

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

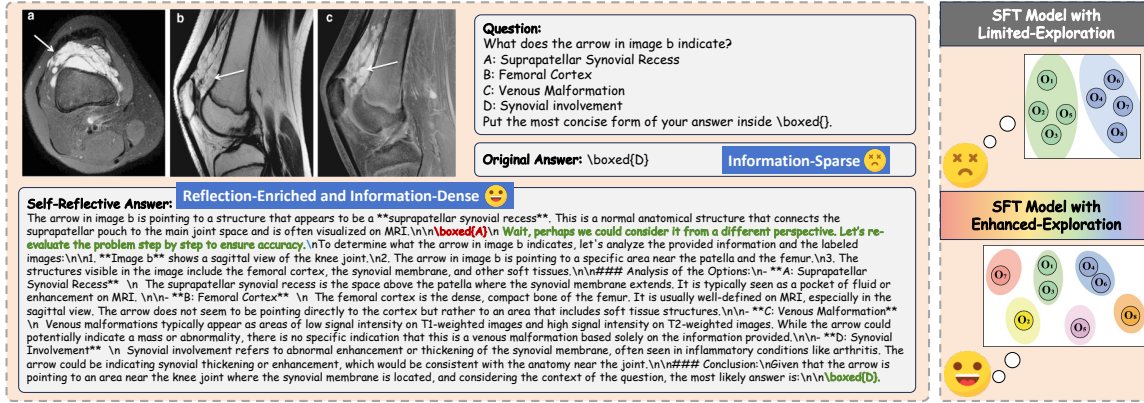


Figure 1: **Left:** Comparison of information-sparse and reflection-enriched, information-dense outputs. **Right:** A model with Enhanced Exploration (bottom) generates a broader, more effective search space, while Limited Exploration (top) results in a narrower, less efficient search.

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

157 Med-PaLM series (Singhal et al., 2023, 2025), re- 208
158 cent efforts have increasingly focused on extending 209
159 LLM capabilities to multimodal medical. LLaVA- 210
160 Med (Li et al., 2023) introduces a biomedical- 211
161 specialized MLLM trained on a curated figure- 212
162 caption dataset with self-instructed instruction- 213
163 following data. The model highlights the poten- 214
164 tial of cost-efficient training strategies for domain- 215
165 specific MLLMs. MedGemma (Sjellergren et al., 216
166 2025) has shown strong generalization across med- 217
167 ical vision-language and text-only tasks, demon- 218
168 strating advanced medical understanding and rea- 219
169 soning on multimodal data. Lingshu (Xu et al., 220
170 2025) proposed a domain-specialized multimodal 221
171 foundation model for medical, supported by a cu- 222
172 rated dataset enriched with medical VQA, CoT rea- 223
173 soning, and report annotations. While prior work 224
174 has made notable progress in medical MLLMs, 225
175 many approaches depend on large model sizes and 226
176 substantial computational resources, which limit 227
177 their accessibility and scalability. 228

178 2.2 Reasoning in Medical LLMs and MLLMs 229

179 Interpretable reasoning remains a central desider- 230
180 atum in medical AI, with recent efforts explor- 231
181 ing general CoT prompting (Wei et al., 2022) and 232
182 program-based logic (Chen et al., 2022) model- 233
183 ing. Although these approaches have shown po- 234
184 tential, they typically rely on costly expert-cu- 235
185 rated annotations (Li et al., 2024b), which limits 236
186 their scalability in real-world clinical settings. RL 237
187 offers a compelling alternative by enabling emerg- 238
188 ent reasoning capabilities without requiring explicit 239
189 supervision, as demonstrated by recent models 240
190 such as DeepSeek-R1 (Guo et al., 2025), which 241
191 achieve notable improvements in reasoning with 242
192 rule-based reward. Building on this paradigm, RLVR 243
193 has been used to improve reasoning reliability, 244
194 with Group Relative Policy Optimization (Shao 245
195 et al., 2024) known for its efficiency and good 246
196 performance. This method is now increasingly 247
197 used to train MLLMs to improve their reasoning 248
198 ability (Meng et al., 2025; Wang et al., 2025; 249
199 Tan et al., 2025). With the success of RLVR, 250
200 several work leverages it on medical MLLMs. 251
201 MedVLM-R1 (Pan et al., 2025) employs RLVR 252
202 to explicit reasoning in medical VQA, achiev- 253
203 ing strong performance and generalization. Its 254
204 emphasis on reasoning highlights the role of RL 255
205 in enhancing transparency and trustworthiness 256
206 in clinical AI systems. GMAI-VL-R1 (Su 257
207 et al., 2025) explores RLVR to enhance reason- 258

208 introducing a multi-agent reasoning data synthesis 209
209 framework, the model outperforms prior models 210
210 on some complex tasks. Lingshu (Xu et al., 2025) 211
211 also leverages an RLVR paradigm, achieving strong 212
212 performance across medical VQA, report genera- 213
213 tion, and text-only QA. Despite these promising 214
214 advances, prior work has been limited in its explo- 215
215 ration of the RLVR stage. 216

216 2.3 Our Distinction 217

217 As discussed above, general CoT consists of multi- 218
218 step natural language reasoning traces that help 219
219 models learn structured reasoning patterns during 220
220 SFT and guide response selection during RLVR. 221
221 However, medical reasoning is far more standard- 222
222 ized and constrained by clinical knowledge than 223
223 open-domain tasks, resulting in limited diversity 224
224 of valid intermediate steps. This restricts the ex- 225
225 ploration space of RLVR and hinders the discov- 226
226 ery of improved reasoning trajectories. To address 227
227 this issue, we introduce reflective-pattern-injected 228
228 CoT data during SFT, equipping the model with 229
229 initial self-reflection and self-correction capabili- 230
230 ties. This design effectively expands the reasoning 231
231 space available to RLVR, enabling more robust im- 232
232 provements on complex medical tasks. In contrast 233
233 to prior work that either targets a narrow set of 234
234 benchmarks or fails to consistently improve SFT 235
235 performance, we not only achieve reliable gains 236
236 during the RLVR stage but also conduct extensive 237
237 experiments to analyze the features in multimodal 238
238 medical tasks, which could provide several mean- 239
239 ingful insights for the future exploration of medical 240
240 MLLMs. 241

241 3 Methodology 242

242 In this section, we outline our methodology for 243
243 advancing multimodal medical understanding and 244
244 reasoning through RLVR with a self-reflective cold 245
245 start, which is depicted in Figure 2. Our approach 246
246 unfolds in two stages: (1) A cold start phase, in 247
247 which we uniquely integrate general multimodal 248
248 data with medical text reasoning data to simultane- 249
249 ously enhance image understanding and restore 250
250 fundamental reasoning skills. Crucially, to ad- 251
251 dress the information sparsity of existing medical 252
252 datasets, we further synthesize both distilled CoT 253
253 and self-reflective CoT for SFT, thereby establish- 254
254 ing a richer and more exploratory reasoning founda- 255
255 tion. (2) A RLVR phase, which enables the model 256
256 to explore a wider spectrum of reasoning trajec- 257

