InfiMed: Low-Resource Medical MLLMs with Advancing Understanding and Reasoning

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) have achieved strong performance in general visual understanding and reasoning; however, their progress in the medical domain remains constrained by the scarcity of informative multimodal medical data and the limited effectiveness of Reinforcement Learning with Verifiable Rewards (RLVR). Moreover, existing work often lacks an in-depth exploration of multimodal medical tasks. To address these issues, during supervised fine-tuning (SFT), we jointly incorporate high-quality textual reasoning data, general multimodal data, and multimodal medical data to enhance foundational medical knowledge while preserving the base model's reasoning capability. Furthermore, to mitigate sparse-information scenarios common in medical datasets, we synthesize reflective-pattern-injected chain-of-thought (CoT) data in addition to standard CoT, endowing the model with structured reflective reasoning and providing a strong initialization for subsequent RLVR training. Based on this training paradigm, we introduce the InfiMed-Series, including InfiMed-SFT-3B and InfiMed-RL-3B, which achieve state-of-the-art performance across seven multimodal medical benchmarks. Notably, InfiMed-RL-3B attains an average accuracy of 59.2%, outperforming larger models such as InternVL3-8B (57.3%), while using only 188K SFT samples and 36K RLVR samples. Finally, we conduct extensive experiments to explore a range of fundamental research questions regarding data composition, reasoning strategies, and training paradigms in multimodal medical models. Our findings provide meaningful insights for the future development of medical MLLMs.

1 Introduction

The rapid development of multimodal large language models (MLLMs) in recent years has marked a transformative phase in artificial intelligence, driving substantial progress across diverse domains. Notably, MLLMs have achieved significant breakthroughs in areas such as object recognition (Yin et al., 2025; Liu et al., 2025d), mathematical reasoning (Zhuang et al., 2025; Peng et al., 2024; Liu et al., 2025b), and graphical user interface (GUI) interaction (Liu et al., 2025a; Luo et al., 2025; Qin et al., 2025), largely attributable to the availability of abundant high-quality multimodal datasets. In contrast, the medical domain remains particularly challenging due to the scarcity of high-quality multimodal data, which severely limits the performance of MLLMs in medical scenarios.

To enhance the medical reasoning capabilities of MLLMs, prior work has primarily relied on large-scale, domain-specific supervised fine-tuning (SFT). For instance, LLaVA-Med (Li et al., 2023) directly utilizes the PMC-15M (Zhang et al., 2023b) dataset for medical concept alignment and instruction following. However, its performance is constrained by the inherent noise of the dataset and the limited amount of reasoning information it provides. Recent studies, such as MedGemma (Sellergren et al., 2025), collect larger and higher-quality medical datasets that cover both textual and multimodal modalities, aiming to further enhance the general medical capabilities of MLLMs. While SFT can be effective, it is highly data-intensive and mainly focuses on memorizing training data (Chu et al., 2025). Building on the success of DeepSeek-R1 (Guo et al., 2025), Reinforcement Learning with Verifiable Rewards (RLVR) has shown significant improvements in exploration and generalization for multimodal tasks (Zhang et al., 2025; Liu et al., 2025a,c). RLVR, which often includes a "cold-start" phase in (MLLMs)(Huang et al., 2025; Peng et al., 2025; Liu et al., 2025a), is also beginning to find notable applications in medical scenarios (Su et al., 2025; Xu et al., 2025).

Despite these ongoing efforts, existing approaches still exhibit notable limitations, which

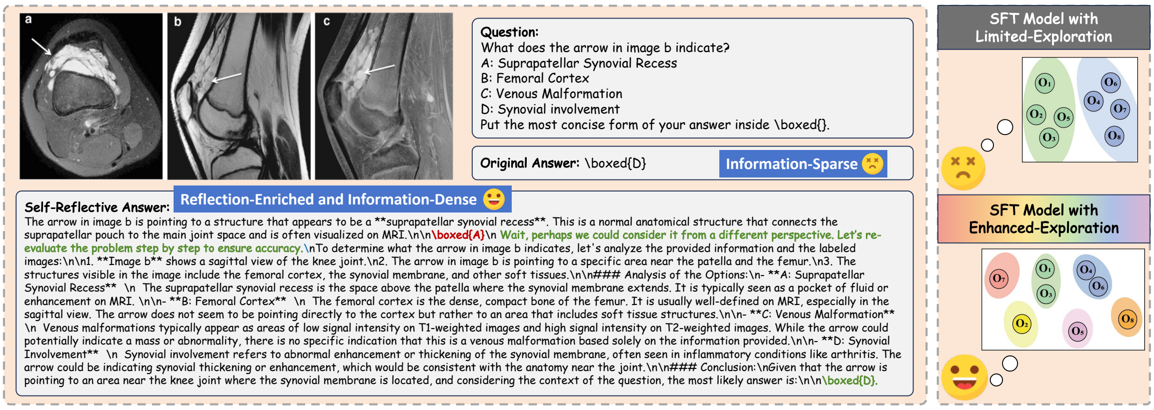


Figure 1: **Left:** Comparison of information-sparse and reflection-enriched, information-dense outputs. **Right:** A model with Enhanced Exploration (bottom) generates a broader, more effective search space, while Limited Exploration (top) results in a narrower, less efficient search.

can be summarized into two key challenges. First, the scarcity of high-quality multimodal medical datasets remains a bottleneck: most existing datasets suffer from sparse information and contain limited explanatory information, which hinders effective model training and results in poor reasoning performance, as shown in Figure 1. Second, although RLVR has been shown to substantially enhance model performance in other domains, its application in medical scenarios remains underexplored. Existing work either lacks extensive exploration across broad benchmarks (Pan et al., 2025; Su et al., 2025) or fails to effectively improve model performance (Xu et al., 2025).

To address the challenges mentioned above, during the SFT stage, we leverage not only multimodal medical data but also general multimodal data to preserve the model's visual perception capabilities, while integrating medical textual data to enhance its domain-specific knowledge. Additionally, we introduce a novel synthesis of reflective-pattern-injected chain-of-thought (CoT) data, effectively addressing the information sparsity present in certain multimodal medical datasets. This approach could also provide a more robust exploratory foundation for subsequent RLVR, enabling a cold-start method with limited resources. Building upon this, we train our InfiMed-SFT-3B model on 188K samples, equipping it with both fundamental reasoning and reflective patterns. We then apply RLVR on top of InfiMed-SFT-3B using 36K samples to obtain InfiMed-RL-3B, further enhancing its exploration capabilities and generalization performance. Extensive experiments show that our InfiMed-series models set new SOTA performance across multiple multimodal medical benchmarks, outperforming similarly-sized models like MedGemma-4B-IT and larger models such as InternVL3-8B, demonstrating the effectiveness of our reflective SFT and RLVR approach. We also investigated the impact of data composition and reasoning strategies through a series of exploratory experiments, yielding valuable insights for the advancement of medical MLLM applications.

In summary, the key contributions of our work are as follows: **(1)** We synthesize reflective-pattern-injected CoT data, equipping the model with initial reflective capabilities and a stronger cold-start foundation for subsequent RLVR. **(2)** We employ a low-resource SFT with 188K samples, enabling the model to develop robust reasoning, comprehension, and reflective patterns. Subsequently, RLVR is applied with 36K samples, effectively boosting the model's exploration capabilities and performance. **(3)** We introduce the InfiMed-series models, **InfiMed-SFT-3B** and **InfiMed-RL-3B**, which achieve SOTA performance among 3B-level MLLMs, with InfiMed-RL-3B outperforming models like MedGemma-4B-IT by 7.64%, and remain competitive even against 7B-level models.

2 Related Work

2.1 Medical MLLMs

In recent years, MLLMs have evolved rapidly and achieved remarkable progress across a wide range of domains, attracting increasing interest in their potential applications within the medical field (Al-Saad et al., 2024). Extensive research efforts have been devoted to enhancing MLLMs' ability to integrate heterogeneous medical data to support critical dimensions in healthcare. Inspired by the success of medical LLMs like HuatuoGPT (Zhang et al., 2023a), Apollo (Wang et al., 2024), and

Med-PaLM series (Singhal et al., 2023, 2025), recent efforts have increasingly focused on extending LLM capabilities to multimodal medical. LLaVA-Med (Li et al., 2023) introduces a biomedical-specialized MLLM trained on a curated figure-caption dataset with self-instructed instruction-following data. The model highlights the potential of cost-efficient training strategies for domain-specific MLLMs. MedGemma (Sellergren et al., 2025) has shown strong generalization across medical vision-language and text-only tasks, demonstrating advanced medical understanding and reasoning on multimodal data. Lingshu (Xu et al., 2025) proposed a domain-specialized multimodal foundation model for medical, supported by a curated dataset enriched with medical VQA, CoT reasoning, and report annotations. While prior work has made notable progress in medical MLLMs, many approaches depend on large model sizes and substantial computational resources, which limit their accessibility and scalability.

