MEDSAGE: Learning Structured Medical Visual Reasoning with a Structured Reasoning Scaffold

Yiru Huo[1*] **Haoran Yu**[1*] **Yichen Shi**[1] **Hongyang Wang**[2]
Changjie Luo[3] **Yao Chen**[3] **Jianzhou Feng**[4†]
[1]Tencent [2]Department of Computer Science and Technology, Tsinghua University
[3]Zhejiang University [4]Singapore University of Technology and Design

Abstract

Reinforcement learning (RL) can improve interpretability in medical vision-language models (VLMs), but medical visual reasoning remains challenging without structured guidance. Existing supervised fine-tuning and reinforcement learning (SFT+RL) approaches often learn task-specific image-to-answer mappings, leading to misalignment between visual evidence and textual reasoning and resulting in shortcut reasoning. To address the above challenges, we propose MEDSAGE, a medical VLMs framework built upon **high-quality structured reasoning sequences.** MEDSAGE introduces a structured path enhancement strategy that formulates medical visual reasoning as a sequence of clinically meaningful stages—localization, visual analysis, knowledge matching, and final decision—thereby guiding models to explore reasonable reasoning paths. We construct two high-quality datasets, **SAGE-sft20K** and **SAGE-rl10K**, to support this training paradigm. Under this framework, SFT induces consistent reasoning structures across tasks, while RL further improves answer correctness and reasoning faithfulness by encouraging self-check guided correction of erroneous predictions. Experiments on five medical benchmark datasets show that MEDSAGE achieves competitive or improved performance while substantially enhancing the robustness, faithfulness, and generalization of medical visual reasoning.

1 Introduction

Medical visual question answering (Med-VQA) aims to generate accurate and clinically meaningful answers based on medical images and natural language questions, and serves as a key task in medical image understanding and intelligent computer-aided diagnosis(Dong et al., 2025). In recent years, medical vision-language models (VLMs) have achieved significant progress in Med-VQA and related tasks(Xu et al., 2024; Yan et al., 2024a). However, most existing medical VLM approaches rely on supervised fine-tuning (SFT) on image–question–answer triplets. Due to the coarse granularity of supervision signals(Wu et al., 2025a; Li et al., 2023a), these models tend to learn task-specific direct mappings from images to answers, making it difficult to model the complex reasoning processes and medical knowledge required for reliable clinical decision-making. Recent studies have introduced reinforcement learning (RL) to enhance reasoning capabilities. Nevertheless, these approaches typically rely on unconstrained free-form language generation, causing visual evidence, medical knowledge, and reasoning steps to become entangled in unstructured text(Lai et al., 2025a; Pan et al., 2025a). This, in turn, limits the stability, transferability, and clinical reliability of the learned reasoning trajectories. These challenges pose significant obstacles for Med-VQA, where generalization and interpretability are critical in real-world clinical settings.

Unlike general-domain visual question answering, medical reasoning follows a progressive and structured diagnostic process. Clinicians typically (i) localize diagnostically relevant Regions of Interest (RoI), (ii) analyze fine-grained visual features, (iii) integrate observations with domain knowledge, and (iv) synthesize evidence into a final diagnostic conclusion. This structured reasoning paradigm is not specific to a single task, but rather forms a shared foundation underlying diverse medical instructions and diagnostic scenarios.

We analysis identifies two fundamental limitations that hinder effective medical reasoning in current SFT+RL based medical VLMs. First, there is a lack of structured multi-stage supervision. Figure 1 illustrates that models trained with structured data tend to accumulate more effective samples than

*The first two authors contribute equally to this work.
†Corresponding Author.

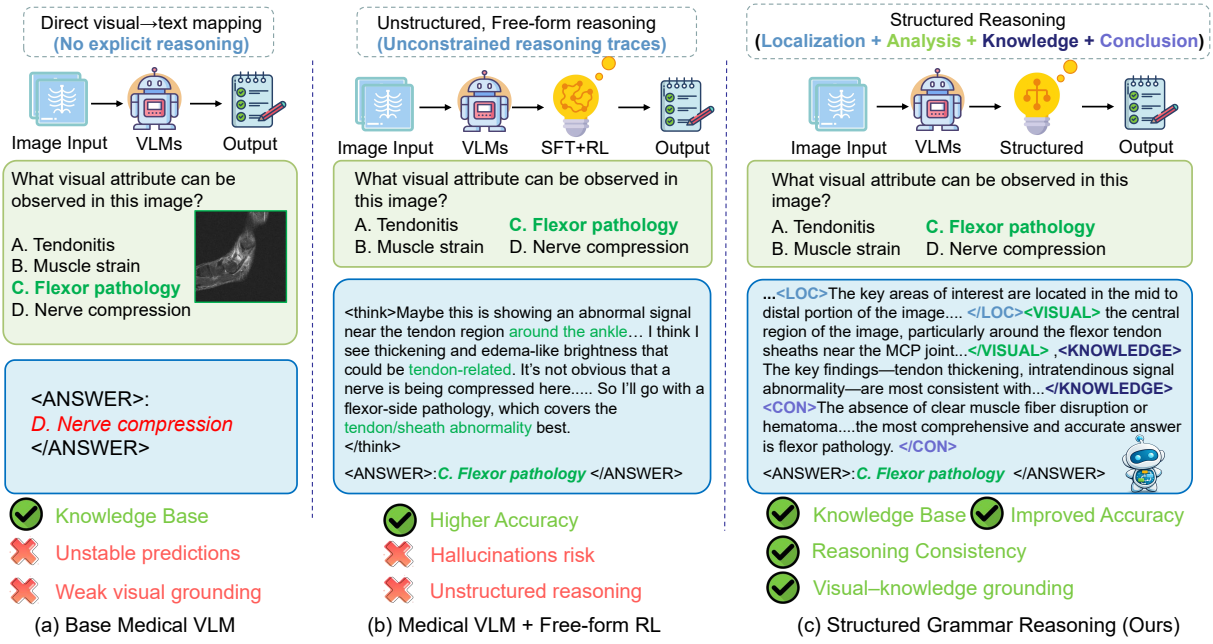


Figure 1: Motivation and overview of structured reasoning in medical vision-language models. We compare (a) base medical VLMs with direct image-to-answer mapping, (b) SFT+RL models with free-form reasoning that often leads to shortcut reasoning, and (c) MEDSAGE, which structures medical visual reasoning into clinically meaningful stages to improve grounding and robustness.

those trained with unstructured data. Existing medical datasets rarely provide supervision signals that reflect how clinicians progressively reason from visual evidence to diagnostic conclusions. SFT tends to learn task-specific input–output correlations, while RL lacks clear guidance for exploring clinically valid reasoning behaviors. Second, there is a misalignment between localized visual evidence and unstructured textual reasoning. Free-form chain-of-thought (CoT) generation collapses global context, localized visual cues, and medical knowledge into unconstrained text, weakening evidence-based reasoning consistency and further encouraging shortcut reasoning.

To address these challenges, we structure medical visual reasoning into four stages—localization, visual analysis, knowledge matching, and decision making—and construct a reasoning-supervised dataset, SAGE-sft20K. We further curate SAGE-rl10K, a 10K-sample visual question answering dataset reformulated as fine-grained visual diagnostic tasks for reinforcement learning.

Building on this, we propose MEDSAGE, a reasoning-guided training framework that integrates SFT and RL. SFT equips the model with multimodal analytical capabilities, while RL further aligns model behavior with the proposed reasoning through self-check–guided correction of erroneous predictions. Extensive experiments show that MEDSAGE consistently outperforms existing methods across multiple medical benchmarks, substantially improving the robustness and interpretability of medical visual reasoning.

Our contributions are threefold:

- **We investigate reasoning limitations of medical VLMs under the SFT+RL paradigm.** Unconstrained free-form reasoning often induces *visual-text misalignment* and *shortcut learning*, hindering interpretability and reliable clinical decision-making.
- **We propose MEDSAGE, a framework enforcing a clinically aligned reasoning scaffold.** By decomposing reasoning into four stages LVKC and curating SAGE-sft20K and SAGE-rl10K, we introduce rigorous process-level supervision through supervised fine-tuning and reinforcement learning.
- **We design a Stage-aware Self-Correction mechanism via GRPO.** We introduce a reflection-based credit assignment strategy that exclusively rewards error correction. This drives the model to internalize self-verification, ensuring adherence to the structured scaffold while enhancing reasoning robustness.

