MEDSAGE: Learning Structured Medical Visual Reasoning via Self-Corrective Reinforcement Learning

Anonymous ACL submission

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 **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 training datasets, **SAGE-sft20K** and **SAGE-rl10K**, to support this training paradigm. Within this framework, SFT induces consistent structured reasoning across tasks, while self-corrective RL further improves answer correctness by enabling the model to revise erroneous predictions during training. encouraging self-check guided correction of erroneous predictions. Experiments on five medical benchmark datasets show that MEDSAGE achieves competitive or improved performance across diverse medical VQA benchmarks. Additional analyses further examine robustness and reasoning faithfulness. Code and data will be publicly released

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(Chen et al., 2024a) 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(Chen et al., 2024b). Due to the coarse granularity of supervision signals(Wu et al., 2025a), 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(Zhou et al., 2025). Nevertheless, these approaches typically rely on the generation of unconstrained free-form language, 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 trajectories of learned reasoning. 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. Existing medical datasets rarely provide supervision

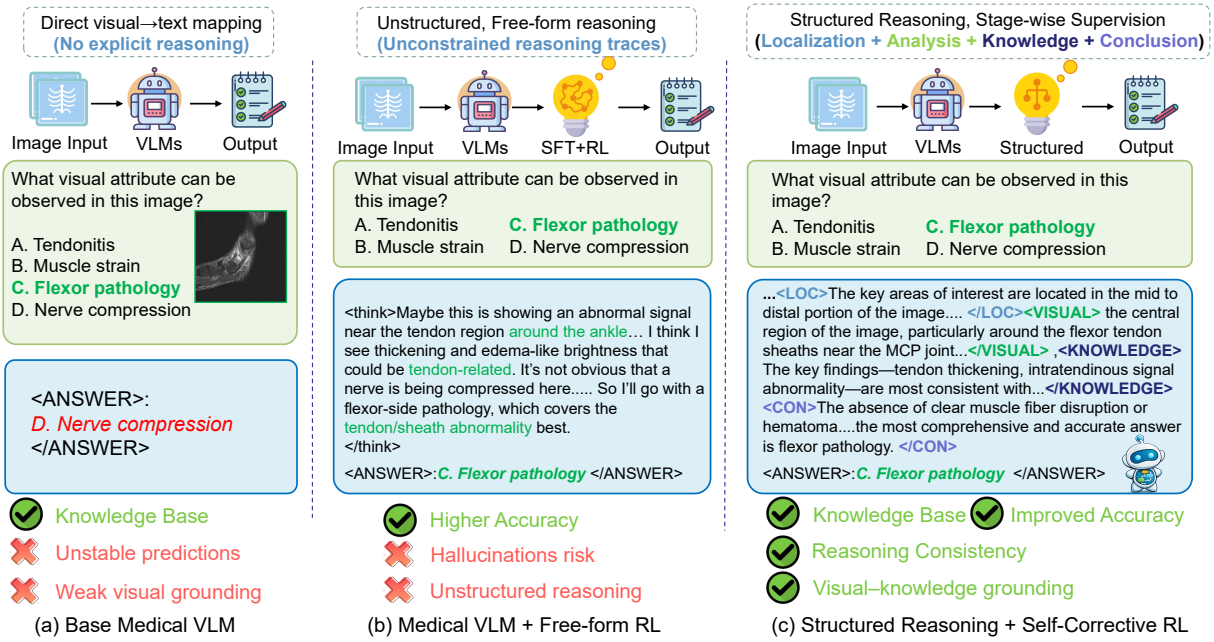


Figure 1: Motivation and overview of structured reasoning in medical vision-language models. We compare (a) base medical VLMs with direct image-to-answer mapping, (b) SFT+RL models with free-form reasoning that often leads to shortcut reasoning, and (c) MEDSAGE, which adopts stage-wise structured reasoning and self-corrective reinforcement learning to promote grounded medical visual reasoning.

signals that reflect how clinicians progressively reason from visual evidence to diagnostic conclusions(Ye and Tang, 2025). 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 conclusion generation (LVKC)—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. Figure 1 contrasts direct, free-form, and structured reasoning paradigms, highlighting the advantages of stage-wise structured supervision. SFT equips the model with multimodal analytical capabilities, while RL further aligns model behavior with the proposed reasoning through self-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 for structured medical visual reasoning.**MEDSAGE formulates reasoning as a sequence of clinically aligned stages, and is supported by two curated datasets, **SAGE-sft20K** and **SAGE-rl10K**, enabling process-level supervision via SFT and RL.