2.2 Reasoning in Medical LLMs and MLLMs

Interpretable reasoning remains a central desideratum in medical AI, with recent efforts exploring general CoT prompting (Wei et al., 2022) and program-based logic (Chen et al., 2022) modeling. Although these approaches have shown potential, they typically rely on costly expert-curated annotations (Li et al., 2024b), which limits their scalability in real-world clinical settings. RL offers a compelling alternative by enabling emergent reasoning capabilities without requiring explicit supervision, as demonstrated by recent models such as DeepSeek-R1 (Guo et al., 2025), which achieve notable improvements in reasoning with rule-based reward. Building on this paradigm, RLVR has been used to improve reasoning reliability, with Group Relative Policy Optimization (Shao et al., 2024) known for its efficiency and good performance. This method is now increasingly used to train MLLMs to improve their reasoning ability (Meng et al., 2025; Wang et al., 2025; Tan et al., 2025). With the success of RLVR, several work leverages it on medical MLLMs. MedVLM-R1 (Pan et al., 2025) employs RLVR to explicit reasoning in medical VQA, achieving strong performance and generalization. Its emphasis on reasoning highlights the role of RL in enhancing transparency and trustworthiness in clinical AI systems. GMAI-VL-R1 (Su et al., 2025) explores RLVR to enhance reasoning and reflection in multimodal medical models. By introducing a multi-agent reasoning data synthesis framework, the model outperforms prior models on some complex tasks. Lingshu (Xu et al., 2025) also leverages an RLVR paradigm, achieving strong performance across medical VQA, report generation, and text-only QA. Despite these promising advances, prior work has been limited in its exploration of the RLVR stage.

2.3 Our Distinction

As discussed above, general CoT consists of multi-step natural language reasoning traces that help models learn structured reasoning patterns during SFT and guide response selection during RLVR. However, medical reasoning is far more standardized and constrained by clinical knowledge than open-domain tasks, resulting in limited diversity of valid intermediate steps. This restricts the exploration space of RLVR and hinders the discovery of improved reasoning trajectories. To address this issue, we introduce reflective-pattern–injected CoT data during SFT, equipping the model with initial self-reflection and self-correction capabilities. This design effectively expands the reasoning space available to RLVR, enabling more robust improvements on complex medical tasks. In contrast to prior work that either targets a narrow set of benchmarks or fails to consistently improve SFT performance, we not only achieve reliable gains during the RLVR stage but also conduct extensive experiments to analyze the features in multimodal medical tasks, which could provide several meaningful insights for the future exploration of medical MLLMs.

3 Methodology

In this section, we outline our methodology for advancing multimodal medical understanding and reasoning through RLVR with a self-reflective cold start, which is depicted in Figure 2. Our approach unfolds in two stages: (1) A cold start phase, in which we uniquely integrate general multimodal data with medical text reasoning data to simultaneously enhance image understanding and restore fundamental reasoning skills. Crucially, to address the information sparsity of existing medical datasets, we further synthesize both distilled CoT and self-reflective CoT for SFT, thereby establishing a richer and more exploratory reasoning foundation. (2) A RLVR phase, which enables the model to explore a wider spectrum of reasoning trajecto-

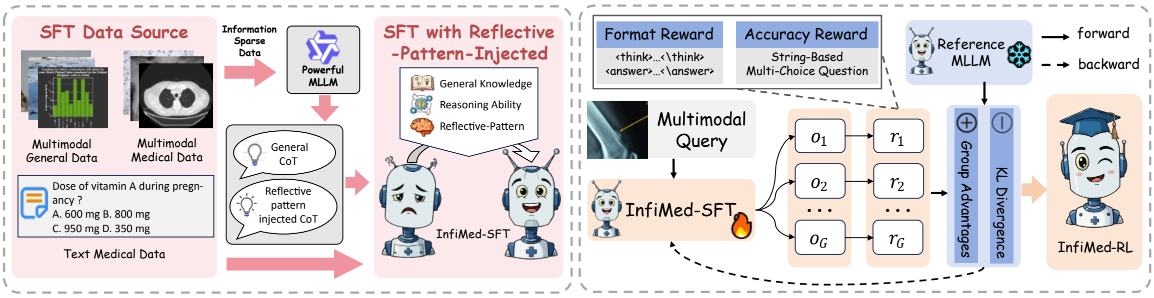


Figure 2: The overall training process of InfiMed-Series models.

ries, thereby producing more robust and clinically faithful multimodal reasoning.

3.1 Reflective-Injected Supervised Fine-Tuning

As mentioned above, since SFT constitutes the foundation for subsequent RLVR, we incorporated not only general multimodal data but also text-based medical reasoning data during SFT to strengthen the model's fundamental multimodal understanding and reasoning capabilities (Sellergren et al., 2025; Xu et al., 2025). However, several existing multimodal medical SFT datasets suffer from insufficient informational richness. For instance, multiple-choice question datasets often only provide the final choice and always lack explicit explanation. To address this limitation, in addition to only generating conventional CoT data to supplement the missing information, we further construct reflective-pattern-injected CoT data, enabling the model to develop more comprehensive and self-corrective reasoning capabilities (Cheng et al., 2024).

The core premise of reflective-pattern-injected is that directly exposing the model to a spectrum of reasoning trajectories, including correct, partially inconsistent, and subtly flawed chains, encourages the development of self-evaluation and error-correction mechanisms.

Formally, given a multimodal medical question q, whose original response consists of insufficient information, consisting of a textual task instruction x and one or more images $\mathcal{I}$, i.e. $q = \{x, \mathcal{I}\}$. We utilize a powerful MLLM (e.g., Qwen2.5-VL-32B-Instruct (Bai et al., 2025)) to generate a batch of candidate responses $\{y_i\}$. Subsequently, leveraging rejection sampling, we partition these candidates into two disjoint subsets: $\{y_i^+\}$, corresponding to correct responses, and $\{y_i^-\}$, corresponding to incorrect responses. For the correct responses $\{y_i^+\}$, we further engage a more advanced MLLM (e.g., Qwen2.5-VL-72B-Instruct (Bai et al., 2025)) to evaluate each response across multiple dimensions, including clinical accuracy, logical reasoning, factual correctness, and completeness.

Finally, we synthesize a reflective-pattern-injected CoT by combining one of the highest-quality responses from $\{y_i^+\}$ with a randomly selected response from $\{y_i^-\}$, thereby creating a novel training instance that emphasizes both reasoning and error-awareness. More details of the reflective-pattern-injected CoT synthesis can be found in the Appendix A.1.

3.2 Reinforcement Learning with Verifiable rewards

3.2.1 Overall Process of RLVR

After the SFT stage with reflective-pattern injection, we use Group Relative Policy Optimization (GRPO) in the RLVR phase to improve stability, building on the method from Deepseek-R1 (Guo et al., 2025). GRPO computes advantages by generating multiple responses for the same query, removing the need for an explicit critic model.

We formally denote the model after the SFT stage with reflective-pattern injection as π_θ, the policy model in RLVR. Given a multimodal medical query q, the policy model $\pi_{\theta_{old}}$ (prior to parameter updates) generates a set of G candidate responses $\{o_i\}_{i=1}^{G}$. For each response o_i, a rule-based reward function $R(o, \text{gt})$ is used to evaluate its quality and assign a score r_i, where gt denotes the ground-truth answer. Based on the collection of rewards $\{r_i\}_{i=1}^{G}$, the group-relative advantages $\{A_i\}_{i=1}^{G}$, which quantify the relative quality of responses within the batch, can be calculated as:

$$A_i = \frac{r_i - mean(\{r_1, r_2, \ldots, r_G\})}{std(\{r_1, r_2, \ldots, r_G\})}, \qquad (1)$$

where $mean(\cdot)$ indicates the average value, and $std(\cdot)$ refers to the standard deviation.

Based on the above group-relative advantages, GRPO updates the policy by maximizing the expected advantage-weighted likelihood ratio. The optimization objective can be formulated as:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{[q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)]}\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}$$
$$\sum_{t=1}^{|o_i|}\left\{\min\left[\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}A_i, \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1-\epsilon, 1+\epsilon\right)A_i\right]\right\}$$
$$-\beta D_{\text{KL}}\left[\pi_\theta \| \pi_{\text{ref}}\right], \quad (2)$$

where the additional Kullback–Leibler term $D_{\text{KL}}[\pi_\theta \| \pi_{\text{ref}}]$ is applied to penalize divergence from the reference policy model π_{ref}, thereby helping to maintain training stability.

3.2.2 Rule-based Reward Construction

Considering the reward function $R(o, \text{gt})$ aims to guide the policy model to learn a suitable and correct reasoning trajectory, we design our total reward R_{total}, which integrates assessments of both output format correctness and accuracy:

$$R_{\text{total}}(o, \text{gt}) = w_{\text{format}} \cdot R_{\text{format}}(o) + w_{\text{acc}} \cdot R_{\text{accuracy}}(o, \text{gt}), \quad (3)$$

where $R_{\text{format}}(o)$ denotes the reward for the correctness of the output format and $R_{\text{accuracy}}(o, \text{gt})$ denotes the reward for the accuracy of the output o relative to the ground-truth result. The non-negative coefficients w_{format} and w_{acc} serve as hyperparameters weighting the relative contribution of the two components, with $w_{\text{format}} + w_{\text{acc}} = 1$.

The format reward $R_{\text{format}}(o)$ assesses whether the output of the policy model π_θ satisfies the predefined format. Notably, $R_{\text{format}}(o) \in \{0, 1\}$, where $R_{\text{format}}(o) = 1$ if all specified format requirements are satisfied; otherwise, $R_{\text{format}}(o) = 0$. Specifically, it verifies two primary aspects:

- **Thinking Progress:** We evaluate whether the model correctly presents its reasoning process according to a predefined format. The model is required to encapsulate its reasoning process and final answer within designated tags.
- **Final Answer Format:** We examine whether the model outputs an explicit final answer, with particular attention to cases where the instructions related to query q require such a response.

