From Reasoning to Pixels: Grounded Medical Multimodal LLMs for VQA and Segmentation

Haowen Gu[1,2], Gensheng Pei[3], Junzhu Mao[1,2], Qiong Wang[1,2], Mingwu Ren[1,2(✉)], and Yazhou Yao[1,2(✉)]

[1] Nanjing University of Science and Technology, Nanjing, China
[2] State Key Laboratory of Intelligent Manufacturing of Advanced Construction Machinery, Nanjing, China
[3] Department of Electrical and Computer Engineering, Sungkyunkwan University, Suwon, Korea
yazhou.yao@njust.edu.cn, renmingwu@mail.njust.edu.cn
https://github.com/NUST-Machine-Intelligence-Laboratory/MedREAL

Abstract. Although Multimodal Large Language Models (MLLMs) have demonstrated impressive performance in Medical Visual Question Answering (Med-VQA), their reliance on global image features often lacks precise pixel-level grounding, thereby limiting clinical trustworthiness. To bridge the semantic gap between high-level clinical reasoning and spatial localization, we propose MedREAL (**Med**ical **RE**asoning-driven **A**nswering and **L**ocalization), a unified framework that seamlessly aligns linguistic reasoning with spatial grounding. Specifically, MedREAL introduces **S**eg **A**nchored **R**easoning **P**ooling (SARP) to distill task-relevant semantic evidence directly from `[SEG]` tokens within the MLLM's hidden states. Furthermore, a **R**easoning-to-**V**isual (R2V) fusion mechanism is proposed to effectively inject these reasoning-aware features into a segmentation pipeline for accurate mask decoding. To facilitate this paradigm, we construct MedRAVS-13K, a comprehensive dataset comprising 13,824 expertly validated samples across four diverse imaging modalities. Extensive experiments demonstrate that MedREAL significantly outperforms state-of-the-arts, achieving 68.49% gIoU and 70.47% cIoU on benchmark evaluations. By generating evidence masks that are strictly consistent with textual diagnoses, MedREAL provides a robust, interpretable framework for reasoning-driven medical image analysis.

Keywords: Medical Visual Question Answering · Multimodal Large Language Models · Visual Grounding · Reasoning-driven Segmentation

1 Introduction

Medical visual question answering (Med-VQA) aims to answer clinically relevant questions about medical images, requiring models to jointly understand visual content and perform domain-specific reasoning. Beyond answer prediction, clinical practice often demands explicit localization of the visual evidence

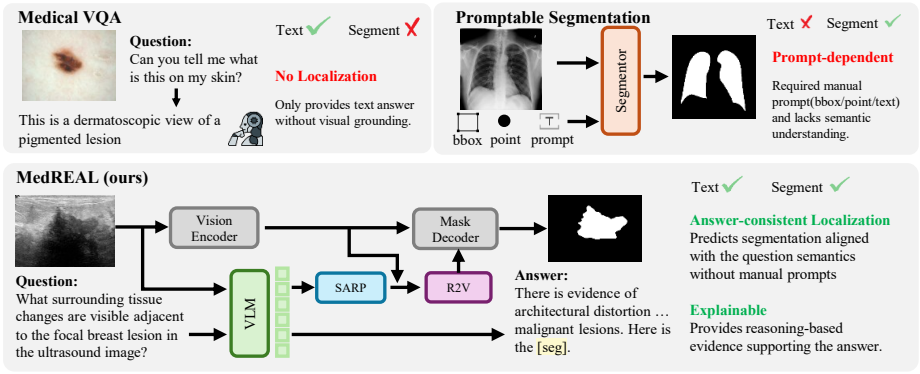

Fig. 1: Comparison between Medical VQA, promptable segmentation, and the proposed MedREAL framework. MedREAL introduces an explicit evidence token to extract reasoning-aligned features, which are fused with visual context to produce answer-consistent evidence localization.

that supports a decision, such as identifying a lesion, anatomical structure, or pathological region. This requirement highlights a critical limitation of existing Med-VQA systems: while recent multimodal large language models (MLLMs) [1, 3, 13, 34, 35, 51] demonstrate strong reasoning capabilities, they typically provide answers without grounding them in pixel-level evidence, thereby reducing interpretability and limiting clinical trust.

Most medical image segmentation methods [6, 14, 38, 39, 52, 61] demonstrate strong capability in spatial localization, yet remain limited in high-level semantic understanding. Although recent promptable foundation models such as SAM3 [10] and its medical variant MedSAM3 [32] enable concept segmentation from text-conditioned inputs, their semantic modeling remains shallow and largely relies on lexical alignment rather than complex clinical reasoning. As a result, these approaches struggle to capture the multi-step diagnostic logic required in real-world scenarios, where segmentation should be guided by question-driven semantic evidence rather than generic visual prompts. This creates a semantic disconnect between high-level diagnostic logic and pixel-level spatial grounding. Overcoming this challenge is essential for building integrated systems capable of joint medical answering and localization, ensuring that every diagnostic response is backed by spatially consistent visual justification.

Simply concatenating MLLMs with segmentation models, however, is suboptimal. Although MLLM hidden states contain rich multimodal information, they are often dominated by global context and broad linguistic patterns, making it difficult to isolate the specific visual features that serve as reasoning evidence. Furthermore, directly feeding text embeddings into a segmentation model leads to poor spatial alignment, as the connection between token-level reasoning and pixel-level localization is not explicitly modeled. As illustrated in Figure 1, an effective solution must bridge this gap by explicitly identifying reasoning-related

evidence within MLLM representations and transforming it into spatially meaningful signals to guide the segmentation process.

To address these challenges, we propose MedREAL (**Med**ical **RE**asoning-driven **A**nswering and **L**ocalization), a unified framework that aligns linguistic reasoning with pixel-level localization within a reasoning-aware latent space. Specifically, we embed an explicit evidence token, `[SEG]`, into the Med-VQA generation sequence, utilizing its hidden representation as a reasoning anchor to distill the critical semantic evidence that justifies the predicted answer. Building upon this design, we introduce the **S**eg-**A**nchored **R**easoning **P**ooling (SARP) module. SARP leverages the `[SEG]` representation as a guiding prior to distill reasoning-relevant features from the MLLM's terminal layers. By effectively suppressing irrelevant global semantics, SARP yields a compact, evidence-aligned representation tailored to the target region. Furthermore, we propose a **R**easoning-to-**V**isual (R2V) Fusion mechanism, which integrates these distilled reasoning features with the global visual context extracted from the segmentor's image encoder. This operation produces a semantically informed conditioning vector that robustly guides the SAM-based mask decoding stage. With this unified architecture, segmentation is driven not only by visual appearance but also by the model's internal diagnostic reasoning, allowing MedREAL to achieve spatially precise and answer-consistent evidence localization.