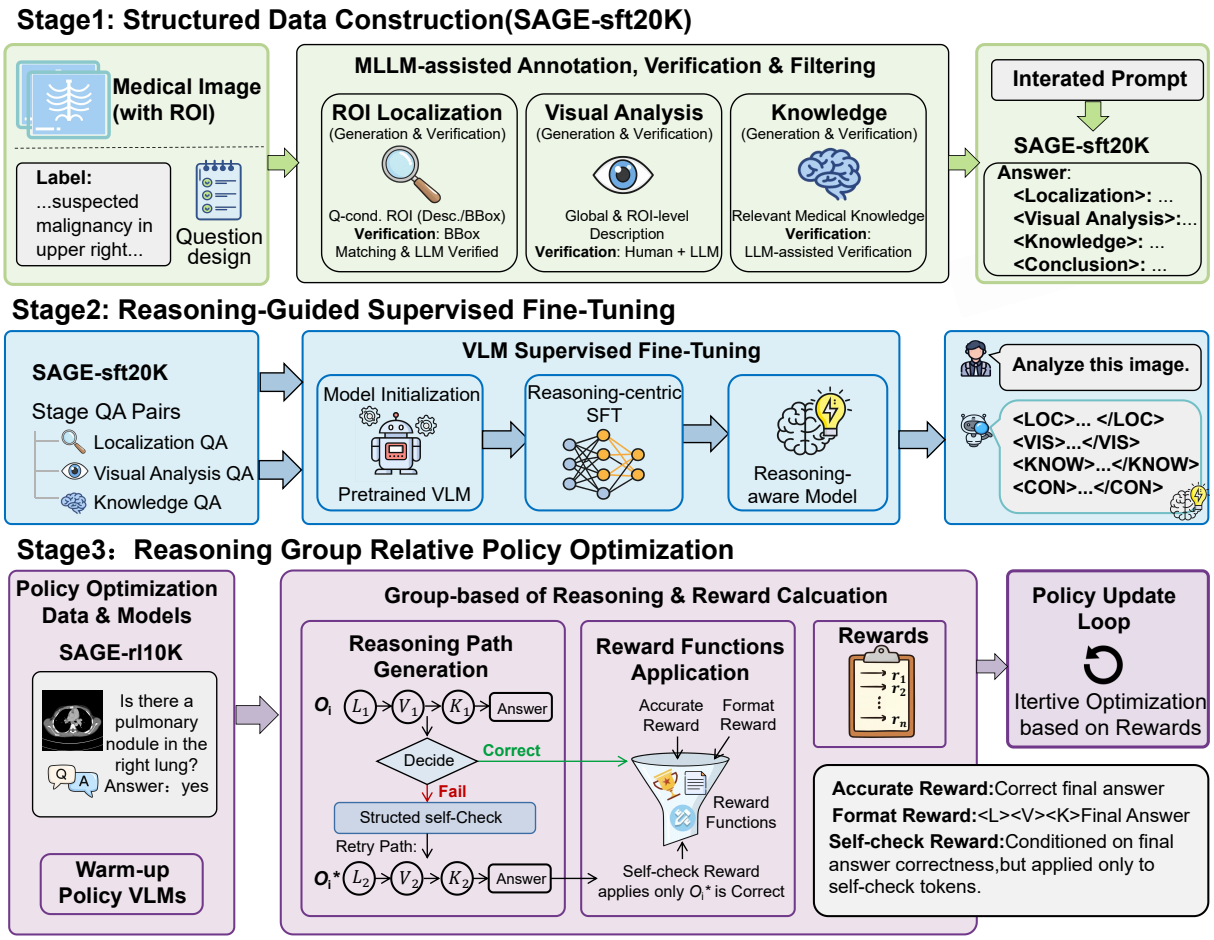


Figure 2: Illustration of the proposed MEDSAGE framework. Stage 1 constructs structured medical reasoning data. Stage 2 performs reasoning-guided supervised fine-tuning to induce stage-wise reasoning behaviors. Stage 3 applies Group Relative Policy Optimization with stage-aware self-check, routing reinforcement signals to self-check tokens upon successful error correction.

2 Related Work

2.1 Medical Vision-Language Models

Medical Visual Question Answering (Med-VQA) serves as a critical benchmark for evaluating multimodal understanding in healthcare. Early approaches primarily relied on discriminative models trained on limited datasets such as VQA-RAD (Lau et al., 2018) and SLAKE (Liu et al., 2021). With the advent of Large Language Models (LLMs), recent works have shifted towards generative paradigms. Models like LLaVA-Med (Li et al., 2023b) and PMC-VQA (Zhang et al., 2023) align visual encoders with LLMs using large-scale biomedical image-text pairs, demonstrating strong capabilities in open-ended generation. More recently, generalist models such as Qwen2-VL (Bai et al., 2023a) and proprietary models like GPT-4o have set new standards for zero-shot performance. However, most existing open-source medical VLMs rely heavily on Supervised Fine-Tuning (SFT) with direct image-to-answer pairs. As noted in recent studies (Wu et al., 2025b; Yan et al., 2024b), this coarse-grained supervision often leads to shortcut learning, where models memorize answer distributions rather than learning diagnostic reasoning, limiting their reliability in complex clinical scenarios.

2.2 Structured Reasoning in Medical AI

Reasoning capabilities are essential for transparent clinical decision-making. Chain-of-Thought (CoT) prompting (Wei et al., 2022) has proven effective in eliciting multi-step reasoning in general LLMs. In the medical domain, recent works have attempted to adapt CoT to enhance interpretability (Singhal et al., 2023). Despite these advances, applying free-form CoT to Medical VLMs remains challenging. Unlike pure text reasoning, medical visual reasoning requires precise grounding of visual evidence

(e.g., lesion localization) before clinical interpretation. Unconstrained free-form reasoning often suffers from visual-text misalignment, leading to hallucinations where the generated reasoning contradicts the visual features (Xu et al., 2024). Recent attempts like VITAR (Chen et al., 2025) focus on improving visual resolution but lack explicit constraints on the reasoning structure. Our work addresses this by introducing a structured scaffold (LVKC) that enforces a clinically aligned workflow—Localization, Visual Analysis, Knowledge, and Conclusion—to ensure grounded and faithful reasoning.

2.3 Reinforcement Learning for Multimodal Reasoning

Reinforcement Learning (RL) has become a key paradigm for aligning LLMs with human preferences, typically using PPO or DPO algorithms (Christiano et al., 2017; Rafailov et al., 2023). Recently, RL has been applied to enhance the reasoning capabilities of Large Language Models, as demonstrated by DeepSeek-R1 (Guo et al., 2025), which utilizes Group Relative Policy Optimization (GRPO) to incentivize self-verification and long-chain reasoning. In the medical multimodal domain, pioneering works like Med-R1 (Lai et al., 2025b) and MedVLM-R1 (Pan et al., 2025b) have begun to explore RL to improve answer accuracy. However, these methods primarily focus on outcome-based rewards (correctness of the final answer) or unconstrained reasoning paths. They do not explicitly penalize structural deviations or incentivize the model to self-correct intermediate visual perception errors. MEDSAGE distinguishes itself by integrating a stage-aware self-correction mechanism within the GRPO framework, explicitly rewarding the model for detecting and correcting its own reasoning flaws during the reinforcement learning phase.

3 Methodology

Our work aims to address three core challenges faced by medical VLMs in structured medical visual reasoning: (1) **Shortcut learning:** Coarse supervision leads to shortcut reasoning. (2) **Visual–text misalignment:** Free-form reasoning entangles visual evidence and medical knowledge. (3) **Limited self-correction:** Standard training paradigms lacks effective error detection and correction.

Figure 2 overviews the proposed MEDSAGE framework, which formulates medical visual reasoning as a unified sequence of clinically meaningful stages. The framework consists of three stages: (1) constructing multi-stage medical reasoning trajectories from existing data; (2) reasoning-guided SFT to induce stage-wise behaviors; and (3) RL with a stage-aware self-correction path to promote structured reasoning and answer correctness.

3.1 Reasoning Trajectory Construction

We construct structured medical reasoning trajectories using a staged data generation pipeline. Starting from publicly available datasets, we collect approximately 60K medical images from multiple sources, primarily including DeepLesion, Roboflow, and PubMedVision (see Appendix A for details). Each image is accompanied by region-level annotations, such as bounding boxes, segmentation masks, or textual region descriptions. Based on these annotations, we design corresponding medical reasoning queries, forming the initial data pool.