The accuracy reward $R_{\text{accuracy}}(o, \text{gt})$ evaluates the correctness of the model output o relative to the ground truth of query q. Importantly, $R_{\text{accuracy}}(o, \text{gt})$ is defined only when the output meets the format constraint, i.e., $R_{\text{format}}(o) = 1$; otherwise, it is zero. This design ensures that the model generates well-structured outputs before being evaluated for correctness. When $R_{\text{format}}(o) = 1$, the computation of $R_{\text{accuracy}}(o, \text{gt})$ depends on the task-specific ground-truth format. The details of task-specific reward functions are presented in the Appendix A.2.

4 Experiment

In this section, we present the experimental setup used to train and evaluate our proposed InfiMed-series models, which are built upon Qwen2.5-VL-3B-Instruct (Bai et al., 2025). We detail the experimental settings and aim to address the following research questions.

- **RQ1:** How do the InfiMed-Series models perform compared with other MLLMs across various medical benchmarks?
- **RQ2:** How do different data types and data numbers influence SFT and RLVR performance?
- **RQ3:** Does reasoning really enhance performance in medical tasks?
- **RQ4:** How do models trained with self-reflection via SFT compare to RLVR-optimized models in their quality and reliability of medical responses?
- **RQ5:** Can our SFT data consistently improve model performance across different base MLLM architectures?
- **RQ6:** Does increasing the amount of RLVR training data lead to further improvements in model performance?

4.1 Experimental Setup

Models. We conduct a comprehensive comparison across a wide range of models. The models include: (1) Proprietary models: GPT-series models (Achiam et al., 2023), Claude Sonnet 4 (Anthropic, 2025), and Gemini-2.5-Flash (Comanici et al., 2025); (2) General open-source models: Qwen2.5-VL series models (Bai et al., 2025), Gemma3 series models (Team et al., 2025) and InternVL series models (Chen et al., 2024c; Zhu et al., 2025); (3) Medical open-source models: MedVLM-R1-2B (Pan et al., 2025), MedGemma-4B-IT (Sellergren et al., 2025), LLaVa-Med-7B (Li et al., 2023), HuatuoGPT-V-7B (Chen et al., 2024b), Lingshu-7B (Xu et al., 2025), BioMediX2-8B (Mullappilly et al., 2024).

Datasets. During the reflective-injected SFT stage, we utilize a total of **188K** samples from three categories: (1) multimodal general data, (2) multimodal medical data, and (3) textual medical data. In the RLVR stage, we utilize **36K** multimodal medical data and general data. An overview of the training datasets is provided in Figure 3.

Evaluation Benchmark. We adopt seven widely

Table 1: **Performance comparison of different MLLMs across various medical vision-language benchmarks.** The best results among models in the 2-4B parameter are **bolded**. MedVLM-R1-2B was trained on OMVQA, and Lingshu was trained on the MMMU- H&M val set.

Model	Size	Accuracy (%)							
		MMMU-H&M	VQA-RAD	SLAKE	PathVQA	PMC-VQA	OMVQA	MedXQA	Avg.
Proprietary Models									
GPT-5	-	83.6	67.8	78.1	52.8	60.0	76.4	71.0	70.0
GPT-4.1	-	75.2	65.0	72.2	55.5	55.2	75.5	45.2	63.4
Claude Sonnet 4	-	74.6	67.6	70.6	54.2	54.4	65.5	43.3	61.5
Gemini-2.5-Flash	-	76.9	68.5	75.8	55.4	55.4	71.0	52.8	65.1
General Open-source Models									
Qwen2.5VL-3B-Instruct	3B	51.3	56.8	63.2	37.1	50.6	64.5	20.7	49.2
Gemma3-4B-IT	4B	34.0	49.9	61.1	43.2	47.9	60.9	20.9	45.4
Qwen2.5VL-7B-Instruct	7B	54.0	65.0	67.6	44.6	51.3	63.5	21.7	52.5
InternVL2.5-8B	8B	53.5	59.4	69.0	42.1	51.3	81.3	21.7	54.0
InternVL3-8B	8B	59.2	65.4	72.8	48.6	53.8	79.1	22.4	57.3
Medical Open-source Models									
MedVLM-R1-2B	2B	35.2	48.6	56.0	32.5	47.6	**77.7**	20.4	45.4
MedGemma-4B-IT	4B	43.7	49.9	76.4	48.8	49.9	69.8	22.3	51.5
LLaVA-Med-7B	7B	29.3	53.7	48.0	38.8	30.5	44.3	20.3	37.8
HuatuoGPT-V-7B	7B	47.3	67.0	67.8	48.0	53.3	74.2	21.6	54.2
Lingshu-7B	7B	-	67.9	83.1	61.9	56.3	82.9	26.7	-
BioMediX2-8B	8B	39.8	49.2	57.7	37.0	43.5	63.3	21.8	44.6
Ours (InfiMed-Series)									
InfiMed-SFT-3B	3B	54.7	58.1	82.0	60.6	53.2	67.0	23.5	57.1
Gemma3-SFT-4B	4B	35.3	59.9	**83.3**	**64.7**	53.3	68.7	21.0	55.2
InfiMed-RL-3B	3B	**55.3**	**60.5**	82.4	62.0	**58.7**	71.7	**23.6**	**59.2**

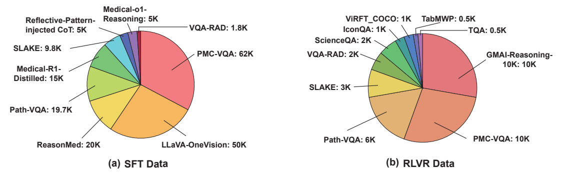


Figure 3: Overview of the training samples for the InfiMed series models in the reflective-injected SFT and RLVR stages.

used multimodal medical benchmarks: MMMU-Health&Medicine (MMMU-H&M) (Yue et al., 2024), VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), PathVQA (He et al., 2020), PMC-VQA (Zhang et al., 2023c), OmniMedVQA (OMVQA) (Hu et al., 2024), and MedXpertQA-Multimodal (MedXQA) (Zuo et al., 2025). These benchmarks span diverse imaging modalities and provide a comprehensive evaluation of image understanding and medical reasoning.

Detailed descriptions of the above settings and additional experimental details are provided in Appendix A.3.

4.2 Results on Various Medical Benchmarks (RQ1 & RQ5)

Table 1 presents a comprehensive comparison of different MLLMs across seven diverse medical vision-language benchmarks. Among all models, proprietary closed-source models consistently outperform both general-purpose and medical-domain open-source models, achieving the highest average accuracy. These models set a strong upper bound, particularly excelling on complex benchmarks such as MMMU-H&M and MedXpertQA, indicating their superior reasoning and image understanding capabilities.

Furthermore, comparisons with existing open-source models show that the InfiMed-series models offer significant performance advantages. Both InfiMed-SFT-3B and InfiMed-RL-3B notably outperform other models of similar scale, achieving average accuracies of 57.1% and 59.2%, respectively, across seven multimodal medical benchmarks.

Notably, our 3B models outperform some larger 7B and 8B models, such as HuatuoGPT-V-7B and InternVL2.5-8B, despite their greater scale. Although a gap remains between InfiMed-RL-3B and Lingshu-7B, our model achieves competitive performance with fewer parameters and using a relatively small amount of data (188K for SFT and 36K for RLVR), compared to Lingshu-7B's 12M samples, HuatuoGPT-V-7B's 1.3M samples, highlighting the efficiency and effectiveness of our training.

Although models such as MedVLM-R1-2B, GMAI-VL-R1-7B (not open-sourced yet), and Lingshu-7B provide some evidence that RLVR can be effective after SFT, they either target only a narrow range of benchmarks or fail to achieve consistent overall gains. By contrast, the 2.1% overall improvement of InfiMed-RL-3B over InfiMed-SFT-3B, along with consistent gains across seven medical benchmarks, clearly demonstrates that RLVR not only could enhance model performance in the medical domain but also complements our SFT phase training, thereby substantiating the effectiveness of RLVR.

To evaluate the model-agnostic robustness of our SFT data, we fine-tuned a fundamentally different backbone, Gemma3-4B-IT, using our training data. Despite substantial architectural differences from previously evaluated models such as Qwen, our data yields significant improvements over the Gemma baseline and achieves state-of-the-art performance within the 3–4B parameter range (excluding InfiMed). Notably, these gains are obtained using only 176K SFT samples, demonstrating that our data is both highly efficient and transferable across heterogeneous MLLM architectures.

4.3 Ablation Study on Data Composition (RQ2 & RQ6)

For both the SFT and RLVR stages, we conduct ablation studies to evaluate how different data types and data scales influence model performance. During SFT, we systematically remove individual data components, including general multimodal data, textual medical data, and reflective-pattern-injected CoT data. The results are summarized in Table 2. Overall, these experiments lead to a key conclusion: ***Unlike general multimodal tasks, medical multimodal problems require the joint integration of visual understanding, textual reasoning, and domain-specific medical knowledge; medical-only training data is insufficient for robust medical MLLM performance.***

Specifically, removing general multimodal data causes substantial performance degradation on visually demanding benchmarks such as OmniMedVQA, highlighting its role in visual-text alignment and general visual reasoning. Excluding textual medical data significantly degrades performance on MMMU-H&M, demonstrating the necessity of domain knowledge, medical terminology, and clinical reasoning patterns. Although reflective-pattern-injected CoT data accounts for only 5K samples, its removal consistently leads to performance drops across most benchmarks, indicating its effectiveness in enhancing multi-step reasoning and self-reflection. Furthermore, replacing reflective CoT with an equal amount of general CoT results in inferior performance, confirming that the reflective formulation itself provides unique benefits.

