluated it on the test split (1,864 samples). Accuracy (Acc) was employed to evaluate the accuracy of diagnosis results. gIoU (Rezatofighi et al. 2019) and cIoU (Zheng et al. 2020) measure the overlap between the predicted segmentation masks and their corresponding ground truth.

Method	overall		
	gIoU	cIoU	Acc
LLaVA-Med (Li et al. 2023a)	-	-	3.48
SAM-Med2D (Cheng et al. 2023)	22.94	51.42	-
READ (Qian, Yin, and Dou 2025)	13.37	25.75	2.52
LISA (Lai et al. 2024)	32.43	31.83	4.71
LISA (ft) (Lai et al. 2024)	44.07	42.89	0.00
LISA (ft-diagnosis) (Lai et al. 2024)	45.87	46.05	53.27
LISA (ft-CoT) (Lai et al. 2024)	45.90	45.92	58.05
Sim4Seg (ft-diagnosis)	51.00	54.06	54.33
Sim4Seg (ft-CoT)	51.86	53.90	69.04
Sim4Seg (ft-CoT) +test-time scaling	**53.11**	**55.83**	**82.63**

Table 2: Main results on test set of M3DS dataset."ft" refers to fine-tuning model with non-diagnostic settings, "ft-diagnosis" indicates fine-tuning model with diagnosis option but without chain-of-thought data, and "ft-CoT" denotes fine-tuning model with chain-of-thought data.

Implementation Details. In the experiments, we employ LISA (Lai et al. 2024) to initialize our model. During training, we trained for four epochs, employed an AdamW optimizer (Loshchilov and Hutter 2019) with a learning rate of 3×10^{-4}, a weight decay of 0.01, a batch size of 2, and a gradient accumulation step of 10. The default loss weights λ_{mask} and λ_{txt} are set to 1.0, the BCE loss weight λ_{bce} is set to 2.0 and the DICE loss weight λ_{dice} is set to 0.5. All experiments were conducted on an NVIDIA H800 GPU.

4.2 Main Results on M3DS Dataset

The proposed Sim4Seg model demonstrates state-of-the-art performance on the M3DS dataset (5 modalities, 10 sub-datasets) through comprehensive experiments. As shown in Table 2, Sim4Seg exhibits significant improvements on evaluation metrics. Sim4Seg enhances medical vision-language models with segmentation capabilities while integrating medical diagnosis into segmentation models. Sim4Seg exceeds reasoning segmentation models by +57.3% segmentation performance and +165.4% diagnosis accuracy.

4.3 Ablation Study

We evaluated Sim4Seg components by testing individual modules and measuring their performance effects. Table 3 compares results w/ and w/o RVLS2M. Overall, the w/ RVLS2M configuration consistently outperforms the w/o RVLS2M setup. FT w/o diagnosis (fine-tuning without diagnosis text like "It is [SEG]") improves medical image segmentation capability but lacks diagnosis capabilities. FT w/ diagnosis (fine-tuning with diagnosis text formatted as "It is [SEG]. + diagnosis result") enhances both segmentation and diagnosis performance. FT w/ diagnosis-CoT (fine-tuning with diagnostic chain-of-thought) adopts format "It is [SEG]. + diagnosis CoT", with CoT generated as described in Section 2.2, significantly improves diagnosis performance. In summary, training on M3DS dataset enhances segmentation and diagnosis performance. Our proposed test-time scaling (TTS) strategy also improves both segmentation