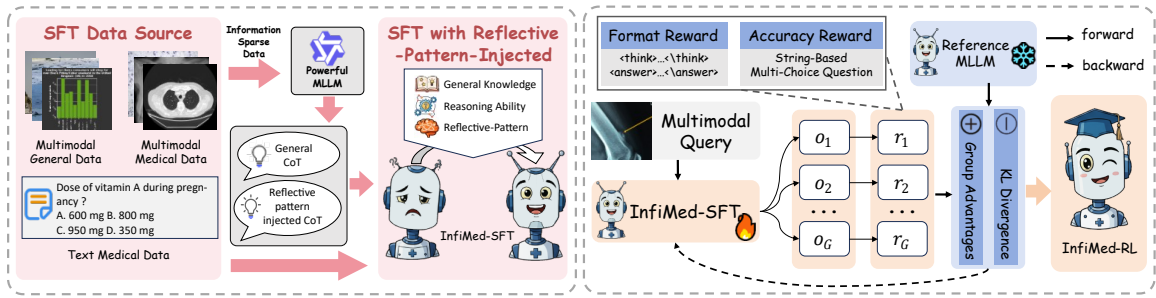


Figure 2: The overall training process of InfiMed-Series models.

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 $R(o, gt)$ 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:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \dots, r_G\})}{\text{std}(\{r_1, r_2, \dots, r_G\})}, \quad (1)$$

where $\text{mean}(\cdot)$ indicates the average value, and $\text{std}(\cdot)$ 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:

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

where the additional Kullback–Leibler term $D_{\text{KL}}[\pi_{\theta} \parallel \pi_{\text{ref}}]$ 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 (Cohanici 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 (Selligren 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 (%)							Avg.
		MMM-U-H&M	VQA-RAD	SLAKE	PathVQA	PMC-VQA	OMVQA	MedXQA	
<i>Proprietary Models</i>									
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
<i>General Open-source Models</i>									
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
<i>Medical Open-source Models</i>									
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
<i>Ours (InfiMed-Series)</i>									
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

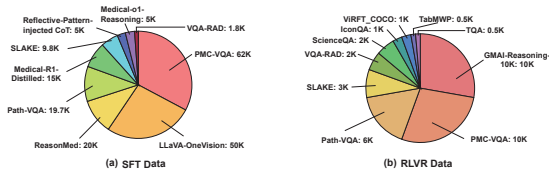


Figure 3: Overview of the training samples for the InfiMed series models in the reflective-injected SFT and RLVR stages.

used multimodal medical benchmarks: MMMU-Health&Medicine (MMM-U-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 OmniMed-VQA, 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

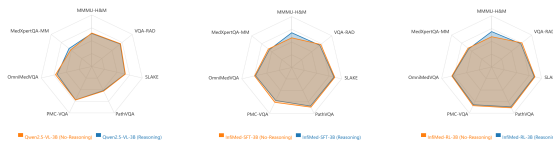


Figure 4: Comparison of direct-answer and reasoning-based prompts on medical benchmarks.

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.

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.** $\Delta|\text{Data}|$ 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	$\Delta \text{Data} $	Accuracy (%)							
		MMM-U-H&M	VQA-RAD	SLAKE	PathVQA	PMC-VQA	OMVQA	MedXQA	Avg.
<i>Base Model</i>									
Qwen2.5VL-3B-Instruct	-	51.3	56.8	63.2	37.1	50.6	64.5	20.7	49.2
<i>Ablation Study in SFT Stage on General Multimodal Data</i>									
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
<i>Ablation Study in SFT Stage on Medical Text Data</i>									
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
<i>Ablation Study in SFT Stage on Reflective CoT Data</i>									
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
<i>Ablation Study in RLVR Stage</i>									
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

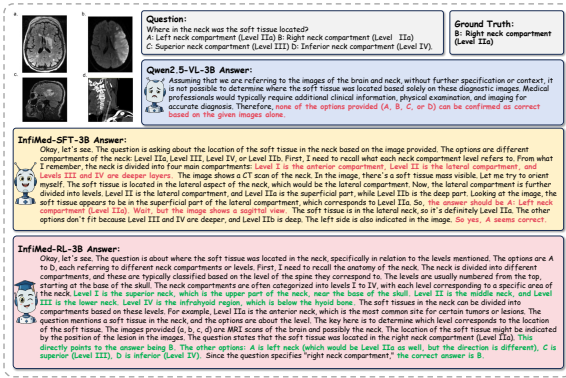


Figure 5: Case study. Red denotes errors or irrelevant content, Green denotes correct or important information.

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.

633 Limitations

634 Although our InfiMed-Series models achieve state-
635 of-the-art (SOTA) performance among MLLMs
636 with a similar number of parameters, they even
637 outperform some MLLMs with larger parameter
638 counts. However, it is undeniable that open-source
639 medical multimodal data often exhibit low qual-
640 ity, including poor image resolution, non-uniform
641 distribution of modalities, and errors introduced
642 during model synthesis. Consequently, some of
643 the results may lack full confidence, and the mod-
644 els' performance on more complex medical down-
645 stream tasks remains to be thoroughly explored.
646 Moreover, how to develop reasoning steps that can
647 be more efficient and accurate in the medical field
648 is a critical issue that needs further study.

649 Ethical considerations

650 Our study is purely empirical and focuses on ad-
651 vancing research in medical multimodal large lan-
652 guage models. We exclusively use standard, pub-
653 licly available medical datasets and open-source
654 models, all of which are accessed and utilized in
655 strict accordance with their respective licenses.

656 We acknowledge the potential broader impacts
657 and risks associated with deploying MLLMs in
658 healthcare, including concerns related to patient
659 safety, clinical accuracy, and potential misuse. Our
660 work recognizes these challenges and aims to con-
661 tribute to the development of more capable medical
662 MLLMs in a manner that benefits society and hu-
663 man well-being. Importantly, this research does not
664 introduce any new or unvalidated clinical applica-
665 tions. Instead, we focus on foundational modeling
666 principles and ethical considerations relevant to the
667 use of MLLMs in medical contexts, and we empha-
668 size that future developments should be guided by
669 careful ethical evaluation.

670 All training data are sourced exclusively from
671 publicly available datasets with appropriate cita-
672 tions, and no private, sensitive, or personally iden-
673 tifiable information is involved. We rely solely
674 on datasets and pretrained models intended for re-
675 search use and adhere to their original licensing
676 terms where specified. For datasets without explicit
677 redistribution licenses, we restrict their use strictly
678 to research purposes. No additional personal data
679 is collected, synthetic data generation is designed
680 to avoid the inclusion of identifiable personal in-
681 formation, and no human annotators are involved
682 in this work. We report model sizes and provide

detailed experimental setups in the Appendix to
support reproducibility.