Another fundamental challenge in this domain is the scarcity of benchmarks that simultaneously evaluate clinical reasoning and spatial grounding. Current Med-VQA datasets [18, 28, 33] predominantly prioritize textual accuracy without pixel-level supervision; conversely, existing medical segmentation datasets [2,5,12,22,44,48] typically lack reasoning-oriented queries. To enable the training and evaluation of reasoning-driven segmentation, we construct MedRAVS-13K (Medical Reasoning, Answering, and Visual Segmentation), a reasoning-aware benchmark comprising 13,824 expertly curated samples. We augment four heterogeneous medical imaging datasets (*i.e.*, BUSI [2], COVID-QU-Ex [48], ISIC-2018 [12], and Kvasir-SEG [22]), spanning ultrasound, X-ray, dermoscopy, and endoscopy modalities. We enrich these datasets with clinically oriented question-answer pairs that explicitly refer to specific regions of interest. Each QA pair is matched with a verified pixel-level segmentation mask, forming perfectly aligned image-text-mask triplets. This provides the structured supervision necessary to link high-level diagnostic reasoning with precise spatial localization.

Extensive experiments on the MedRAVS-13K dataset demonstrate that our MedREAL significantly enhances segmentation quality. Compared to existing representative methods, MedREAL achieves state-of-the-art performance with an overall gIoU of 68.49% and cIoU of 70.47%. Crucially, by aligning linguistic reasoning with pixel-level grounding, our method ensures that the generated evidence masks are semantically consistent with the predicted clinical answers. This synergy effectively mitigates the "black-box" nature of traditional MLLMs, offering substantially improved interpretability and reliability for clinical decision support systems. In summary, our primary contributions are fourfold:

- We propose MedREAL, a unified framework that couples Med-VQA reasoning with pixel-level grounding by embedding explicit `[SEG]` tokens as reasoning anchors to distill semantic evidence from MLLM hidden states.
- We introduce the SARP module and R2V fusion mechanism to effectively extract reasoning-relevant features and systematically inject them into a SAM-based pipeline for precise mask decoding.
- We construct MedRAVS-13K, a comprehensive benchmark dataset of 13,824 samples across four diverse imaging modalities, providing a robust foundation of clinical QA pairs and verified pixel-level masks.
- Extensive evaluations show that MedREAL significantly outperforms existing methods, achieving **68.49% gIoU** and **70.47% cIoU**, while ensuring high semantic consistency between linguistic reasoning and spatial evidence.

2 Related Work

Medical Visual Question Answering. Medical visual question answering aims to empower models to resolve clinically relevant queries by synthesizing visual content with domain-specific medical knowledge. While early efforts primarily focused on fusion strategies between static image features and linguistic encodings [16,17,54,62], the field has recently been transformed by MLLMs. These advanced architectures [11,27,29,47,49] leverage massive pretraining to demonstrate sophisticated cross-modal reasoning, enabling them to interpret complex clinical descriptions and generate nuanced natural language responses. However, a critical gap remains as these models predominantly operate on global semantic representations. Despite their linguistic fluency, they typically fail to provide explicit pixel-level grounding for their predictions. This absence of spatial evidence makes it difficult to verify the decision rationale against specific pathological regions, thereby limiting their interpretability and clinical trustworthiness in high-stakes diagnostic environments. To overcome this, MedREAL embeds an `[SEG]` token as a reasoning anchor to distill local semantic evidence, successfully bridging the gap between linguistic answers and precise spatial grounding.

Medical Image Segmentation. Medical image segmentation is a fundamental task in medical image analysis, aiming to accurately delineate lesions or anatomical structures to support diagnosis and treatment planning. Traditional methods [9,20,41,46,53,59] rely on U-Net [45] and its variants, which perform pixel-wise classification through fully convolutional networks. More recently, promptable segmentation approaches, such as SAM [39], have demonstrated strong generalization across multiple modalities and large-scale datasets, enabling segmentation guided by points, bounding boxes, or textual prompts. However, these methods [23, 25, 52, 60, 63] typically rely solely on visual information and lack language-based reasoning constraints, making it challenging to perform question-driven segmentation. In clinical diagnostic workflows, it is essential for models to provide localized evidence that directly corresponds to a specific query; therefore, the effective integration of linguistic reasoning with spatial segmentation remains an open and critical challenge. Unlike conventional visual-centric models, our MedREAL framework explicitly injects MLLM-derived clinical logic

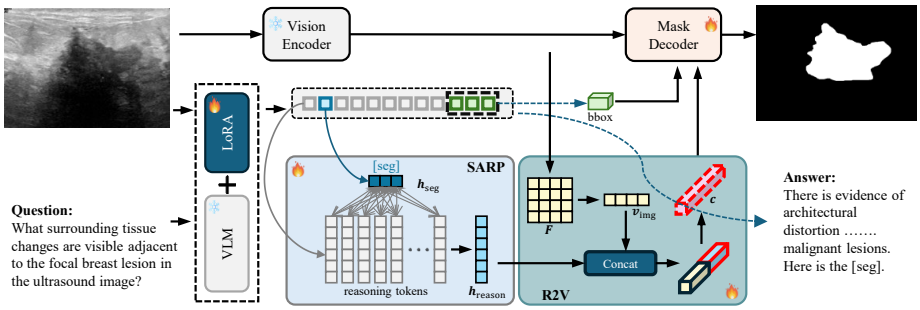

Fig. 2: Overview of MedREAL. A medical image and question are processed by a VLM producing answer tokens including [SEG]. SARP distills reasoning-relevant features from [SEG], which are fused with global visual features via R2V Fusion to condition the mask decoder, producing an evidence-aligned segmentation mask guided by both reasoning and visual context.

into the segmentation pipeline via the R2V mechanism, ensuring masks are directly driven by diagnostic reasoning.

Reasoning Guided Segmentation. Reasoning-guided segmentation aims to incorporate high-level semantic reasoning into pixel-level predictions. Existing studies [7, 8, 15, 19, 26, 36, 43, 50, 55–58] have largely focused on natural scenes, predominantly relying on a single `[seg]` token to represent the entire target semantic mask. While this design enables coarse alignment between language and the predicted mask, it represents the target as a holistic concept and overlooks fine-grained evidence that is critical for precise localization. In medical imaging, this line of research is still in its early stages. Some approaches [21, 32] attempt to fuse attention maps or intermediate features from question-answering models with segmentation networks to enhance semantic consistency, but they typically rely on global language representations or simple feature concatenation, lacking explicit anchoring mechanisms to localize regions relevant to the question. Moreover, the scarcity of high-quality reasoning-annotated datasets for medical tasks makes joint training and evaluation challenging. MedREAL uniquely addresses these bottlenecks by utilizing the SARP module to extract localized, fine-grained semantic evidence rather than holistic representations, supported by our comprehensive MedRAVS-13K benchmark for rigorous evaluation.

