RadDiagSeg-M: A Vision Language Model for Joint Diagnosis and Multi-Target Segmentation in Radiology

Chengrun Li[1] **Corentin Royer**[3] **Haozhe Luo**[2] **Bastian Wittmann**[1] **Xia Li**[3]
Ibrahim Hamamci[1] **Sezgin Er**[1] **Anjany Sekuboyina**[1] **Bjoern Menze**[1] [†]

[1] University of Zurich, Switzerland
[2] University of Bern, Switzerland
[3] ETH Zurich, Switzerland
[†] Corresponding Author

Abstract

Most current medical vision language models struggle to jointly generate diagnostic text and pixel-level segmentation masks in response to complex visual questions. This represents a major limitation towards clinical application, as assistive systems that fail to provide both modalities simultaneously offer limited value to medical practitioners. To alleviate this limitation, we first introduce *RadDiagSeg-D*, a dataset combining abnormality detection, diagnosis, and multi-target segmentation into a unified and hierarchical task. *RadDiagSeg-D* covers multiple imaging modalities and is precisely designed to support the development of models that produce descriptive text and corresponding segmentation masks in tandem. Subsequently, we leverage the dataset to propose a novel vision-language model, *RadDiagSeg-M*, capable of joint abnormality detection, diagnosis, and flexible segmentation. *RadDiagSeg-M* provides highly informative and clinically useful outputs, effectively addressing the need to enrich contextual information for assistive diagnosis. Finally, we benchmark *RadDiagSeg-M* and showcase its strong performance across all components involved in the task of multi-target text-and-mask generation, establishing a robust and competitive baseline. Code for this work is published at: github.com/RadDiagSeg

Introduction

Advances in medical Vision Language Models (VLMs), such as LLaVA-Med, Med-PaLM M, MedGemma, have demonstrated vast assistive potentials for several medical tasks (Li et al. 2023; Tu et al. 2024; Sellergren et al. 2025). As one of the most important diagnostic tools, radiological images (*e.g.*, X-ray, CT, and MRI) offer a high amount of clinical insights. Whilst demonstrating strong capabilities in understanding radiological images and answering questions, medical VLMs unanimously fail to accurately reflect their findings through an accurate pixel-level segmentation mask. This renders their results less reliable, given the known problem of LM hallucination (Liu et al. 2024a). To effectively assist clinicians, a model should be able to provide textual answers and accurate segmentation masks in tandem.

The emergence of promptable segmentation foundation models (FMs) in the medical field, such as Biomed-Parse (Zhao et al. 2025) and MedSAM (Ma et al. 2024), enables the segmentation of varying medical targets with

Figure 1: Overview. *RadDiagSeg-M* is capable of jointly detecting and diagnosing abnormality, and providing multi-target segmentation masks.

user-defined prompts, *e.g.*, points, boxes, and text labels. Architectures such as LISA (Lai et al. 2024) and Sa2VA (Yuan et al. 2025) provide ways to connect powerful pre-trained VLMs with the segmentation FMs, enabling segmentation with free-form text prompts. Early endeavors in the medical field followed the idea of LISA, such as VividMed (Luo et al. 2025) and MedPLIB (Huang et al. 2025). However, these models only work with the Referring Segmentation (Ref-Seg) or the Visual Question Answering (VQA) task, thus failing at more complex tasks requiring both textual answers and masks at the same time. Furthermore, current models are unable to generate multiple masks for a given image with one prompt, partially compromising flexibility and clinical utility. Narrowing this gap, our model can answer complex questions with text and segmentation masks of abnormalities and the corresponding infected organs.

Given the complexity and novelty of our task, we identify the absence of datasets, an effective benchmark, and suitable models to serve as a baseline. In this paper, we address these limitations: First, we propose the *RadDiagSeg-D* dataset consisting of more than 28k high-quality data samples covering major radiological modalities, *i.e.*, X-ray and CT, by aggregating and processing several public datasets (Tahir et al. 2021; Zhao et al. 2025; Antonelli et al. 2022). Each sample in *RadDiagSeg-D* consists of 3-step hierarchical questions: a close-ended VQA for abnormality detection, an open-ended VQA for diagnosis, and a segmentation task for one or multiple objects. The questions get more difficult with progression, and failing an earlier step will lead to automatic failure for the rest. The design of this task

fosters explicit, step-by-step answers that are easier to inspect and offer more granularity, while maintaining extensive coverage of the VLM capabilities. Second, we propose the *RadDiagSeg-M* (Radiological Diagnostic Segmentation) VLM. *RadDiagSeg-M* is built upon the state-of-the-art architecture proposed by Lai et al. (2024), where we expand the vocabulary of our medical VLM with special segmentation generation tokens to trigger mask generation. Notably, the overall model is trained end-to-end with a unified training process. Unlike existing models, which only support single mask generation, our model inherently supports a flexible number of mask generations. Finally, we design the evaluation process and publish the benchmarking tool for the research community to enable effective and thorough evaluation for multi-target text-and-mask generation tasks.

In summary, our contributions are as follows:

1. We introduce *RadDiagSeg-D*, a dataset comprising over 28k samples. Each sample includes three-step hierarchical questions covering VQA and segmentation.

2. We propose *RadDiagSeg-M*, a radiological VLM that is capable of joint abnormality detection, diagnosis, and flexible multi-target segmentation.

3. We design and publish a benchmarking tool for the effective evaluation of *RadDiagSeg-D*.

4. Experimental results indicate that *RadDiagSeg-M* achieves start-of-the-art results on the VQA sub-tasks of the *RadDiagSeg-D* benchmark, while establishing a competitive baseline for the complete task.

Related Methods

Segmentation Models in Medical Images

Specialist models targeting a specific organ of a specific modality have been the dominating approach regarding medical image segmentation over the past decade. CNN-based architectures like the U-Net (Ronneberger, Fischer, and Brox 2015) and its variants, such as Swin-UNet, ResUNet++, and TransUnet (Cao et al. 2022; Jha et al. 2019; Chen et al. 2021) have achieved competitive results on many specific segmentation tasks. Whilst offering robust performance on the trained modality, such specialist models, however, have shown limited generalization across modalities. Recent universal segmentation paradigms (Kirillov et al. 2023; Zou et al. 2023) enabled the emergence of many generalist models for medical images, such as MedSAM (Ma et al. 2024), SAM-Med2D (Cheng et al. 2023), and BiomedParse (Zhao et al. 2025). Unlike the SAM-like models (Kirillov et al. 2023; Ma et al. 2024; Cheng et al. 2023), requiring explicit positional prompts such as dots or boxes, Biomed-Parse (Zhao et al. 2025) works solely with textual labels. Despite its novelty, we argue that the text encoder employed by BiomedParse inherently lacks the ability to understand free-form text prompts. Therefore, it cannot handle the complex VQA tasks. *RadDiagSeg-M* overcomes this limitation with Multimodal LM to enable language comprehension and complex question-answering behavior.