- **We introduce a self-corrective reinforcement learning mechanism.** By rewarding successful self-check-based error correction during training, MEDSAGE guides exploration toward structured and consistent reason-

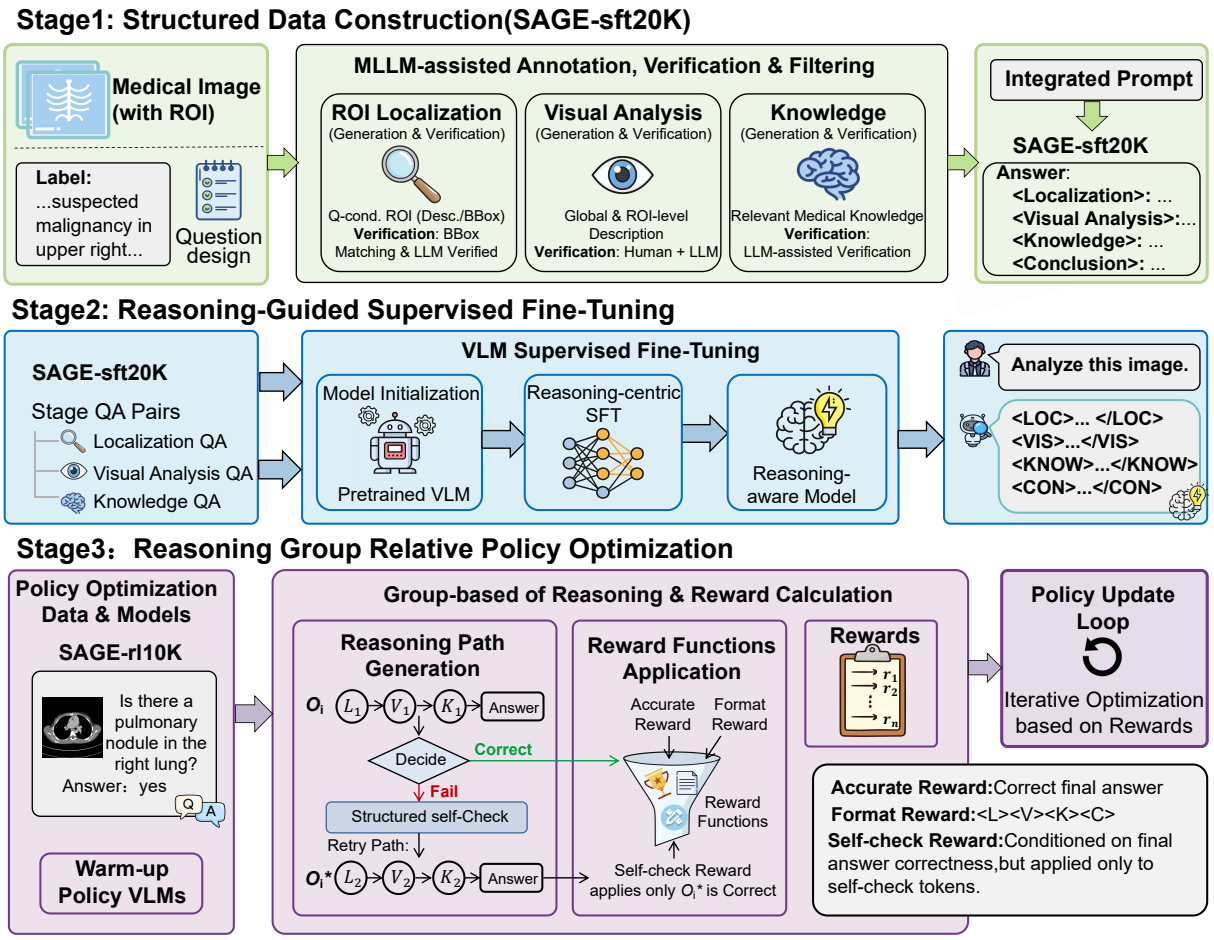


Figure 2: Illustration of the proposed MEDSAGE framework. Stage 1 constructs structured medical reasoning data. Stage 2 performs reasoning-guided supervised fine-tuning to induce stage-wise reasoning behaviors. Stage 3 applies GRPO with a **self-corrective** signal, assigning reinforcement credit to the self-check step only when it successfully corrects an initial error.

ing without relying on free-form reasoning at inference time.

2 Related Work

2.1 Medical Vision-Language Models

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., 2018a) 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., 2023a) 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 HuatuoGPT-Vision (Chen et al., 2024c) and proprietary models like Lingshu (Xu et al., 2025) have set new standards for zero-shot performance. However, most existing open-source medical VLMs rely heavily on 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 Reasoning in Medical VLMs

Reasoning is essential for transparent clinical decision-making. CoT prompting (Wei et al., 2022) has been shown to elicit multi-step reasoning in large language models and has been adapted to medical applications to improve interpretability (Singhal et al., 2023). However, directly applying free-form CoT to medical VLMs is challenging, as medical visual reasoning requires precise grounding of visual evidence prior

to clinical interpretation(Le-Duc et al., 2025; Kim et al., 2025). Unconstrained reasoning often leads to visual-text misalignment and hallucinated explanations (Xu et al., 2024). RL has been increasingly adopted to enhance reasoning in large language models, including Group Relative Policy Optimization(GRPO)-based optimization frameworks such as DeepSeek-R1 (Guo et al., 2025), which encourage extended reasoning through group-wise relative comparison. In the medical multimodal domain, several RL-based methods have been proposed. VITAR (Chen et al., 2025) and MedEYES (Zhu et al., 2025a) apply RL to improve visual perception and answer accuracy, while Med-R1 (Lai et al., 2025b) and MedVLM-R1 (Pan et al., 2025b) extend RL-based alignment to medical VQA settings. Despite these advances, existing approaches primarily define rewards at the outcome level, focusing on final predictions, without explicitly constraining the structure of intermediate reasoning.