3 Method

3.1 Overall Pipeline

We propose MedREAL, a unified framework that couples medical visual question answering with pixel-level segmentation by explicitly injecting multimodal reasoning semantics into the localization process. The overall architecture is illustrated in Figure 2.

Unlike standard pipeline models that treat text generation and segmentation as isolated steps, MEDREAL unifies them within a reasoning-aware latent space. Given a medical image $\boldsymbol{I}$ and a clinical text query $\boldsymbol{Q}$, the multimodal large language model (MLLM) autoregressively generates a textual response $\boldsymbol{A}$ alongside a sequence of latent hidden states $\boldsymbol{H}$. To explicitly bridge language and vision, we prompt the MLLM to output a dedicated `[SEG]` token when localization is required. The language generation process can be formulated as:

$$\boldsymbol{A}, \boldsymbol{H} = \mathcal{F}_{\text{MLLM}}(\boldsymbol{I}, \boldsymbol{Q}) \tag{1}$$

where $\boldsymbol{H} \in \mathbb{R}^{B \times L \times D}$ encapsulates the dense semantic representations of the generated reasoning process.

Rather than relying solely on generic visual prompts, MEDREAL leverages this textual reasoning to guide the spatial decoding. We introduce the Seg-Anchored Reasoning Pooling (**SARP**) module to extract a compact reasoning representation $\boldsymbol{h}_{\text{reason}}$ from $\boldsymbol{H}$. Subsequently, the Reasoning-to-Visual (**R2V**) Fusion module dynamically integrates $\boldsymbol{h}_{\text{reason}}$ with the visual feature maps $\boldsymbol{F}$ extracted by the segmentor's image encoder. This yields a semantically informed prompt $\boldsymbol{c}$, which is fed into the mask decoder $\mathcal{D}_{\text{mask}}$ to predict the final evidence mask $\boldsymbol{M}$:

$$\boldsymbol{M} = \mathcal{D}_{\text{mask}}(\boldsymbol{F}, \boldsymbol{c}). \tag{2}$$

By formulating the pipeline in this cohesive manner, the segmentation mask is strictly conditioned on the diagnostic logic formulated by the MLLM, ensuring semantic alignment between the textual answer and the localized visual evidence.

3.2 Seg-Anchored Reasoning Pooling

While the `[SEG]` token acts as a trigger for the segmentation task, extracting features exclusively from its corresponding hidden state is suboptimal. Clinical reasoning is inherently distributed across the preceding linguistic sequence; a single token lacks the capacity to encapsulate complex diagnostic multi-step logic. To address this, we propose the SARP module, which utilizes the `[SEG]` token as an attention anchor to aggregate reasoning-relevant semantics from the broader context.

Let $\boldsymbol{H} \in \mathbb{R}^{B \times L \times D}$ denote the terminal-layer hidden states of the MLLM. We first isolate the representation of the `[SEG]` token using a binary positional mask $\boldsymbol{m}^s \in \{0,1\}^L$. To ensure robustness against potential sequence shifts, the segmentation anchor $\boldsymbol{h}_{\text{seg}} \in \mathbb{R}^D$ is computed via masked average pooling:

$$\boldsymbol{h}_{\text{seg}} = \frac{\sum_{i=1}^{L} m_i^s \boldsymbol{H}_i}{\sum_{i=1}^{L} m_i^s + \epsilon}, \tag{3}$$

where ϵ is a small constant to prevent zero division. This anchor explicitly captures the model's semantic state at the exact moment spatial grounding is invoked. To adaptively filter out irrelevant linguistic noise (*e.g.*, grammatical structures or generic clinical boilerplate) and concentrate on specific diagnostic

evidence, we project $\boldsymbol{h}_{\text{seg}}$ into a query vector $\boldsymbol{q}$. Concurrently, the preceding reasoning sequence within $\boldsymbol{H}$ is linearly mapped into key $\boldsymbol{K}$ and value $\boldsymbol{V}$ matrices. An attention-based pooling mechanism is then applied to measure the relevance of each reasoning token to the segmentation trigger:

$$\boldsymbol{\alpha} = \text{Softmax}\left(\frac{\boldsymbol{q}\boldsymbol{K}^\top}{\sqrt{d_k}}\right), \tag{4}$$

where d_k is the scaling factor based on the hidden dimension. The final reasoning representation $\boldsymbol{h}_{\text{reason}}$ is derived through a weighted summation:

$$\boldsymbol{h}_{\text{reason}} = \boldsymbol{\alpha}\boldsymbol{V}. \tag{5}$$

This operation distills a compact, high-density semantic vector that explicitly encodes the clinical rationale necessitating the segmentation, providing a highly informative prior for the subsequent localization phase.

3.3 Reasoning-to-Visual Fusion

Although $\boldsymbol{h}_{\text{reason}}$ is semantically rich, it exists entirely within the linguistic latent space and lacks the spatial inductive biases required for precise pixel-level mask generation. To translate abstract diagnostic logic into an explicit spatial guidance signal, we introduce the R2V Fusion module, which bridges the extracted reasoning with global anatomical context.

Given the multi-scale visual feature map $\boldsymbol{F} \in \mathbb{R}^{B\times C\times H'\times W'}$ from the segmentor's image encoder, we first apply global average pooling to collapse the spatial dimensions. A linear projection followed by a non-linear activation σ maps this representation into a global visual context vector $\boldsymbol{v}_{\text{img}} \in \mathbb{R}^{B\times D_c}$:

$$\boldsymbol{v}_{\text{img}} = \sigma\Big(\text{Linear}\big(\text{AvgPool}(\boldsymbol{F})\big)\Big). \tag{6}$$

This vector provides a structural anatomical prior. To enable deep multimodal interaction, we concatenate $\boldsymbol{v}_{\text{img}}$ with the reasoning representation $\boldsymbol{h}_{\text{reason}}$ along the channel dimension. The fused tensor is then projected through a two-layer multi-layer perceptron (MLP) parameterized by ϕ:

$$\boldsymbol{c} = \phi\big([\boldsymbol{v}_{\text{img}}, \boldsymbol{h}_{\text{reason}}]\big), \tag{7}$$

where $[\cdot,\cdot]$ denotes the concatenation operation. The resulting conditional vector $\boldsymbol{c}$ fundamentally alters the segmentation paradigm: rather than relying on geometric prompts (*e.g.*, points or boxes), $\boldsymbol{c}$ acts as a semantically-informed pseudo-prompt for the SAM-based mask decoder. This ensures the decoded mask strictly adheres to the diagnostic evidence articulated by the MLLM.