Method		FracAtlas	BF	BFD	ISBI	ISIC	KS	BUSI	TN3K	CXD	FIVES	Avg.
w/o RVLS2M												
zero-shot	gIoU	1.42	3.42	6.26	81.25	47.71	45.49	41.70	23.56	5.93	7.47	26.42
	cIoU	1.43	2.92	5.99	76.69	49.21	46.74	32.20	21.67	6.99	7.21	25.11
	Acc	28.36	10.45	0.00	9.9	0.00	32.00	36.67	0.00	0.00	1.98	11.94
+FT w/o diagnosis	gIoU	1.63	5.15	8.35	85.11	63.40	47.03	42.58	37.79	11.99	27.90	33.09
	cIoU	1.50	3.31	10.00	82.26	62.71	46.08	45.21	33.90	18.84	28.90	33.27
	Acc	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
+FT w/ diagnosis	gIoU	1.50	5.23	10.28	85.93	63.59	56.44	41.91	40.01	14.69	32.55	**35.21**
	cIoU	1.69	7.10	11.00	83.50	63.25	54.48	46.46	39.62	21.34	33.86	**36.23**
	Acc	88.06	79.10	48.19	69.31	64.50	82.00	63.33	34.20	36.63	32.67	59.80
+FT w/ diagnosis-CoT	gIoU	1.58	5.15	8.35	85.09	63.97	54.29	45.91	40.12	15.26	31.89	35.16
	cIoU	1.71	3.31	10.00	81.85	64.85	49.79	47.30	38.93	22.49	33.22	35.35
	Acc	88.52	81.39	57.83	73.27	64.50	84.00	80.00	45.60	39.60	31.68	**64.64**
w/ RVLS2M												
+FT w/ diagnosis	gIoU	3.39	6.99	10.85	86.26	69.49	65.08	50.72	46.38	19.17	32.44	39.08
	cIoU	3.99	7.72	11.51	83.74	71.88	59.17	52.23	51.36	35.84	34.17	41.16
	Acc	88.06	82.09	42.17	54.46	63.83	57.00	58.33	46.09	34.65	31.68	55.84
+FT w/ diagnosis-CoT	gIoU	6.41	7.07	11.97	86.96	68.37	67.23	49.42	48.31	20.98	36.70	40.34
	cIoU	5.42	5.46	13.86	85.47	67.02	60.17	49.49	54.18	37.13	38.64	41.68
	Acc	89.55	86.57	63.86	74.26	68.00	87.00	81.67	72.80	41.58	33.66	69.90
+FT w/ diagnosis-CoT (TTS)	gIoU	6.98	7.48	13.61	87.44	69.33	68.28	49.28	50.03	23.53	37.77	**41.37**
	cIoU	6.21	9.40	15.90	85.48	69.72	63.04	50.10	55.66	39.58	39.75	**43.48**
	Acc	98.51	95.52	80.72	80.20	76.83	92.00	78.33	93.81	45.54	68.32	**80.98**

Table 3: Detailed ablation result on each sub-set of M3DS dataset. BF refers to bone fracture dataset, BFD indicates Bone Fracture Detection dataset, KS denotes Kvasir-SEG dataset, and CXD represents ChestX-Det dataset.

Method	gIoU	cIoU	Avg.
LISA (Lai et al. 2024)	**32.43**	31.83	32.13
LISA w/ RVLS2M module	31.82	**39.88**	**35.85**

Table 4: Zero-shot effectiveness of RVLS2M module. The proposed RVLS2M module also serves as an efficient prompt creator in zero-shot scenarios.

Modality	LISA		Sim4Seg	
	gIoU	cIoU	gIoU	cIoU
X-Ray	6.34	14.01	**10.45**	**19.07**
Dermoscopy	43.51	27.10	**44.97**	**34.96**
Endoscopy	36.13	**35.86**	**41.38**	31.12
Ultrasound	26.17	23.42	**27.08**	**23.69**
Fundus Photography	13.37	13.56	**15.76**	**16.17**

Table 5: Cross-modality generalization via testing on untrained modalities. Sim4Seg achieves superior performance, demonstrating robust adaptation to diverse modality of medical data.

and diagnosis performance, indicating that every component contributes critically to the MDS task.

4.4 Analysis

Impact of RVLS2M Module Under Zero-Shot Setting. The proposed RVLS2M module demonstrates significant improvements in segmentation performance, as validated in ablation studies. Notably, it also serves as an effective prompt creator in zero-shot scenarios. As shown in Table 4, using LISA (Lai et al. 2024) as the base model and incorporating the RVLS2M module without any training improves performance by 11.6%. This confirms the plug-and-play effectiveness and flexibility of the RVLS2M module.

Impact of Different τ Strategies. Figure 6 reveals an inverted U-shaped relationship between segmentation performance and RVLSM granularity, controlled by the τ strategy illustrated in Equation 8. Performance improves when selecting 12 to 36 grid cells with highest similarity, but declines at 48 grid cells. Similarly, optimal results emerge when selecting the top 36 grid cells at 16×16 grid resolution, with performance increasing from 8×8 to 16×16 grid resolution and decreasing from 16×16 to 64×64 grid resolution. This indicates that excessively coarse grids produce blurred regions, while overly fine grids retain spurious correlations from high-similarity points. These findings demonstrate the critical role of optimal τ strategy selection in RVLSM effectiveness.