3 Methodology

Our work addresses three challenges in structured medical visual reasoning: (1) **Coarse image–answer supervision** that induces shortcut learning; (2) **Visual–text misalignment** caused by free-form reasoning; and (3) the **lack of effective self-correction mechanisms** in standard SFT and RL training.

Figure 2 overviews the proposed MEDSAGE framework, which formulates medical visual reasoning as a sequence of clinically meaningful stages. Figure 3 provides an intuitive comparison between structured and unstructured training paradigms. 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) self-corrective reinforcement learning to improve answer correctness while encouraging grounded, structured reasoning.

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 three sources, including DeepLesion(Yan et al., 2018), Roboflow(Alexandrova et al., 2015), and PubMedVision(Chen et al., 2024c) (see Appendix A for details). Each image is accompanied by region-level

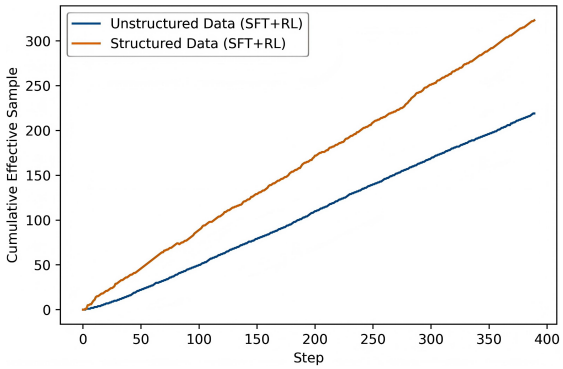


Figure 3: Cumulative effective samples versus training steps under the SFT+RL paradigm with 700 training samples.

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.

RoI normalization. To increase the diversity of region specifications, we 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 (Stage 1), 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, GPT-4o(Hurst et al., 2024), to jointly analyze the RoI and the global image context:

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

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.

Reasoning Path 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 (Stage 2), we perform reasoning-based SFT to initialize the model with stable and explicit LVKC-structured medical reasoning.

For SAGE-sft20K, 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 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:

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

where the target sequence explicitly follows the fixed stage order. In addition, a small number of stage-level QA pairs are included to stabilize stage-specific supervision.

This SFT warm-up stage enables the model to produce complete, stage-aligned medical reasoning paths with clear structural boundaries and flexible content realization.

3.3 RL with Self-Corrective Structured Exploration

To further refine LVKC-structured medical reasoning beyond supervised imitation, we adopt a GRPO-based reinforcement learning framework with a self-corrective auxiliary signal to improve exploration under sparse rewards. The self-corrective mechanism is used only during training to shape reasoning behavior and introduces no inference-time overhead.

Format Reward. To enforce the fixed LVKC stage order, we introduce a rule-based format reward. Let $\mathcal{S}^*$ denote the ordered sequence of required stage tokens and $\phi(o)$ extract special tokens from output o. The format reward is defined as

$$r_{\mathrm{fmt}} = \mathbb{I}[\phi(o) = \mathcal{S}^*], \tag{8}$$

which provides dense structural supervision during RL training.

Self-Check-Guided Retry. Given an image $\mathcal{V}$ and query q, the policy π_θ first samples a reasoning trajectory τ_1 with answer a_1. If a_1 is incorrect, the same model performs a structured self-check over τ_1 and samples a revised trajectory τ_2 with answer a_2. We define a binary self-check success reward

$$r_{\mathrm{sc}} = \mathbb{I}[a_1 \neq y \wedge a_2 = y], \tag{9}$$

which is non-zero only when self-checking successfully corrects an error.

Self-Check Reward under GRPO. The self-check reward is applied exclusively to tokens generated in the self-check step. Let m_t indicate self-check tokens. The corresponding advantage is

$$A^{(\mathrm{sc})} = \alpha r_{\mathrm{sc}}, \tag{10}$$

and the GRPO objective for self-checking is

$$\mathcal{L}_{\mathrm{sc}} = -A^{(\mathrm{sc})} \sum_t m_t \log \pi_\theta(o_t \mid o_{<t}, \mathcal{V}, q). \tag{11}$$

The revised trajectory τ_2 is optimized separately using standard outcome-based rewards, while all other tokens receive zero advantage.