Medical Vision Language Models

Vision Language Models (VLMs) in the general vision domains have demonstrated their promising abilities in the vision-related understanding and answering tasks, *e.g.*, VQA and image captioning. (Wang et al. 2024a; Team et al. 2025; Chen et al. 2024). As medical images play a vital role in clinical practices, medical VLMs such as RadFM (Wu et al. 2023), LlaVA-Med (Li et al. 2023), MedPaLM M (Tu et al. 2024), Med-flamingo (Moor et al. 2023) were developed with large-scale radiology datasets and with techniques from the general domain, *e.g.*, instruction fine-tuning (Liu et al. 2023). Despite strong performance on downstream tasks, existing models lack the ability to segment abnormalities—a critical requirement in radiological practice.

Medical VLMs with Segmentation Mask Outputs

The success of models such as LISA (Lai et al. 2024), LLM-Seg (Wang and Ke 2024), SegLLM (Wang et al. 2024b), and Sa2VA (Yuan et al. 2025) demonstrates the potential of connecting VLMs with generalist segmentation models. In the medical domain, several VLMs (Luo et al. 2025; Huang et al. 2025) inspired by LISA have been developed to perform segmentation tasks. However, these models cannot simultaneously answer complex questions and generate segmentation masks for relevant findings. Concurrent work of UniBiomed (Wu et al. 2025) attempts to address this challenge. Given the critical role of radiological imaging in medical diagnosis, a model that can detect, diagnose, and deliver pixel-level segmentations is essential for advancing toward a reliable radiological AI assistant. To this end, we propose *RadDiagSeg-M*, a VLM enhanced with the capability to answer complex questions alongside accurate, pixel-level segmentation across major radiological imaging modalities.

Methods

RadDiagSeg-M: Model Architecture

Our model generally follows the embedding-as-prompt architecture proposed in LISA (Lai et al. 2024), which has been widely adopted by VLMs with segmentation capabilities in the medical field (Luo et al. 2025; Huang et al. 2025). However, most of the above-mentioned models utilize their components trained on in-house data or train a decoder module from scratch. Besides, training of these models involves multiple stages and (un-)freezing different parts at different stages. In our work, we propose a model structure built entirely with open-source components, trained in a simplistic yet elegant two-stage end-to-end process. Our model surpasses LISA-like models regarding flexibility in mask generation and the joint language-segmentation capability, as is demonstrated in Table 1.

RadDiagSeg-M consists of three main components: a vision backbone, a multimodal language model (multimodal LM), and a mask decoder, as is shown in Figure 2. The multimodal LM processes a user text prompt together with an image to generate a text answer. Following LISA, we re-purpose a series of $<seg>$ tokens to guide the segmentation process. If the multimodal LM chooses to carry out the segmentation task, a special segmentation token, *e.g.*,

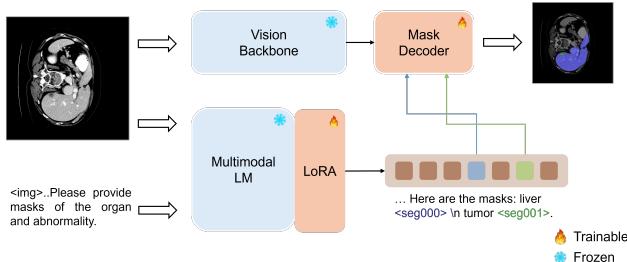Figure 2: Architecture of *RadDiagSeg-M*. Our model supports multi-target flexible segmentation.

	Dect	Diag	Seg	Mul-Seg	VQA-Seg
BiomedParse	✗	✗	✓	✓	✗
LISA	✓	✗	✓	✗	✗
MedGemma	✓	✓	✗	✗	✗
MedPLIB	✓	✓	✓	✗	✗
UniBiomed	✓	✓	✓	✗	✓
RadDiagSeg-M	✓	✓	✓	✓	✓

Table 1: Comparison of model capabilities on detection, diagnosis, segmentation, multi-target segmentation, and VQA segmentation. Mul-Seg refers to the ability to generate multiple clearly-referred masks for different targets. *RadDiagSeg-M* is the first model capable of all tasks.

<seg000>, is generated. The last layer hidden embedding of the special token is passed through the mask decoder to create a binary segmentation mask. These special tokens also differentiate the normal VQA behavior from tasks requiring segmentation output, since only the answer containing *<seg>* will activate the segmentation process.

Vision backbone. The vision backbone F_{enc} extracts pixel-level visual features from the input medical image x_{image} to support mask generation. We adopt the image encoder from MedSAM (Ma et al. 2024) to leverage pretrained model knowledge. Given a batch of b input images $x_{\text{image}} \in \mathbb{R}^{b \times 3 \times W \times H}$, the images are transformed by the vision backbone into image embeddings $z_{\text{image}} \in \mathbb{R}^{b \times 256 \times \frac{W}{16} \times \frac{H}{16}}$.

Multimodal LM. Many general domain multimodal LMs demonstrate strong question answering capabilities when directly applied to medical tasks (see LLaVA (Liu et al. 2023), Qwen-VL (Wang et al. 2024a), and InternVL (Chen et al. 2024)). However, due to the unique properties of radiological images, Multimodal LM's internal image encoders trained on natural images typically fail to generalize. Therefore, to create a powerful multimodal LM for radiological images, we substitute the multimodal LM's native image encoder with a pretrained medical CLIP-based variant (Zhang et al. 2024). Additionally, we apply LoRA (Hu et al. 2022) for parameter-efficient fine-tuning of LM. We discuss design choices in the ablation studies.

Every sample in a batch of b consists of an image prompt and a text prompt: $(x_{\text{image}}, x_{\text{text}})$. We feed the image-text pair to the multimodal LM, which in turn outputs a text response $\hat{y}_{\text{text}}$. The corresponding last-layer hidden state can be described as z_{emb}, the process of which can be formulated as

$$z_{\text{emb}} = \text{LM}(x_{\text{image}}, x_{\text{text}}). \tag{1}$$

Mask decoder. The mask decoder module consists of the mask decoder from MedSAM (Ma et al. 2024) and a linear projection layer to align the embedding shape. When the multimodal LM decides to generate segmentation mask(s), $\hat{y}_{\text{text}}$ will contain one or multiple segmentation control tokens. Therefore the last-layer embedding z_{emb} contains a non-empty subset of segmentation token embeddings h_{seg}, expressed as $z_{\text{emb}} \supseteq \{h_{\text{seg}}^{(i)}\}_{i=1}^{k}, 0 < k \le n$.