Impact of Different Test-Time Scaling Strategies. As shown in Figure 7 (a), we investigate the impact of the pa-

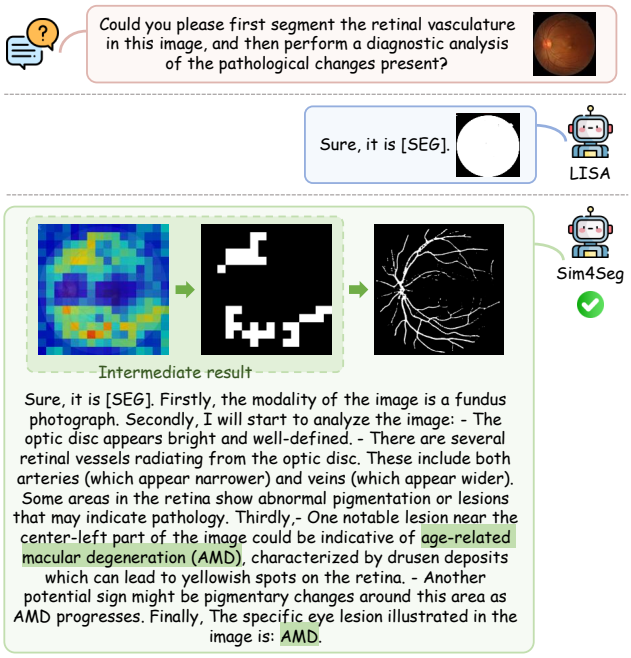

Figure 5: Case Study.

dataset	LISA		Sim4Seg	
	gIoU	cIoU	gIoU	cIoU
ISIC	58.54	54.12	**60.44**	**56.83**
Kvasir-SEG	36.13	**35.86**	**41.38**	31.12
TN3K	29.06	25.40	**33.32**	**33.62**
ChestX-Det	6.21	7.24	**7.78**	**9.13**
FIVES	13.37	13.56	**15.76**	**16.17**

Table 6: Cross-dataset generalization performance of Sim4Seg when evaluated on datasets excluded from training.

rameter m denoted in Equation 11 on model performance. The results demonstrate that higher values of m (generating more CoTs) enhance the performance of medical diagnosis. As illustrated in Figure 7 (b), we explore the effect of parameter n defined in Equation 12. The findings reveal that increasing n (producing more segmentation masks) improves medical segmentation performance.

Generalization Capability for Cross-Modality. To validate the modality generalization capability of Sim4Seg, we cyclically excluded the training data of each modality one at a time, trained the model on remaining modalities, and tested on the excluded modality. As shown in Table 5, our Sim4Seg model demonstrates superior modality generalization performance compared to others. This indicates the robustness and flexibility of our model in adapting to diverse medical imaging modalities.

Generalization Capability for Untrained Dataset. We iteratively excluded the ISIC, Kvasir-SEG, TN3K, ChestX-Det, and FIVES datasets from training data and tested our Sim4Seg model on excluded datasets to evaluate its generalization performance across various medical datasets. As presented in Table 6, Sim4Seg achieves better performance in cross-dataset generalization compared to other method, which enhances its real-world applicability in clinical deployment scenarios.

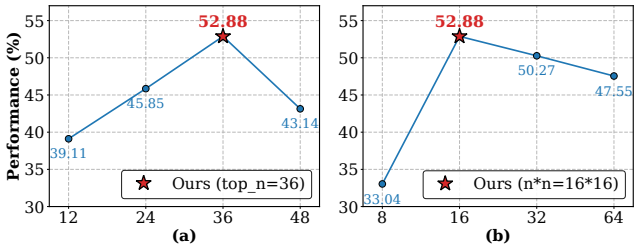

Figure 6: Impact of Different τ Strategies. An inverted U-shaped relationship exists between segmentation performance and RVLSM granularity, controlled by τ. Selecting 36 grid cells (a) and using a 16×16 grid resolution (b) achieve peak performance.

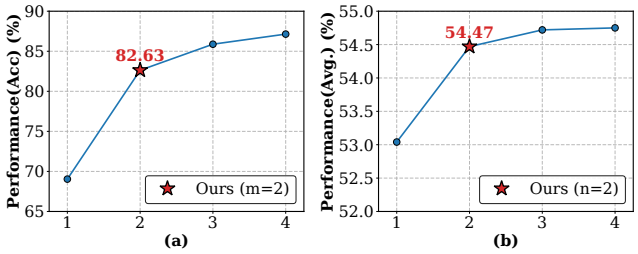

Figure 7: Performance under different test-time scaling parameters. (a) Increasing CoT paths m improves diagnosis accuracy, and (b) more segmentation masks n enhances segmentation performance.