Method	RAD.	SLAKE	PathVQA	PMC.	MMMU	Average	Δ (%)
Proprietary models							
GPT-4.1	65.0	72.2	55.5	55.2	75.2	64.6	-4.2
Claude-Sonnet-4	67.6	70.6	54.2	54.4	74.6	64.2	-4.6
Gemini-2.5-Flash	68.5	75.8	55.4	55.5	76.9	66.4	-2.4
General-purpose Models							
LLaVA-v1.5-8B(Liu et al., 2023)	54.2	59.4	54.1	36.4	38.2	48.4	-20.4
LLaVA-Next-7B(Liu et al., 2024)	52.6	57.9	47.9	35.5	33.1	45.4	-23.4
LLaVA-Next-13B(Liu et al., 2024)	55.8	58.9	51.9	36.6	39.3	48.5	-20.3
Qwen2.5-VL-7B(Bai et al., 2025)	67.3	69.5	63.9	50.4	56.7	61.6	-7.2
InternVL3-8B(Zhu et al., 2025b)	67.3	69.7	65.7	52.7	–	63.8	-5.0
Medical-specific non-reasoning VLMs							
Med-Flamingo(Moor et al., 2023)	45.4	43.5	54.7	23.3	28.3	39.0	-29.8
RadFM(Wu et al., 2025a)	50.6	34.4	38.7	25.9	27.0	35.3	-33.5
LLaVA-Med-7B(Li et al., 2023b)	51.4	48.6	56.2	24.7	36.9	43.5	-25.3
HuatuoGPT-V-8B(Chen et al., 2024c)	59.4	66.8	59.8	51.4	56.7	58.8	-10.0
Lingshu-7B(Xu et al., 2025)	62.2	78.9	**71.7**	55.6	70.0	67.6	-1.2
Medical-specific reasoning VLMs							
Med-R1(Lai et al., 2025a)	55.9	55.1	53.3	45.8	32.7	48.5	-20.3
MedVLM-R1(Pan et al., 2025a)	61.4	56.1	55.2	44.8	35.5	50.6	-18.2
ViTAR(Chen et al., 2025)	70.1	**80.8**	67.0	57.2	**72.0**	**69.4**	+0.6
MedEyes(Zhu et al., 2025a)	70.7	79.1	64.8	55.3	59.7	65.9	-2.9
MEDSAGE (Ours)	**70.4**	79.8	68.7	**58.3**	67.2	68.8	**0.0**

Table 1: Comparison across five medical benchmarks. Δ denotes the *absolute difference in average accuracy (percentage points)* compared to MEDSAGE. Bold numbers indicate the best result among open-source VLMs, and gray numbers indicate that the model has been trained on the corresponding dataset.

Total Reward Composition. The final reward is computed using dynamic weight normalization:

$$R_{\text{total}} = \frac{w_{\text{acc}} r_{\text{acc}} + w_{\text{fmt}} r_{\text{fmt}} + \mathbb{I}_{\text{sc}} w_{\text{sc}} r_{\text{sc}}}{w_{\text{acc}} + w_{\text{fmt}} + \mathbb{I}_{\text{sc}} w_{\text{sc}}}, \quad (12)$$

where $\mathbb{I}_{\text{sc}}$ indicates whether self-checking is triggered. This normalization keeps reward scales comparable across direct and retry paths and implicitly favors efficient reasoning without additional length penalties.

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., 2018b) 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., 2025) and adopt a standard SFT+RL training paradigm. SFT is performed using LLaMA-Factory (Zheng et al., 2024) 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(Yaowei Zheng, 2025). 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 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 (Rasley et al., 2020) acceleration. More implementation details are provided in the supplementary materials.

4.3 Main Results

Table 1 reports the performance of MEDSAGE across five medical VQA benchmarks. MEDSAGE achieves an average accuracy of 68.8%, demonstrating competitive performance among open-source medical VLMs and outperforming the majority of general-purpose and medical-specific baselines. Compared to prior reasoning-oriented models,

Model	Dataset	Localization	Visual Analysis	Knowledge	Reasoning	Average
Baseline	PathVQA(He et al., 2020)	2.706	2.412	3.059	2.529	2.676
	SLAKE(Liu et al., 2021)	3.014	2.350	3.623	2.614	2.900
	VQA-RAD(Lau et al., 2018b)	2.751	2.320	2.911	2.421	2.600
SFT	PathVQA(He et al., 2020)	2.961	2.625	3.612	3.365	3.140
	SLAKE(Liu et al., 2021)	3.446	2.629	3.965	3.036	3.269
	VQA-RAD(Lau et al., 2018b)	3.256	2.927	3.422	3.140	3.186
SFT+RL	PathVQA(He et al., 2020)	3.146	3.020	3.948	3.644	**3.439**
	SLAKE(Liu et al., 2021)	3.637	2.921	4.096	3.347	**3.500**
	VQA-RAD(Lau et al., 2018b)	3.830	3.447	3.652	3.453	**3.595**

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

Method	VQA-RAD	SLAKE	PathVQA	Average
Vanilla QA SFT	36.7	45.9	35.3	39.3
Vanilla QA SFT+RL	36.2	43.2	46.3	41.9
Reasoned QA SFT	67.4	72.8	64.7	68.3
Reasoned QA SFT+RL	**68.4**	77.8	66.1	70.7
RPA QA SFT	67.3	77.4	66.1	70.2
RPA QA SFT+RL	68.1	**78.8**	**67.0**	**71.3**

Table 3: Main ablation results on representative medical VQA benchmarks. Vanilla QA uses original QA supervision, Reasoned QA incorporates annotated reasoning paths, and RPA QA further augments reasoning paths via Reasoning Path Augmentation.