3.4 Optimization Objective

The entire MedREAL framework is trained end-to-end, jointly optimizing both the text generation quality and the spatial segmentation accuracy. The total loss

$\mathcal{L}$ is formulated as a weighted combination:

$$\mathcal{L} = \lambda_{\text{txt}}\mathcal{L}_{\text{CE}}(\hat{\boldsymbol{Y}}_{\text{txt}}, \boldsymbol{Y}_{\text{txt}}) + \lambda_{\text{mask}}\Big[\lambda_{\text{bce}}\mathcal{L}_{\text{BCE}}(\hat{\boldsymbol{M}}, \boldsymbol{M}) + \lambda_{\text{dice}}\mathcal{L}_{\text{DICE}}(\hat{\boldsymbol{M}}, \boldsymbol{M})\Big]. \quad (8)$$

Here, $\mathcal{L}_{\text{CE}}$ represents the standard auto-regressive cross-entropy loss applied to the textual response. For the spatial localization, the mask loss is computed as a weighted sum of the per-pixel Binary Cross-Entropy ($\mathcal{L}_{\text{BCE}}$) and the Dice loss ($\mathcal{L}_{\text{DICE}}$), which handles class imbalance in medical lesions. This composite objective effectively bridges the MLLM and the visual segmentor, forcing the shared latent space to align linguistic logic with pixel-level boundaries.

4 Dataset

4.1 Data Source

As summarized in Figure 3, we construct **MedRAVS-13K** (Medical Reasoning, Answering, and Visual Segmentation) to rigorously evaluate the MEDREAL framework. We integrate four diverse publicly available sources: BUSI [2], COVID-QU-Ex [48], ISIC-2018 [12], and Kvasir-SEG [22]. Unlike conventional benchmarks that treat text and masks in isolation, MedRAVS-13K explicitly aligns fine-grained clinical question-answering with precise visual segmentation masks. This design enables a dual assessment of sophisticated diagnostic reasoning and localized evidence generation.

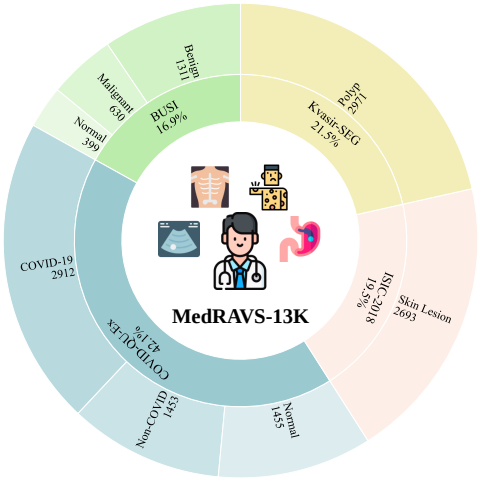

Fig. 3: Overview of MedRAVS-13K.

The dataset spans multiple modalities, *e.g.*, ultrasound, X-ray, dermoscopy, and endoscopy, covering critical organs such as the breast, lung, skin, and colon. These sources exhibit substantial heterogeneity in imaging physics and lesion morphology, ranging from low-contrast sonographic boundaries to intricate endoscopic mucosal textures. Such multi-modality composition provides a robust and challenging testbed for evaluating reasoning-driven segmentation across clinically distinct scenarios.

4.2 Data Generation and Curation

Figure 4 details the multi-stage generation and curation pipeline designed to align clinical reasoning with pixel-level annotations. For each source dataset, we

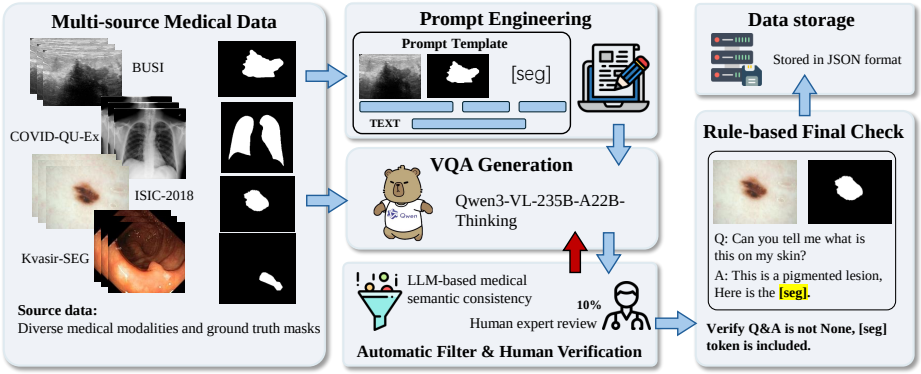

Fig. 4: Dataset Generation Pipeline of MedRAVS-13K. The process illustrates the integration of raw images and masks to generate reasoning-aware text, followed by rigorous quality assurance.

develop modality- and organ-specific prompts that incorporate the raw medical images alongside their ground-truth masks. We leverage the advanced reasoning capabilities of Qwen3-VL-235B-A22B-Thinking to generate clinically accurate question-answering (QA) pairs. Crucially, a `[SEG]` token is explicitly embedded within each generated sequence to denote the region of interest. This mechanism ensures that the textual rationale is strictly grounded in the annotated anatomy or lesion, compelling the model to bridge abstract diagnostic logic with concrete visual evidence.

To guarantee data fidelity, we implement a two-tier quality assurance protocol. First, an automated filtering stage employs a powerful large language model to evaluate the medical plausibility, semantic consistency, and visual relevance of each generated pair, discarding low-quality outputs. Second, we perform meticulous manual verification on a 10% subset of the data to confirm clinical accuracy and verify the spatial alignment between the textual queries and the segmentation masks. To mitigate potential biases arising from class and modality imbalances, we apply a controlled sampling strategy across the pathological categories during the final split construction, ensuring a representative benchmark. Ultimately, MedRAVS-13K comprises over 13,000 expertly curated samples. Each instance seamlessly integrates a medical image, a `[SEG]`-augmented QA pair, and a corresponding pixel-level mask. These queries encompass diverse diagnostic intents, including polar (yes/no), descriptive, and localization-focused questions. Because the masks exhibit substantial variation in scale and morphology across the four modalities, the dataset presents highly challenging visual grounding scenarios that demand the joint optimization of reasoning, segmentation, and answer prediction.

Table 1: Referring phrases used for different datasets and semantic classes. These ground-truth derived templates serve as explicit prompts for non-reasoning baselines.