Iteratively for each of the segmentation token embedding $h_{\text{seg}}^{(i)}$, the mask decoder MD will process the image embed-ding z_{image} to generate the binary segmentation mask $\hat{M}$. The decoding process for a mask can be formalized as:

$$\hat{M} = \text{MD}(h_{\text{seg}}^{(i)}, z_{\text{image}}). \tag{2}$$

Training objectives. The training objective is to jointly minimize the auto-regressive LM loss $\mathcal{L}_{\text{text}}$ and the segmentation loss $\mathcal{L}_{\text{seg}}$. As the segmentation loss, we adopt a combination of pixel-level binary cross-entropy (BCE) loss and Dice loss, following MedSAM (Ma et al. 2021). Overall, the composition objective $\mathcal{L}$ can be framed as:

$$\mathcal{L} = \lambda_{\text{text}} \cdot \mathcal{L}_{\text{text}} + \lambda_{\text{seg}} \cdot \mathcal{L}_{\text{seg}}, \tag{3}$$

where

$$\mathcal{L}_{\text{seg}} = \lambda_{\text{bce}} \cdot \mathcal{L}_{\text{bce}}(\hat{M}, M) + \lambda_{\text{dice}} \cdot \mathcal{L}_{\text{dice}}(\hat{M}, M). \tag{4}$$

RadDiagSeg-M: Training Data and Tasks

As is illustrated in Figure 3, the training process of our model involves three different tasks, all of which are derived from widely-adopted public datasets. In the following, we describe these three tasks in detail:

Referring segmentation task (Ref-Seg). A sample contains an image, a binary segmentation mask, and a text label for the segmentation target. A template to normalize our data points is: **USER**: "<img> Please segment *Target* in the *Image Modality*." **ASSISTANT**: "Here is the mask for *Target* <seg>." The text label can either directly refer to the segmentation target, *e.g.*, liver, or by its functionality, *e.g.*, hepatic organ. During training, a label will be randomly sampled to ensure diversity in training data. We adopt subsets of BiomedParseData (Zhao et al. 2025) as the training set.

Visual question answering task (VQA). VQA task involves the generation of accurate natural language answers to visually-based questions. To preserve the visual-language capability of pretrained multimodal LM according to McKinzie et al. (2024), we incorporate public radiological VQA datasets throughout the training. We utilize VQA-RAD (Lau et al. 2018) and SLAKE (Liu et al. 2021), both of which provide high-quality visually-based question answer pairs in the radiology domain.
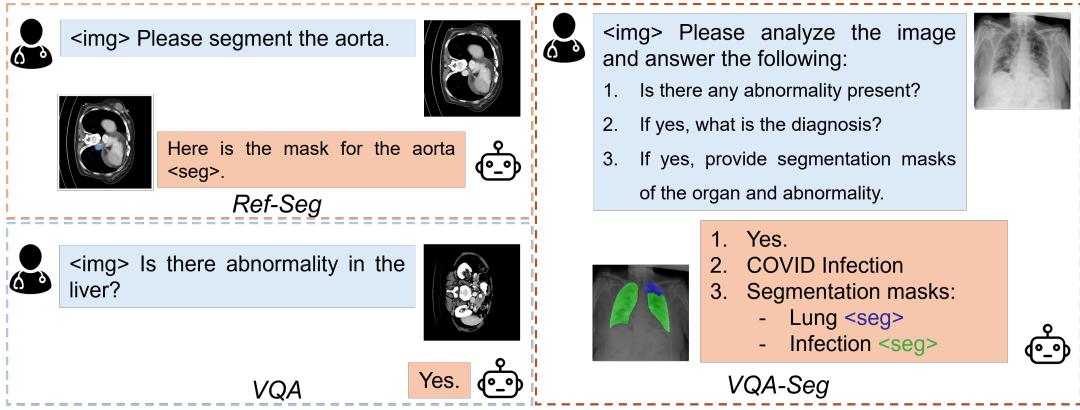
Figure 3: Three types of tasks in training. Notably, the VQA Segmentation (VQA-Seg) is a novel complex task that involves three steps: textual tasks of detection and diagnosis, and a multi-target segmentation task.

VQA segmentation task (VQA-Seg). To foster *RadDiagSeg-M*'s ability in answering complex questions while providing segmentation mask(s), we proposed a complex task comprised of hierarchical VQA tasks and segmentation tasks. Data is structured in a unified format, where a data sample with positive findings is formatted as shown in Figure 3. For data samples with negative findings (no visible abnormality), a negative text answer is utilized. Every question includes three steps: a close-ended question with a binary answer for detection, an open-ended question for diagnosis, and a subsequent segmentation task requiring one or multiple masks. Specifically, we process a subset from BiomedParseData (Zhao et al. 2025; Antonelli et al. 2022) and COVID-QU-Ex (Tahir et al. 2021) as the *RadDiagSeg-D* dataset used in experiments. Notably, our proposed VQA-Seg task guides the model to think and respond in a step-by-step manner, producing both coherent diagnoses and clinically meaningful segmentation masks.

RadDiagSeg-D Benchmark

Our proposed VQA-Seg task is complex and challenging in its structured steps of questions. For the text part, we consider the right combination of detection and diagnosis as a correct prediction. Since the mask generation is conditioned on text embedding, we argue that a correct text answer is a prerequisite for meaningful segmentation masks.

The hierarchical complexity of the task also presents challenges to the effective evaluation, since there hasn't been an established benchmark for this composite task. We address this deficiency by extending a widely adopted medical benchmarking tool. The evaluation process treats the first step of the task as a close-ended question and computes the F1 score. For the second step, we treat the diagnosis problem as an open-ended VQA and report the overall F1, *i.e.*, only the correct combination of detection and diagnosis is considered a success. A failure in the earlier stage will automatically stop the evaluation, leading to zero results in the following stages. For example, if the model answers, *e.g.*, *"1. Yes. 2. There is ..."* while the ground truth is *"1. No"*. The answer fails at the detection level, leading automatically

to the failure of the following steps and the whole task. We document details of the evaluation process in the appendix.

Experiments

Model Implementation

Model details. Unless otherwise specified, the implementation of *RadDiagSeg-M* relies on the following components: We adopted the respective MedSAM components as the vision backbone and mask decoder. For the multimodal LM, there are three key components: the LM, the image encoder, and the multimodal projector. We use the LM from *PaliGemma2-3b-pt-224* (Steiner et al. 2024). The image encoder is BiomedCLIP (Zhang et al. 2024) due to its existing knowledge on medical images and CLIP-style pretraining (Radford et al. 2021). A multimodal projector projects the raw image embedding to the LM embedding (Tolstikhin et al. 2021). In comparison to previous works, all of our components are open-sourced, underpinning the effectiveness and flexibility of our design. The choices of components are discussed in the ablation studies. Additional details can be found in the appendix.

Training details. We utilize two NVIDIA H100 (80 GB) GPUs for training and leverage the DeepSpeed ZeRO engine (Rasley et al. 2020) for efficient distributed computation. We adopt a two-stage training process, which comprises initial pre-training followed by fine-tuning. The pre-training stage aims at aligning different components and activating the full model capability for segmentation, whereas the fine-tuning stage optimizes model performance on the specific VQA-Seg task. As indicated in Figure 2, the trainable components are the LoRA, the multimodal projector within the multimodal LM, the mask decoder, together with the text embeddings of segmentation tokens. Complete specification is documented in the appendix.