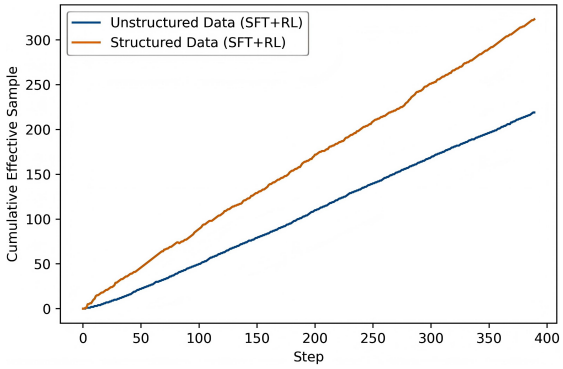


Figure 3: Cumulative effective samples versus training steps under the SFT+RL paradigm with 700 training samples.

ROI normalization. To increase the diversity of region specifications, we deliberately adopt two complementary ROI representations. Coordinate-based ROIs are expressed using bounding box coordinates, while text-based ROIs are specified via natural-language location descriptions, both treated as valid region representations. For each image, we construct a set of plausible ROIs

$$\mathcal{R} = \{r_1, r_2, \ldots, r_M\}, \tag{1}$$

by applying simple transformations—such as coordinate perturbation, context expansion, and sub-region sampling—to coordinate-based ROIs, in order to improve robustness to variations in region

localization. Text-based ROIs are retained without perturbation. The full image is additionally included as a global ROI to support holistic visual reasoning.

Stage-wise reasoning generation. As shown in Figure 2, we generate structured reasoning trajectories after ROI normalization. For each region $r \in \mathcal{R}$, reasoning proceeds in fixed stages, starting with *Visual Analysis* followed by *Knowledge Matching*.

For visual analysis, we use a multimodal large language model (MLLM) to jointly analyze the ROI and the global image context:

$$\mathrm{Vis}(r) = f_{\text{vis}}(I, r, q), \tag{2}$$

where the model is conditioned on the cropped region together with the full image.

Knowledge matching with diversity. Conditioned on the visual analysis, knowledge matching aligns observed visual evidence with clinically relevant medical concepts. We perform knowledge retrieval using a large language model augmented with a medical knowledge base. To introduce controlled diversity while preserving a fixed stage order, we define a small set of clinically common interpretation templates

$$\mathcal{T} = \{t_1, t_2, \ldots, t_T\}, \tag{3}$$

For each region r and template t, we generate knowledge matching and the corresponding reasoning output as

$$\mathrm{Know}(r, t) = f_{\text{know}}(\mathrm{Vis}(r), t), \tag{4}$$

$$\mathrm{Out}(r, t) = f_{\text{out}}(\mathrm{Vis}(r), \mathrm{Know}(r, t), y), \tag{5}$$

where all trajectories are constrained to share the same ground-truth answer y.

Trajectory definition and structured data augmentation. A structured reasoning trajectory is defined as

$$\tau(r, t) = \big[\mathrm{Vis}(r), \mathrm{Know}(r, t), \mathrm{Out}(r, t)\big]. \tag{6}$$

We perform Reasoning Path Augmentation(RPA) by pairing multiple plausible ROIs with different knowledge interpretation templates, generating multiple answer-consistent trajectories for the same image. Although the stage order remains fixed, this strategy increases data diversity by varying region specifications and clinically plausible reasoning patterns, thereby improving robustness.

3.2 Reasoning-Supervised SFT Warm-up

As illustrated in Figure 2. In the second stage, we perform reasoning SFT to initialize the model with stable and explicit Localization–Visual Analysis–Knowledge Matching–Conclusion (LVKC) structured medical reasoning.

Each training instance is represented as $\tau = (\mathcal{V}, \mathcal{Q}, y^L, y^V, y^K, y^C, \mathcal{A})$, where $\mathcal{V}$ denotes the medical image, $\mathcal{Q}$ is the input question, (y^L, y^V, y^K, y^C) corresponds to a full four-stage reasoning sequence following the LVKC order of Localization, Visual Analysis, Knowledge Matching, and Conclusion, and $\mathcal{A}$ is the final answer.

During SFT, the model is trained to maximize the likelihood of generating the entire structured reasoning sequence together with the final answer:

$$\mathcal{L}_{\text{SFT}} = -\mathbf{E}_{\tau \sim \mathcal{D}} \sum_{t=1}^{T} \log \pi_\theta(y_t \mid \mathcal{V}, \mathcal{Q}, y_{<t}), \tag{7}$$

where the target sequence explicitly follows the fixed stage order.

This SFT warm-up stage teaches the model to directly produce complete, stage-aligned medical reasoning paths, ensuring clear structural boundaries between reasoning stages while allowing flexible content realization within each stage.

3.3 RL with Self-Check-Guided Structured Exploration

To refine LVKC-structured medical reasoning beyond supervised imitation, we use a reinforcement learning framework that combines reflection-based retry with group-wise relative policy optimization (GRPO). Learning is driven by answer correctness, while the self-check mechanism biases exploration toward detecting and correcting potential reasoning errors rather than passively relying on retries.

Self-Check-Guided Retry. Given a medical image $\mathcal{V}$ and a query q, the policy π_θ first samples a reasoning trajectory

$$\tau_1 = o^{(1)}_{1:T} \sim \pi_\theta(o \mid \mathcal{V}, q), \tag{8}$$

which encodes a complete LVKC reasoning process and an answer a_1. If a_1 is correct, we accept τ_1. Otherwise, the same model produces a structured self-check that reviews the trajectory along fixed dimensions (visual grounding, visual analysis, and medical knowledge consistency), and then samples a revised trajectory

$$\tau_2 = o^{(2)}_{1:T'} \sim \pi_\theta(o \mid \mathcal{V}, q, \tau_1), \tag{9}$$

Method	VQA-RAD	SLAKE	PathVQA	PMC-VQA	MMMU	Average	Δ (%)
Proprietary models							
GPT-4.1	65.0	72.2	55.5	55.2	75.2	64.6	-5.2
Claude-Sonnet-4	67.6	70.6	54.2	54.4	74.6	64.3	-5.5
Gemini-2.5-Flash	68.5	75.8	55.4	55.5	76.9	66.4	-3.4
General-purpose Models							
LLaVA-v1.5-8B	54.2	59.4	54.1	36.4	38.2	48.5	-21.3
LLaVA-Next-7B	52.6	57.9	47.9	35.5	33.1	41.4	-28.4
LLaVA-Next-13B	55.8	58.9	51.9	36.6	39.3	44.5	-25.3
Qwen2.5-VL-7B	67.3	69.5	63.9	50.4	56.7	61.6	-8.2
InternVL3-8B	67.3	69.7	65.7	52.7	–	63.9	-5.9
Medical-specific non-reasoning VLMs							
Med-Flamingo	45.4	43.5	54.7	23.3	28.3	39.0	-30.8
RadFM	50.6	34.4	38.7	25.9	27.0	35.3	-34.5
LLaVA-Med-7B	51.4	48.6	56.2	24.7	36.9	43.6	-26.2
HuatouGPT-V-8B	59.4	66.8	59.8	51.4	56.7	58.8	-11.0
Lingshu-7B	62.2	78.9	**71.7**	55.6	70.0	67.7	-2.1
Medical-specific reasoning VLMs							
Med-R1	55.9	55.1	53.3	45.8	32.7	46.6	-23.2
MedVLM-R1	61.4	56.1	55.2	44.8	35.5	50.6	-19.2
ViTAR	70.1	**80.8**	67.0	57.2	**72.0**	69.4	-0.4
MedEyes	70.7	79.1	64.8	55.3	59.7	65.9	-3.9
MEDSAGE (Ours)	**70.4**	79.8	66.7	**58.3**	65.8	**69.8**	**0.0**

Table 1: Comparison across five medical benchmarks. Δ indicates the performance gap (%) compared to our method. Bold numbers indicate the best result in open-source VLMs and gray numbers indicate that the model has been trained on the corresponding dataset.

yielding an answer a_2.

We define a binary self-check success reward:

$$r_{\text{sc}} = \mathbb{I}[a_1 \neq y \ \wedge \ a_2 = y], \tag{10}$$

self-checking receives reward only when it converts an incorrect first attempt into a correct retry, and 0 otherwise.

Self-Check Reward under GRPO. We apply the self-check reward only to tokens generated in the self-check step. Let $m_t \in \{0, 1\}$ be a mask indicating whether token t belongs to the self-check span. The scaled self-check advantage is

$$A^{(\text{sc})} = \alpha \, r_{\text{sc}}, \tag{11}$$

where α controls the weight of the self-check reward. The GRPO objective for self-check tokens is

$$\mathcal{L}_{\text{sc}} = -A^{(\text{sc})} \sum_t m_t \log \pi_\theta(o_t \mid o_{<t}, \mathcal{V}, q). \tag{12}$$

In addition, the revised trajectory τ_2 is optimized with standard outcome-based rewards applied to tokens in τ_2. These outcome rewards are not applied to self-check tokens, and all remaining tokens receive zero advantage.