We further observe that moderate variations in data proportions lead to performance fluctuations across benchmarks, suggesting complementary effects between general multimodal and textual medical data. In the RLVR stage, removing general multimodal data causes performance on MMMU-H&M to fall below the SFT baseline, indicating that RLVR relying solely on medical multimodal data may reduce reasoning capability rather than enhance it. Motivated by this observation, we further investigate how the initialization and data configuration of the RLVR stage affect performance, including the impact of different starting points and increased RLVR training data. Additional ablation results about SFT and detailed description are provided in Appendix A.4.

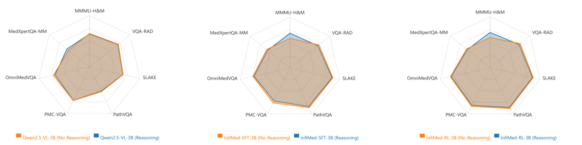


Figure 4: Comparison of direct-answer and reasoning-based prompts on medical benchmarks.

4.4 Analysis of reasoning effectiveness in Medical Scenarios (RQ3)

To explore reasoning effectiveness in medical scenarios, we evaluated the model using two prompts: (1) a direct-answer prompt, where the model is asked to output only the final prediction; (2) a reasoning-augmented prompt, where the model is encouraged to generate intermediate reasoning steps before providing the answer. This setup allows us to examine the impact of explicit reasoning. Results are shown in Figure 4.

Our experiments show that explicit reasoning prompts generally degrade performance on most medical benchmarks, with two exceptions: both InfiMed-SFT and InfiMed-RL benefit on MMMU-H&M, and the general-purpose Qwen2.5-VL-3B-Instruct improves on MMMU-H&M and MedXpertQA. This indicates that explicit reasoning is not universally beneficial for medical-focused MLLMs, even with RLVR. The gains on MMMU-H&M and MedXpertQA can be attributed to their reasoning-intensive feature, which requires multi-step deduction and cross-modal integration. For Qwen2.5-VL-3B-Instruct, reasoning prompts help organize latent knowledge and reduce uncertainty. In contrast, InfiMed models rely on efficient, domain-specific direct answering strategies, and enforcing explicit reasoning can interfere with these pathways, leading to performance degradation. Moreover, most other benchmarks are primarily knowledge- or recognition-driven, where answers can be derived directly from visual cues or domain expertise. In such cases, explicit reasoning introduces unnecessary complexity and increases the risk of hallucina-

Table 2: **Ablation study examining data composition during the training stage.** Δ|Data| denotes the amount of data change applied to the training set. w/o-general, w/o-text, and w/o-refcot refer to training configurations where the general multimodal data, textual medical data, and reflective-pattern-injected CoT data are removed, respectively. gen_mm, text, and general_cot denote the general multimodal data, medical textual data, and general CoT data components included in the training corpus.

Model	Δ|Data|	Accuracy (%)							
		MMMU-H&M	VQA-RAD	SLAKE	PathVQA	PMC-VQA	OMVQA	MedXQA	Avg.
Base Model									
Qwen2.5VL-3B-Instruct	-	51.3	56.8	63.2	37.1	50.6	64.5	20.7	49.2
Ablation Study in SFT Stage on General Multimodal Data									
InfiMed-SFT-3B	-	54.7	58.1	82.0	60.6	53.2	67.0	23.5	57.1
InfiMed-SFT-3B+gen_mm	+20K	54.0	60.7	81.8	55.8	55.5	67.5	22.3	56.8
InfiMed-SFT-3B−gen_mm	−20K	48.0	58.1	82.1	58.3	51.4	66.9	22.6	55.4
Ablation Study in SFT Stage on Medical Text Data									
InfiMed-SFT-3B-w/o-general	−50K	50.0	60.5	80.5	60.4	51.6	59.7	22.6	55.1
InfiMed-SFT-3B+text	+20K	50.7	63.4	80.4	57.3	54.4	67.3	21.4	56.4
InfiMed-SFT-3B−text	−20K	50.7	60.3	82.2	58.1	53.8	67.3	22.6	56.4
InfiMed-SFT-3B-w/o-text	−40K	44.0	61.0	81.6	60.4	51.1	64.7	21.9	54.9
Ablation Study in SFT Stage on Reflective CoT Data									
InfiMed-SFT-3B-w-generalcot	–	50.7	60.5	81.9	57.7	52.7	66.4	23.4	56.2
InfiMed-SFT-3B-w/o-refcot	−5K	50.0	60.1	81.3	60.5	53.1	64.7	22.8	56.1
Ablation Study in RLVR Stage									
InfiMed-RL-3B	-	55.3	60.5	82.4	62.0	58.7	71.7	23.6	59.2
InfiMed-RL-3B-w/o-general	−10K	53.3	60.7	81.9	61.6	58.3	70.0	23.6	58.4

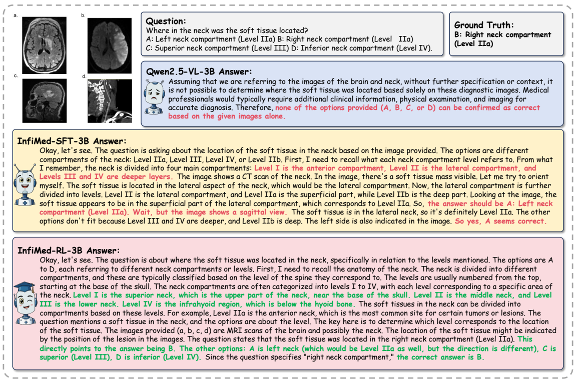


Figure 5: Case study. Red denotes errors or irrelevant content, Green denotes correct or important information.

tion, ultimately harming performance.

To further examine the impact of model scale, we conduct additional experiments with larger models and different model families, including the Lingshu (Xu et al., 2025) and HUATUO (Chen et al., 2024b) series. As shown in Figure 7 in the Appendix, performance degradation caused by explicit reasoning is still observed on several benchmarks at the 7B and 32B scales. These results reinforce our conclusion that ***explicit reasoning is not universally beneficial for medical multimodal tasks and suggest that, even in larger models, explicit reasoning may interfere with performance on medical multimodal tasks.***

4.5 Case Study (RQ4)

Our case study reveals distinct response behaviors across models. Qwen2.5-VL-3B-Instruct adopts a conservative strategy, explicitly acknowledging insufficient medical knowledge and ultimately failing to provide a definitive answer. In contrast, InfiMed-SFT-3B can generate a reasoning chain and exhibit a reflective pattern; however, it still converges on an incorrect conclusion. This indicates that SFT primarily enables the model to imitate the form of reflection, without fully achieving genuine understanding or effective application of reflective reasoning. InfiMed-RL-3B, by comparison, demonstrates a more structured and deliberate reasoning process. Beyond identifying the correct option, it actively examines and evaluates choices, highlighting the role of RLVR in encouraging systematic reasoning rather than reliance on memorized patterns. More case studies are presented in the Appendix A.5.

5 Conclusion

We introduce the InfiMed-Series models, a set of MLLMs specialized for medical tasks. To address the scarcity and sparsity of multimodal medical data, we augmented the training sets with general multimodal and textual medical data and synthesized reflective-pattern-injected CoT data, enabling the models to acquire initial exploratory capabilities and providing a structured foundation for subsequent RLVR. Experimental results across diverse reasoning-intensive and understanding-oriented medical benchmarks show that the InfiMed-Series models achieve SOTA accuracy among models with similar parameter counts and even surpass some larger models. Beyond performance gains, our analysis provides new insights into the behavior and potential of MLLMs in medical scenarios.

Limitations

Although our InfiMed-Series models achieve state-of-the-art (SOTA) performance among MLLMs with a similar number of parameters, they even outperform some MLLMs with larger parameter counts. However, it is undeniable that open-source medical multimodal data often exhibit low quality, including poor image resolution, non-uniform distribution of modalities, and errors introduced during model synthesis. Consequently, some of the results may lack full confidence, and the models' performance on more complex medical downstream tasks remains to be thoroughly explored. Moreover, how to develop reasoning steps that can be more efficient and accurate in the medical field is a critical issue that needs further study.

Ethical considerations

Our study is purely empirical and focuses on advancing research in medical multimodal large language models. We exclusively use standard, publicly available medical datasets and open-source models, all of which are accessed and utilized in strict accordance with their respective licenses.

We acknowledge the potential broader impacts and risks associated with deploying MLLMs in healthcare, including concerns related to patient safety, clinical accuracy, and potential misuse. Our work recognizes these challenges and aims to contribute to the development of more capable medical MLLMs in a manner that benefits society and human well-being. Importantly, this research does not introduce any new or unvalidated clinical applications. Instead, we focus on foundational modeling principles and ethical considerations relevant to the use of MLLMs in medical contexts, and we emphasize that future developments should be guided by careful ethical evaluation.

All training data are sourced exclusively from publicly available datasets with appropriate citations, and no private, sensitive, or personally identifiable information is involved. We rely solely on datasets and pretrained models intended for research use and adhere to their original licensing terms where specified. For datasets without explicit redistribution licenses, we restrict their use strictly to research purposes. No additional personal data is collected, synthetic data generation is designed to avoid the inclusion of identifiable personal information, and no human annotators are involved in this work. We report model sizes and provide detailed experimental setups in the Appendix to support reproducibility.