Datasets. Given the efforts required in the annotation of radiological images and the scarcity of medical data in general, there are few datasets providing both pixel-level masks for the organ and abnormality. We transformed and

	VQA-RAD				SLAKE			
	F1	Recall	OpenQ-Acc	OpenQ-Recall	F1	Recall	OpenQ-Acc	OpenQ-Recall
RadFM	<u>0.442</u>	**0.474**	<u>0.335</u>	<u>0.407</u>	0.714	0.695	<u>0.725</u>	<u>0.758</u>
LlaVA-Med	0.069	0.372	0.140	0.246	0.075	0.443	0.362	0.492
MedGemma	0.164	0.449	**0.415**	**0.518**	0.066	0.565	0.593	0.664
LISA	0.052	0.229	0.080	0.132	0.070	0.314	0.207	0.244
UniBiomed	0.020	0.084	0.045	0.023	0.106	0.011	0.076	0.088
PaliGemma2	0.352	0.350	0.150	0.169	0.337	0.336	0.245	0.253
RadDiagSeg-M-PT	**0.460**	<u>0.458</u>	0.238	0.291	<u>0.718</u>	<u>0.716</u>	0.666	0.705
RadDiagSeg-M-FT	0.351	0.353	0.306	0.245	**0.774**	**0.778**	**0.754**	**0.779**

Table 2: VQA results for *RadDiagSeg-M*. "OpenQ" denotes open-ended questions, while "Acc" denotes accuracy. *RadDiagSeg-M* outperforms most baselines by a wide margin and achieves state-of-the-art performance on SLAKE.

aggregated COVID-QU-Ex (Tahir et al. 2021) and subsets of MSD (Antonelli et al. 2022; Zhao et al. 2025) into *RadDiagSeg-D*, containing 22k samples for training and 6.8k for testing. For X-ray, potential abnormalities are COVID-19 and non-COVID infection, where healthy samples represent negatives. For CT, potential abnormalities are liver and pancreas tumors, where the slices with no visible tumor are considered as negative samples. We format the samples according to the example shown in Figure 3.

The training process involves three types of tasks, where a different combination is adopted for the two stages. For the pre-training stage, we adopt Ref-Seg task with 199k samples from BiomedParseData (Zhao et al. 2025) covering all three key modalities of radiology images, *i.e.*, X-ray, CT and MRI. 5k VQA samples are taken from VQA-RAD (Lau et al. 2018) and SLAKE (Liu et al. 2021) to maintain the joint understanding capability. In the fine-tuning stage, we adopt a mixture of the VQA and the *RadDiagSeg-D* datasets, resulting in a total of 32k high-quality samples. *RadDiagSeg-D* constitutes 22k samples, and the rest 10k are VQA samples. We discuss the effect of the data mix of the fine-tuning stage in the ablation studies. For 3D modalities like CT and MRI, the slices from the same volume don't infiltrate across data partitions (Zhao et al. 2025).

Evaluation metrics. For evaluation, we adopt F1 and Recall as metrics for the VQA tasks. We additionally document the Recall and Accuracy for open-ended questions. Following common practices, the Dice score is used to benchmark the quality of segmentation. For *RadDiagSeg-D*, given the label imbalance, we use F1 as the metric for detection and diagnosis.

Experiment Results

RadDiagSeg-M is novel in its ability to answer complex questions with structured text answers and well-referred segmentation mask(s). We first present the model's capability on downstream tasks of Ref-Seg and VQA. Subsequently, we present the results on the complex task of VQA-Seg.

Ref-Seg results. To evaluate the effectiveness of the pre-training stage, we benchmarked on the Ref-Seg task across all three modalities, *i.e.*, X-ray, CT, and MRI. As shown

in Figure 5, we compare our pre-trained (PT) model with three approaches: the vanilla approach combining Grounding DINO with MedSAM (Liu et al. 2024b; Ma et al. 2024), LISA (Lai et al. 2024), and UniBiomed (Wu et al. 2025). In the first approach, Grounding-DINO processes a text label and generates a bounding box to prompt MedSAM for mask generation. LISA achieves strong performance across many general-domain benchmarks. We consider these two approaches as our baselines.

RadDiagSeg-M consistently outperforms both baselines across all modalities with a margin of over 0.2 in Dice. Our improvement confirms the advantages of the joint embedding space and the benefits of domain-specific pre-training. We acknowledge the gap between our model and the concurrent method UniBiomed. Despite the difference in training scale, we aim to create a model capable of joint text and multi-target segmentation generation (see Figure 4).

VQA results. To maintain and improve the capabilities of image understanding and question answering, we have included proportions of VQA data in every training stage. Table 2 presents the evaluation results on the test sets of two radiology VQA datasets: VQA-RAD (Lau et al. 2018) and SLAKE (Liu et al. 2021). We benchmarked two variants of our model (pre-trained variant (PT), fine-tuned variant (FT)) against state-of-the-art (SOTA) medical VLMs (RadFM, LLaVA-Med, and MedGemma) and VLMs with segmentation capability (LISA, UniBiomed). Both variants of our model show improved performance over the base model PaliGemma2 (Steiner et al. 2024). We highlight that our FT model achieves state-of-the-art results on SLAKE. We observe a decline in part of the metrics from the PT to the FT variant, which we attribute to the joint learning objective of the *RadDiagSeg-D* task.

Notably, we observe the collective poor performance of current SOTA models with segmentation capabilities, LISA, and UniBiomed. Their scores indicate failures to correctly answer the majority of questions. Especially, UniBiomed was trained on both datasets, yet it demonstrates the worst performance reported. We present qualitative examples of close- and open-ended questions in Figure 4. We analyze and discuss the results further in the qualitative analysis.

Figure 4: Qualitative results for the VQA-Seg and VQA tasks. *RadDiagSeg-M* is the only model answering all three questions correctly while providing multiple well-referred masks following the instructions. The results underpin the capability of *RadDiagSeg-M* at multi-target text-and-mask generation.
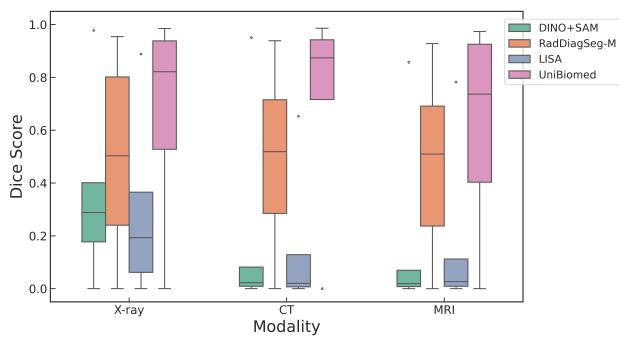

Figure 5: Box plot comparing performances of models on the Ref-Seg task across three modalities. *RadDiagSeg-M* significantly outperforms the baseline methods.

VQA-Seg results. Given the novelty and complexity of the task, Table 1 indicates that *RadDiagSeg-M* is the only model capable of this task. To guarantee a fair comparison, if another tested model failed to generate meaningful results for the full task, we amputated the task to retrieve meaningful scores. Despite the efforts, empty fields in Table 3 indicate the failure of existing models to perform all the required sub-tasks. *RadDiagSeg-M* is the first model capable of joint detection, diagnosis, and multi-target segmentation.