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

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama
Ahmad, Ilge Akkaya, Florencia Leoni Aleman,
Diogo Almeida, Janko Altschmidt, Sam Altman,
Shyamal Anadkat, and 1 others. 2023. Gpt-4 techni-
cal report. [arXiv preprint arXiv:2303.08774](https://arxiv.org/abs/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](https://doi.org/10.1016/j.jmir.2024.e59505),
26:e59505.
- Anthropic. 2025. Introducing Claude 4. <https://www.anthropic.com/news/claude-4>.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wen-
bin 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](https://arxiv.org/abs/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](https://arxiv.org/abs/2412.18925). [Preprint](https://arxiv.org/abs/2412.18925),
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 mul-
timodal llms at scale](https://arxiv.org/abs/2406.19280). [Preprint](https://arxiv.org/abs/2406.19280), arXiv:2406.19280.
- Wenhu Chen, Xueguang Ma, Xinyi Wang, and
William W Cohen. 2022. Program of thoughts
prompting: Disentangling computation from reason-
ing for numerical reasoning tasks. [arXiv preprint
arXiv:2211.12588](https://arxiv.org/abs/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](https://arxiv.org/abs/2405.11540), 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](https://arxiv.org/abs/2411.00855).

736	Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Sheng-	Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and	792
737	bang Tong, Saining Xie, Dale Schuurmans, Quoc V	Xiao-Ming Wu. 2021. Slake: A semantically-labeled	793
738	Le, Sergey Levine, and Yi Ma. 2025. Sft mem-	knowledge-enhanced dataset for medical visual ques-	794
739	orizes, rl generalizes: A comparative study of	tion answering. In <u>2021 IEEE 18th international</u>	795
740	foundation model post-training. <u>arXiv preprint</u>	<u>symposium on biomedical imaging (ISBI)</u> , pages	796
741	<u>arXiv:2501.17161</u> .	1650–1654. IEEE.	797
742	Gheorghe Comanici, Eric Bieber, Mike Schaekermann,	Yuhang Liu, Pengxiang Li, Congkai Xie, Xavier Hu, Xi-	798
743	Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Mar-	aotian Han, Shengyu Zhang, Hongxia Yang, and Fei	799
744	cel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and	Wu. 2025a. Infigui-r1: Advancing multimodal gui	800
745	1 others. 2025. Gemini 2.5: Pushing the frontier with	agents from reactive actors to deliberative reasoners.	801
746	advanced reasoning, multimodality, long context, and	<u>arXiv preprint arXiv:2504.14239</u> .	802
747	next generation agentic capabilities. <u>arXiv preprint</u>	Zeyu Liu, Yuhang Liu, Guanghao Zhu, Congkai Xie,	803
748	<u>arXiv:2507.06261</u> .	Zhen Li, Jianbo Yuan, Xinyao Wang, Qing Li, Shing-	804
749	Daya Guo, Dejian Yang, Haowei Zhang, Junxiao	Chi Cheung, Shengyu Zhang, and 1 others. 2025b.	805
750	Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shi-	Infi-mmr: Curriculum-based unlocking multimodal	806
751	rong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025.	reasoning via phased reinforcement learning in mul-	807
752	Deepseek-r1: Incentivizing reasoning capability in	timodal small language models. <u>arXiv preprint</u>	808
753	llms via reinforcement learning. <u>arXiv preprint</u>	<u>arXiv:2505.23091</u> .	809
754	<u>arXiv:2501.12948</u> .	Zihan Liu, Zhuolin Yang, Yang Chen, Chankyu Lee,	810
755	Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and	Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping.	811
756	Pengtao Xie. 2020. Pathvqa: 30000+ questions for	2025c. Acereason-nemotron 1.1: Advancing math	812
757	medical visual question answering. <u>arXiv preprint</u>	and code reasoning through sft and rl synergy. <u>arXiv</u>	813
758	<u>arXiv:2003.10286</u> .	<u>preprint arXiv:2506.13284</u> .	814
759	Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Jun-	Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang	815
760	jun He, Yu Qiao, and Ping Luo. 2024. Omnimed-	Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang.	816
761	vqa: A new large-scale comprehensive evaluation	2025d. Visual-rft: Visual reinforcement fine-tuning.	817
762	benchmark for medical lvlm. In <u>Proceedings of</u>	<u>arXiv preprint arXiv:2503.01785</u> .	818
763	<u>the IEEE/CVF Conference on Computer Vision and</u>	Ziyu Liu, Yuhang Zang, Yushan Zou, Zijian Liang, Xi-	819
764	<u>Pattern Recognition</u> , pages 22170–22183.	aoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin,	820
765	Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao,	and Jiaqi Wang. 2025e. <u>Visual agentic reinforcement</u>	821
766	Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui	<u>fine-tuning</u> . <u>Preprint</u> , arXiv:2505.14246.	822
767	Lin. 2025. Vision-r1: Incentivizing reasoning capa-	Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei	823
768	bility in multimodal large language models. <u>arXiv</u>	Chang, Song-Chun Zhu, Oyvind Taffjord, Peter Clark,	824
769	<u>preprint arXiv:2503.06749</u> .	and Ashwin Kalyan. 2022a. Learn to explain: Multi-	825
770	Jason J Lau, Soumya Gayen, Asma Ben Abacha, and	modal reasoning via thought chains for science ques-	826
771	Dina Demner-Fushman. 2018. A dataset of clini-	tion answering. In <u>The 36th Conference on Neural</u>	827
772	cally generated visual questions and answers about	<u>Information Processing Systems (NeurIPS)</u> .	828
773	radiology images. <u>Scientific data</u> , 5(1):1–10.	Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu,	829
774	Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng	Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark,	830
775	Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang,	and Ashwin Kalyan. 2022b. Dynamic prompt learn-	831
776	Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-	ing via policy gradient for semi-structured mathemat-	832
777	onevision: Easy visual task transfer. <u>arXiv preprint</u>	ical reasoning. <u>arXiv preprint arXiv:2209.14610</u> .	833
778	<u>arXiv:2408.03326</u> .	Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou	834
779	Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto	Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and	835
780	Usuyama, Haotian Liu, Jianwei Yang, Tristan Nau-	Song-Chun Zhu. 2021. Iconqa: A new benchmark	836
781	mann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-	for abstract diagram understanding and visual lan-	837
782	med: Training a large language-and-vision assis-	guage reasoning. In <u>The 35th Conference on Neural</u>	838
783	tant for biomedicine in one day. <u>arXiv preprint</u>	<u>Information Processing Systems (NeurIPS) Track on</u>	839
784	<u>arXiv:2306.00890</u> .	<u>Datasets and Benchmarks</u> .	840
785	Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro RAS	Run Luo, Lu Wang, Wanwei He, and Xiaobo Xia.	841
786	Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue,	2025. Gui-r1: A generalist r1-style vision-language	842
787	Yixiong Chen, Xiaorui Lin, and 1 others. 2024b. Ab-	action model for gui agents. <u>arXiv preprint</u>	843
788	domenatlas: A large-scale, detailed-annotated, &	<u>arXiv:2504.10458</u> .	844
789	multi-center dataset for efficient transfer learning	Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang	845
790	and open algorithmic benchmarking. <u>Medical Image</u>	Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han,	846
791	<u>Analysis</u> , 97:103285.		