Dataset	Class	Train Samples	Val Samples	Referring Phrase
BUSI [2]	Benign	915	396	benign breast mass
	Malignant	441	189	malignant breast tumor
	Normal	276	123	normal breast tissue
COVID-QU-Ex [48]	COVID-19	2330	582	COVID-19 lung infection
	Non-COVID	1162	291	non-COVID lung infection
	Normal	1164	291	normal lung region
ISIC-2018 [12]	Skin Lesion	2593	100	skin lesion region
Kvasir-SEG [22]	Polyp	2615	356	polyp region

5 Experiments

5.1 Experiment Setup

Datasets. To evaluate the performance of MedREAL, all experiments are conducted on the MedRAVS-13K dataset as described in Sec. 4. For comparative analysis involving baseline models that lack inherent reasoning capabilities, we utilize standard category labels to implement the segmentation task. Detailed specifications regarding the prompting mechanisms for various architectures are provided in the subsequent sections.

Evaluation Metrics. Following established protocols in reasoning segmentation, we adopt gIoU and cIoU as primary metrics. gIoU is defined as the arithmetic mean of per image IoU scores, while cIoU represents the ratio of total intersection to total union across the dataset. Since cIoU is heavily biased toward large area objects and prone to instability, gIoU is preferred in medical contexts where lesion scales vary significantly, as it provides a more balanced measure of segmentation quality.

Implementation Details. Our model is trained on two NVIDIA 48G A6000 GPUs using the DeepSpeed [42] engine to enable memory-efficient optimization and stable large scale multimodal training. We adopt the AdamW [37] optimizer with an initial learning rate of 2×10^{-4} and apply weight decay for regularization. Unless otherwise specified, we adopt Qwen3-VL-2B-Instruct [4] as the MLLM and SAM [24] as the segmentor for all experiments. Following the optimization strategy in LISA [26], we set the loss weights λ_{txt} and λ_{mask} to 1.0. For the mask supervision components, λ_{bce} and λ_{dice} are assigned values of 2.0 and 0.5.

5.2 Comparison with State-of-the-Art Methods

To evaluate MedREAL, we benchmark it against three distinct paradigms: referring-based segmentation (*i.e.*, OVSeg [30], SAM3 [10], MedSAM3 [32]), agent-based iterative systems (*i.e.*, SAM3-Agent [10], MedSAM3-Agent [32]), and end-to-end reasoning segmentation (*i.e.*, LISA [26]).

Table 2: Comparison of gIoU and cIoU across different datasets. Best results are highlighted in **bold**, and second-best are underlined.

Method	BUSI		COVID-QU-Ex		ISIC-2018		Kvasir-SEG		Overall	
	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	gIoU	cIoU
Refering-based Segmentation										
OVSeg [30]	4.72	4.79	17.59	16.75	26.19	24.56	13.88	13.58	13.48	18.75
SAM3 [10]	17.69	0.68	0.00	0.00	1.04	1.78	0.00	0.00	5.43	1.39
MedSAM3 [32]	62.93	58.97	**71.56**	**75.03**	**83.03**	80.43	**81.82**	**82.02**	**70.99**	**78.28**
Agent-based Segmentation										
SAM3-Agent [10]	41.61	19.67	5.24	8.12	7.17	0.92	4.08	3.76	16.21	4.20
MedSAM3-Agent [32]	58.29	55.10	46.88	46.23	82.65	**80.88**	71.89	62.84	55.71	74.33
Reasoning-based Segmentation										
LISA [26]	41.76	48.84	62.30	63.88	71.15	56.06	57.50	56.30	55.70	56.05
MedREAL (**Ours**)	**69.15**	**62.52**	64.51	66.92	82.21	71.74	76.35	69.73	68.49	70.47

Table 3: Comparison of text generation performance between LISA and our method.

Method	BLEU					ROUGE		
	BLEU	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
LISA [26]	24.04	56.94	29.69	18.88	13.15	59.81	33.41	48.11
MedREAL (**Ours**)	**89.12**	**98.75**	**97.74**	**97.00**	**96.26**	**96.69**	**95.84**	**96.69**

Because referring-based purely structural models lack intrinsic reasoning capabilities, they cannot process the complex VQA queries in MedRAVS-13K directly. To accommodate them and establish an upper-bound for structural decoding, we provide explicit, oracle-like text prompts derived from the ground-truth category of each query, with detailed templates summarized in Table 1. For the agent-based approaches, we deploy Qwen3-VL-8B-Thinking as the reasoning engine to autonomously generate prompts for the segmentors (capped at 20 rounds). Finally, LISA is evaluated as the primary reasoning-segmentation baseline, utilizing LLaVA-7B [29] as its multimodal backbone.

Quantitative Segmentation Results. Table 2 presents the quantitative localization performance. When compared to the direct end-to-end reasoning baseline (LISA), MedREAL demonstrates overwhelming superiority, improving the overall gIoU from 55.70% to 68.49%. This substantial margin validates that our SARP and R2V modules extract far more precise spatial priors than LISA's holistic `[SEG]` embedding approach. This advantage is particularly evident in high-noise environments like the BUSI ultrasound dataset, where MedREAL achieves 69.15% gIoU compared to LISA's 41.76%, proving our method's robustness against low-contrast boundaries and speckle noise.

It is crucial to correctly contextualize the performance of MedSAM3. While it reaches high metrics (e.g., 70.99% overall gIoU), this purely structural model benefits heavily from explicit ground-truth category prompts (as defined in Table 1) and extensive pretraining on the original source datasets. However, this

label-driven paradigm fails entirely in real-world diagnostic workflows that require multi-step logic rather than explicit naming. This vulnerability is starkly exposed when evaluating MedSAM3-Agent: once forced to autonomously deduce the prompt from complex VQA queries rather than relying on oracle labels, its overall gIoU drastically collapses to 55.71%. The performance degradation is exceptionally severe on the COVID-QU-Ex dataset (dropping from 71.56% to 46.88%), as the decoupled reasoning engine struggles to comprehend complex pathological descriptions, generating inaccurate textual prompts that mislead the segmentation backbone. In contrast, MedREAL operates entirely on natural language questions without predefined category hints or oracle labels, yet achieves highly competitive performance (68.49% overall gIoU). By seamlessly bridging the gap between abstract diagnostic reasoning and spatial grounding within a unified latent space, MedREAL establishes a new state-of-the-art for reasoning-driven medical architectures.