Notably, comparable VLMs with segmentation capabilities demonstrate difficulties in following instructions and answering questions step-by-step. For example, even adding an explicit prompt requiring a binary *yes/no* answer in the

detection task, UniBiomed still fails to generate a binary answer. Similarly, LISA fails at generating meaningful text after the detection step. Towards the objective of building an assistive model for clinicians, such models' failure hinders the potential of clinical application. The result also verifies the necessity of more complex tasks like *RadDiagSeg-D*.

Besides being the first model capable of performing the whole task, we establish a robust baseline for all the tasks. On X-ray, *RadDiagSeg-M* shows leading performance in all the task categories. On CT, *RadDiagSeg-M* achieves comparable performance on detection, and an improved performance in diagnosis and organ segmentation. Collectively, we outperform existing methods with improved metrics and set a competent baseline for the task.

Qualitative analysis. Figure 4 illustrates the joint complex question answering and flexible segmentation abilities of *RadDiagSeg-M*. For the VQA-Seg task, other comparable models, except MedGemma, fail to follow the instructions to answer the questions. More importantly, for LISA and UniBiomed, the failure in diagnosis subsequently leads to ambiguity in the segmentation target. As reported in UniBiomed (Wu et al. 2025), its improvement of performance is partly attributed to the mask generation process conditioned on the textual output and input. Therefore, if the textual answer is ambiguous, the mask will also be less credible. Similarly, the complete devoid of useful textual information from LISA leads to ambiguity of the segmentation target. Furthermore, the example of the VQA task demonstrates the deterioration of language capability from UniBiomed, answering different questions with the same irrelevant and incorrect answer. Collectively, these examples confirm our claim that a more complex task with both text

	X-ray				CT			
	Detection F1	Diagnosis F1	Segmentation Dice-Org	Dice-Abn	Detection F1	Diagnosis F1	Segmentation Dice-Org	Dice-Abn
MedGemma	<u>0.904</u>	0.625	–	–	**0.526**	0.336	–	–
LISA*	0.857	–	–	–	<u>0.521</u>	–	–	–
UniBiomed*	0.838	<u>0.724</u>	–	0.410	0.502	<u>0.627</u>	–	**0.214**
RadDiagSeg-M	**0.912**	**0.864**	**0.833**	**0.541**	0.506	**0.657**	**0.670**	<u>0.103</u>

*Amputation and adaptation of questions needed for meaningful value.
– Model is *not* capable of performing the task.

Table 3: Performance on *RadDiagSeg-D* task. "Org" abbreviates organ, and "Abn" abnormality. *RadDiagSeg-M* demonstrates state-of-the-art results on X-ray, and achieves comparably competitive results on CT, setting a robust baseline for the task.

Encoder	LM params	ImgTok #	SLAKE F1	Ref-Seg CT Dice
SigLIP	3b	256	0.740	0.227
B-CLIP	3b	256	0.744	**0.488**
MedSAM	3b	256	0.693	0.263
B-CLIP	10b	1024	**0.763**	<u>0.484</u>
B-CLIP	10b	256	<u>0.759</u>	0.483

Table 4: Ablation of components in the multimodal LM, including image encoder, language model parameter count, and number of image tokens. Notably, the image encoder has the strongest overall impact on the downstream tasks.

Seg %	Text %	SLAKE F1	*RadDiagSeg-D* Diagnosis F1
0.8	0.2	0.694	0.812
0.6	0.4	**0.743**	**0.885**

Table 5: Ablation of fine-tuning data composition. "Seg" denotes the proportion of VQA-Seg samples, and "Text" that of VQA samples. A higher proportion of "Text" samples improves the model's performance on downstream tasks.

and segmentation, *e.g.*, *RadDiagSeg-D*, is needed towards building an assistive VLM for radiological diagnosis.

Through the qualitative analysis, we demonstrate the practical significance of a complex VQA task like *RadDiagSeg-D* for the assistive diagnosis in the radiological image field. We confirm through the success of *RadDiagSeg-M* that the joint ability to answer complex questions and generate multiple masks is both important and learnable.

Ablation Studies

Effect of component choice in multimodal LM. Starting with the image encoder in Table 4, using a medical-aware vision encoder B-CLIP (BiomedCLIP) significantly improves segmentation performance and moderately enhances performance on the visual-language understanding tasks, compared to the baseline of SigLIP. This highlights the importance of a vision encoder pre-trained with domain-specific images. While MedSAM offers a structurally simpler alternative, its lack of language awareness during pretraining appears to limit performance, particularly in cross-modal tasks.

Increasing the language model size and the number of image tokens further boosts performance following the scaling law (Kaplan et al. 2020). Comparing variants with Biomed-CLIP as encoder, we find that increasing the LM size and the image tokens yields consistent improvements on VQA tasks. However, scaling up doesn't result in any improvement on the Ref-Seg task.

Effect of data composition in fine-tuning stage. The objective of the fine-tuning stage is to jointly improve the performance on text generation and segmentation capabilities for the *RadDiagSeg-D* task. We ran the ablation study on the X-ray portion of *RadDiagSeg-D* and VQA datasets. Given the composite loss function of Equation 3 used in our training process, the task composition has a direct impact on the flow of gradients, thus directly influencing the outcomes. We ablated the effect in Table 5, where a higher proportion of VQA data with pure-text output mitigates the model collapsing on the language abilities, thus benefiting the joint improvement of all downstream tasks with a 0.07 increase on the *RadDiagSeg-D* Diagnosis F1 score.

Conclusion

In this paper, we focus on the capabilities of radiological VLMs to jointly generate high-quality diagnostic text and clearly referred segmentation masks. To this end, we introduce *RadDiagSeg-D*, a dataset spanning major radiology modalities. The complex task of *RadDiagSeg-D* is designed to improve the joint question answering and flexible segmentation abilities of medical VLMs. We further present *RadDiagSeg-M*, a radiological VLM capable of joint abnormality detection, diagnosis, and flexible multi-target segmentation. Given the novelty of *RadDiagSeg-D*, we additionally release a benchmarking tool to support standardized evaluation. Experiments demonstrate that our model achieves competitive performance on downstream tasks and SOTA performance on SLAKE. Furthermore, *RadDiagSeg-M* is the first model capable of tackling the full complex task of *RadDiagSeg-D*, setting benchmarks on text-based tasks

and establishing a strong baseline for the whole task.

References

Antonelli, M.; Reinke, A.; Bakas, S.; Farahani, K.; Kopp-Schneider, A.; Landman, B. A.; Litjens, G.; Menze, B.; Ronneberger, O.; Summers, R. M.; et al. 2022. The medical segmentation decathlon. *Nature communications*, 13(1): 4128.

Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; and Wang, M. 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In *ECCV*, 205–218. Springer.

Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A. L.; and Zhou, Y. 2021. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 24185–24198.

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

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Huang, X.; Shen, L.; Liu, J.; Shang, F.; Li, H.; Huang, H.; and Yang, Y. 2025. Towards a Multimodal Large Language Model with Pixel-Level Insight for Biomedicine. In *AAAI*, volume 39, 3779–3787.

Jaegle, A.; Borgeaud, S.; Alayrac, J.-B.; Doersch, C.; Ionescu, C.; Ding, D.; Koppula, S.; Zoran, D.; Brock, A.; Shelhamer, E.; Henaff, O. J.; Botvinick, M.; Zisserman, A.; Vinyals, O.; and Carreira, J. 2022. Perceiver IO: A General Architecture for Structured Inputs & Outputs. In *ICLR*.

Jha, D.; Smedsrud, P. H.; Riegler, M. A.; Johansen, D.; De Lange, T.; Halvorsen, P.; and Johansen, H. D. 2019. Resunet++: An advanced architecture for medical image segmentation. In *ISM*, 225–2255. IEEE.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *ICLR*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *ICCV*, 4015–4026.

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*, 9579–9589.

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

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *NeurIPS*, 36: 28541–28564.

Liu, B.; Zhan, L.-M.; Xu, L.; Ma, L.; Yang, Y.; and Wu, X.-M. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *ISBI*, 1650–1654. IEEE.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *NeurIPS*, 36: 34892–34916.

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

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 38–55. Springer.

Luo, L.; Tang, B.; Chen, X.; Han, R.; and Chen, T. 2025. VividMed: Vision Language Model with Versatile Visual Grounding for Medicine. In *NAACL*, 1800–1821.

Ma, J.; Chen, J.; Ng, M.; Huang, R.; Li, Y.; Li, C.; Yang, X.; and Martel, A. L. 2021. Loss odyssey in medical image segmentation. *Medical image analysis*, 71: 102035.

Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024. Segment anything in medical images. *Nature Communications*, 15(1): 654.

McKinzie, B.; Gan, Z.; Fauconnier, J.-P.; Dodge, S.; Zhang, B.; Dufter, P.; Shah, D.; Du, X.; Peng, F.; Belyi, A.; et al. 2024. Mm1: methods, analysis and insights from multimodal llm pre-training. In *ECCV*, 304–323. Springer.

Moor, M.; Huang, Q.; Wu, S.; Yasunaga, M.; Dalmia, Y.; Leskovec, J.; Zakka, C.; Reis, E. P.; and Rajpurkar, P. 2023. Med-flamingo: a multimodal medical few-shot learner. In *ML4H*, 353–367. PMLR.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*, 8748–8763. PmLR.

Rasley, J.; Rajbhandari, S.; Ruwase, O.; and He, Y. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD*, 3505–3506.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241. Springer.

Sellergren, A.; Kazemzadeh, S.; Jaroensri, T.; Kiraly, A.; Traverse, M.; Kohlberger, T.; Xu, S.; Jamil, F.; Hughes, C.; Lau, C.; et al. 2025. MedGemma Technical Report. *arXiv preprint arXiv:2507.05201*.

Steiner, A.; Pinto, A. S.; Tschannen, M.; Keysers, D.; Wang, X.; Bitton, Y.; Gritsenko, A.; Minderer, M.; Sherbondy, A.; Long, S.; et al. 2024. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*.

Tahir, A. M.; Chowdhury, M. E. H.; Qiblawey, Y.; Khandakar, A.; Rahman, T.; Kiranyaz, S.; Khurshid, U.; Ibtehaz, N.; Mahmud, S.; and Ezeddin, M. 2021. COVID-QU-Ex. https://doi.org/10.34740/kaggle/dsv/3122898. Kaggle Dataset.

Team, G.; Kamath, A.; Ferret, J.; Pathak, S.; Vieillard, N.; Merhej, R.; Perrin, S.; Matejovicova, T.; Ramé, A.; Rivière, M.; et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Tolstikhin, I. O.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A.; Keysers, D.; Uszkoreit, J.; et al. 2021. Mlp-mixer: An all-mlp architecture for vision. *NeurIPS*, 34: 24261–24272.

Tu, T.; Azizi, S.; Driess, D.; Schaekermann, M.; Amin, M.; Chang, P.-C.; Carroll, A.; Lau, C.; Tanno, R.; Ktena, I.; et al. 2024. Towards generalist biomedical AI. *Nejm Ai*, 1(3): AIoa2300138.

Wang, J.; and Ke, L. 2024. Llm-seg: Bridging image segmentation and large language model reasoning. In *CVPR*, 1765–1774.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, X.; Zhang, S.; Li, S.; Kallidromitis, K.; Li, K.; Kato, Y.; Kozuka, K.; and Darrell, T. 2024b. SegLLM: Multi-round Reasoning Segmentation. *arXiv preprint arXiv:2410.18923*.

Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Towards Generalist Foundation Model for Radiology. *CoRR*.

Wu, L.; Nie, Y.; He, S.; Zhuang, J.; Luo, L.; Mahboobani, N.; Vardhanabhuti, V.; Chan, R. C. K.; Peng, Y.; Rajpurkar, P.; et al. 2025. UniBiomed: A Universal Foundation Model for Grounded Biomedical Image Interpretation. *arXiv preprint arXiv:2504.21336*.

Yuan, H.; Li, X.; Zhang, T.; Huang, Z.; Xu, S.; Ji, S.; Tong, Y.; Qi, L.; Feng, J.; and Yang, M.-H. 2025. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*.

Zhang, S.; Xu, Y.; Usuyama, N.; Xu, H.; Bagga, J.; Tinn, R.; Preston, S.; Rao, R.; Wei, M.; Valluri, N.; Wong, C.; Tupini, A.; Wang, Y.; Mazzola, M.; Shukla, S.; Liden, L.; Gao, J.; Crabtree, A.; Piening, B.; Bifulco, C.; Lungren, M. P.; Naumann, T.; Wang, S.; and Poon, H. 2024. A Multimodal Biomedical Foundation Model Trained from Fifteen Million Image–Text Pairs. *NEJM AI*, 2(1).

Zhao, T.; Gu, Y.; Yang, J.; Usuyama, N.; Lee, H. H.; Kiblawi, S.; Naumann, T.; Gao, J.; Crabtree, A.; Abel, J.; et al. 2025. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature methods*, 22(1): 166–176.

Zou, X.; Yang, J.; Zhang, H.; Li, F.; Li, L.; Wang, J.; Wang, L.; Gao, J.; and Lee, Y. J. 2023. Segment everything everywhere all at once. *NeurIPS*, 36: 19769–19782.

Appendix

Dataset Details

We provide more details on the three tasks we adopted in the training stages: Ref-Seg, VQA, and VQA-Seg.

Ref-Seg Dataset.

Consisting of 199,829 samples, Ref-Seg includes diverse samples from three different radiological modalities. Table 6 presents the composition of the dataset.