Finally, large language models are used exclusively for language editing, such as correcting grammatical and typographical errors. They are not involved in any core research activities, including idea formulation, methodological design, experimentation, or result interpretation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Rawan AlSaad, Alaa Abd-Alrazaq, Sabri Boughorbel, Arfan Ahmed, Max-Antoine Renault, Rafat Damseh, and Javaid Sheikh. 2024. Multimodal large language models in health care: applications, challenges, and future outlook. Journal of medical Internet research, 26:e59505.

Anthropic. 2025. Introducing Claude 4. https://www.anthropic.com/news/claude-4.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. 2024a. Huatuogpt-o1, towards medical complex reasoning with llms. Preprint, arXiv:2412.18925.

Junying Chen, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. 2024b. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. Preprint, arXiv:2406.19280.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. arXiv preprint arXiv:2211.12588.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 24185–24198.

Kanzhi Cheng, Yantao Li, Fangzhi Xu, Jianbing Zhang, Hao Zhou, and Yang Liu. 2024. Vision-language models can self-improve reasoning via reflection. arXiv preprint arXiv:2411.00855.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. arXiv preprint arXiv:2501.17161.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948.

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

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 22170–22183.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749.

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

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. arXiv preprint arXiv:2306.00890.

Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro RAS Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, and 1 others. 2024b. Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. Medical Image Analysis, 97:103285.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. 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.

Yuhang Liu, Pengxiang Li, Congkai Xie, Xavier Hu, Xiaotian Han, Shengyu Zhang, Hongxia Yang, and Fei Wu. 2025a. Infigui-r1: Advancing multimodal gui agents from reactive actors to deliberative reasoners. arXiv preprint arXiv:2504.14239.

Zeyu Liu, Yuhang Liu, Guanghao Zhu, Congkai Xie, Zhen Li, Jianbo Yuan, Xinyao Wang, Qing Li, Shing-Chi Cheung, Shengyu Zhang, and 1 others. 2025b. Infi-mmr: Curriculum-based unlocking multimodal reasoning via phased reinforcement learning in multimodal small language models. arXiv preprint arXiv:2505.23091.

Zihan Liu, Zhuolin Yang, Yang Chen, Chankyu Lee, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025c. Acereason-nemotron 1.1: Advancing math and code reasoning through sft and rl synergy. arXiv preprint arXiv:2506.13284.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025d. Visual-rft: Visual reinforcement fine-tuning. arXiv preprint arXiv:2503.01785.

Ziyu Liu, Yuhang Zang, Yushan Zou, Zijian Liang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025e. Visual agentic reinforcement fine-tuning. Preprint, arXiv:2505.14246.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022a. Learn to explain: Multimodal reasoning via thought chains for science question answering. In The 36th Conference on Neural Information Processing Systems (NeurIPS).

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2022b. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. arXiv preprint arXiv:2209.14610.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks.

Run Luo, Lu Wang, Wanwei He, and Xiaobo Xia. 2025. Gui-r1: A generalist r1-style vision-language action model for gui agents. arXiv preprint arXiv:2504.10458.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, and 1 others. 2025. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. arXiv preprint arXiv:2503.07365.

Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Sara Pieri, Saeed Yahya Alseiari, Shanavas Cholakkal, Khaled Aldahmani, Fahad Khan, Rao Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. 2024. Bimedix2: Bio-medical expert lmm for diverse medical modalities. Preprint, arXiv:2412.07769.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. arXiv preprint arXiv:2502.19634.

Shuai Peng, Di Fu, Liangcai Gao, Xiuqin Zhong, Hongguang Fu, and Zhi Tang. 2024. Multimath: Bridging visual and mathematical reasoning for large language models. arXiv preprint arXiv:2409.00147.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. 2025. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. arXiv preprint arXiv:2503.07536.

Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, and 1 others. 2025. Ui-tars: Pioneering automated gui interaction with native agents. arXiv preprint arXiv:2501.12326.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. arXiv preprint arXiv:2507.05201.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. Hybridflow: A flexible and efficient rlhf framework. arXiv preprint arXiv: 2409.19256.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. Nature, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. Nature Medicine, 31(3):943–950.

Yanzhou Su, Tianbin Li, Jiyao Liu, Chenglong Ma, Junzhi Ning, Cheng Tang, Sibo Ju, Jin Ye, Pengcheng Chen, Ming Hu, and 1 others. 2025. Gmai-vl-r1: Harnessing reinforcement learning for multimodal medical reasoning. arXiv preprint arXiv:2504.01886.

Yu Sun, Xingyu Qian, Weiwen Xu, Hao Zhang, Chenghao Xiao, Long Li, Yu Rong, Wenbing Huang, Qifeng Bai, and Tingyang Xu. 2025. Reasonmed: A 370k multi-agent generated dataset for advancing medical reasoning. Preprint, arXiv:2506.09513.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. 2025. Reason-rft: Reinforcement fine-tuning for visual reasoning. arXiv preprint arXiv:2503.20752.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. arXiv preprint arXiv:2503.19786.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. 2025. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. arXiv preprint arXiv:2504.08837.

Xidong Wang, Nuo Chen, Junyin Chen, Yidong Wang, Guorui Zhen, Chunxian Zhang, Xiangbo Wu, Yan Hu, Anningzhe Gao, Xiang Wan, and 1 others. 2024. Apollo: A lightweight multilingual medical llm towards democratizing medical ai to 6b people. arXiv preprint arXiv:2403.03640.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, and 1 others. 2025. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044.

Heng Yin, Yuqiang Ren, Ke Yan, Shouhong Ding, and Yongtao Hao. 2025. Rod-mllm: Towards more reliable object detection in multimodal large language models. In Proceedings of the Computer Vision and Pattern Recognition Conference, pages 14358–14368.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024.

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.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, and 1 others. 2023a. Huatuogpt, towards taming language model to be a doctor. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 10859–10885.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. arXiv preprint arXiv:2503.12937.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, and 1 others. 2023b. Biomedclip: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arxiv 2023. arXiv preprint arXiv:2303.00915.

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

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. 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. Association for Computational Linguistics.

Zijie Zhou. 2025. The table qa dataset distilled from deepseek-r1. `https://huggingface.co/datasets/jared-zhou/TQA-Distill-R1`.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479.

Wenwen Zhuang, Xin Huang, Xiantao Zhang, and Jin Zeng. 2025. Math-puma: Progressive upward multimodal alignment to enhance mathematical reasoning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, pages 26183–26191.

Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. In Forty-second International Conference on Machine Learning.

A Appendix

A.1 Construction of reflective-pattern-injected CoT

In this section, we present the detailed construction process of the reflective-pattern-injected CoT.

For multimodal datasets with sparse information (e.g., multiple-choice questions), each query is defined as $q = \{x, \mathcal{I}\}$, where x denotes the textual task instruction and $\mathcal{I}$ represents one or more images. We first employ Qwen2.5-VL-32B (Bai et al., 2025) to generate 10 candidate responses $\{y_i\}_{i=1}^{10}$ for each query q. Through rejection sampling, we divide these into two subsets: $\{y_i^+\}_{i=1}^{m}$ and $\{y_i^-\}_{i=1}^{n}$, where $m + n = 10$.

For each response in $\{y_i^+\}_{i=1}^{m}$, we apply the following prompt to generate a score:

> **Prompt for CoT Quality Evaluation**
>
> You are a medical reasoning evaluator. Assess the following response based on these criteria:
> **1. Clinical accuracy:** Correct incorporation of medical facts, clinical guidelines, and evidence-based practices. Accuracy, relevance, and appropriateness of clinical details.
> **2. Logical reasoning:** Coherent reasoning process, logically leading to the answer, well-supported by clinical knowledge.
> **3. Factual correctness:** All statements are factually correct and consistent with established medical knowledge.
> **4. Completeness:** Thorough coverage of all necessary aspects without missing critical information.
> **Question:** $\{q\}$
> **Response:** $\{y_i^+\}$
> Please evaluate the response on the above criteria and ONLY provide the `Dict` object with two keys:
> {`'score'`: integer between 1 and 10, `'justification'`: concise explanation of the score. }

After that, we compute the pass@10 for each query q, which corresponds to the number of correct responses among the 10 generated candidates, i.e., m. For $m \geq 6$, we directly select the y_i^+ with the highest score as the generated CoT. If multiple y_i^+ share the highest score, we randomly choose one.

For queries with $1 \leq m \leq 5$, we synthesize a reflective-pattern-injected CoT. Specifically, we first select one of the correct responses y_i^+ with the highest score and then randomly select one of the incorrect responses y_i^-. The reflective-pattern-injected CoT is subsequently synthesized through the following operation:

Synthesis of the reflective-pattern-injected CoT

$\{y_i^-\}$ Wait, perhaps we could consider it from a different perspective. Let's re-evaluate the problem step by step to ensure accuracy. $\{y_i^+\}$

Finally, we obtain CoT data enriched with reflective patterns through the integration of the aforementioned data, and we will release it once it is ready.

In summary, in our data construction pipeline, Qwen2.5-VL-32B-Instruct is first employed to generate multiple candidate responses for each input. Incorrect responses are then filtered using ground-truth answers. The remaining correct candidates are subsequently evaluated and scored by Qwen2.5-VL-72B-Instruct across multiple dimensions. This multi-stage generation, filtering, and scoring procedure is designed to mitigate biases introduced by reliance on a single model. Nevertheless, we acknowledge that some inherent bias is unavoidable, as synthetic medical data construction necessarily involves large-model generation. Accordingly, our pipeline emphasizes reducing the impact of such bias rather than eliminating it entirely.