Text Generation Quality. Beyond spatial localization, clinically integrated systems must maintain high diagnostic articulation. Table 3 compares the textual VQA generation quality of MedREAL against LISA. Our framework achieves near-perfect BLEU [40] and ROUGE [31] scores, outperforming the baseline by massive margins. These results convincingly demonstrate that extracting reasoning semantics for visual guidance does not compromise the MLLM's inherent linguistic capabilities. Instead, our unified architecture successfully harmonizes high-fidelity text generation with precise pixel-level evidence grounding, offering a complete and trustworthy diagnostic response.

5.3 Ablation Studies

To rigorously investigate the individual contributions of Box Guidance, the Seg-Anchored Reasoning Pooling (SARP) module, and the Reasoning-to-Visual (R2V) fusion mechanism, we conduct comprehensive ablation experiments. To ensure fair comparison and preserve parameter parity, ablated modules are substituted with equivalent linear projections rather than simply being bypassed. The quantitative impacts across gIoU and cIoU metrics are detailed in Table 4.

The Necessity of Spatial Priors. Removing all three key components degrades performance to a baseline of 11.35% gIoU, reflecting the inherent difficulty of zero-shot spatial grounding without explicit localization priors. Reintroducing only Box Guidance provides a sharp increase to 45.02% gIoU. While coarse spatial anchors offer strong

Table 4: Ablation study isolating the contributions of bounding-box guidance, SARP, and R2V modules.

Box Guide	SARP	R2V	gIoU (%)	cIoU (%)
✗	✗	✗	11.35	19.69
✓	✗	✗	45.02	52.84
✗	✓	✓	36.24	37.74
✓	✗	✓	23.22	47.12
✓	✓	✗	45.50	46.96
✓	✓	✓	**68.49**	**70.47**

regional cues, this configuration still lacks the fine-grained semantic alignment required to delineate precise lesion boundaries.

The Interdependence of SARP and R2V. The interplay between reasoning extraction and visual fusion proves critical. When SARP is active but R2V is replaced by a linear mapping, the model achieves 45.50% gIoU, comparable to the Box Guidance baseline, yet struggles with cIoU. This reveals that while SARP successfully distills concentrated evidence from the reasoning sequence, these semantic features remain bottlenecked without a dedicated fusion mechanism to project them into the visual space. Conversely, enabling R2V without SARP yields a detrimental 23.22% gIoU; forcefully injecting unrefined, global linguistic representations directly into the visual pipeline introduces severe semantic noise, actively disrupting spatial localization.

Synergistic Integration. As shown in Table 4, enabling both SARP and R2V without Box Guidance yields 36.24% gIoU, demonstrating that reasoning-driven conditioning can partially compensate for missing spatial priors, but remains insufficient on its own. The full MedREAL architecture seamlessly integrates all three components to establish the upper bound (68.49% gIoU / 70.47% cIoU). This confirms a highly complementary relationship: Box Guidance anchors the general region, SARP distills the exact diagnostic rationale, and R2V effectively translates this rationale into actionable spatial conditioning.

Scaling Behavior. Finally, we explore the scaling behavior of the underlying multimodal foundation model. As reported in Table 5, upgrading the backbone from Qwen3-VL-2B-Instruct to the 4B-Instruct variant, while freezing the rest of the MedREAL architecture, triggers a substantial gIoU improvement from 68.49% to 75.96%. This empirically validates that our framework efficiently leverages the enhanced reasoning capabilities of larger MLLMs, dynamically translating superior linguistic logic into higher-quality evidence masks.

Table 5: Effect of MLLM scaling on reasoning-driven segmentation performance. Upgrading the backbone directly translates to superior localization.

Backbone MLLM	gIoU (%)	cIoU (%)
Qwen3-VL-2B-Instruct	68.49	**70.47**
Qwen3-VL-4B-Instruct	**75.96**	70.26

5.4 Qualitative Analysis

Robustness across Modalities. Figure 5 evaluates segmentation across four challenging clinical scenarios. General-domain models (*e.g.*, OVSeg [30], SAM3 [10]) fail on low-contrast ultrasound (Row 1) and diffuse X-ray infections (Row 2) due to limited domain knowledge. While MedSAM3 [32] improves boundary prediction through domain pretraining, its decoupled reasoning makes it sensitive to visual noise, such as endoscopic blood artifacts (Row 4). LISA [26] narrows the semantic gap but relies on a single, global `[seg]` token. This design limits

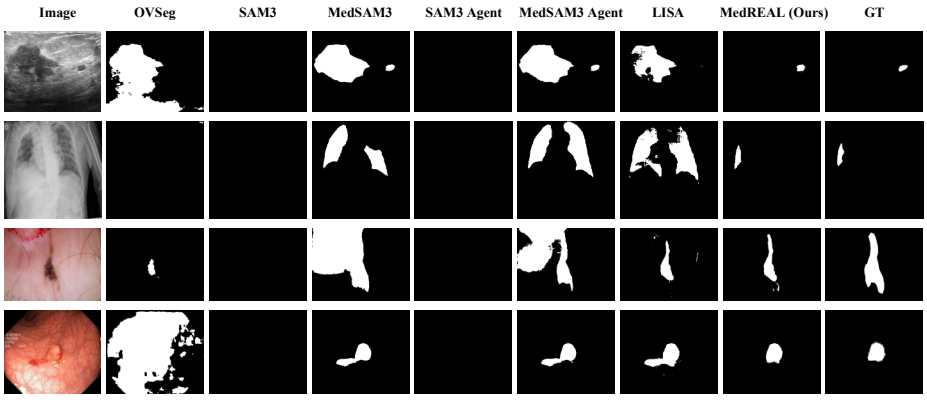

Fig. 5: Qualitative comparison with competing methods across four diverse medical imaging modalities (Top to Bottom: BUSI [2], COVID-QU-Ex [48], ISIC-2018 [12], and Kvasir-SEG [22]). MedREAL consistently produces precise evidence masks that are robust to structural artifacts and highly aligned with the underlying clinical reasoning.