847	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 .	903
848		904
849		905
850		
851	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 .	906
852		907
853		908
854		909
855		910
856		
857		
858	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 .	911
859		912
860		913
861		914
862		915
863		
864	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 .	916
865		917
866		918
867		919
868		920
869	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 lmm with strong reasoning abilities through two-stage rule-based rl. arXiv preprint arXiv:2503.07536 .	921
870		922
871		923
872		924
873		925
874	Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, and 1 others. 2025. Uitars: Pioneering automated gui interaction with native agents. arXiv preprint arXiv:2501.12326 .	926
875		927
876		928
877		929
878		930
879	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 .	931
880		932
881		933
882		934
883		935
884	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 .	936
885		937
886		938
887		939
888		940
889		941
890	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 .	942
891		943
892		944
893		945
894		946
895	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. <i>Nature</i> , 620(7972):172–180.	947
896		948
897		949
898		950
899		951
900	Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and	952
901		953
902		954
	1 others. 2025. Toward expert-level medical question answering with large language models. <i>Nature Medicine</i> , 31(3):943–950.	955
		956
		957
	Yanzhou Su, Tianbin Li, Jiyao Liu, Chenglong Ma, Junzhi Ning, Cheng Tang, Siboj 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 Wenhui Chen. 2025. VI-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. <i>Advances in neural information processing systems</i> , 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-mlm: Towards more reliable object detection in multimodal large language models. In <i>Proceedings of the Computer Vision and Pattern Recognition Conference</i> , 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.	

958 Mmmu: A massive multi-discipline multimodal un-
959 derstanding and reasoning benchmark for expert
960 agi. In Proceedings of the IEEE/CVF Conference
961 on Computer Vision and Pattern Recognition, pages
962 9556–9567.

963 Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu,
964 Zhihong Chen, Guiming Chen, Jianquan Li, Xi-
965 angbo Wu, Zhang Zhiyi, Qingying Xiao, and 1 oth-
966 ers. 2023a. Huatuogpt, towards taming language
967 model to be a doctor. In Findings of the Association
968 for Computational Linguistics: EMNLP 2023, pages
969 10859–10885.

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

975 Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen
976 Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Ra-
977 jesh Rao, Mu Wei, Naveen Valluri, and 1 others.
978 2023b. Biomedclip: A multimodal biomedical foun-
979 dation model pretrained from fifteen million scien-
980 tific image-text pairs. arxiv 2023. arXiv preprint
981 arXiv:2303.00915.

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

987 Yaowei Zheng, Richong Zhang, Junhao Zhang, Yan-
988 han Ye, Zheyang Luo, Zhangchi Feng, and Yongqiang
989 Ma. 2024. Llamafactory: Unified efficient fine-
990 tuning of 100+ language models. In Proceedings
991 of the 62nd Annual Meeting of the Association
992 for Computational Linguistics (Volume 3: System
993 Demonstrations), Bangkok, Thailand. Association
994 for Computational Linguistics.

995 Zijie Zhou. 2025. The table qa dataset distilled
996 from deepseek-r1. [https://huggingface.co/](https://huggingface.co/datasets/jared-zhou/TQA-Distill-R1)
997 [datasets/jared-zhou/TQA-Distill-R1](https://huggingface.co/datasets/jared-zhou/TQA-Distill-R1).

998 Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu,
999 Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan,
1000 Weijie Su, Jie Shao, and 1 others. 2025. Internvl3:
1001 Exploring advanced training and test-time recipes
1002 for open-source multimodal models. arXiv preprint
1003 arXiv:2504.10479.

1004 Wenwen Zhuang, Xin Huang, Xiantao Zhang, and Jin
1005 Zeng. 2025. Math-puma: Progressive upward multi-
1006 modal alignment to enhance mathematical reasoning.
1007 In Proceedings of the AACL Conference on Artificial
1008 Intelligence, volume 39, pages 26183–26191.

1009 Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen,
1010 Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding,
1011 and Bowen Zhou. 2025. Medxpertqa: Benchmarking
1012 expert-level medical reasoning and understanding. In
1013 Forty-second International Conference on Machine
1014 Learning.

A Appendix 1015

A.1 Construction of 1016 reflective-pattern-injected CoT 1017

In this section, we present the detailed construction 1018
process of the reflective-pattern-injected CoT. 1019

For multimodal datasets with sparse informa- 1020
tion (e.g., multiple-choice questions), each query 1021
is defined as $q = \{x, \mathcal{I}\}$, where x denotes the tex- 1022
tual task instruction and \mathcal{I} represents one or more 1023
images. We first employ Qwen2.5-VL-32B (Bai 1024
et al., 2025) to generate 10 candidate responses 1025
 $\{y_i\}_{i=1}^{10}$ for each query q . Through rejection sam- 1026
pling, we divide these into two subsets: $\{y_i^+\}_{i=1}^m$ 1027
and $\{y_i^-\}_{i=1}^n$, where $m + n = 10$. 1028

For each response in $\{y_i^+\}_{i=1}^m$, we apply the fol- 1029
lowing prompt to generate a score: 1030

Prompt for CoT Quality Evaluation

You are a medical reasoning evaluator. As- 1031
sess the following response based on these 1032
criteria:

1. Clinical accuracy: Correct incorpora- 1033
tion of medical facts, clinical guidelines, 1034
and evidence-based practices. Accuracy, 1035
relevance, and appropriateness of clinical 1036
details.

2. Logical reasoning: Coherent reason- 1037
ing process, logically leading to the answer, 1038
well-supported by clinical knowledge.

3. Factual correctness: All statements are 1039
factually correct and consistent with estab- 1040
lished medical knowledge.

4. Completeness: Thorough coverage of all 1041
necessary aspects without missing critical 1042
information.

Question: $\{q\}$

Response: $\{y_i^+\}$

Please evaluate the response on the above 1043
criteria and ONLY provide the Dict object 1044
with two keys:

`'score': integer between 1 and 10,` 1045
`'justification': concise explanation of` 1046
`the score. }` 1047

After that, we compute the pass@10 for each 1048
query q , which corresponds to the number of cor- 1049
rect responses among the 10 generated candidates, 1050
i.e., m . For $m \geq 6$, we directly select the y_i^+ with 1051
the highest score as the generated CoT. If multiple 1052
 y_i^+ share the highest score, we randomly choose 1053

1038 one.

1039 For queries with $1 \leq m \leq 5$, we synthesize
1040 a reflective-pattern-injected CoT. Specifically, we
1041 first select one of the correct responses y_i^+ with
1042 the highest score and then randomly select one of
1043 the incorrect responses y_i^- . The reflective-pattern-
1044 injected CoT is subsequently synthesized through
1045 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. $\{y_i^+\}$

1046
1047 Finally, we obtain CoT data enriched with reflect-
1048 ive patterns through the integration of the afore-
1049 mentioned data, and we will release it once it is
1050 ready.

1051 In summary, in our data construction pipeline,
1052 Qwen2.5-VL-32B-Instruct is first employed to gener-
1053 ate multiple candidate responses for each input.
1054 Incorrect responses are then filtered using ground-
1055 truth answers. The remaining correct candidates
1056 are subsequently evaluated and scored by Qwen2.5-
1057 VL-72B-Instruct across multiple dimensions. This
1058 multi-stage generation, filtering, and scoring proce-
1059 dure is designed to mitigate biases introduced by
1060 reliance on a single model. Nevertheless, we ac-
1061 knowledge that some inherent bias is unavoidable,
1062 as synthetic medical data construction necessarily
1063 involves large-model generation. Accordingly, our
1064 pipeline emphasizes reducing the impact of such
1065 bias rather than eliminating it entirely.