A.2 Reward Function

The task-specific reward functions are as follows:

- **String-based Tasks:** For textual answers, $R_{\text{accuracy}}(o, \text{gt})$ is computed by normalizing both the model output and the ground truth (e.g., lowercasing, removing redundant spaces). This function evaluates the extracted answer from the output o, denoted as o_{ans}, by comparing it to the ground truth answer gt. We use the Jaccard function to measure the similarity between o_{ans} and gt. The Jaccard function can be formulated as: $Jaccard(o_{\text{ans}}, \text{gt}) = \frac{|o_{\text{ans}} \cap \text{gt}|}{|o_{\text{ans}} \cup \text{gt}|}$.
- **Multiple-Choice Questions:** For tasks that require selecting an option from a predefined set, $R_{\text{accuracy}}(o, \text{gt})$ is calculated by directly comparing the model's extracted predicted answer, o_{ans}, with the correct ground truth option, gt. A match results in a reward of 1, while a mismatch yields a reward of 0.
- **Mathematical Tasks:** For tasks involving mathematical expressions or numerical answers, $R_{\text{accuracy}}(o, \text{gt})$ is determined by a specialized verification function, denoted `math_verify`$(o_{\text{ans}}, \text{gt})$. This function evaluates the extracted answer from the output o, denoted o_{ans}, against the ground truth answer gt. The `math_verify` function is designed to handle nuances of mathematical evaluation, potentially allowing for symbolic equivalence or specified numerical tolerances. A successful verification yields a reward of 1; otherwise, 0.
- **Grounding Tasks:** For tasks where a model predicts a bounding box, we use the Intersection over Union (IoU) as the reward. This score measures the overlap between the predicted and ground-truth bounding boxes.

A.3 Details of the Experimental Setup

Training Datasets. In the SFT stage, we use a total of **188K** training samples from three categories: (1) multimodal general data (LLaVA-OneVision (Li et al., 2024a)), (2) multimodal medical data (VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), PathVQA (He et al., 2020), PMC-VQA (Zhang et al., 2023c), and our synthetic reflective-pattern-injected CoT), and (3) text-based medical data (ReasonMed (Sun et al., 2025), Medical-R1-Distill (Chen et al., 2024a), and Medical-o1-Reasoning (Chen et al., 2024a)). During the RLVR stage, we employ **36K** samples from multimodal general and medical datasets, including VQA-RAD, SLAKE, PathVQA, PMC-VQA, GMAI-Reasoning (Su et al., 2025), IconQA (Lu et al., 2021), ScienceQA (Lu et al., 2022a), TabMWP (Lu et al., 2022b), TQA (Zhou, 2025), and ViRFT_COCO (Liu et al., 2025e).

A detailed description of each dataset is provided as follows:

- LLaVA-OneVision (Li et al., 2024a): LLaVA-OneVision is a large-scale multimodal dataset comprising 4.8 million samples collected from diverse sources. It includes single-image, multi-image, and video modalities, and is specifically designed to train vision-language models for unified visual and textual understanding.
- VQA-RAD (Lau et al., 2018): VQA-RAD is a medical visual question answering dataset con-

structed for assessing multimodal understanding of radiology. It consists of radiological images paired with manually curated question-answer pairs authored by clinical experts. The dataset includes both open-ended and binary (yes/no) questions.

- SLAKE (Liu et al., 2021): SLAKE is a medical visual question answering dataset comprising 642 annotated radiological images spanning 39 anatomical structures and 12 disease categories. The dataset includes conditions such as various cancers (e.g., brain, liver, kidney, lung) and thoracic diseases (e.g., atelectasis, pleural effusion, pulmonary masses, and pneumothorax).
- PathVQA (He et al., 2020): PathVQA is a large-scale dataset developed for medical visual question answering tasks in the domain of pathology. It comprises 32,799 expert-annotated question-answer pairs spanning seven question categories, grounded in 4,998 high-resolution pathology images. The dataset includes both binary (yes/no) and open-ended questions.
- PMC-VQA (Zhang et al., 2023c) PMC-VQA is a large-scale medical visual question answering dataset designed to facilitate research on multimodal understanding in the medical domain. It comprises 227K VQA pairs grounded in 149K medical images, covering a wide range of imaging modalities and disease types.
- ReasonMed (Sun et al., 2025): ReasonMed is the largest open-source medical textual reasoning dataset containing 370K QA examples, which is distilled and filtered from three competitive large-language models (Qwen-2.5-72B, DeepSeek-R1-Distill-Llama-70B, and HuatuoGPT-o1-70B).
- Medical-R1-Distill (Chen et al., 2024a): Medical-R1-Distill-Data is an SFT dataset distilled from the DeepSeek-R1, constructed on verifiable medical questions from HuaTuoGPT-o1. It provides reasoning chains for medical problems, enabling the initialization and supervision of models' reasoning processes in the medical domain.
- Medical-o1-Reasoning (Chen et al., 2024a): medical-o1-reasoning-SFT is a SFT dataset focused on verifiable medical problems, where candidate solutions are generated by GPT-4o and validated by a medical verifier, providing high-quality reasoning chains and answers for training medical reasoning models.
- GMAI-Reasoning (Su et al., 2025) is a high-quality medical visual reasoning dataset comprising 10K curated multiple-choice questions constructed from 95 publicly available medical datasets spanning 12 imaging modalities (e.g., X-ray, CT, MRI). Each question is paired with standardized visual inputs and metadata, and generated using GPT-based prompting, following rigorous preprocessing and quality control procedures.
- IconQA (Lu et al., 2021): IconQA is a large-scale dataset containing 107,439 questions designed to assess abstract icon image understanding and visual language reasoning abilities.
- ScienceQA (Lu et al., 2022a): ScienceQA contains 21k multimodal questions, which align with California Common Core Content Standards, covering diverse science domains, many enriched with images, lectures, and explanations to support reasoning-oriented training.
- TabMWP (Lu et al., 2022b): Tabular Math Word Problems (TabMWP) is a multimodal dataset designed for training models to solve math word problems using both textual and tabular data. It contains 38,431 problems spanning elementary to high school levels, including both free-text and multiple-choice questions.
- TQA (Zhou, 2025): Textbook Question Answering (TQA) is a multimodal dataset designed for training models to answer questions using both textual and visual content from middle school science textbooks. Each sample provides a question, relevant textual context, and associated images, enabling models to learn to reason over multimodal inputs and generate accurate answers.
- ViRFT_COCO (Liu et al., 2025e) is a vision-language dataset derived from COCO, containing around 6,000 samples. It aims to enhance models' ability to detect all instances of a given category within an image and output the corresponding bounding boxes with confidences under strict formatting constraints.

Implementation Details. Our InfiMed-Series models include InfiMed-SFT-3B and InfiMed-RL-3B.

- InfiMed-SFT-3B, which is built upon Qwen2.5-VL-3B (Bai et al., 2025), is trained using LLaMA-Factory (Zheng et al., 2024). We utilize 8 NVIDIA H800 GPUs. The vision tower and multimodal projector are frozen during training, while the language model remains fully trainable. We use a cosine learning rate scheduler with an initial learning rate of 5×10^{-6}, a warmup ratio of 0.1, and train for 5 epochs. The batch size is set to 4 per device. Furthermore, we set the maximum input resolution to 262,144 pixels for images, while text inputs are truncated to a maximum length of 4,096 tokens.

- InfiMed-RL-3B is built upon InfiMed-SFT-3B via EasyR1 (Sheng et al., 2024). For the RLVR reward function $R_{\text{total}}(o, \text{gt}) = w_{\text{format}} \cdot R_{\text{format}}(o) + w_{\text{acc}} \cdot R_{\text{accuracy}}(o, \text{gt})$, we set the weights $w_{\text{format}} = 0.1$ and $w_{\text{acc}} = 0.9$. All experiments were conducted using 16 NVIDIA H800 GPUs. For each phase, we used a learning rate of 1.0×10^{-6}, a batch size of 256 for training updates, a rollout batch size of 256, and generated 16 rollouts per sample during policy exploration.

Evaluation Framework To ensure consistency with prior work and a comprehensive, standardized evaluation, we adopt MedEvalKit (Xu et al., 2025), a systematic framework that integrates mainstream medical benchmarks and task types, supporting a range of question formats, including multiple-choice questions, open-ended questions, and closed-ended questions. We adopt the multimodal evaluation component of the framework, combining rule-based methods with the LLM-as-a-Judge strategy.

Evaluation Benchmarks We evaluate our InfiMed-Series models on seven widely used multimodal medical benchmarks, assessing both their reasoning ability and their understanding of medical knowledge. The detailed description of the benchmarks is as follows:

- MMMU (Yue et al., 2024): MMMU is a benchmark designed to assess the capabilities of multimodal models on large-scale, multidisciplinary tasks. It comprises 11.5K meticulously curated multimodal questions drawn from university exams, quizzes, and textbooks, covering six core disciplines, including Health & Medicine. The Health & Medicine includes 1,752 test questions—accounting for 17% of the entire benchmark—and is further subdivided into five specialized domains: Basic Medical Science, Clinical Medicine, Diagnostics and Laboratory Medicine, Pharmacy, and Public Health.
- VQA-RAD (Lau et al., 2018): VQA-RAD is a dataset consisting of question–answer pairs grounded in radiological medical images, intended for training and evaluating medical visual question answering systems. It includes both open-ended questions and binary yes/no questions. In total, the dataset comprises 2,248 QA pairs linked to 315 medical images, with all annotations manually curated by a team of clinicians to ensure clinical relevance and accuracy.