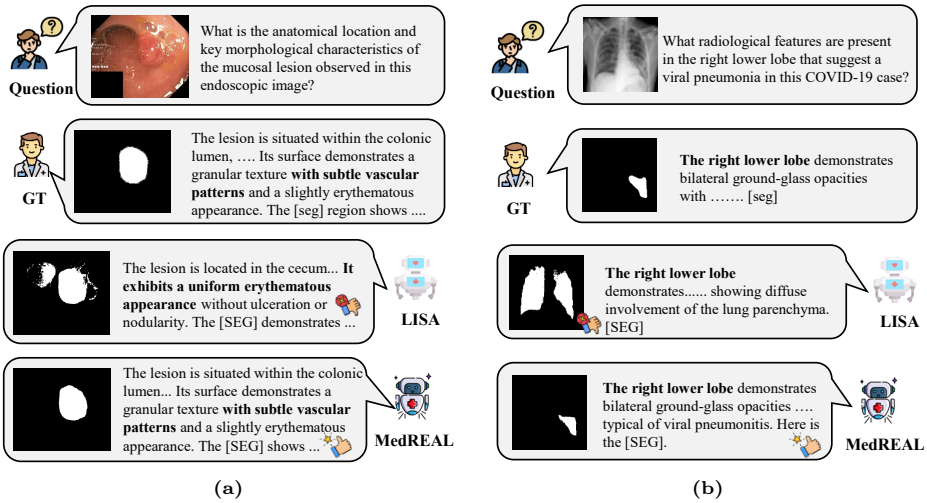

Fig. 6: Qualitative comparison between MedREAL and LISA. MedREAL demonstrates superior semantic consistency between the generated VQA text and the spatial segmentation mask.

precise spatial localization, causing LISA to over-segment irregular skin lesion boundaries with small color changes (Row 3). In contrast, MedREAL directly feeds diagnostic reasoning into the visual decoder via SARP and R2V. This rich semantic information ensures robust artifact removal and accurate boundary segmentation across all modalities. Unlike baseline methods that segment

the most visually dominant structures, our approach localizes targets guided by underlying clinical logic. Consequently, MEDREAL reliably distinguishes true pathological regions from structurally similar healthy tissues.

Reasoning-to-Mask Consistency. Figure 6 analyzes the internal consistency between generated VQA responses and spatial masks. LISA [26] frequently exhibits a serious mismatch: in the endoscopic case (Figure 6a), its incorrect text prediction of the polyp leads to a false-positive mask, while in the X-ray case (Figure 6b), it correctly localizes the infection in text but completely fails to locate it spatially. Such unpredictable misalignment greatly reduces user trust in critical medical scenarios, as the visual evidence fails to support the diagnostic text. MEDREAL resolves this mismatch. By conditioning the spatial decoding on extracted reasoning features rather than general prompts, our predicted mask serves as an accurate, pixel-level representation of the correct diagnostic text. This reliable alignment turns the MLLM from a black-box predictor into an interpretable and verifiable clinical tool.

6 Conclusion

In this paper, we present MEDREAL, a unified framework that bridges the semantic gap between high-level diagnostic reasoning and pixel-level spatial grounding in medical multimodal large language models. By introducing the Seg-Anchored Reasoning Pooling (SARP) and Reasoning-to-Visual (R2V) Fusion mechanisms, our approach effectively distills localized semantic evidence from the model's internal reasoning sequence to guide precise mask decoding. Supported by our newly curated MedRAVS-13K benchmark, extensive evaluations across four diverse imaging modalities confirm that MEDREAL achieves superior spatial alignment (68.49% gIoU) compared to existing reasoning-segmentation architectures. Crucially, by strictly coupling linguistic outputs with visual evidence, our method significantly enhances the interpretability and trustworthiness of AI-assisted clinical decision support.

Limitations and Future Work. Currently, MEDREAL operates exclusively on 2D imaging modalities and incurs computational overhead due to the autoregressive nature of the MLLM backbone. To address these constraints, our future work focuses on integrating 3D visual encoders into the R2V module to support volumetric data (*e.g.*, CT and MRI) and exploring knowledge distillation to accelerate inference for real-time clinical deployment.

7 Acknowledgement

This work was supported by the National Defense Science and Technology Industry Bureau Technology Infrastructure Project (JSZL2024606C001).

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in brief **28**, 104863 (2020)
3. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
4. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., , et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
5. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al.: The liver tumor segmentation benchmark (lits). Medical image analysis **84**, 102680 (2023)
6. Butoi, V.I., Ortiz, J.J.G., Ma, T., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Universeg: Universal medical image segmentation. In: CVPR. pp. 21438–21451 (2023)
7. Cai, X., Li, L., Pei, G., Chen, T., Pan, J., Yao, Y., Wang, W.: Unbiased object detection beyond frequency with visually prompted image synthesis. In: The Fourteenth International Conference on Learning Representations (2026)
8. Cai, X., Pei, G., Sun, Z., Yao, Y., Shen, F., Wang, W.: Iris: Bringing real-world priors into diffusion model for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26909–26919 (2026)
9. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: ECCV. pp. 205–218. Springer (2022)
10. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. In: ICLR (2026)
11. Chen, J., Gui, C., Ouyang, R., Gao, A., Chen, S., Chen, G.H., Wang, X., Cai, Z., Ji, K., Wan, X., et al.: Towards injecting medical visual knowledge into multimodal llms at scale. In: EMNLP. pp. 7346–7370 (2024)
12. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)
13. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
14. Fan, W., Fang, H., Li, R., Lin, Y., An, C., Luo, X.: Anatomy-Aware Frequency-Attention Transformer Networks for Liver Couinaud CT/MR Segmentation . In: MICCAI. vol. LNCS 15960, pp. 55 – 65. Springer Nature Switzerland (October 2025)
15. Ghezloo, F., Seyfioglu, M.S., Soraki, R., Ikezogwo, W.O., Li, B., Vivekanandan, T., Elmore, J.G., Krishna, R., Shapiro, L.: Pathfinder: A multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23431–23441 (2025)

16. Gong, H., Chen, G., Liu, S., Yu, Y., Li, G.: Cross-modal self-attention with multi-task pre-training for medical visual question answering. In: ICMR. pp. 456–460 (2021)
17. Gu, H., Pei, G., Sun, Z., Ren, M., Shu, X., Yao, Y., Shen, F.: Medfg-vqa: Low-frequency memory and graph attention for lightweight medical vqa. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 42755–42764 (2026)
18. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
19. Howlader, P., Nguyen-Canh, H., Das, S., Xu, J., Le, H., Samaras, D.: Cora: Consistency-guided semi-supervised framework for reasoning segmentation. In: WACV. pp. 5934–5944 (2026)
20. Huang, H., Lin, L., Tong, R., Hu, H., Zhang, Q., Iwamoto, Y., Han, X., Chen, Y.W., Wu, J.: Unet 3+: A full-scale connected unet for medical image segmentation. In: ICASSP. pp. 1055–1059. Ieee (2020)
21. Huang, S., Liang, H., Wang, Q., Zhong, C., Zhou, Z., Shi, M.: Seg-sam: Semantic-guided sam for unified medical image segmentation. arXiv preprint arXiv:2412.12660 (2024)
22. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., De Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: MMM. pp. 451–462 (2019)
23. Jiang, C., Ding, T., Song, C., Tu, J., Yan, Z., Shao, Y., Wang, Z., Shang, Y., Han, T., Tian, Y.: Medical sam3: A foundation model for universal prompt-driven medical image segmentation. arXiv preprint arXiv:2601.10880 (2026)
24. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything (2023), https://arxiv.org/abs/2304.02643
25. Konwer, A., Yang, Z., Bas, E., Xiao, C., Prasanna, P., Bhatia, P., Kass-Hout, T.: Enhancing sam with efficient prompting and preference optimization for semi-supervised medical image segmentation. In: CVPR. pp. 20990–21000 (2025)
26. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: CVPR. pp. 9579–9589 (2024)
27. Lai, Y., Zhong, J., Li, M., Zhao, S., Li, Y., Psounis, K., Yang, X.: Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. TMM (2026)
28. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 180251 (2018)
29. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. NeurIPS **36**, 28541–28564 (2023)
30. Liang, F., Wu, B., Dai, X., Li, K., Zhao, Y., Zhang, H., Zhang, P., Vajda, P., Marculescu, D.: Open-vocabulary semantic segmentation with mask-adapted clip. In: CVPR. pp. 7061–7070 (2023)
31. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
32. Liu, A., Xue, R., Cao, X.R., Shen, Y., Lu, Y., Li, X., Chen, Q., Chen, J.: Medsam3: Delving into segment anything with medical concepts. arXiv preprint arXiv:2511.19046 (2025)

33. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: ISBI. pp. 1650–1654. IEEE (2021)
34. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/
35. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023)
36. Liu, Y., Peng, B., Zhong, Z., Yue, Z., Lu, F., Yu, B., Jia, J.: Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. arXiv preprint arXiv:2503.06520 (2025)
37. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization (2019), https://arxiv.org/abs/1711.05101
38. Lv, X., Dong, X., Wang, L., Yang, J., Zhao, L., Pu, B., Jin, Z., Li, X.: Test-time domain generalization via universe learning: A multi-graph matching approach for medical image segmentation. In: CVPR. pp. 15621–15631 (2025)
39. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature communications **15**(1), 654 (2024)
40. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL. pp. 311–318 (2002)
41. Pei, G., Chen, T., Wang, Y., Cai, X., Shu, X., Zhou, T., Yao, Y.: Seeing what matters: Empowering clip with patch generation-to-selection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24862–24872 (2025)
42. Rasley, J., Rajbhandari, S., Ruwase, O., He, Y.: Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In: ACM SIGKDD. pp. 3505–3506 (2020)
43. Ren, Z., Huang, Z., Wei, Y., Zhao, Y., Fu, D., Feng, J., Jin, X.: Pixellm: Pixel reasoning with large multimodal model. In: CVPR. pp. 26374–26383 (2024)
44. Riedel, E.O., de la Rosa, E., Baran, T.A., Petzsche, M.H., Baazaoui, H., Yang, K., Musio, F.A., Huang, H., Robben, D., Seia, J.O., et al.: Isles'24–a real-world longitudinal multimodal stroke dataset. arXiv preprint arXiv:2408.11142 (2024)
45. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
46. Sun, S., Gu, H., Xie, C., Ren, Y., Ren, M., Zhang, H.: Bridging granularity gaps: Hierarchical semantic learning for cross-domain few-shot segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 9215–9223 (2026)
47. Sun, Z., Yao, Y., Liu, T., Li, Z., Shen, F., Tang, J.: Jo-snc: Combating noisy labels through fostering self-and neighbor-consistency. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
48. Tahir, A.M., Chowdhury, M.E., Khandakar, A., Rahman, T., Qiblawey, Y., Khurshid, U., Kiranyaz, S., Ibtehaz, N., Rahman, M.S., Al-Maadeed, S., et al.: Covid-19 infection localization and severity grading from chest x-ray images. Computers in biology and medicine **139**, 105002 (2021)
49. Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al.: Towards generalist biomedical ai. Nejm Ai **1**(3), AIoa2300138 (2024)
50. Wang, J., Ke, L.: Llm-seg: Bridging image segmentation and large language model reasoning. In: CVPR. pp. 1765–1774 (2024)
51. Wei, H., Sun, Y., Li, Y.: Deepseek-ocr: Contexts optical compression. arXiv preprint arXiv:2510.18234 (2025)

52. Wei, X., Cao, J., Jin, Y., Lu, M., Wang, G., Zhang, S.: I-medsam: Implicit medical image segmentation with segment anything. In: ECCV. pp. 90–107. Springer (2024)
53. Wu, J., Zhang, Y., Tang, X.: Simultaneous tissue classification and lateral ventricle segmentation via a 2d u-net driven by a 3d fully convolutional neural network. In: EMBC. pp. 5928–5931. IEEE (2019)
54. Xu, J., Pei, G., Liu, H., Yao, Y.: Gsv2x: Geometry-aware uncertainty modeling and orthogonal fusion for robust roadside perception. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21409–21419 (2026)
55. Yan, C., Wang, H., Yan, S., Jiang, X., Hu, Y., Kang, G., Xie, W., Gavves, E.: Visa: Reasoning video object segmentation via large language models. In: ECCV. pp. 98–115. Springer (2024)
56. Yang, Z., Pei, G., Chen, T., Yuan, X., Zhang, H., Shu, X., Yao, Y.: Beyond quadratic: Linear-time change detection with rwkv. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 11811–11819 (2026)
57. Yang, Z., Pei, G., Chen, T., Zhou, Y., Zhou, T., Yao, Y., Shen, F.: Efficiency follows global-local decoupling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25524–25535 (2026)
58. Yang, Z., Pei, G., Yao, Y., Zhou, T., Ding, L., Shen, F.: ChangeTitans: Toward remote sensing change detection with neural memory. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–14 (2025)
59. Yin, J., Chen, T., Chen, Y., Pei, G., Shu, X., Yao, Y., Shen, F.: Pca-seg: Revisiting cost aggregation for open-vocabulary semantic and part segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 27633–27643 (June 2026)
60. Zhang, W., Wu, H., Qin, J.: Domesticating sam for breast ultrasound image segmentation via spatial-frequency fusion and uncertainty correction. In: ECCV. pp. 20–37. Springer (2024)
61. Zhang, Z., Yin, G., Zhang, B., Liu, W., Zhou, X., Wang, W.: A semantic knowledge complementarity based decoupling framework for semi-supervised class-imbalanced medical image segmentation. In: CVPR. pp. 25940–25949 (2025)
62. Zhou, Y., Kang, X., Ren, F.: Employing inception-resnet-v2 and bi-lstm for medical domain visual question answering. In: CLEF. pp. 1–11 (2018)
63. Zhu, J., Hamdi, A., Qi, Y., Jin, Y., Wu, J.: Medical sam 2: Segment medical images as video via segment anything model 2. arXiv preprint arXiv:2408.00874 (2024)