1066 A.2 Reward Function

1067 The task-specific reward functions are as follows:

- 1068 • **String-based Tasks:** For textual answers,
1069 $R_{\text{accuracy}}(o, \text{gt})$ is computed by normalizing both
1070 the model output and the ground truth (e.g., low-
1071 ercasing, removing redundant spaces). This func-
1072 tion evaluates the extracted answer from the out-
1073 put o , denoted as o_{ans} , by comparing it to the
1074 ground truth answer gt . We use the Jaccard func-
1075 tion to measure the similarity between o_{ans} and
1076 gt . The Jaccard function can be formulated as:
1077
$$\text{Jaccard}(o_{\text{ans}}, \text{gt}) = \frac{|o_{\text{ans}} \cap \text{gt}|}{|o_{\text{ans}} \cup \text{gt}|}.$$

1078 • **Multiple-Choice Questions:** For tasks that re-
1079 quire selecting an option from a predefined set,
1080 $R_{\text{accuracy}}(o, \text{gt})$ is calculated by directly compar-

ing 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.

- 1081 • **Mathematical Tasks:** For tasks involv-
1082 ing mathematical expressions or numerical
1083 answers, $R_{\text{accuracy}}(o, \text{gt})$ is determined by
1084 a specialized verification function, denoted
1085 $\text{math_verify}(o_{\text{ans}}, \text{gt})$. This function evaluates
1086 the extracted answer from the output o , denoted
1087 o_{ans} , against the ground truth answer gt . The
1088 math_verify function is designed to handle nu-
1089 ances of mathematical evaluation, potentially
1090 allowing for symbolic equivalence or specified
1091 numerical tolerances. A successful verification
1092 yields a reward of 1; otherwise, 0.
1093 • **Grounding Tasks:** For tasks where a model
1094 predicts a bounding box, we use the Intersec-
1095 tion over Union (IoU) as the reward. This score
1096 measures the overlap between the predicted and
1097 ground-truth bounding boxes.
1098
1099
1100
1101

1102 A.3 Details of the Experimental Setup

1103 **Training Datasets.** In the SFT stage, we use
1104 a total of **188K** training samples from three cat-
1105 egories: (1) multimodal general data (LLaVA-
1106 OneVision (Li et al., 2024a)), (2) multimodal
1107 medical data (VQA-RAD (Lau et al., 2018),
1108 SLAKE (Liu et al., 2021), PathVQA (He et al.,
1109 2020), PMC-VQA (Zhang et al., 2023c), and
1110 our synthetic reflective-pattern-injected CoT), and
1111 (3) text-based medical data (ReasonMed (Sun
1112 et al., 2025), Medical-R1-Distill (Chen et al.,
1113 2024a), and Medical-o1-Reasoning (Chen et al.,
1114 2024a)). During the RLVR stage, we employ
1115 **36K** samples from multimodal general and medical
1116 datasets, including VQA-RAD, SLAKE, PathVQA,
1117 PMC-VQA, GMAI-Reasoning (Su et al., 2025),
1118 IconQA (Lu et al., 2021), ScienceQA (Lu et al.,
1119 2022a), TabMWP (Lu et al., 2022b), TQA (Zhou,
1120 2025), and ViRFT_COCO (Liu et al., 2025e).

1121 A detailed description of each dataset is provided
1122 as follows:

- 1123 • **LLaVA-OneVision (Li et al., 2024a):** LLaVA-
1124 OneVision is a large-scale multimodal dataset
1125 comprising 4.8 million samples collected from
1126 diverse sources. It includes single-image, multi-
1127 image, and video modalities, and is specifically
1128 designed to train vision-language models for uni-
1129 fied visual and textual understanding.
- 1130 • **VQA-RAD (Lau et al., 2018):** VQA-RAD is a
1131 medical visual question answering dataset con-

1132 structured for assessing multimodal understanding
1133 of radiology. It consists of radiological images
1134 paired with manually curated question-answer
1135 pairs authored by clinical experts. The dataset
1136 includes both open-ended and binary (yes/no)
1137 questions.

- 1138 • SLAKE (Liu et al., 2021): SLAKE is a medi- 1184
1139 cal visual question answering dataset comprising 1185
1140 642 annotated radiological images spanning 39 1186
1141 anatomical structures and 12 disease categories. 1187
1142 The dataset includes conditions such as various 1188
1143 cancers (e.g., brain, liver, kidney, lung) and tho- 1189
1144 racic diseases (e.g., atelectasis, pleural effusion, 1190
1145 pulmonary masses, and pneumothorax). 1191
1146 • PathVQA (He et al., 2020): PathVQA is a large- 1192
1147 scale dataset developed for medical visual ques- 1193
1148 tion answering tasks in the domain of pathology. 1194
1149 It comprises 32,799 expert-annotated question- 1195
1150 answer pairs spanning seven question categories, 1196
1151 grounded in 4,998 high-resolution pathology im- 1197
1152 ages. The dataset includes both binary (yes/no) 1198
1153 and open-ended questions. 1199
1154 • PMC-VQA (Zhang et al., 2023c) PMC-VQA is 1200
1155 a large-scale medical visual question answering 1201
1156 dataset designed to facilitate research on multi- 1202
1157 modal understanding in the medical domain. It 1203
1158 comprises 227K VQA pairs grounded in 149K 1204
1159 medical images, covering a wide range of imag- 1205
1160 ing modalities and disease types. 1206
1161 • ReasonMed (Sun et al., 2025): ReasonMed is the 1207
1162 largest open-source medical textual reasoning 1208
1163 dataset containing 370K QA examples, which is 1209
1164 distilled and filtered from three competitive large- 1210
1165 language models (Qwen-2.5-72B, DeepSeek-R1- 1211
1166 Distill-Llama-70B, and HuatuoGPT-o1-70B). 1212
1167 • Medical-R1-Distill (Chen et al., 2024a): Medical- 1213
1168 R1-Distill-Data is an SFT dataset distilled from 1214
1169 the DeepSeek-R1, constructed on verifiable medi- 1215
1170 cal questions from HuaTuoGPT-o1. It provides 1216
1171 reasoning chains for medical problems, enabling 1217
1172 the initialization and supervision of models’ rea- 1218
1173 soning processes in the medical domain. 1219
1174 • Medical-o1-Reasoning (Chen et al., 2024a): 1220
1175 medical-o1-reasoning-SFT is a SFT dataset fo- 1221
1176 cused on verifiable medical problems, where can- 1222
1177 didate solutions are generated by GPT-4o and 1223
1178 validated by a medical verifier, providing high- 1224
1179 quality reasoning chains and answers for training 1225
1180 medical reasoning models. 1226
1181 • GMAI-Reasoning (Su et al., 2025) is a high- 1227
1182 quality medical visual reasoning dataset com- 1228
1183 prising 10K curated multiple-choice questions 1229

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 gen-
erated using GPT-based prompting, following
rigorous preprocessing and quality control proce-
dures.