4.5 Case Study

Figure 5 presents a multi-modal input case from the M3DS dataset, comparing outputs derived from the baseline model and our approach. Compared with the baseline model, our method generated precise segmentation masks along with diagnosis reasoning CoT. This validates the effectiveness of our proposed Sim4Seg model in MDS tasks.

5 Conclusion

In this paper, we introduced a new task named Medical Diagnosis Segmentation (MDS). To enable research in MDS, we constructed the Multimodal Multi-disease Medical Diagnosis Segmentation (**M3DS**) dataset. We further proposed **Sim4Seg**, an effective model leveraging a novel Region-Aware Vision-Language Similarity to Mask (**RVLS2M**) module. Additionally, we explored a test-time scaling strategy for MDS task to improve overall performance.

References

Abedeen, I.; Rahman, M. A.; Prottyasha, F. Z.; Ahmed, T.; Chowdhury, T. M.; and Shatabda, S. 2023. FracAtlas: A Dataset for Fracture Classification, Localization and Segmentation of Musculoskeletal Radiographs. *Scientific Data*, 10(1).

Al-Dhabyani, W.; Gomaa, M.; Khaled, H.; and Fahmy, A. 2020. Dataset of breast ultrasound images. *Data in Brief*, 28: 104863.

Chen, J.; Gui, C.; Ouyang, R.; Gao, A.; Chen, S.; Chen, G.; Wang, X.; Cai, Z.; Ji, K.; Wan, X.; and Wang, B. 2024. Towards Injecting Medical Visual Knowledge into Multimodal LLMs at Scale. In *EMNLP 2024*, 7346–7370. Association for Computational Linguistics.

Cheng, J.; Fu, B.; Ye, J.; Wang, G.; Li, T.; Wang, H.; Li, R.; Yao, H.; Chen, J.; Li, J.; Su, Y.; Zhu, M.; and He, J. 2025. Interactive Medical Image Segmentation: A Benchmark Dataset and Baseline. In *CVPR 2025*, 20841–20851. Computer Vision Foundation / IEEE.

Cheng, J.; Ye, J.; Deng, Z.; Chen, J.; Li, T.; Wang, H.; Su, Y.; Huang, Z.; Chen, J.; Jiang, L.; Sun, H.; He, J.; Zhang, S.; Zhu, M.; and Qiao, Y. 2023. SAM-Med2D. *CoRR*, abs/2308.16184.

Codella, N. C. F.; Gutman, D. A.; Celebi, M. E.; Helba, B.; Marchetti, M. A.; Dusza, S. W.; Kalloo, A.; Liopyris, K.; Mishra, N. K.; Kittler, H.; and Halpern, A. 2018. 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 *ISBI 2018*, 168–172. IEEE.

Darabi, P. 2024. Bone Fracture Detection: Computer Vision Project.

Gong, H.; Chen, J.; Chen, G.; Li, H.; Li, G.; and Chen, F. 2023. Thyroid region prior guided attention for ultrasound segmentation of thyroid nodules. *Comput. Biol. Medicine*, 155: 106389.

Guo, X.; Chai, W.; Li, S.; and Wang, G. 2024. LLaVA-Ultra: Large Chinese Language and Vision Assistant for Ultrasound. In *ACMMM 2024*, 8845–8854. ACM.

Gutman, D. A.; Codella, N. C. F.; Celebi, M. E.; Helba, B.; Marchetti, M. A.; Mishra, N. K.; and Halpern, A. 2016. Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC). *CoRR*, abs/1605.01397.

Hu, R.; Rohrbach, M.; and Darrell, T. 2016. Segmentation from Natural Language Expressions. In *ECCV 2016*, volume 9905 of *Lecture Notes in Computer Science*, 108–124. Springer.

Jha, D.; Smedsrud, P. H.; Riegler, M. A.; Halvorsen, P.; de Lange, T.; Johansen, D.; and Johansen, H. D. 2020. Kvasir-SEG: A Segmented Polyp Dataset. In *MMM 2020*, volume 11962 of *Lecture Notes in Computer Science*, 451–462. Springer.

Kai, J.; Xingru, H.; Jingxing, Z.; Yunxiang, L.; Yan, Y.; Yibao, S.; Qianni, Z.; Yaqi, W.; and Juan, Y. 2022. FIVES: A Fundus Image Dataset for Artificial Intelligence based Vessel Segmentation. *Scientific data*, 9(1): 475.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.; Dollár, P.; and Girshick, R. B. 2023. Segment Anything. In *ICCV 2023*, 3992–4003. IEEE.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. LISA: Reasoning Segmentation via Large Language Model. In *CVPR 2024*, 9579–9589. IEEE.