4 Experiments and Results

4.1 Benchmarks

We evaluate MEDSAGE on five medical VQA benchmarks. PathVQA (He et al., 2020), SLAKE (Liu et al., 2021), and VQA-RAD (Lau et al., 2018) are standard datasets for medical visual question answering. PMC-VQA (Zhang et al., 2023) contains 2,000 expert-annotated medical QA pairs. MMMU-Med (Yue et al., 2024), a medical subset of the multimodal reasoning benchmark MMMU, focuses on higher-level medical reasoning. Together, these datasets span diverse medical imaging modalities, including CT, MRI, X-ray, pathology slides, and multimodal clinical scenarios.

4.2 Implementation Details

We build our framework on Qwen2.5-VL-7B (Bai et al., 2023b) and adopt a standard SFT+RL training paradigm. SFT is performed using LLaMA-Factory and conducted for 4 epochs with AdamW (1×10^{-5}), a maximum sequence length of 4096, and bfloat16 precision. RL is implemented with Easy-R1. We train the model for 8 epochs with a learning rate of 1×10^{-5}, a maximum sequence length of 2048, and a maximum generation length

Model	Dataset	Localization	Visual Analysis	Knowledge	Reasoning	Average
Baseline	PathVQA	2.706	2.412	3.059	2.529	3.033
	SLAKE	3.014	2.350	3.623	2.614	3.033
	VQA-RAD	2.751	2.320	2.911	2.421	3.033
SFT	PathVQA	2.961	2.625	3.612	3.365	3.037
	SLAKE	3.446	2.629	3.965	3.036	3.073
	VQA-RAD	3.256	2.927	3.422	3.140	3.033
SFT+RL	PathVQA	3.146	3.020	3.948	3.644	3.033
	SLAKE	3.637	2.921	4.096	3.347	3.034
	VQA-RAD	3.830	3.447	3.652	3.453	3.033

Table 2: GPT-score (1–5) based evaluation of model response quality across multiple datasets and dimensions. The first decimal place is emphasized, while thesecond and third decimal decimal places are de-emphasized in gray for reference.

Method	SLAKE	VQA-RAD	PathVQA	Overall
Baseline-sft	45.9	36.7	35.3	39.3
Baseline-sft+RL	43.2	36.2	46.3	41.9
Low-level Aug-sft	72.8	67.4	66.1	68.8
Low-level Aug-sft+RL	77.8	68.4	66.1	70.8
RPA-sft	77.4	67.3	64.7	69.8
RPA-sft+RL	**78.8**	68.1	**67.0**	**71.3**

Table 3: Main ablation results on representative medical benchmarks. Methods are incrementally augmented to assess the contribution of structured reasoning, curriculum learning, and reinforcement learning.Reasoning Path Aug-RPA

Config.	RAD	SLAKE	Path	Ave
Reward configuration (Base + Self-Check bonus α)				
Base only ($\alpha = 0$)	69.9	78.3	63.8	70.7
Base + SC ($\alpha = 0.15$)	70.3	78.8	64.4	71.2
Base + SC ($\alpha = 0.20$)	**70.7**	**79.1**	**64.8**	**71.5**
Base + SC ($\alpha = 0.25$)	69.8	78.5	64.2	70.8

Table 4: Ablation on the self-check bonus α. The base reward consists of accuracy and format rewards applied to the final answer. An additional self-check bonus α is assigned exclusively to reflection tokens when self-checking successfully corrects an initial error. Moderate values of α yield the best overall performance.

Method	RAD	SLAKE	Path	Ave
SFT	67.3	77.4	64.7	69.8
SFT + RL	68.1	78.8	67.0	71.3
SFT + RL + Self	70.4	79.8	66.7	72.3

Table 5: Ablation on SFT, SFT+RL, and SFT+RL augmented with stage-aware self-correction.

of 1024 tokens. For each instance, 6 candidate responses are sampled to estimate policy gradients. All experiments are conducted on four NVIDIA A100 GPUs with DeepSpeed acceleration. More implementation details are provided in the supplementary materials.

4.3 Main Results

MEDSAGE is evaluated across five medical VQA benchmarks, as shown in Table 1. It achieves the best overall performance among open-source models, with an average accuracy of 69.8, consistently outperforming existing general-purpose and medical-specific VLMs. Compared to prior medical reasoning–oriented models, MEDSAGE surpasses MedVLM-R1 by a large margin and achieves competitive performance with ViTAR, while maintaining stronger generalization across datasets.

On VQA-RAD and SLAKE, MEDSAGE achieves accuracies of 70.4 and 79.8, respectively, demonstrating robust performance on radiology-focused benchmarks. On PMC-VQA, MEDSAGE attains the best result of 58.3, indicating its effectiveness on expert-annotated medical QA. On MMMU-Med, MEDSAGE achieves competitive performance under challenging multimodal medical reasoning scenarios. Overall, these results suggest that stage-aware self-correction under the SFT+RL paradigm provides more effective learning signals for medical visual reasoning than standard reinforcement learning approaches.

4.4 Analysis on Structured Thinking

Table 2 analyzes the effect of structured thinking under different training stages. Compared to the baseline, SFT consistently improves response quality across localization, visual analysis, knowledge, and reasoning. Further gains are observed after SFT+RL, with the most notable improvements in localization and reasoning. For example, on VQA-RAD, the reasoning score increases from 3.14 (SFT) to 3.45 (SFT+RL), while localization improves from 3.26 to 3.83. This trend indicates that SFT+RL encourages more structured and coherent reasoning.

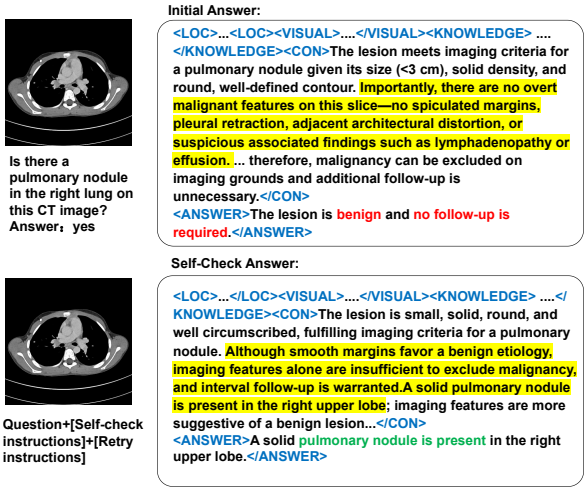


Figure 4: Stage-aware self-correction example during the RL stage. The model first generates an initial reasoning trajectory that leads to an incorrect answer. A reflection-based self-correction step revises the erroneous intermediate reasoning.

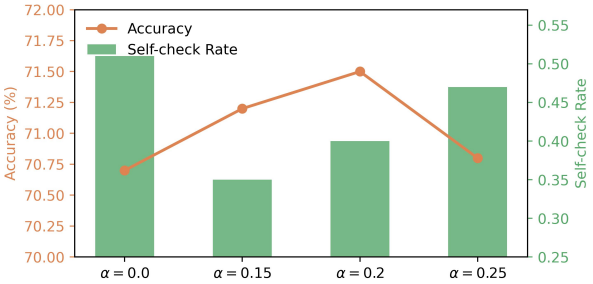


Figure 5: Effect of the weighting coefficient α on accuracy and self-check rate. Accuracy is reported on the left y-axis, and the self-check rate on the right y-axis, indicating the fraction of instances where self-correction is applied.

4.5 Ablation Study

Effectiveness of Reasoning Path Augmentation. Table 3 shows that reasoning path augmentation consistently improves performance over both the baseline and low-level augmentation. Under supervised fine-tuning, RPA achieves higher accuracy than low-level augmentation (69.8 vs. 68.8), and further gains are observed when combined with reinforcement learning, with RPA-SFT+RL reaching 71.3 average accuracy. This indicates that RPA provides more effective reasoning supervision, especially under the SFT+RL setting.

Reward Function Design. Table 4 examines the effect of the self-check bonus α. Introducing a self-check bonus consistently improves performance over the base reward, with the best results achieved at a moderate value of $\alpha = 0.20$ (71.5 average accuracy). Larger α values degrade performance, suggesting that excessive emphasis on self-checking can hinder stable optimization.