MEDSAGE substantially outperforms MedVLM-R1 and Med-R1, and achieves performance comparable to ViTAR, while showing more consistent generalization across datasets. On radiology-focused benchmarks, MEDSAGE attains 70.4 on VQA-RAD and 79.8 on SLAKE. It also achieves the best open-source result on PMC-VQA (58.3) and competitive performance on MMMU-Med, demonstrating robustness under challenging multimodal medical reasoning scenarios.

Overall, MEDSAGE delivers competitive accuracy across five benchmarks, indicating good cross-dataset generalization under the structured reasoning training paradigm.

4.4 Analysis on Structured Thinking

We follow LLaVA-Med (Li et al., 2023b) to use the GPT API for automated response scoring. Table 2 analyzes the effect of structured reasoning across different training stages, where the baseline is evaluated using few-shot prompting. Compared to the baseline, SFT consistently improves response quality across all dimensions. Further gains are observed with SFT+RL, particularly in localization and reasoning (e.g., on VQA-RAD, reasoning improves from 3.1 to 3.4, and localization from 3.2 to 3.8). This trend indicates that SFT+RL encourages more structured and coherent reasoning.

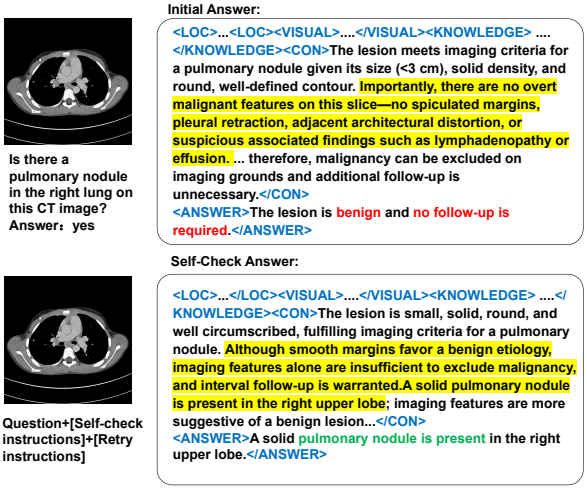


Figure 4: Self-corrective example during the RL stage. The model first generates an initial reasoning trajectory that leads to an incorrect answer, and then revises the erroneous intermediate reasoning through a self-check-guided correction step.

4.5 Ablation Study

Effectiveness of Reasoning Path Augmentation. Table 3 presents an ablation study comparing different reasoning supervision schemes. Compared to Vanilla QA SFT, incorporating explicit reason-

ing supervision (Reasoned QA SFT) leads to substantial performance gains across all benchmarks, improving the average accuracy from 39.3% to 68.3%. Building upon this, RPA further improves the overall performance under SFT, achieving a higher average accuracy than Reasoned QA SFT (70.0 vs. 68.3). While the improvements vary across benchmarks, RPA shows particularly strong gains on SLAKE, suggesting that augmented reasoning paths provide more effective supervision for complex medical reasoning. When combined with reinforcement learning, RPA exhibits more consistent benefits, with RPA SFT+RL achieving the best average accuracy of 71.3%. These results indicate that RPA complements reinforcement learning by providing higher-quality reasoning supervision, leading to more stable performance improvements.

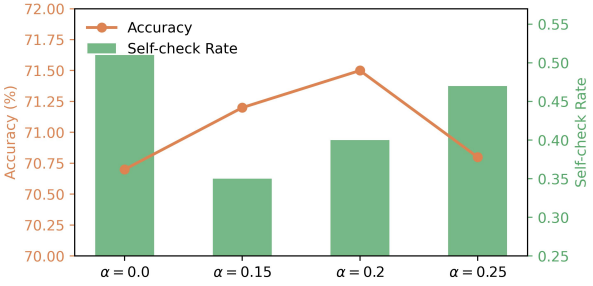


Figure 5: Effect of the weighting coefficient α on accuracy and self-check rate. Accuracy is reported on the left y-axis, and the self-check rate on the right y-axis, indicating the fraction of instances where self-correction is applied.

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 $\alpha = 0.20$ (73.0 average accuracy). Larger α values degrade performance, suggesting that excessive emphasis on self-checking can hinder stable optimization. This indicates that self-checking is most effective when it acts as an auxiliary signal to refine reasoning, rather than dominating the optimization objective.

Effectiveness of Self-Corrective Reinforcement Learning. Table 5 evaluates the effectiveness of the proposed **self-corrective reinforcement learning** under different training settings. Compared to SFT and SFT+RL, incorporating self-corrective learning yields consistent performance gains, with notable improvements on RAD and SLAKE. Figure 4 provides a qualitative example

Config.	RAD	SLAKE	Path	Ave
Reward configuration (Base + Self-Check bonus α)				
Base only ($\alpha = 0$)	69.9	78.3	65.8	71.3
Base + SC ($\alpha = 0.15$)	70.3	78.8	67.4	72.1
Base + SC ($\alpha = 0.20$)	**70.4**	**79.8**	**68.7**	**72.9**
Base + SC ($\alpha = 0.25$)	69.8	78.5	66.2	71.5

Table 4: Ablation study on the self-check bonus α. The base reward includes accuracy and format rewards, while the self-check bonus is applied when self-checking successfully corrects an initial error. Moderate values of α achieve the best overall performance.