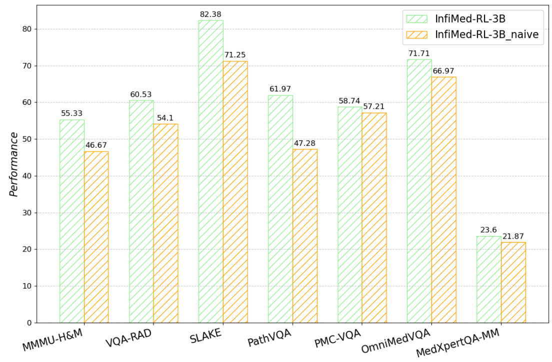


Figure 6: Performance comparison of InfiMed-RL-3B and InfiMed-RL-3B_naive on medical benchmarks. InfiMed-RL-3B_naive denotes directly utilizing RLVR upon Qwen2.5-VL-3B.

- SLAKE (Liu et al., 2021): SLAKE is a bilingual (Chinese-English) dataset specifically designed for medical visual question answering systems. It consists of 642 medical images paired with 14,028 question-answer instances.
- PathVQA (He et al., 2020): PathVQA is designed for visual question answering in the field of pathology. It comprises 4,998 pathology images collected from two pathology textbooks and the PEIR digital library, accompanied by a total of 32,799 question-answer pairs.
- PMC-VQA (Zhang et al., 2023c): PMC-VQA is a large-scale multimodal dataset constructed for medical visual question answering. It contains 227,000 VQA questions grounded in 149,000 medical images spanning a wide range of imaging modalities and disease types.
- OmniMedVQA (Hu et al., 2024): OmniMedVQA is a large-scale and comprehensive visual question answering benchmark tailored specifically for the medical domain. It aggregates data from 73 distinct medical datasets, comprising 118,010 images and 127,995 question-answer pairs. The benchmark encompasses 12 different medical imaging modalities and covers more than 20 anatomical regions of the human body.
- MedXpertQA (Zuo et al., 2025): MedXpertQA is a benchmark specifically designed to evaluate professional medical knowledge. It comprises 4,460 questions spanning 17 medical specialties and 11 organ systems. In our experiments, we utilize only the multimodal subset of the dataset.

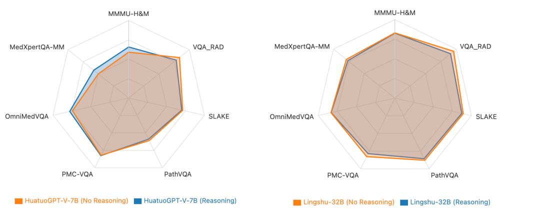
A.4 Supplementary Analysis for Ablation Experiments

In addition to the ablations on individual data components, we conduct supplementary experiments to examine the effect of moderate changes in data proportions. Specifically, we increase and decrease the amounts of general multimodal data and textual medical data by 20K samples. As shown in Table 2, these adjustments result in performance variations across benchmarks, reflecting the complementary roles of the two data types. Our final data composition is determined based on empirical observations from preliminary experiments, which indicate that this setting provides a stable balance between visual understanding and medical-domain reasoning. While alternative ratios may yield further improvements, an exhaustive search for optimal data proportions is beyond the scope of this work.

To further isolate the contribution of reflective-pattern-injected CoT, we conduct an additional ablation in which reflective CoT data is replaced with an equal amount of general CoT samples. Models trained with reflective CoT consistently outperform those trained with general CoT, demonstrating that the reflective formulation itself provides unique benefits by encouraging intermediate reasoning, error identification, and self-correction.

The lower half of Table 2 reports RLVR ablation results focusing on variants trained without general multimodal data. Removing this component causes performance on MMMU-H&M to fall below that of InfiMed-SFT-3B, indicating that RLVR relying solely on medical multimodal data—typically less reasoning-intensive—can reduce overall reasoning capability and lead to performance degradation.

In this subsection, we further present RLVR ablation studies that explore two different training initializations: one starting from the Qwen2.5-VL-3B-Instruct model (Bai et al., 2025), and the other from our InfiMed-SFT-3B. The results are shown in Figure 6. Notably, on benchmarks requiring substantial domain-specific knowledge, such as VQA-RAD, PathVQA, and SLAKE, the InfiMed-RL-3B_naive model significantly underperforms InfiMed-RL-3B. This indicates that directly applying RLVR to Qwen2.5-VL-3B-Instruct without domain-specific SFT leads to inferior performance, particularly on tasks that demand the understanding and memorization of medical knowledge. These results underscore the importance of a proper cold-start phase, where injecting knowledge-rich data during SFT is critical for establishing a solid foundation for subsequent RLVR.



(a) Huatuo-GPT-V-7B (b) Lingshu-32B

Figure 7: Comparison of direct-answer and reasoning-based prompts on medical benchmarks with larger models.

Consistent trends are also observed on MMMU-H&M, where InfiMed-RL-3B achieves a substantially higher score (55.33) than InfiMed-RL-3B_naive (46.67). Given that MMMU-H&M requires both complex reasoning and comprehensive multimodal understanding, this further highlights the essential role of the SFT phase in enabling effective knowledge integration. Similar improvements on PMC-VQA and MedXpertQA-MM further confirm that prior domain-specific fine-tuning facilitates more effective RLVR training.

Finally, we investigate whether enlarging the RLVR dataset leads to additional performance gains. We construct an expanded RLVR set by incorporating 16K additional medical questions from SLAKE (5K), PathVQA (9K), and VQA-RAD (2K), prioritizing cases where InfiMed-SFT-3B exhibits the lowest accuracy across multiple attempts. As shown in Table 3, simply increasing the number of RLVR training questions does not consistently improve performance. This suggests that while additional RLVR data may offer marginal benefits, the gains are limited relative to the significant effort required to curate high-quality multimodal medical RLVR data.

A.5 Case Studies on Qwen2.5-VL-3B-Instruct, InfiMed-SFT-3B, and InfiMed-RL-3B

In this section, we present additional case studies to illustrate the distinct responses of Qwen2.5-VL-3B, InfiMed-SFT-3B, and InfiMed-RL-3B. In summary, our analysis reveals that the InfiMed-SFT-3B model already obtains self-reflective ability but is prone to delivering redundant responses, whereas the InfiMed-RL-3B model minimizes verbosity while ensuring accuracy in its answers.

This change is consistent with the characteris-

Table 3: **Ablation study examining data composition during the RLVR stage.** Δ|Data| denotes the changes of the training dataset.

Model	**Δ|Data|**	**Accuracy (%)**							
		MMMU-H&M	**VQA-RAD**	**SLAKE**	**PathVQA**	**PMC-VQA**	**OMVQA**	**MedXQA**	**Avg.**
Base Model									
Qwen2.5VL-3B	-	51.3	56.8	63.2	37.1	50.6	64.5	20.7	49.2
Ablation Study in RLVR Stage									
InfiMed-RL-3B	-	55.3	60.5	82.4	62.0	58.7	71.7	23.6	59.2
InfiMed-RL-3B+medical_mm	+16K	56.0	60.9	82.2	61.5	58.4	70.0	23.6	58.9
InfiMed-RL-3B-w/o-general	−10K	53.3	60.7	81.9	61.6	58.3	70.0	23.6	58.4

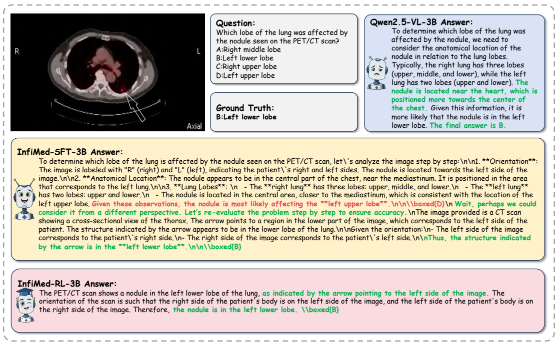


Figure 8: Case study on Qwen2.5-VL-3B, InfiMed-Series models in medical VQA. Red denotes errors or irrelevant content, whereas Green denotes correct or important information. All of the models answered correctly; InfiMed-SFT-3B answered correctly after reflection.

tics of the GRPO optimization mechanism. During GRPO training, the gradient update for each rollout response is normalized by its length, which amplifies the learning signal for shorter and correct responses. Consequently, the reinforcement learning stage naturally encourages the model to prefer concise and accurate answers rather than extended reflective chains. The reflective supervision used during SFT primarily serves to enrich the model's reasoning search space for RLVR, rather than to enforce reflective verbosity in the final model. After RLVR, the model shifts from imitating reflection-style formats to producing more efficient outputs, which explains the observed reduction in explicit reflective content.

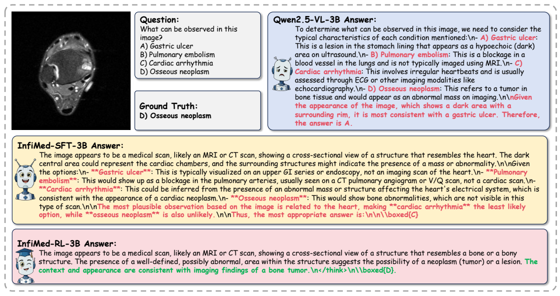


Figure 9: Case study on Qwen2.5-VL-3B, InfiMed-Series models in medical VQA. Red denotes errors or irrelevant content, whereas Green denotes correct or important information. Only InfiMed-RL-3B answered correctly, providing a direct and accurate response by correctly identifying the key visual information and selecting the correct answer with a concise and focused explanation.