Effectiveness of Stage-aware Self-Correction. Table 5 evaluates stage-aware self-correction under different training settings. Compared to SFT and SFT+RL, incorporating self-correction yields consistent performance gains, with notable improvements on RAD and SLAKE. This suggests that explicitly correcting intermediate reasoning errors leads to more reliable final predictions.

Effect of the Self-Check Weighting Coefficient α. Figure 5 illustrates the effect of α on accuracy and the self-check rate. Accuracy peaks at an intermediate α while the self-check rate remains moderate, whereas larger α increases self-check frequency without further accuracy gains. This indicates that selective self-checking is more effective than excessive retries.

5 Conclusion

In this work, we present MEDSAGE, a structured reasoning–guided framework for medical vision–language models under the SFT+RL paradigm. By organizing medical visual reasoning into clinically aligned stages and introducing stage-aware self-correction, MEDSAGE improves reasoning faithfulness and answer correctness without relying on free-form reasoning at inference time. Extensive experiments demonstrate that MEDSAGE achieves competitive or improved performance while substantially enhancing the robustness and interpretability of medical visual reasoning.

6 Limitations

MEDSAGE relies on high-quality structured reasoning annotations, which are costly to obtain and may limit scalability. In addition, this work does not conduct an in-depth study on how localization information is explicitly modeled and utilized during structured reasoning.

References

Wenjie Dong, Shuhao Shen, Yuqiang Han, Tao Tan, Jian Wu, and Hongxia Xu. Generative models in medical visual question answering: A survey. *Applied Sciences*, 15(6):2983, 2025.

Dexuan Xu, Yanyuan Chen, Jieyi Wang, Yue Huang, Hanpin Wang, Zhi Jin, Hongxing Wang, Weihua Yue, Jing He, Hang Li, et al. Mlevlm: Improve multi-level progressive capabilities based on multimodal large language model for medical visual question answering. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4977–4997, 2024.

Quan Yan, Junwen Duan, and Jianxin Wang. Multimodal concept alignment pre-training for generative medical visual question answering. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5378–5389, 2024a.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1): 7866, 2025a.

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

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*, 2025a.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 337–347. Springer, 2025a.

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

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

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

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023a.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1): 7866, 2025b.

Quan Yan, Junwen Duan, and Jianxin Wang. Multimodal concept alignment pre-training for generative medical visual question answering. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5378–5389, 2024b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Kaitao Chen, Shaohao Rui, Yankai Jiang, Jiamin Wu, Qihao Zheng, Chunfeng Song, Xiaosong Wang, Mu Zhou, and Mianxin Liu. Think twice to see more: Iterative visual reasoning in medical vlms. *arXiv preprint arXiv:2510.10052*, 2025.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*, 2025b.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 337–347. Springer, 2025b.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023b.

Source	Data Size
DeepLesion	–
Roboflow	–
PubMedVision	–
Total	**60K**

Table 6: Statistics of the initial data pool (QA pairs), collected from multiple public medical image sources.

A Dataset Construction Details

A.1 Source Datasets

Our dataset is constructed by integrating multiple publicly available medical imaging datasets with heterogeneous forms of spatial and semantic supervision, including **DeepLesion**, **Roboflow**-based medical detection datasets, and **PubMedVision**.

DeepLesion and Roboflow provide explicit region-of-interest (ROI) annotations in the form of bounding boxes, along with corresponding clinical labels. For these datasets, we construct task-specific medical questions (e.g., multiple-choice and open-ended formats) based on the annotated regions and labels, and treat the provided bounding boxes as strong ROI grounding signals that identify clinically relevant image regions.

In contrast, PubMedVision does not include explicit ROI annotations. However, many samples contain implicit localization cues embedded in natural language descriptions within the original questions or answers (e.g., anatomical locations, laterality, or arrow-referenced regions). For these samples, we extract and normalize such language-based localization cues and treat them as weak ROI grounding signals, which specify approximate regions relevant to the clinical question without requiring precise spatial annotations.

Table 7: Dataset statistics of the initial data pool (QA pairs).

Dataset	Source	Data Size
Initial Data Pool	**DeepLesion**	–
	Roboflow	–
	PubMedVision	–
Total		**60K**

A.2 Data Format

We constructed the dataset based on the LLAVA framework, as shown in Fig. 6.

```
"conversations": [
    {
        "from": "human" ,
        "value":  <image>\n<Instructions>
    } , {
        "from": "gpt" ,
        "value": <Response>
    } ]
```

Figure 6: Data format

A.3 Data Construction Pipeline

Starting from approximately 60K initial samples, we construct explicit multi-stage reasoning trajectories aligned with the proposed medical visual reasoning framework (*Localization* $\rightarrow$ *Visual Analysis* $\rightarrow$ *Knowledge Matching* $\rightarrow$ *Reasoning*). The overall data construction pipeline consists of six stages: metadata extraction, multi-task instruction set construction, high-quality task seed data construction, stage-wise reasoning trajectory synthesis, trajectory consistency verification, and manual inspection.

A.3.1 Metadata Extraction

In medical multimodal datasets, aligned image-label pairs constitute the most fundamental sources of information. We treat each pair as an atomic unit, enabling the synthesis of complex and realistic task scenarios through controlled composition.

To address the heterogeneity of raw data, we employ a stratified processing strategy. High-fidelity images with precise ROI annotations and comprehensive textual labels undergo direct metadata extraction with minimal preprocessing. For samples with reliable labels but varying visual quality, we apply a strict filtration process: images below a visual quality threshold are discarded, while those maintaining diagnostic value are retained.

For weakly labeled or unlabeled data, we perform a joint evaluation of visual integrity and label completeness. In cases where visual quality is sufficient but labels are ambiguous or sparse, we leverage high-capacity Visual Large Language Models (VLLMs) for semantic enrichment and label imputation. Samples failing to meet reliability standards in both modalities are either repurposed as negative examples or excluded.

A.3.2 Multi-Task Instruction Set Construction

We systematically analyze high-frequency and representative task types encountered in real clinical

scenarios. For each task category, we design a diverse set of instruction templates that vary in input format, task objective, and reasoning requirement. These instructions guide the model through multi-stage processes such as localization, visual analysis, knowledge invocation, and integrated reasoning, resulting in a multi-task instruction set with strong generalization capability.

A.3.3 High-Quality Task Seed Data Construction

Prior to large-scale data synthesis, we select a subset of high-quality samples from the metadata pool to serve as seed data. These samples satisfy strict criteria in terms of image quality, label accuracy, and task relevance. Based on these seed samples, we employ high-capacity multimodal models to generate structured, stage-wise reasoning trajectories, which are further refined and expanded to provide a reliable foundation for subsequent large-scale synthesis.

A.3.4 Stage-Wise Reasoning Trajectory Synthesis

We synthesize reasoning trajectories by treating the *image + (original or enhanced) label* pair as the minimal trusted input and explicitly generating four-stage trajectories consistent with the medical visual reasoning framework. This design decomposes the otherwise black-box, end-to-end prediction process into interpretable, verifiable, and controllable intermediate steps, closely reflecting real clinical image interpretation workflows.

Localization. Medical image decision-making is predominantly driven by local visual evidence. The localization stage narrows the model's attention from the entire image to clinically relevant regions, reducing background noise and providing traceable evidence anchors for downstream analysis. In practice, ROI annotations and labels from the original data are used as initialization, and GPT-4o is prompted to identify and describe the corresponding critical regions in textual form.

Visual Analysis. This stage explicitly describes observable image characteristics, analogous to radiological findings in clinical practice. It prevents unsupported leaps in reasoning and enhances interpretability and auditability. Using the localization output as a starting point, GPT-4o is guided to produce detailed descriptions and analyses of visual features, forming a comprehensive understanding of the image.

Knowledge Matching. We construct medical knowledge bases covering disease spectra, imaging sign terminology, differential diagnosis rules, and guideline- or commonsense-based associations. During trajectory construction, image content, labels, and outputs from the previous stages are used to retrieve relevant knowledge, after which GPT-4o synthesizes coherent and contextually appropriate medical knowledge descriptions.

Reasoning. The reasoning stage enforces a strict evidence-to-conclusion principle: final conclusions must be supported by information from the preceding stages. This closes the reasoning loop and avoids trajectory contamination such as unsupported correct answers or logically sound explanations leading to incorrect conclusions. All prior stage outputs are provided as input, and GPT-4o integrates them into a concise, label-consistent reasoning process while compressing earlier-stage content for efficient information utilization.