Method	RAD	SLAKE	Path	Ave
SFT	67.3	77.4	66.1	70.2
SFT + RL	68.1	78.8	67.0	71.3
SFT + RL + SC	**70.4**	**79.8**	**68.7**	**72.9**

Table 5: Ablation on SFT, SFT+RL, and SFT+RL augmented with Self-Corrective(SC) RL.

of self-corrective behavior during the RL stage, illustrating how the model revises an initial incorrect reasoning trajectory through self-check-guided correction. This suggests that explicitly correcting intermediate reasoning errors during training 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 under the proposed **self-corrective** training scheme.

5 Conclusion

In this work, we present MEDSAGE, a structured reasoning–guided framework for medical vision–language models trained under supervised fine-tuning and reinforcement learning. By organizing medical visual reasoning into clinically aligned stages and introducing a **self-corrective reinforcement learning mechanism**, 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

Jiawei Chen, Dingkang Yang, Yue Jiang, Yuxuan Lei, and Lihua Zhang. Miss: A generative pre-training and fine-tuning approach for med-vqa. In *International Conference on Artificial Neural Networks*, pages 299–313. Springer, 2024a.

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.

Qi Chen, Ruoshan Zhao, Sinuo Wang, Vu Minh Hieu Phan, Anton van den Hengel, Johan Verjans, Zhibin Liao, Minh-Son To, Yong Xia, Jian Chen, et al. A survey of medical vision-and-language applications and their techniques. *arXiv preprint arXiv:2411.12195*, 2024b.

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.

Guanghao Zhou, Panjia Qiu, Cen Chen, Jie Wang, Zheming Yang, Jian Xu, and Minghui Qiu. Reinforced mllm: A survey on rl-based reasoning in multimodal large language models. *arXiv preprint arXiv:2504.21277*, 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*, 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.

Jiarui Ye and Hao Tang. Multimodal large language models for medicine: A comprehensive survey. *arXiv preprint arXiv:2504.21051*, 2025.

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, 2018a.

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, 2023a.

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.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, et al. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 7346–7370, 2024c.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

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.

Khai Le-Duc, Duy MH Nguyen, Phuong TH Trinh, Tien-Phat Nguyen, Nghiem T Diep, An Ngo, Tung Vu, Trinh Vuong, Anh-Tien Nguyen, Mau Nguyen, et al. S-chain: Structured visual chain-of-thought for medicine. *arXiv preprint arXiv:2510.22728*, 2025.

Soo Yong Kim, Suin Cho, Vincent-Daniel Yun, and Gyeongyeon Hwang. Medclm: Learning to localize and reason via a cot-curriculum in medical vision-language models. *arXiv preprint arXiv:2510.04477*, 2025.

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.

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.

Chunzheng Zhu, Yangfang Lin, Shen Chen, Yijun Wang, and Jianxin Lin. Medeyes: Learning dynamic visual focus for medical progressive diagnosis. *arXiv preprint arXiv:2511.22018*, 2025a.

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.

Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging*, 5(3): 036501–036501, 2018.

Sonya Alexandrova, Zachary Tatlock, and Maya Cakmak. Roboflow: A flow-based visual programming language for mobile manipulation tasks. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 5537–5544. IEEE, 2015.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025b.

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

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.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

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, 2018b.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the*

62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), Bangkok, Thailand, 2024. Association for Computational Linguistics. URL http://arxiv.org/abs/2403.13372.

Shenzhi Wang Zhangchi Feng Dongdong Kuang Yuwen Xiong Yaowei Zheng, Junting Lu. Easyr1: An efficient, scalable, multi-modality rl training framework. https://github.com/hiyouga/EasyR1, 2025.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3505–3506, 2020.

A Dataset Construction Details

A.1 Source Datasets

Our dataset integrates multiple publicly available medical imaging sources with heterogeneous forms of region supervision, including **DeepLesion**(Yan et al., 2018), **Roboflow**(Alexandrova et al., 2015), and **PubMedVision**(Chen et al., 2024c).

DeepLesion and Roboflow provide explicit Region of Interest (RoI) annotations in the form of bounding boxes with corresponding clinical labels. These annotations are treated as strong RoI grounding signals and are used to construct region-aware medical reasoning questions.

PubMedVision does not provide explicit RoI annotations. Instead, we extract implicit localization cues from natural language descriptions (e.g., anatomical locations or laterality) and treat them as weak RoI grounding signals, specifying approximate regions relevant to the clinical question.

Furthermore, recognizing that the large-scale nature of PubMedVision might introduce overlaps with clinical evaluation benchmarks (e.g., VQA-RAD(Lau et al., 2018b), SLAKE(Liu et al., 2021), and PathVQA(He et al., 2020)), we implemented a rigorous de-duplication pipeline. Specifically, we employed the Perceptual Hashing (pHash) algorithm to generate visual fingerprints for all 1.3M images in PubMedVision and cross-referenced them against the evaluation sets. Any samples exhibiting high perceptual similarity were systematically filtered out. This process effectively mitigates the risk of data contamination, ensuring that the SOTA performance reported in our results reflects genuine medical reasoning and generalization capabilities rather than data leakage or memorization.