- 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 con-
tains 21k multimodal questions, which align with
California Common Core Content Standards,
covering diverse science domains, many enriched
with images, lectures, and explanations to sup-
port reasoning-oriented training.
- TabMWP (Lu et al., 2022b): Tabular Math Word
Problems (TabMWP) is a multimodal dataset de-
signed 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 Answer-
ing (TQA) is a multimodal dataset designed for
training models to answer questions using both
textual and visual content from middle school sci-
ence textbooks. Each sample provides a question,
relevant textual context, and associated images,
enabling models to learn to reason over multi-
modal 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

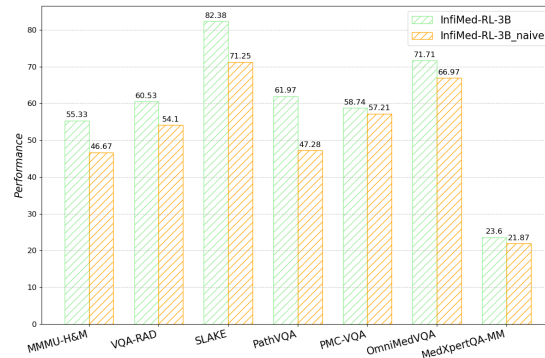


Figure 6: Performance comparison of InfiMed-RL-3B and InfiMed-RL-3B_naive on medical benchmarks. InfiMed-RL-3B_naive denotes directly utilizing RLVR upon Qwen2.5-VL-3B.

to ensure clinical relevance and accuracy.

- 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.

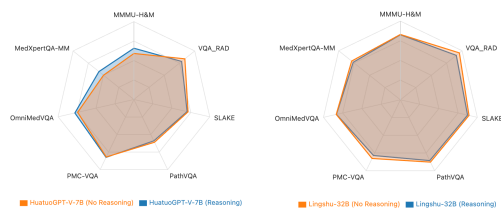
A.4 Supplementary Analysis for Ablation Experiments

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-



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

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

start phase, where injecting knowledge-rich data during SFT is critical for establishing a solid foundation for subsequent RLVR.

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. $\Delta|Data|$ denotes the changes of the training dataset.

Model	$\Delta Data $	Accuracy (%)							Avg.
		MMMU-H&M	VQA-RAD	SLAKE	PathVQA	PMC-VQA	OMVQA	MedXQA	
<i>Base Model</i>									
Qwen2.5-VL-3B	-	51.3	56.8	63.2	37.1	50.6	64.5	20.7	49.2
<i>Ablation Study in RLVR Stage</i>									
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

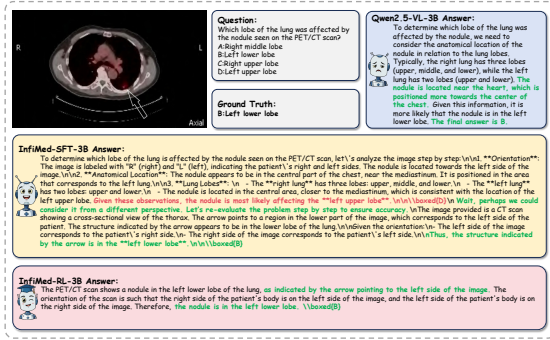


Figure 8: Case study on Qwen2.5-VL-3B, InfiMed-Series models in medical VQA. Red denotes errors or irrelevant content, whereas Green denotes correct or important information. All of the models answered correctly; InfiMed-SFT-3B answered correctly after reflection.

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.

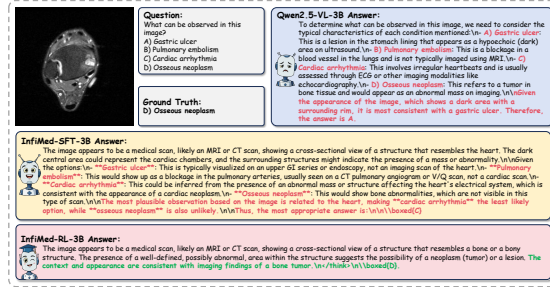


Figure 9: Case study on Qwen2.5-VL-3B, InfiMed-Series models in medical VQA. Red denotes errors or irrelevant content, whereas Green denotes correct or important information. Only InfiMed-RL-3B answered correctly, providing a direct and accurate response by correctly identifying the key visual information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

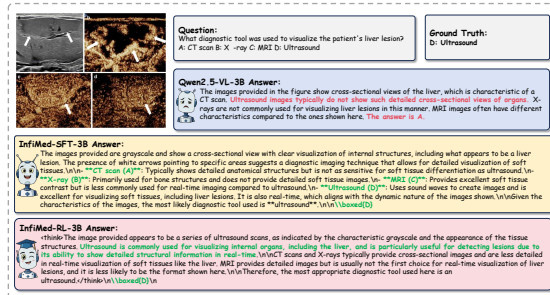


Figure 10: Case study on Qwen2.5-VL-3B, InfiMed-Series models in medical VQA. Red denotes errors or irrelevant content, whereas Green denotes correct or important information. Qwen2.5-VL-3B’s response is incorrect because it fundamentally fails to recognize the visual characteristics of an ultrasound 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.

Question:
What is being displayed in this ultrasound image?

A) Esophagus B) uterus C) Testis D) Prostate

Ground Truth:
(B) uterus

Qwen2.5-VL-3B Answer:
The ultrasound image shows a structure with a central area that appears to be fluid-filled, surrounded by a more echogenic (brighter) area. This is characteristic of a testicular ultrasound. The central area could represent the epididymis or spermatic cord, while the surrounding area represents the testis. Therefore, the correct answer is C.

InfiMed-SFT-3B Answer:
To determine what is being displayed in the ultrasound image, let's analyze the visual characteristics. The image shows a structure that appears to be a solid, oval-shaped mass with some internal echogenicity. This suggests it could be a gland or organ. The location is in a region that is consistent with the location of the prostate gland in males. Therefore, the most likely structure being displayed is the prostate gland.

InfiMed-RL-3B Answer:
The ultrasound image shows a structure that appears to be a solid, oval-shaped mass with some internal echogenicity. This suggests it could be a gland or organ. The location is in a region that is consistent with the location of the prostate gland in males. Therefore, the most likely structure being displayed is the prostate gland.

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.

Question:
An 18-year-old female athlete who plays basketball comes to the clinic with persistent knee pain that has been unresponsive to anti-inflammatory medication. The plain radiograph demonstrates a lesion located in the femoral metaphysis (figure shown). What is the most probable diagnosis?

A. Ewing sarcoma
B. Osteosarcoma
C. Eosinophilic granuloma
D. Enchondroma
E. Osteoid osteoma

Ground Truth: (B) Osteosarcoma

Qwen2.5-VL-3B Answer:
The metaphysis of the femur, which is the area of the bone between the epiphysis and the diaphysis, is the most common site for Ewing sarcoma. Ewing sarcoma is a type of cancer that typically presents with bone pain and swelling. The other options are less likely to present with a lesion in the metaphysis of the femur.

InfiMed-SFT-3B Answer:
The plain radiograph shows a lesion in the femoral metaphysis of an 18-year-old female athlete. Let's evaluate the options. The presence of a patient pattern in the image would support the diagnosis of Ewing sarcoma. Let's re-evaluate the problem step by step to ensure accuracy. To determine the most probable diagnosis for the lesion in the femoral metaphysis of an 18-year-old female athlete, let's analyze the provided information and the options. The patient's demographics and activity level are consistent with Ewing sarcoma. The lesion's location and characteristics are also consistent with Ewing sarcoma. Therefore, the most probable diagnosis is Ewing sarcoma.