A.3.5 Trajectory Consistency Verification

To prevent cross-stage contradictions, insufficient evidential support, and concept mismatches, we employ Qwen2.5-VL-72B to perform consistency verification for each synthesized trajectory. The verification focuses on detecting logical conflicts across stages, missing or unsupported reasoning steps, incorrect knowledge associations, and internally contradictory statements. Consistency checking serves as a filtering mechanism: samples exhibiting clear conflicts or broken evidence chains are either discarded or regenerated, ensuring strong traceability and cross-stage coherence in the final training data.

A.3.6 Manual Inspection

Finally, we conduct multiple rounds of manual inspection as the last quality control step to address subtle errors that automated verification may miss, such as clinically imprecise phrasing, overly assertive conclusions, or improper handling of borderline cases. Manual inspection is performed via repeated random sampling across batches and task types. Inspectors evaluate stage completeness and coherence, alignment between visual descriptions and image evidence, medical validity of knowledge matching, and whether final conclusions are properly derived from evidence. Inspection outcomes are used to iteratively refine synthesis templates,

expand consistency-checking rules, and adjust filtering thresholds and negative sample strategies.

A.4 RL Data

After constructing the supervised training data, we further develop reinforcement learning (RL) data and reward mechanisms to more effectively strengthen the model's reasoning capability and adherence to structured output formats. Specifically, we perform balanced sampling from the previously curated dataset according to instruction types and select approximately 10K samples as the basis for RL training. For each sample, the complete multi-stage reasoning trajectory is explicitly segmented into four stages—**localization**, **visual analysis**, **knowledge matching**, and **reasoning**—and annotated with predefined structural tokens. These annotations are used to impose structural constraints during model rollouts and to define a **format reward**, which encourages the model to generate reasoning trajectories with clear stage boundaries and stable structure. Importantly, the format reward does not enforce strict textual overlap between the model-generated content and the human-constructed ground truth at each stage; instead, it only requires stage completeness, structural correctness, and preservation of stage-specific key information, thereby avoiding overfitting to fixed expressions.

During RL training, for each rollout trajectory produced by the model, we introduce an automatic evaluation via a **trajectory quality scoring model**, which provides a **trajectory consistency reward**. This scoring model is trained using contrastive learning, where positive samples consist of high-quality reasoning trajectories that have passed prior consistency checks and manual inspection, and negative samples are drawn from low-quality trajectories that were filtered out or discarded during data curation. The scoring process focuses on detecting logical inconsistencies across stages, verifying whether conclusions are sufficiently supported by key information from preceding stages, and assessing the overall semantic coherence and medical plausibility of the reasoning chain. Similar to the format reward, the consistency reward is defined under a key-information-preserving criterion, allowing diverse yet reasonable reasoning paths rather than enforcing exact reproduction of reference trajectories.

In addition to trajectory-level supervision, we use the **final answer** of each sample as the target signal to construct an **answer reward**, which evaluates whether the model's final output is semantically consistent with the reference answer or falls within an acceptable range of correctness. By jointly optimizing the format reward, trajectory consistency reward, and outcome reward, the model is encouraged to maintain structured, interpretable reasoning while progressively improving the robustness, stability, and explainability of multi-stage medical visual reasoning.

B Experiment Details

B.1 Detailed Experimental Settings

All experiments were conducted on 4×A100 GPUs. Supervised fine-tuning and reinforcement learning were implemented using different frameworks, each optimized for its respective training paradigm. Specifically, we adopt LLaMA-Factory for supervised fine-tuning and Easy-R1 for reinforcement learning. All experiments were performed with bfloat16 precision and DeepSpeed acceleration.

Supervised Fine-Tuning (SFT) Supervised fine-tuning is performed using the **LLaMA-Factory** framework. We fine-tune all parameters of the language model while freezing the vision encoder to reduce memory consumption and stabilize training. The model is trained for 4 epochs with a learning rate of 1×10^{-5} using the AdamW optimizer.

The maximum sequence length is set to 4096. The per-device batch size is 4, with gradient accumulation over 16 steps, resulting in an effective batch size of 1024. All experiments are conducted in bfloat16 precision, and gradient checkpointing is enabled to further reduce memory usage.

Reinforcement Learning (R-GRPO) Reinforcement learning is conducted using the **Easy-R1** framework, with the R-GRPO algorithm applied under answer-level supervision. Similar to SFT, we optimize all parameters of the language model while keeping the vision encoder frozen throughout RL training.

The model is trained for 8 epochs with a learning rate of 1×10^{-5}. The maximum sequence length is set to 2048, and the maximum generation length is limited to 1024 tokens. The per-device batch size is 2 with gradient accumulation over 8 steps.

During sampling, we generate 6 candidate responses per instance to estimate the policy gradient. Rewards are computed based on the outcome reward and trajectory-level consistency signals, and

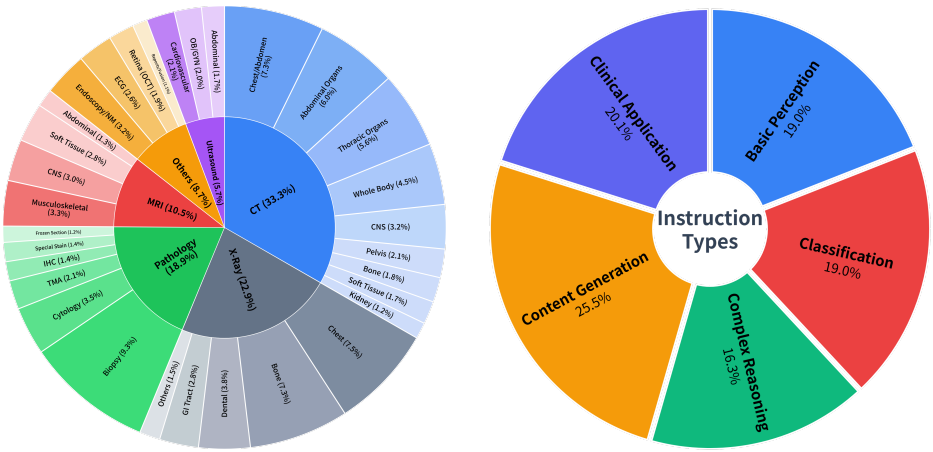


Figure 7: The MCoT generators distribution of each splits.

are normalized within each batch to stabilize training.

B.2 Detailed Results

Experiments on OmniMedVQA. We evaluate our model on the OmniMedVQA benchmark to assess its robustness across diverse medical imaging modalities. Following the standard evaluation protocol, OmniMedVQA covers eight modalities, including CT, MRI, X-ray, ultrasound, and several specialized imaging types.

As shown in Table X, our model achieves competitive overall performance across modalities, demonstrating stable accuracy without relying on modality-specific adaptation. Notably, performance variations across modalities remain moderate, suggesting that the proposed reasoning-guided training framework generalizes beyond the imaging distributions emphasized during training.

Compared with both medical-specific and general-purpose vision–language models, our approach maintains a consistent advantage in average accuracy. These results indicate that explicitly structured reasoning supervision contributes to improved cross-modality robustness, even when the evaluation domains differ from the primary training data.

Experiments on MMMU-Med. We further evaluate our model on the MMMU-Med benchmark, which spans multiple medical subdomains, including basic medical science, clinical medicine, diagnostics, pharmacy, and public health. This benchmark tests the ability of models to integrate visual and textual information across heterogeneous medical knowledge domains.

Results in Table Y show that our model achieves strong average accuracy across all subcategories, with particularly balanced performance between clinically oriented and knowledge-intensive questions. Compared to existing medical and general-purpose VLM baselines, our approach demonstrates improved consistency across subdomains rather than excelling in a single category.

These findings suggest that the proposed task-agnostic reasoning Scaffold supports generalization across diverse medical disciplines, enabling stable performance on both factual and reasoning-intensive medical queries.

B.3 Additional Analysis

Llama under Few-shot Evaluation. We observed that Llama-3.2-11B-Vision-Instruct performed significantly worse in few-shot evaluation compared to zero-shot evaluation. Therefore, we conducted a statistical analysis of the outputs from Llama-3.2 on several tasks where its performance was particularly poor in the few-shot evaluation. We found that the proportion of outputs labeled as "Both" for the step type was 99.9%. Additionally, in terms of informativeness, all outputs of Llama-3.2 were "Uninformative". In the Description Error Types task, 74.0% of its outputs were invalid, meaning they fell outside the defined label range. Overall, Llama-3.2 exhibited a decline in performance across every task in the few-shot evaluation, which may be related to its training process, suggesting that Llama-3.2 might lack multi-round and multi-image training.

Description Relevance in Pairwise Comparison.