A.2 Metadata Extraction

We treat aligned image–label pairs as the basic units for data construction. To handle data heterogeneity, we apply basic quality filtering and normalization across sources. Samples with reliable visual content and labels are retained directly, while weakly labeled cases are selectively enriched using high-capacity vision-language models for semantic completion. Samples failing basic reliability checks are excluded.

A.3 Data Format

We follow the LLaVA-style (Li et al., 2023b) conversational data format, as shown in Fig. 6. Each sample consists of a user message containing the medical image and task instructions, and a corresponding assistant response that encodes the structured output.

Dataset	Source	Data Size
Initial Data Pool	DeepLesion	26851
	Roboflow	15057
	PubMedVision	19097
Total		**61005**

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

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

Figure 6: LLaVA-style conversational data format used for structured medical reasoning supervision.

A.4 Details of SAGE-SFT20K Construction

Starting from approximately 60K initial samples, we construct a curated SFT dataset (SAGE-SFT20K) with explicit multi-stage medical reasoning trajectories aligned with our framework (*Localization* → *Visual Analysis* → *Knowledge Matching* → *Conclusion*). The construction pipeline focuses on generating high-quality, stage-consistent reasoning traces suitable for supervised learning.

Stage-wise Trajectory Generation. For each image–label pair, we employ a high-capacity multimodal large language model (GPT-4o) (Hurst et al., 2024) to synthesize a four-stage medical reasoning trajectory. Each stage is generated using a dedicated system prompt that specifies the role, inputs, and constraints of the corresponding reasoning step.

Specifically, the localization stage produces a concise anatomical region description relevant to the clinical question (Figure 8). The visual analysis stage describes observable visual evidence within the specified region in a radiology-style manner (Figure 9). The knowledge matching stage summarizes clinically relevant medical knowledge conditions guided by visual evidence and the answer label for supervision (Figure 10). Finally, the conclusion stage assembles all stage outputs into a structured, label-consistent reasoning trajectory with-

Algorithm 1: Structured Reasoning Trajectory Construction

Input: Medical images I, questions q, answers y, region annotations; Interpretation templates $\mathcal{T}$; Multimodal LLM.

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

foreach (I, q, y) **do**
  Normalize region annotations to obtain a set of RoIs $\mathcal{R}$;
  foreach $r \in \mathcal{R}$ **do**
    Generate visual analysis $\mathrm{Vis}(r)$;
    foreach $t \in \mathcal{T}$ **do**
      Generate knowledge matching $\mathrm{Know}(r, t)$ and reasoning output $\mathrm{Out}(r, t)$;
      Add $\tau(r, t) = [\mathrm{Vis}(r), \mathrm{Know}(r, t), \mathrm{Out}(r, t)]$ to $\mathcal{D}_{struct}$;

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

out introducing new information (Figure 11). This prompt-driven, stage-wise generation decomposes end-to-end prediction into interpretable and verifiable intermediate steps, closely reflecting clinical reasoning workflows.

Quality Control and Consistency Verification. To ensure cross-stage coherence, we apply automated consistency verification to all synthesized trajectories. A large vision-language model is used to detect logical conflicts, unsupported conclusions, incorrect knowledge associations, and inconsistencies across stages. Samples failing verification are discarded or regenerated, resulting in a filtered set of reasoning trajectories with strong evidence traceability.

Manual Review. We further conduct targeted manual inspection on randomly sampled trajectories to address subtle issues not captured by automated checks, such as imprecise medical phrasing or borderline diagnostic cases. Feedback from manual review is used to iteratively refine synthesis prompts and verification rules. Figure 7 presents the statistical overview of the resulting dataset, illustrating the distribution of imaging modalities, anatomical regions, and instruction types.

A.5 Details about SAGE-rl10K

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.

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 (Yaowei Zheng, 2025) for reinforcement learning. All experiments were performed with bfloat16 precision and DeepSpeed (Rasley et al., 2020) acceleration.