VQA Dataset

VQA-RAD (Lau et al. 2018) consists of 1,790 samples in the training set and 451 samples in the test set. We adopt the English subset of SLAKE (Liu et al. 2021), which comprises 4,919 samples in the training set and 1,061 samples in the test set.

VQA-Seg Dataset

We processed two subsets of MSD (Antonelli et al. 2022; Zhao et al. 2025) and COVID-QU-EX (Tahir et al. 2021) into the *RadDiagSeg-D* dataset. The resulting dataset contains over 28,000 samples, with 22k used for training and approximately 6k for testing. Figure 6 illustrates the detailed composition of the *RadDiagSeg-D* training set.

Label imbalance is observed in both imaging modalities. In the X-ray subset, positive labels are more prevalent than negative ones, whereas in the CT subset, the opposite trend is present. This imbalance increases the difficulty of the task, as models cannot rely on overfitting to a dominant class to achieve strong performance. Accordingly, we account for this factor in the evaluation and report the F1 score to enable a fair comparison across methods.

Evaluation of *RadDiagSeg-D*

RadDiagSeg-D consists of three-step questions: a close-ended VQA detection question, an open-ended VQA diagnosis question, and a multi-target segmentation task. During the evaluation, we evaluate the answers following the steps.

Detection F1

Detection task evaluates on a binary basis, with positive (*yes*) and negative (*no*) findings. For the predicted results, we explicitly map the answer to the binary labels. The computation of the F1 score can be formalized as follows: Given a set of binary true labels $y_{\text{true}} \in \{0, 1\}^n$ and processed predicted labels $y_{\text{pred}} \in \{0, 1\}^n$, we compute the F1 score:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

where the binary F1 score is the harmonic mean of precision and recall:

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{Recall} = \frac{TP}{TP + FN}$$

	Dataset Name	Num Samples	Segmentation Target
X-ray	Chest Xray Masks and Labels Dataset	1,632	healthy lung, tuberculosis lung
	SIIM-ACR Pneumothorax Segmentation	2379	pneumothorax
	COVID-19 Radiography Database	39,014	COVID, lung, viral pneumonia, lung opacity
CT	MSD	27,699	liver, liver tumour, pancreas, pancreas tumour, lung tumour, colon cancer primaries, spleen
	amos22	108,704	abdominal organs: spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, pancreas, right adrenal gland, left adrenal gland, duodenum, bladder, prostate/uterus
	COVID-19 CT	1,187	lungs, covid-19 infections
	LIDR-IDRI	7,389	lung nodule
MRI	amos22	11,825	abdominal organs as above

Table 6: Overview of Ref-Seg Dataset. The majority of data points originate from the CT modality, followed by X-ray and MRI.

Figure 6: *RadDiagSeg-D* overview and label distribution. The label imbalance within the training set poses another challenge to the task.

Diagnosis F1

The second open-ended question, which pertains to diagnosis, presents a greater challenge. Our evaluation focuses on the correctness of the diagnostic answer, hence the overall correct text answer to the whole task. We consider a diagnosis correct if it meets one of the following two situations:

- Ground truth is negative; processed detection prediction is also negative.
- Ground truth is positive; processed detection prediction is positive, *and* diagnostic label is present in the predicted text.

Otherwise, an answer is considered wrong. For each instance i, we formalize the process as a binary correctness indicator:

$$c^{(i)} = \begin{cases} 1 & \text{if } y_{\text{pred}} = y_{\text{true}} \\ 0 & \text{otherwise} \end{cases}$$

The F1 score is then computed between the predicted correctness $\{c^{(i)}\}$ and a ground truth vector of ones, reflecting whether the model answers both the detection and diagnosis questions correctly.

Segmentation Dice

The last step of the task is the evaluation of segmentation masks. We report the mean Dice score of true positive valid predictions in the paper (see Table 3). The Dice score aims to evaluate the overlap between predicted segmentation and ground truth, with a range from 0 (no overlap) to 1 (perfect overlap). Mathematically, for a binary-class segmentation, the Dice score is defined as:

$$\text{Dice} = 2 \cdot \frac{|P \cap G|}{|P| + |G|}$$

where:

- P is the set of predicted foreground pixels,
- G is the set of ground truth foreground pixels,
- $|P \cap G|$ is the number of correctly predicted foreground pixels (i.e., true positives).

Amputated Evaluation

Amputated evaluation is employed when no meaningful results can be observed by directly applying the standardized evaluation procedure as introduced above. In such cases, we decompose the question into three evaluation steps: *Detection*, *Diagnosis*, and *Diagnosis + Segmentation*. For numbers reported in Table 3, LISA failed after the *Detection* task. UniBiomed struggled to follow instructions to answer *yes/no* in the *Detection* step and generated diagnostic text, thus requiring extra steps for meaningful data analysis.

Implementation Details

Details of Multimodal LM

Figure 7 illustrates the architecture of the Multimodal Language Model (Multimodal LM) and the flow of multimodal information. As the core component responsible for integrating diverse input modalities and making crucial decisions, a robust Multimodal LM serves as the foundation

of *RadDiagSeg-M*. While previous sections have detailed the image encoder and language model, we now provide a more focused discussion on the multimodal projector, which bridges vision and language modalities.

The multimodal projector performs two key functions: aligning the embedding dimensions and reducing the number of image tokens. Given an image input x_{image}, its encoded representation is denoted as $z'_{\text{image}} \in \mathbb{R}^{N_{\text{image}} \times D_{\text{image}}}$, where N is the number of tokens and D the embedding dimension. Similarly, the text input x_{text} yields an embedding $z_{\text{text}} \in \mathbb{R}^{N_{\text{text}} \times D_{\text{text}}}$ after the language model's embedding layer.

We denote the projection function as f_{proj} and the target number of image tokens as $\hat{N}_{\text{image}}$. The transformation of the multimodal projector is given by:

$$\hat{z}'_{\text{image}} = f_{\text{proj}}(z'_{\text{image}}) \in \mathbb{R}^{\hat{N}_{\text{image}} \times D_{\text{text}}} \qquad (5)$$

In our implementation, we draw inspiration from the architecture of MLP-Mixer (Tolstikhin et al. 2021), applying projection along both the token and embedding dimensions, with an intermediate transposition step. Specifically:

- **Embedding dimension alignment** (D): Following prior works (Liu et al. 2023; McKinzie et al. 2024), we employ a two-layer multilayer perceptron (MLP) to map visual features into the language embedding space.
- **Token reduction** (N): To condense image information, we adopt a Perceiver-style cross-attention module (Jaegle et al. 2022), using a fixed set of learnable queries to extract a compressed representation.

Training Parameters

Table 7 summarizes the complete training configuration employed in our experiments. For both the pre-training and fine-tuning stages, we validate on the test set every 100 steps and select the best-performing checkpoint based on validation performance as the final model. Specifically, the *Pre-training (PT)* variant corresponds to the final checkpoint at the end of training, while the *Fine-tuning (FT)* variant is taken from the checkpoint at step 1000.