We observed that on both MiCEval-HARD and MiCEval-NORMAL, all MLLMs performed worse in description relevance during few-shot evaluation compared to zero-shot evaluation. We analyzed each model's accuracy across the labels in the description relevance task, with results shown in Tables **??** and **??**. Our analysis revealed that the prediction distribution of each model shifted in different ways during few-shot evaluation. We believe this performance drop may be due to the larger number of label classes in description relevance compared to other tasks, as well as the generally weaker in-context learning abilities of MLLMs for this task.

B.4 Ablation Study

C Case Study

In this chapter, we present examples based on two different MiCEval metrics: the high-scoring and low-scoring cases. Figures **??** to **??** show the best cases based on the *Correctness*$_{type}$-based MiCEval metric, while Figures **??** to **??** illustrate the worst cases. The high-scoring cases for the *Correctness*$_{all}$-based MiCEval metric are shown in Figures **??** to **??**, and the low-scoring cases are depicted in Figures **??** to **??**.

Algorithm 1: Construction of the Structured Medical Visual Reasoning Dataset

Input: Initial data pool $\mathcal{D}_{init}$ consisting of medical images with associated questions and answers, and optional ROI annotations and clinical labels; Medical Visual Reasoning Scaffold (MVRS) $\mathcal{S}$; Auxiliary large language models $\mathcal{M}$.

Output: Structured reasoning dataset $\mathcal{D}_{struct}$.

Initialize $\mathcal{D}_{struct} \leftarrow \emptyset$;

foreach *sample* $(I, q, a) \in \mathcal{D}_{init}$ **do**

  if *ROI annotation is available* **then**

    Generate localization description $y^{(L)}$ conditioned on the annotated ROI;

  else

    Infer candidate ROI using LLM-assisted proposal and multi-model verification;

    Generate localization description $y^{(L)}$;

  Generate visual analysis description $y^{(V)}$ conditioned on $y^{(L)}$ and image I;

  Generate knowledge matching description $y^{(K)}$ by associating visual evidence in $y^{(V)}$ with relevant medical concepts;

  Generate final reasoning $y^{(R)}$ that integrates $(y^{(L)}, y^{(V)}, y^{(K)})$ to support the answer a;

  if *stage-wise validity checks and global consistency verification are satisfied* **then**

    Add $(I, q, a, y^{(L)}, y^{(V)}, y^{(K)}, y^{(R)})$ to $\mathcal{D}_{struct}$;

Perform additional automatic filtering and manual review on a subset of $\mathcal{D}_{struct}$;

return $\mathcal{D}_{struct}$;

Role
You are an experienced radiologist.

Task
Given a medical image and its associated clinical question, generate a concise localization description that specifies the image region relevant to answering the question.

Inputs
1. Medical image
2. Clinical question:{*prompt*}
3. Region of interest cue (internal reference, not to be mentioned explicitly):{*roi_hint*}

Instructions
- Describe the approximate anatomical location in human-readable medical terms.
- May refer to anatomical structures, laterality, and relative position (e.g., upper/lower, medial/lateral).
- Do NOT include pixel coordinates or bounding box values.
- Do NOT describe visual appearance.
- Do NOT provide diagnosis, interpretation, or answer to the question.
- Do NOT mention how the region was obtained.

Figure 8: Different system prompts of Scoring Evaluation. **System Prompt for Step with MiCEval** is the system prompt for MLLM-based MiCEval evaluation metrics.

Role
You are a medical imaging observer.

Task
Given a medical image and a specified region of interest, describe the visual evidence observable in the image that is relevant to the clinical question.

Inputs
1. Medical image
2. Clinical question:{*prompt*}
3. Region of interest summary (for grounding only):{*roi_summary*}

Description Guidelines
Generate a visual analysis with two clearly separated parts:

(A) GLOBAL CONTEXT
- Briefly describe 1–2 observable background features outside the ROI that are relevant for context.
- Do NOT speculate about imaging modality details unless clearly visible.

(B) ROI EVIDENCE
- Describe only visually observable features within the ROI.
- Focus on objective attributes: location (anatomical), shape, margin, size (relative), intensity/density, internal pattern, spatial relationship to nearby structures.
- Use neutral, descriptive language.
- Do NOT interpret findings or imply diagnosis.
- Do NOT mention how the ROI was obtained.
- Do NOT include pixel coordinates or bounding box values.

Figure 9: Different system prompts of Scoring Evaluation. **System Prompt for Step with MiCEval** is the system prompt for MLLM-based MiCEval evaluation metrics.

Role
You are a medical reasoning assistant.

Task
Given visual evidence from a medical image and the ground-truth answer label (for guidance only), summarize the key medical knowledge conditions that must be satisfied to support the correct answer.

Inputs
1. Clinical question:{*prompt*}
2. Visual evidence summary: {*visual_summary*}
3. Ground-truth answer label (internal guidance, MUST NOT be mentioned): {*label*}

Instructions
- Use the provided label only to guide which medical knowledge is relevant.
- Do NOT mention the label, diagnosis name, or final answer explicitly.
- Do NOT restate the visual evidence verbatim.
- Express knowledge as abstract medical conditions, criteria, or principles.
- Focus on what must be true for the answer to be correct, not on explaining why.
- Keep the output concise and suitable for supervision.

Figure 10: Different system prompts of Scoring Evaluation. **System Prompt for Step with MiCEval** is the system prompt for MLLM-based MiCEval evaluation metrics.

Role
You are a content integrator.

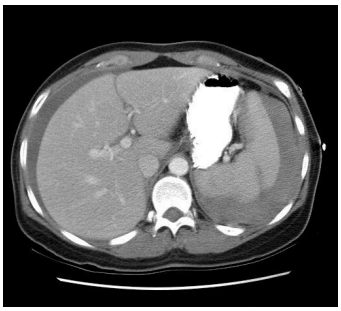
Task
You are given structured components of a medical reasoning example: `[LOC]`, `[VIS]`, `[KNO]`, `[ANS]`.
Your task is to assemble them in order without adding new information, new interpretations, or narrative content.

Do NOT:
- Add transitional or explanatory text
- Expand or rewrite the content
- Introduce first-person narration

Output
Output the components in the following order with their tags preserved:

```
[LOC]
[VIS]
[KNO]
[ANS]
```

Figure 11: Different system prompts of Scoring Evaluation. **System Prompt for Step with MiCEval** is the system prompt for MLLM-based MiCEval evaluation metrics.



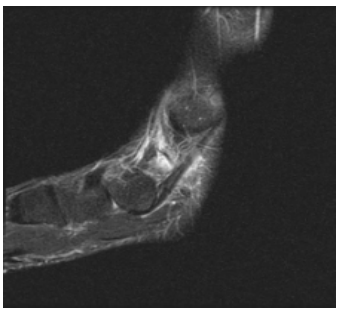
Is there intraperitoneal hemorrhage present in this image?

Yes No

CT Sample

This is a non-contrast axial CT image of the upper abdomen. Abdominal organs show largely normal attenuation. A focal area of hyperdense free fluid is identified in the left upper quadrant, prompting focused evaluation of the peritoneal cavity.<LOC>The region of interest is the left upper quadrant near the spleen, where hyperdense free fluid is present {"box_2d": [315, 742, 785, 968], "label": "perisplenic hyperdense fluid"}. No definite abnormal fluid is seen in the right upper quadrant {"point": [420, 215], "label": "right upper quadrant"}. </LOC><VISUAL>Hyperdense free fluid adjacent to the spleen is noted in the left upper quadrant, with an estimated attenuation of 40–60 HU, consistent with acute blood. The fluid is free within the peritoneal cavity and tracks along the left paracolic gutter. The spleen appears grossly intact on this slice, though an occult splenic or vascular injury cannot be excluded.</VISUAL><KNOWLEDGE>- Acute intraperitoneal hemorrhage appears as hyperdense free fluid (≈30–70 HU) on non-contrast CT.- Perisplenic free fluid raises suspicion for splenic or left upper abdominal vascular injury.</KNOWLEDGE><CON>Hyperdense, free intraperitoneal fluid with attenuation consistent with acute blood supports the diagnosis of hemoperitoneum, likely originating from the left upper abdomen. </CON><ANSWER>Findings are consistent with acute intraperitoneal hemorrhage (hemoperitoneum).</ANSWER>

(a) CT Modality



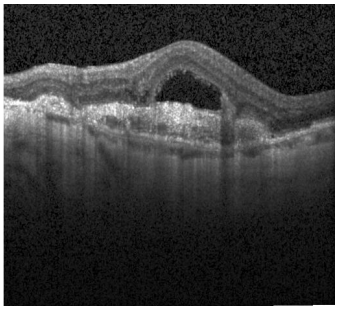
What attribute can be observed in this image?
Options:
A. Tendonitis
B. Muscle strain
C.Flexor pathology
D. Nerve compression