InfiMed-RL-3B Answer:
The patient is an 18-year-old female athlete with persistent knee pain unresponsive to medication. The plain radiograph shows a lesion in the femoral metaphysis. The options provided include osteosarcoma, Ewing sarcoma, eosinophilic granuloma, enchondroma, and osteoid osteoma. Ewing sarcoma is a type of cancer that typically presents with bone pain and swelling. The other options are less likely to present with a lesion in the metaphysis of the femur.

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.

Question:
A 55-year-old male patient is admitted to the hospital with symptoms including bilateral muscle aches and weakness affecting the anterior thighs and lower posterior legs, persisting for one month. He reports a 10 lb weight loss. Physical examination reveals numbness in the anterior thighs and posterior lower legs, without any skin abnormalities or abnormal reflexes. Laboratory results indicate elevated inflammatory markers. CT scans of chest, abdomen, and pelvis show no significant findings, and EKG testing is negative. After performing abdominal angiography (image shown), which of the following conditions is most commonly associated with the patient's underlying diagnosis?

(A) Systemic lupus erythematosus
(B) Chronic lymphocytic leukemia
(C) HIV
(D) Tuberculosis
(E) Hepatitis B

Ground Truth: (E) Hepatitis B

Qwen2.5-VL-3B Answer:
The patient's symptoms include bilateral muscle aches and weakness, weight loss, and numbness in the anterior thighs and posterior lower legs. These symptoms are consistent with vasculitis, which can be caused by various conditions such as systemic lupus erythematosus (SLE), chronic lymphocytic leukemia (CLL), HIV, tuberculosis (TB), and hepatitis B. However, the abdominal angiography images show multiple small vessel occlusions, which is more suggestive of a vasculitis that affects small vessels. The correct answer is A.

InfiMed-SFT-3B Answer:
The images provided are angiograms showing abnormal vascular structures, likely indicative of a vascular malformation or abnormal blood flow. Given the clinical presentation of bilateral muscle aches and weakness, numbness, weight loss, and elevated inflammatory markers, the most likely diagnosis is Systemic Lupus Erythematosus (SLE). SLE is a systemic autoimmune disease that can affect various organs and tissues. The angiographic findings are consistent with SLE, as it is commonly associated with other symptoms such as joint pain, rash, and serositis. The angiographic findings do not strongly suggest CLL, HIV, TB, or Hepatitis B. CLL is a cancer of the white blood cells and does not typically present with the vascular abnormalities seen here. The angiographic findings are consistent with SLE, as it is commonly associated with other symptoms such as joint pain, rash, and serositis. The angiographic findings do not strongly suggest CLL, HIV, TB, or Hepatitis B. CLL is a cancer of the white blood cells and does not typically present with the vascular abnormalities seen here. The angiographic findings are consistent with SLE, as it is commonly associated with other symptoms such as joint pain, rash, and serositis.

InfiMed-RL-3B Answer:
The patient's symptoms and clinical presentation suggest a vasculitis, likely involving the small to medium-sized vessels. The abdominal angiography findings in this context are highly suggestive of a vasculitis, such as vasculitis associated with Systemic Lupus Erythematosus (SLE). The most common vasculitis associated with abdominal angiography findings in this context is SLE. The angiographic findings are most consistent with SLE. Therefore, the most likely underlying diagnosis is hepatitis B.

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.

Question:
A 6-year-old boy presents with a forearm deformity, and radiographs in Figure A confirm the diagnosis. What additional diagnostic tests should be performed for this patient?

(A) CBC, renal ultrasound, and echocardiogram
(B) CBC and MRI
(C) CBC and renal ultrasound only
(D) No additional testing needed
(E) CBC

Ground Truth:
(A) CBC, renal ultrasound, and echocardiogram

Qwen2.5-VL-3B Answer:
The patient has a forearm deformity that is confirmed by the radiographs. The most common causes of forearm deformities include radial head dislocation, radial head fracture, and radial head aplasia. These conditions can be diagnosed based on the clinical presentation and physical examination findings. However, if there is any suspicion of a more serious underlying condition such as radial head dislocation or fracture, further imaging studies like MRI may be necessary to rule out other potential diagnoses. Therefore, the answer is B.

InfiMed-SFT-3B Answer:
To determine the appropriate additional diagnostic tests for a 6-year-old boy presenting with a forearm deformity, let's analyze the situation step by step. The patient is a young child with a forearm deformity. This could be due to various reasons such as congenital conditions, trauma, or metabolic disorders. The provided image shows radiographs of the forearm and hand. The specific findings would help in diagnosing conditions like rickets, osteogenesis imperfecta, or other bone disorders. However, without specific details from the radiographs, we cannot pinpoint the exact condition. To ensure comprehensive evaluation, the following tests are recommended: Complete Blood Count (CBC) to detect anemia, infection, or other hematological abnormalities; Renal Ultrasound to evaluate kidney function and look for structural abnormalities that might be related to renal issues; Echocardiogram to assess cardiac function and detect any underlying cardiovascular issues that could be related to metabolic disorders. Additional Tests: In addition to the above tests, it is essential to rule out systemic conditions like anemia or infection. A renal ultrasound might be necessary to check for renal abnormalities, especially if there is suspicion of renal tubular acidosis or other renal metabolic disorders. ECG/Echocardiogram: Necessary to evaluate cardiac function and detect any underlying cardiovascular issues related to metabolic disorders. Conclusion: Given the clinical presentation and the need to rule out systemic conditions, the most appropriate additional tests would include a CBC, renal ultrasound, and echocardiogram. Therefore, the correct answer is A.

InfiMed-RL-3B Answer:
The patient is a 6-year-old boy with a forearm deformity, and the radiographs suggest a potential bone or joint issue. Given the age and presentation, it's important to rule out any underlying systemic conditions or complications. A complete blood count (CBC) is a basic test that can help identify infections, anemia, or other systemic issues. A renal ultrasound might be necessary to check for renal abnormalities or conditions that could affect bone health. An echocardiogram could be important if there are concerns about cardiac involvement, especially if the deformity is related to a systemic condition. Therefore, the most appropriate additional tests would include a CBC, renal ultrasound, and echocardiogram.

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.

Question:
What does Figure 8 show?

(A) The concentration of CA after 7 hours of administration.
(B) The concentration of CA after 14 hours of administration.
(C) The concentration of CA after 21 hours of administration.
(D) The concentration of CA after 28 hours of administration.


Ground Truth:
(B) The concentration of CA after 14 hours of administration.

Qwen2.5-VL-3B Answer:
In the figure, the x-axis is labeled 'Time [h]' and the y-axis is labeled 'Conc. [nM]', indicating that it shows the concentration of CA over time. The graphs in panel D correspond to different time points, with the graph labeled 'a' showing data from 0 to 7 hours, 'b' from 0 to 14 hours, 'c' from 0 to 21 hours, and 'd' from 0 to 28 hours. Therefore, the correct answer is (C) The concentration of CA after 21 hours of administration.