Discussion on Ambiguity of Model Answers

Figure 4 presents the outputs of comparable models, highlighting the deteriorated language capabilities of UniBiomed. As noted in the UniBiomed paper (Wu et al. 2025), its performance improvements are partly attributed to the generation of segmentation masks conditioned on both textual output and user input. However, when the textual response is ambiguous or misleading, the reliability of the generated mask correspondingly degrades. In such cases, as exemplified in Figure 4, potential users may find it difficult to interpret the reference of the predicted mask, reducing its clinical utility. Similarly, the complete lack of informative text in LISA's output results in ambiguity regarding the segmentation target. Collectively, these examples substantiate our claim that complex multimodal tasks, such as those presented in *RadDiagSeg-D*, are essential for developing truly assistive VLMs for radiological diagnosis.
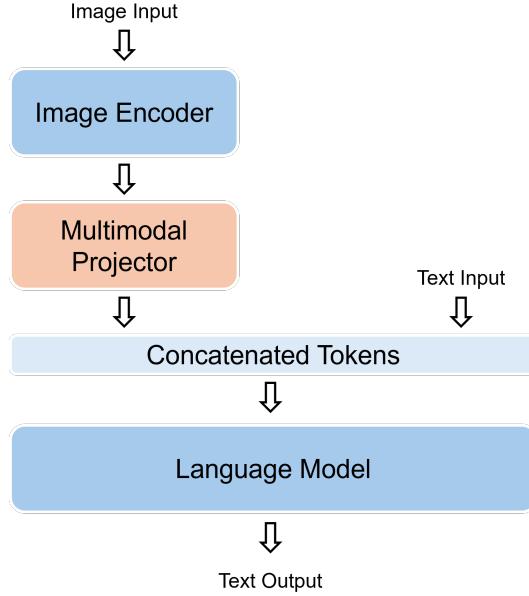
Figure 7: Structure of Multimodal LM. Multimodal projector aligns the visual and textual embedding space and reduces the number of image tokens.

	Pre-training	Fine-tuning
Scheduler	Warmup + Cosine	
Optimizer	AdamW(Kingma and Ba 2015)	
Loss	$\lambda_{\text{text}} = 1.0$ $\lambda_{\text{seg}} = 1.0$	
	$\lambda_{\text{dice}} = 0.5$ $\lambda_{\text{bce}} = 2.0$	
Num Trainable Params	625M	
epochs	3	5
learning rate	2e-4	2e-5
batch size	256	64
training time (hrs)	70	20

Table 7: Training specification in the pre-training and fine-tuning stages.

From the perspective of clinical assistance, UniBiomed's response in Figure 4 demonstrates that even if the segmentation mask appears accurate, it holds limited value for clinicians if the textual reference is unclear. In contrast, *RadDiagSeg-M* not only provides explicit labels for the predicted mask but also includes contextual information, such as an additional organ mask, offering greater potential for clinical support and interpretability.

More Qualitative Examples

We present more qualitative examples of Ref-Seg (Figure 8) and *RadDiagSeg-D* (Figure 9 for CT and Figure 10 for X-ray). We observe that *RadDiagSeg-M* generally provides accurate and context-aware segmentation masks in the Ref-Seg task. Examples from *RadDiagSeg-D* demonstrate further the *RadDiagSeg-M*'s capability in answering complex questions and providing multi-target segmentation. We also present examples where our model fails in both detection and diagnosis. These failure cases highlight the current limitations of the model and offer insights for future improvement.

Limitations

While we believe the introduction of *RadDiagSeg-D* and *RadDiagSeg-M* represents a significant step toward the development of assistive radiological VLMs that provide meaningful clinical support, we acknowledge certain limitations. In particular, *RadDiagSeg-D* is subject to label variability, primarily due to the limited availability of open-source datasets that provide both diagnostic text and multi-target segmentation masks. Additionally, there remains room for improvement in segmentation performance, especially for small or subtle anatomical targets. Addressing these challenges—particularly enhancing joint complex question-answering and fine-grained segmentation—constitutes a key direction for our future work.
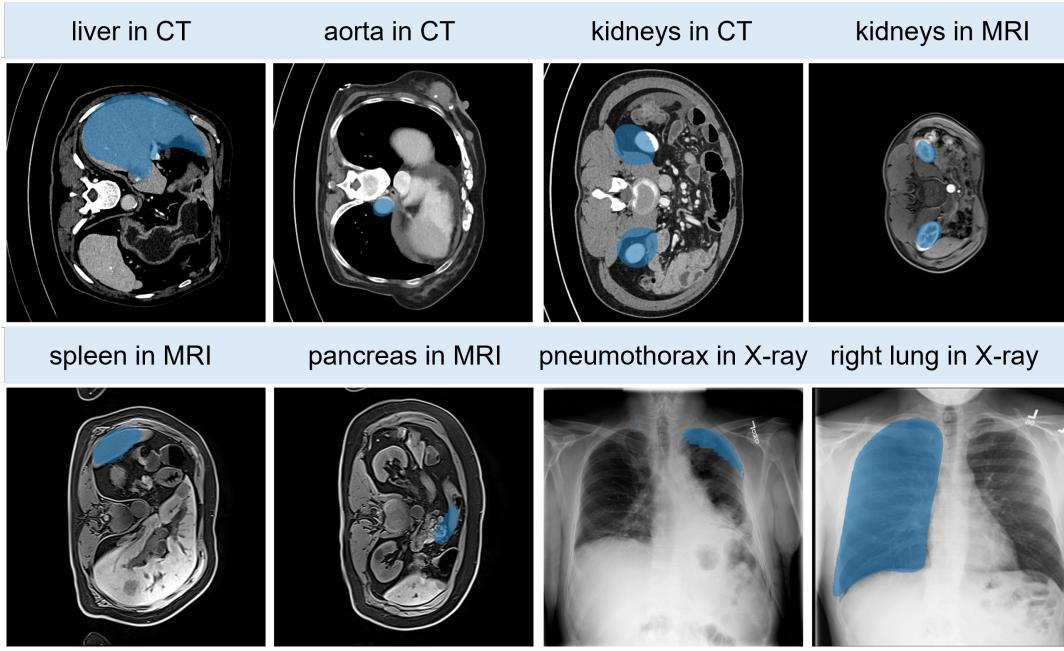
Figure 8: Qualitative Examples of Ref-Seg. *RadDiagSeg-M* provides accurate segmentation masks across radiological modalities: X-ray, CT, and MRI. However, the model struggles to accurately segment targets with irregular shapes, e.g. pancreas.

Figure 9: Qualitative Examples of *RadDiagSeg-D* in CT. Answers are from *RadDiagSeg-M* with the organ mask visualized. *green* marks correct textual answer, while *red* the wrong answer. Notably, if the model fails at the detection step, as in the right-bottom example, evaluation will automatically end.

Figure 10: Qualitative Examples of *RadDiagSeg-D* in X-ray. Answers are from *RadDiagSeg-M* with the organ mask visualized. *green* marks correct textual answer, while *red* the wrong answer.