MRI Sample

This is a sagittal MRI scan of the hand and distal forearm, likely acquired with a fluid-sensitive sequence (T2-weighted or STIR). Osseous structures, flexor tendons, and surrounding soft tissues are well visualized. Abnormal hyperintense signal is present around the flexor tendons, suggesting inflammatory or degenerative pathology.<LOC>The primary region of interest is the flexor tendon compartment near the metacarpophalangeal (MCP) joint {"point": [588, 415], "label": "flexor tendon at MCP joint"}, extending proximally toward the carpal tunnel {"box_2d": [485, 235, 845, 665], "label": "flexor tendon compartment with abnormal signal"}, where the flexor digitorum tendons demonstrate abnormal signal intensity and morphology.</LOC><VISUAL>The flexor tendons appear thickened with irregular contours and increased intratendinous signal. Hyperintense fluid surrounds the tendon sheaths, consistent with tenosynovitis. Mild adjacent soft tissue edema is also present. No definite muscle fiber disruption or focal hematoma is identified.</VISUAL><KNOWLEDGE> - Flexor tendon pathology on MRI includes tendon thickening, intratendinous signal abnormality, and surrounding tenosynovial fluid. - Muscle strain primarily affects muscle bellies rather than tendons. - Nerve compression typically presents with nerve signal changes rather than primary tendon abnormalities. </KNOWLEDGE> <CON> The combination of tendon thickening, intratendinous hyperintensity, and surrounding tenosynovial fluid is characteristic of flexor tendon pathology. The absence of muscle injury or primary nerve abnormalities supports this diagnosis over alternative options. </CON> <ANSWER>Findings are most consistent with flexor pathology.</ANSWER>

(b) MRI Modality

Figure 12: We visualize sample cases from five distinct imaging modalities. For each case, the input image, question, ground truth, and the response predicted by our method are presented.(part 1.)

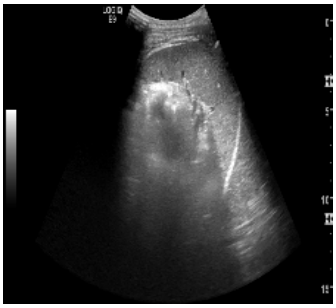


What is the specific condition portrayed in this image?
Options:
A. Conjunctivitis
B. Cataracts
C. Choroidal Neovascularization (CNV)
D.Retinitis pigmentosa

OCT Sample

This is a high-resolution OCT scan of the macular region, showing preserved retinal layer stratification overall. A focal dome-shaped elevation of the retinal pigment epithelium (RPE) and overlying neurosensory retina is present centrally, with disruption of normal architecture limited to this area. No diffuse retinal thinning, widespread cystoid changes, or outer retinal atrophy are observed. <LOC>The primary region of interest is the central macula, where a dome-shaped elevation of the RPE and neurosensory retina is present {"point": [235, 608], "label": "focal RPE elevation"}, accompanied by a hyporeflective sub-RPE space suggestive of fluid {"box_2d": [118, 436, 355, 785], "label": "sub-RPE fluid and neovascular material"}. Surrounding retinal layers appear largely normal.</LOC> <VISUAL>A localized dome-shaped elevation of the RPE is seen centrally, creating a hyporeflective space beneath it consistent with serous detachment. Hyperreflective material is present in the sub-RPE space, suggestive of neovascular tissue. The overlying neurosensory retina is elevated but otherwise intact, with no diffuse intraretinal cysts or widespread retinal thickening. </VISUAL><KNOWLEDGE>PED: Separation of RPE from Bruch's membrane, appearing as a dome-shaped elevation.Signal Markers: Sub-RPE hyperreflectivity (SHRM) indicates neovascular/fibrovascular tissue; hyporeflectivity indicates exudative fluid.Clinical Context: These features are hallmarks of active exudative maculopathy</KNOWLEDGE><CON> The focal PED and sub-RPE hyporeflective space confirm serous detachment. The presence of internal hyperreflective material (SHRM) indicates that the detachment is driven by neovascular tissue rather than simple fluid. Preserved surrounding retinal architecture further supports a localized, active CNV process. </CON> <ANSWER>Findings are most consistent with choroidal neovascularization (CNV).</ANSWER>

(c) MRI Modality



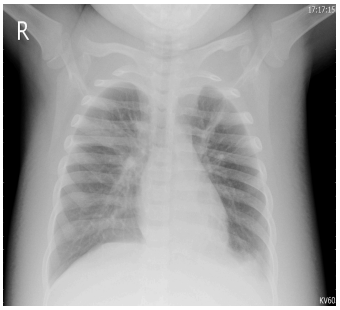
What part is shown in this ultrasound image?
Options:
A. Muscle
B. Small intestine
C.spleen
D. Adrenal gland

Ultrasound Sample

This is a grayscale (B-mode) ultrasound image obtained from the left upper quadrant using a subcostal or intercostal approach. A large, solid organ with homogeneous echotexture occupies much of the image. Rib shadowing is seen superiorly, and overall image quality is adequate with normal depth and penetration. No surrounding free fluid or mass effect is identified.<LOC> The primary region of interest is a large, homogeneous organ in the left upper quadrant {"box_2d": [180, 400, 860, 730], "label": "spleen"} with a smooth, crescent-shaped contour, located beneath the rib cage {"point": [105, 395], "label": "rib shadowing"} and oriented obliquely from upper left to lower right. </LOC> <VISUAL>The organ demonstrates uniform hypoechoic to isoechoic echotexture with a smooth convex outer margin. Mild posterior acoustic enhancement is present, consistent with normal solid organ appearance. No focal lesions, cysts, or parenchymal heterogeneity are observed. The margins are well defined, and adjacent structures show no abnormal findings.</VISUAL> <KNOWLEDGE> - The spleen is visualized in the left upper quadrant on abdominal ultrasound and typically has a crescent (“boomerang”) shape. - It shows homogeneous echotexture and mild posterior acoustic enhancement. - Other structures such as bowel, adrenal gland, or muscle do not share this size, shape, or echotexture. </KNOWLEDGE> <CON> The location, size, crescent shape, and homogeneous echotexture with posterior acoustic enhancement are characteristic of the spleen. Alternative structures do not match these combined ultrasound features. </CON><ANSWER>The imaged structure is the spleen. </ANSWER>

(d) OCT Modality

Figure 12: We visualize sample cases from five distinct imaging modalities. For each case, the input image, question, ground truth, and the response predicted by our method are presented.(part 2.)



What content appears in this image?
Options:
A. Arm
B.Chest
C. Liver
D. Heart

X-Ray Sample

This is a posteroanterior (PA) chest X-ray. The bony thorax, lungs, mediastinum, heart, and diaphragm are clearly visualized. Lung fields appear clear with normal vascular markings. The cardiac silhouette is centrally located and within normal size limits. The diaphragm is smooth, with the right hemidiaphragm slightly higher than the left, which is physiologic. No acute abnormalities are identified.
<LOC>The primary regions of interest include the bilateral lung fields {"box_2d": [310, 145, 875, 885], "label": "bilateral lung fields"}, mediastinum and heart {"box_2d": [450, 435, 800, 630], "label": "heart and mediastinum"}, diaphragm {"box_2d": [830, 150, 965, 870], "label": "diaphragm"}, and thoracic cage (ribs, clavicles, and spine). The right-side orientation marker (“R”) {"point": [115, 85], "label": "right-side marker"} aids anatomic localization.<\LOC>
<VISUAL>The ribs, clavicles, and thoracic spine are intact and symmetrically aligned. Both lungs show normal aeration without consolidation, effusion, or pneumothorax. The trachea is midline. The cardiac silhouette demonstrates a normal cardiothoracic ratio. Costophrenic angles are sharp, and the diaphragm contours are smooth.</VISUAL><KNOWLEDGE>
- A PA chest X-ray visualizes the lungs, heart, mediastinum, diaphragm, and thoracic cage.- Normal findings include clear lung fields, sharp costophrenic angles, and a normal cardiothoracic ratio.</KNOWLEDGE><CON>
The image contains the full thoracic cavity with characteristic structures of a chest radiograph, including lungs, heart, mediastinum, and bony thorax. Other options represent partial structures rather than the complete imaging field.
</CON><ANSWER>This image represents the chest.</ANSWER>

(e) X-Ray Modality

Figure 12: We visualize sample cases from five distinct imaging modalities. For each case, the input image, question, ground truth, and the response predicted by our method are presented.(part 3.)