InfiMed-SFT-3B Answer:
Figure 8 shows the concentration of CA (likely referring to a compound like curcumin or another substance) over time, with labeled points a, b, c, and d indicating different regions or stages of concentration change. The graphs in Figure 8 correspond to these labeled points, showing the concentration over time. Since the graphs in Figure 8 represent data points at regular intervals (likely hours), and considering the options provided, the correct answer is (A) The concentration of CA after 7 hours of administration.

InfiMed-RL-3B Answer:
The question is asking what Figure 8 shows. The image is divided into parts A, B, C, and D. Figure 8 is the middle panel. I think in the context of Figure 8, it has a color bar that goes from 0 to 0.5 nM, and there are labeled points a, b, c, d. These points are likely indicating different locations in the sample. The graph in Figure 8 corresponds to these points, with time on the x-axis and concentration on the y-axis. The graphs in D (a, b, c, d) show concentration over time. The concentration starts at around 0.1 nM and increases over 15 hours. Since the graphs are for the same time period, and the concentration increases linearly, it suggests that Figure 8 is showing the concentration profile of a substance over time. In the context of the question, the question is about Figure 8 specifically. Figure 8 is a map or a distribution of concentration across the sample. The labeled points a, b, c, d correspond to the locations in the sample where the concentration is being measured. Since the graphs in D are for these points, and they show the concentration increasing over time, Figure 8 must be the concentration map after some time of administration. The options given are about the time of administration. The graphs in D show that the concentration increases over 15 hours. Figure 8, which is the concentration map, would be showing the distribution of the substance after a certain period. If the graphs in D are for 15 hours, then Figure 8 must be the concentration after 14 hours, as the graphs continue to the 15-hour mark. Therefore, the answer should be B: The concentration of CA after 14 hours of administration.

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.



Question:
Does this image show any variation from normal?
A. It's difficult to determine if there is an abnormality based on this image.
B. There is a possibility of abnormality, but further tests are needed for confirmation.
C. I'm not sure, but there might be a slight abnormality.
D. No, It's normal.

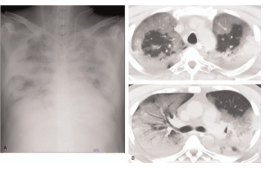
Ground Truth:
D. No, It's normal.

Qwen2.5-VL-3B Answer:
To determine if there is an abnormality in the image, we need to consider the typical appearance of normal tissues and structures. The image shows a cross-sectional view of what appears to be a body part, possibly the neck or throat area, with various structures visible. **Identify Normal Structures:** In a normal CT scan of the neck or throat, you would expect to see structures such as the thyroid gland, larynx, trachea, and other soft tissues. These structures should have specific shapes and densities that are characteristic of their normal state. **Analyze Abnormalities:** If there are any deviations from these expected shapes or densities, it could indicate an abnormality. For example, an enlarged thyroid gland, a mass in the larynx, or changes in the tracheal wall could suggest pathology. **Consider the Image:** Based on the provided image, there are no obvious signs of gross abnormalities like masses, calcifications, or unusual shapes that would typically indicate pathology. However, without a detailed analysis by a radiologist who specializes in interpreting such images, it is challenging to definitively say whether there is an abnormality. **Given the options:** Option A suggests it's difficult to determine if there is an abnormality based on this image alone. Option B implies there is a possibility of abnormality but requires further testing. Option C suggests there might be a slight abnormality, which is less definitive than option A. Option D states it's normal, which is too definitive given the lack of clear indicators of abnormality. **Considering all factors, the most appropriate answer is: A**

InfiMed-SFT-3B Answer:
To determine if the image shows any variation from normal, let's analyze the visual characteristics. **Overall Structure:** The image appears to show a cross-sectional view of a structure, possibly a part of the body like the brain or spinal cord. The symmetry and structure seem intact. **Anatomical Features:** There are no obvious structural abnormalities such as holes, irregularities, or unusual shapes that would suggest a pathological condition. **Color and Texture:** The colors and textures appear consistent with what would be expected in a normal anatomical structure. There are no areas that stand out as abnormal in terms of color or texture. **Comparison to Normal Expectations:** Without a reference image or additional context, it is challenging to definitively say whether there is an abnormality. **However, the image does not show any clear signs of pathology.** **Given these observations, the most appropriate answer is:** **A**. Wait, perhaps we could consider it from a different perspective. **Let's re-evaluate the problem step by step to ensure accuracy.** To determine if the image shows any variation from normal, let's analyze the provided MRI scan step by step. **General Observations:** The image appears to be an MRI scan of the abdomen or pelvis. There are structures visible, including what looks like the spine, surrounding tissues, and possibly organs. **Specific Features:** The central area appears to have a relatively uniform structure, which is typical for normal tissue. There are no obvious signs of abnormal masses, lesions, or distortions in the immediate vicinity of the spine or surrounding structures. **Comparison to Normal Anatomy:** Normal anatomy in this region would typically show well-defined structures without significant abnormalities. The image does not show any clear signs of pathology such as tumors, inflammation, or other anomalies. **Conclusion:** Based on the uniform appearance and lack of visible abnormalities, the image does not show any variation from normal. **Therefore, the most appropriate answer is:** **D**

InfiMed-RL-3B Answer:
The image appears to be a medical scan, likely an MRI or CT scan, showing a cross-sectional view of the body. The presence of contrast and the structure suggest it could be a diagnostic image. However, without specific medical training or additional context, it's challenging to definitively identify any abnormalities. The appearance could be normal, but the presence of contrast and the complexity of the image might indicate the presence of some structural or physiological changes. **D**

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.



Question:
What is the cardiac size on the chest X-ray in panel (A)?
A: Enlarged B: Normal
C: Small D: Cannot be determined

Ground Truth:
B: Normal

Qwen2.5-VL-3B Answer:
The chest X-ray in panel (A) shows a normal-sized heart, which is indicated by the absence of any signs of enlargement such as cardiomegaly. The answer is B.

InfiMed-RL-3B Answer:
The chest X-ray in panel (A) shows a clear view of the heart. The cardiac silhouette appears to be within the normal range for the patient's thoracic cavity, suggesting a normal cardiac size. Enlarged or abnormal cardiac size would typically be indicated by a heart that appears disproportionately large relative to the thoracic cavity. **B**

InfiMed-SFT-3B Answer:
To determine the cardiac size on the chest X-ray in panel (A), let's analyze the image step by step. **Chest X-ray (Panel A):** The image shows a frontal view of the chest. The heart shadow is visible, and its size relative to the rest of the thoracic cavity can be assessed. There is no clear indication of significant enlargement or reduction in the size of the heart compared to the surrounding structures. **Comparison with Normal Cardiac Size:** In a normal chest X-ray, the heart should not appear disproportionately large or small compared to the thoracic cavity. The heart should occupy a proportionate portion of the thoracic cavity. **Analysis:** The heart appears to be of normal size in relation to the thoracic cavity. There is no evidence of cardiomegaly (enlargement) or microcardia (smallness). **Conclusion:** Based on this analysis, the cardiac size on the chest X-ray in panel (A) appears to be normal. **B**

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.