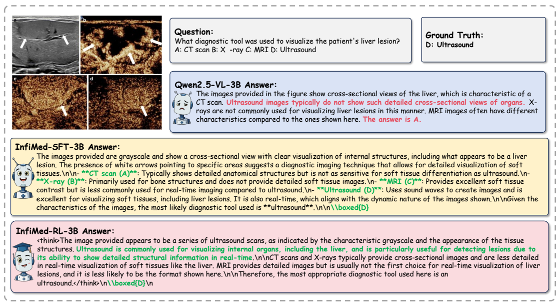


Figure 10: Case study on Qwen2.5-VL-3B, InfiMed-Series models in medical VQA. Red denotes errors or irrelevant content, whereas Green denotes correct or important information. Qwen2.5-VL-3B's response is incorrect because it fundamentally fails to recognize the visual characteristics of an ultrasound scan, leading to a flawed conclusion. InfiMed-SFT-3B provides a detailed, step-by-step reasoning process, while InfiMed-RL-3B offers a more direct and accurate answer, showcasing its improved ability to instantly recognize diagnostic imaging types.

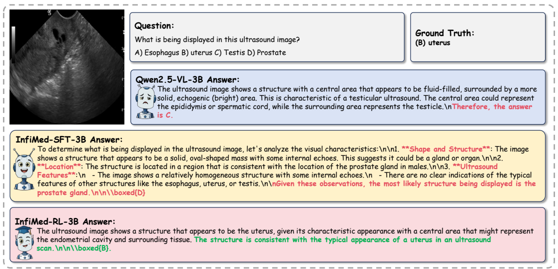


Figure 11: Case study on Qwen2.5-VL-3B, InfiMed-Series models in medical VQA. Red denotes errors or irrelevant content, whereas Green denotes correct or important information. Incorrectly identifies the image and chooses the wrong answer. InfiMed-SFT-3B provides a detailed analysis that correctly rules out most options but ultimately guesses the wrong answer. InfiMed-RL-3B is the only model that correctly identifies the organ shown, providing the most direct and accurate response.

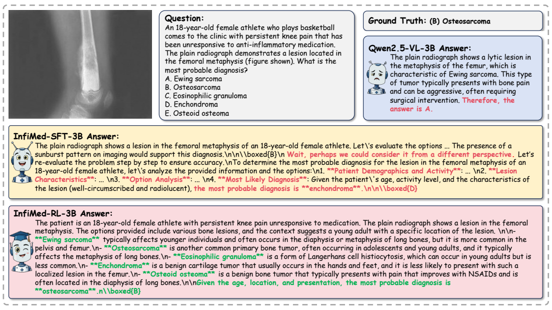


Figure 12: Case study on Qwen2.5-VL-3B, InfiMed-Series models in medical VQA. Red denotes errors or irrelevant content, whereas Green denotes correct or important information. Qwen2.5-VL-3B incorrectly identifies the lesion. InfiMed-SFT-3B initially makes a correct deduction but then overthinks the problem, leading to a self-correction that results in the wrong answer. InfiMed-RL-3B provides a detailed and accurate analysis.

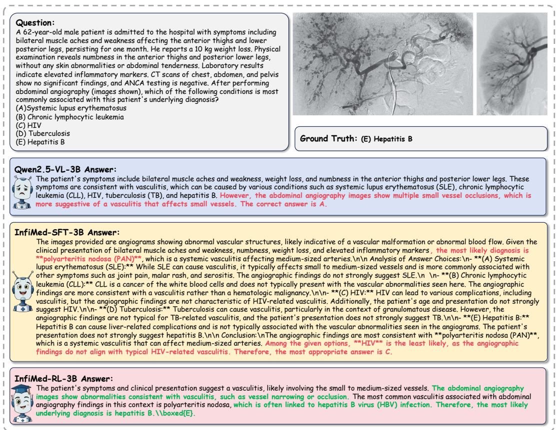


Figure 13: Case study on Qwen2.5-VL-3B, InfiMed-Series models in medical VQA. Red denotes errors or irrelevant content, whereas Green denotes correct or important information. Only InfiMed-RL-3B successfully synthesizes the angiographic findings and patient symptoms to provide the correct diagnosis.

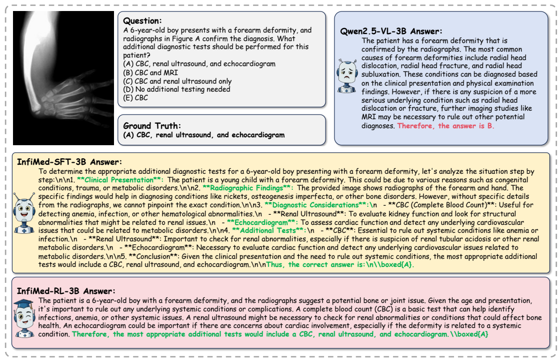


Figure 14: Case study on Qwen2.5-VL-3B, InfiMed-Series models in medical VQA. Red denotes errors or irrelevant content, whereas Green denotes correct or important information. Qwen2.5-VL-3B incorrectly identifies the necessary tests. Both InfiMed-SFT-3B and InfiMed-RL-3B correctly solve this case, but InfiMed-RL-3B provides a more direct and efficient answer.

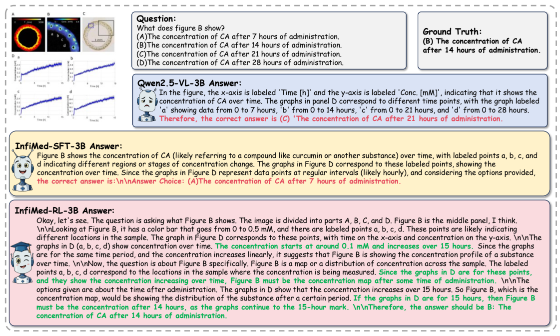


Figure 15: Case study on Qwen2.5-VL-3B, InfiMed-Series models in medical-related VQA (MMMU-H&M). Red denotes errors or irrelevant content, whereas Green denotes correct or important information. Qwen2.5-VL-3B incorrectly identifies the relevant data series. InfiMed-SFT-3B fails to correctly interpret the graphs and provides an incorrect analysis. Only InfiMed-RL-3B correctly synthesizes the information from both the concentration map and the graphs to provide the right answer.

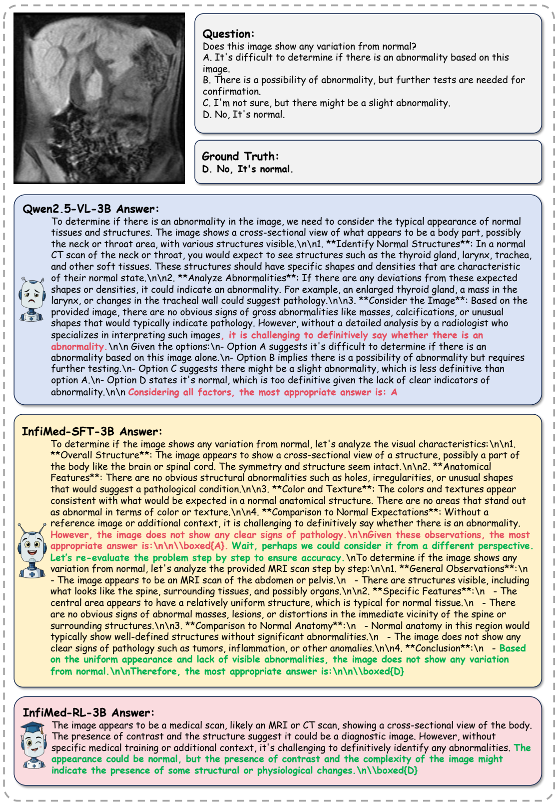


Figure 16: Case study on Qwen2.5-VL-3B, InfiMed-Series models in medical VQA. Red denotes errors or irrelevant content, whereas Green denotes correct or important information. Qwen2.5-VL-3B's response is incorrect, while InfiMed-SFT-3B correctly answers after a detailed analysis and reflection, and InfiMed-RL-3B provides the most direct and accurate correct answer.

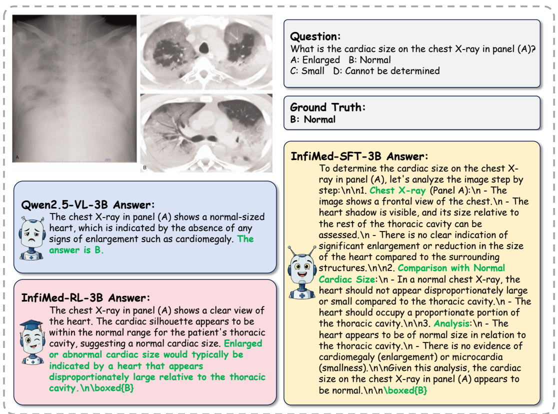


Figure 17: Case study on Qwen2.5-VL-3B, InfiMed-Series models in medical VQA. Red denotes errors or irrelevant content, whereas Green denotes correct or important information. All three models correctly identify the cardiac size as normal, with Qwen2.5-VL-3B providing a concise answer, InfiMed-SFT-3B offering a detailed, step-by-step analysis, and InfiMed-RL-3B giving a direct and well-reasoned response.