Lan, M.; Chen, C.; Zhou, Y.; Xu, J.; Ke, Y.; Wang, X.; Feng, L.; and Zhang, W. 2025. Text4Seg: Reimagining Image Segmentation as Text Generation. In *ICLR 2025*. OpenReview.net.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023a. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. In *NeurIPS 2023*.

Li, Y.; Wang, H.; Duan, Y.; and Li, X. 2023b. CLIP Surgery for Better Explainability with Enhancement in Open-Vocabulary Tasks. *CoRR*, abs/2304.05653.

Lian, J.; Liu, J.; Zhang, S.; Gao, K.; Liu, X.; Zhang, D.; and Yu, Y. 2021. A Structure-Aware Relation Network for Thoracic Diseases Detection and Segmentation. *IEEE Transactions on Medical Imaging*, 40(8): 2042–2052.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *NeurIPS 2023*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *ICLR 2019*. OpenReview.net.

Qian, R.; Yin, X.; and Dou, D. 2025. Reasoning to Attend: Try to Understand How Token Works. In *CVPR 2025*, 24722–24731. Computer Vision Foundation / IEEE.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML 2021*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Ramesh, K. K. D.; Kumar, G. K.; Swapna, K.; Datta, D.; and Rajest, S. S. 2021. A Review of Medical Image Segmentation Algorithms. *EAI*, 7(27): e6.

Rasheed, H. A.; Maaz, M.; Mullappilly, S. S.; Shaker, A. M.; Khan, S. H.; Cholakkal, H.; Anwer, R. M.; Xing, E. P.; Yang, M.; and Khan, F. S. 2024. GLaMM: Pixel Grounding Large Multimodal Model. In *CVPR 2024*, 13009–13018. IEEE.

Ren, Z.; Huang, Z.; Wei, Y.; Zhao, Y.; Fu, D.; Feng, J.; and Jin, X. 2024. PixelLM: Pixel Reasoning with Large Multimodal Model. In *CVPR 2024*, 26364–26373. IEEE.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I. D.; and Savarese, S. 2019. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In *CVPR 2019*, 658–666. Computer Vision Foundation / IEEE.

Roboflow100. 2023. bone fracture Dataset. https://universe.roboflow.com/roboflow-100/bone-fracture-7fylg. Visited on 2025-07-17.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, 234–241. Springer.

Saha, A.; Hosseinzadeh, M.; and Huisman, H. J. 2021. End-to-end prostate cancer detection in bpMRI via 3D CNNs: Effects of attention mechanisms, clinical priori and decoupled false positive reduction. *Medical Image Anal.*, 73: 102155.

Tong, Q.; Lu, Z.; Liu, J.; Zheng, Y.; and Lu, Z. 2025. MediSee: Reasoning-based Pixel-level Perception in Medical Images. *CoRR*, abs/2504.11008.

Wang, R.; Lei, T.; Cui, R.; Zhang, B.; Meng, H.; and Nandi, A. K. 2022. Medical image segmentation using deep learning: A survey. *IET Image Process.*, 16(5): 1243–1267.

Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; and Summers, R. M. 2017. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *CVPR 2017*, 3462–3471. IEEE Computer Society.

Xia, Z.; Han, D.; Han, Y.; Pan, X.; Song, S.; and Huang, G. 2024. GSVA: Generalized Segmentation via Multimodal Large Language Models. In *CVPR 2024*, 3858–3869. IEEE.

Zhang, X.; Wu, C.; Zhao, Z.; Lin, W.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering. *CoRR*, abs/2305.10415.

Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; and Ren, D. 2020. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. In *AAAI 2020*, 12993–13000. AAAI Press.

Zhou, Y.; Song, L.; and Shen, J. 2025a. Improving Medical Large Vision-Language Models with Abnormal-Aware Feedback. In *ACL 2025*, 12994–13011. Association for Computational Linguistics.

Zhou, Y.; Song, L.; and Shen, J. 2025b. MAM: Modular Multi-Agent Framework for Multi-Modal Medical Diagnosis via Role-Specialized Collaboration. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, 25319–25333. Association for Computational Linguistics.

Zhou, Z.; Siddiquee, M. M. R.; Tajbakhsh, N.; and Liang, J. 2018. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In *MICCAI 2018*, volume 11045 of *Lecture Notes in Computer Science*, 3–11. Springer.