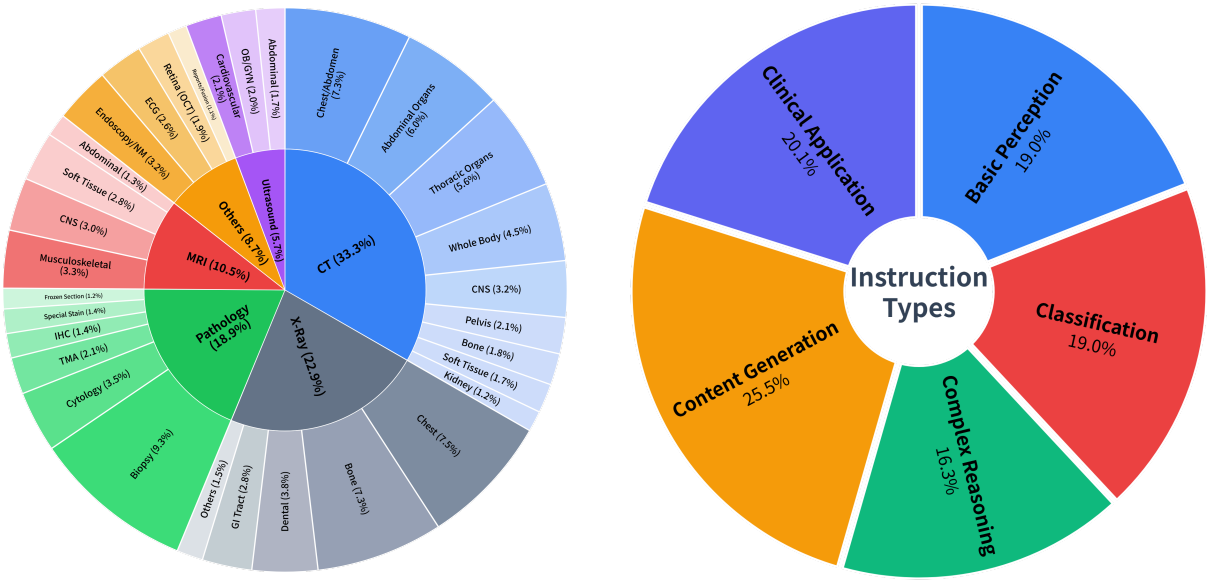
Supervised Fine-Tuning (SFT) Supervised fine-tuning is performed using the **LLaMA-Factory** framework. We fine-tune all parameters of the language model while freezing the vision encoder to reduce memory consumption and stabilize training. The model is trained for 4 epochs with a learning rate of 1×10^{-5}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



(a) Distribution of imaging modalities and anatomical regions.

(b) Distribution of instruction types.

Figure 7: Dataset composition overview.

Rewards are computed based on the outcome reward and trajectory-level consistency signals, and are normalized within each batch to stabilize training.

C Case Study

We present qualitative case studies to illustrate the behavior of our method across different medical imaging modalities. Representative examples are selected to highlight how the model performs stage-wise medical reasoning under diverse visual and clinical conditions.

Figures 12 show sample cases from five imaging modalities, including CT, MRI, OCT, Ultrasound, and X-ray. For each case, we visualize the input image, the associated clinical question, the ground-truth answer, and the response generated by our method.

These examples illustrate how the model produces structured intermediate reasoning and final answers across different imaging modalities.

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: System prompt used for the *Localization* stage of reasoning trajectory generation.

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: System prompt used for the *Visual Analysis* stage of reasoning trajectory generation.

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: System prompt used for the *Knowledge Matching* stage of reasoning trajectory generation.

Role
You are a content integrator.

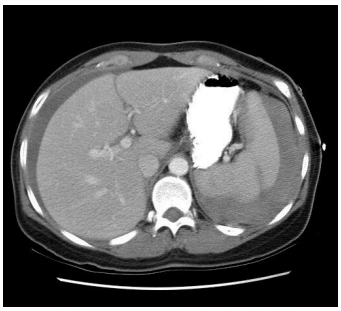
Task
You are given structured components of a medical reasoning example: `[LOC]`, `[VIS]`, `[KNO]`, `[ANS]`.
Your task is to assemble them in order without adding new information, new interpretations, or narrative content.

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: System prompt used for the *Conclusion* stage of reasoning trajectory generation.



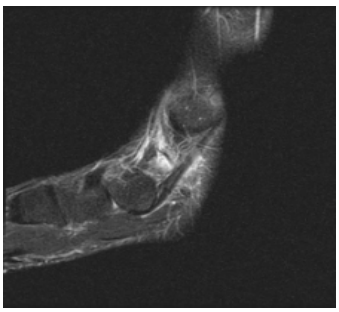
Is there intraperitoneal hemorrhage present in this image?

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



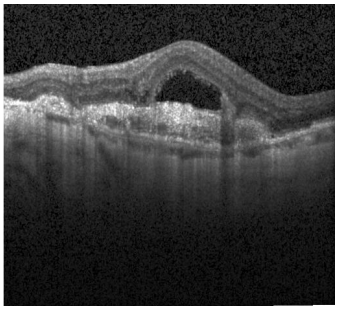
What attribute can be observed in this image?
Options:
A. Tendonitis
B. Muscle strain
C.Flexor pathology
D. Nerve compression

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.)

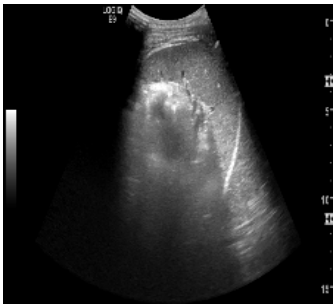


What is the specific condition portrayed in this image?
Options:
A. Conjunctivitis
B. Cataracts
C. Choroidal Neovascularization (CNV)
D.Retinitis pigmentosa

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) OCT Modality



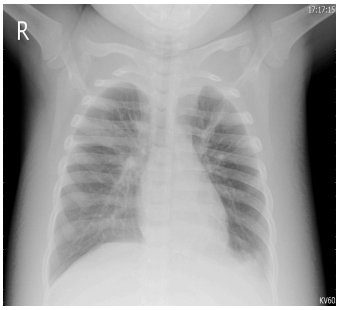
What part is shown in this ultrasound image?
Options:
A. Muscle
B. Small intestine
C.spleen
D. Adrenal gland

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) Ultrasound 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.)