GMAI-VL-R1: Harnessing Reinforcement Learning for Multimodal Medical Reasoning

Anonymous ACL submission

Abstract

Recent advances in general medical AI have made significant strides, but existing models often lack the reasoning capabilities needed for complex medical decision-making. This paper presents GMAI-VL-R1, a multimodal medical reasoning model enhanced by reinforcement learning (RL) to improve its reasoning abilities. Through iterative training, GMAI-VL-R1 optimizes decision-making, significantly boosting diagnostic accuracy and clinical support. We also develop a reasoning data synthesis method, generating step-by-step reasoning data via rejection sampling, which further enhances the model's generalization. Experimental results show that after RL training, GMAI-VL-R1 excels in tasks such as medical image diagnosis and visual question answering. While the model demonstrates basic memorization with supervised fine-tuning, RL is crucial for true generalization. Our work establishes new evaluation benchmarks and paves the way for future advancements in medical reasoning models.

1 Introduction

Integrating multimodal medical data such as images, clinical records, and patient histories is crucial for comprehensive decision making, diagnostic accuracy, and clinical outcomes (Acosta et al., 2022; Lipkova et al., 2022). However, Developing large-scale multimodal models that can reason over complex medical evidence remains challenging, especially when reliable reasoning and reflection are required for high-stakes clinical decisions (Kline et al., 2022).

Existing medical multimodal models have advanced through fine tuning on carefully constructed multimodal instruction data (Li et al., 2024c; Chen et al., 2024a; Moor et al., 2023; Tu et al., 2024; Li et al., 2024a; Liu et al., 2023; Zhang et al., 2024), improving tasks such as image captioning, visual question answering, medical report generation, and disease detection (Li et al., 2024b; Thawakar et al., 2024). However, a key limitation is their limited reasoning capability. Most existing models depend heavily on supervised fine tuning (SFT), emphasizing memorization of input and output mappings (Chu et al., 2025), rather than developing deeper reasoning abilities. Although these models perform well on familiar tasks, they lack sufficient flexibility when encountering novel or complex scenarios. In medical contexts, where data complexity and uncertainty are prevalent, relying only on pattern recognition is inadequate for effective clinical decision making.

Inspired by DeepSeek-R1 (DeepSeek-AI, 2025), we develop GMAI-VL-R1, a general purpose multimodal medical model leveraging reinforcement learning tuning (RLT) to enhance Chain of Thought (CoT) reasoning and reflection (Wei et al., 2022). Unlike traditional models that depend on supervised fine tuning (SFT), GMAI-VL-R1 directly applies RLT to the base model, specifically Qwen VL 7b (Bai et al., 2023). The RLT training procedure is illustrated in Fig. 1-a). Given a medical image, GMAI-VL-R1 generates multiple responses, each including explicit reasoning steps and final answers produced by large vision language models (LVLMs). We then utilize rule based rewards, including accuracy, format correctness, and redundancy avoidance, to guide policy gradient optimization (Shao et al., 2024) during model updates. Through reinforcement learning tuning, the model gains practical reasoning experience via repeated training and self correction, surpassing superficial pattern recognition. This strengthened reasoning capability enables GMAI-VL-R1 to provide more reliable results in high risk clinical decisions and equips it to manage complex and previously unseen medical scenarios effectively, as demonstrated in Fig. 1-c).

To develop our RLT approach, we construct GMAI-Reasoning10K, a high-quality medical visual question answering (VQA) dataset, as shown

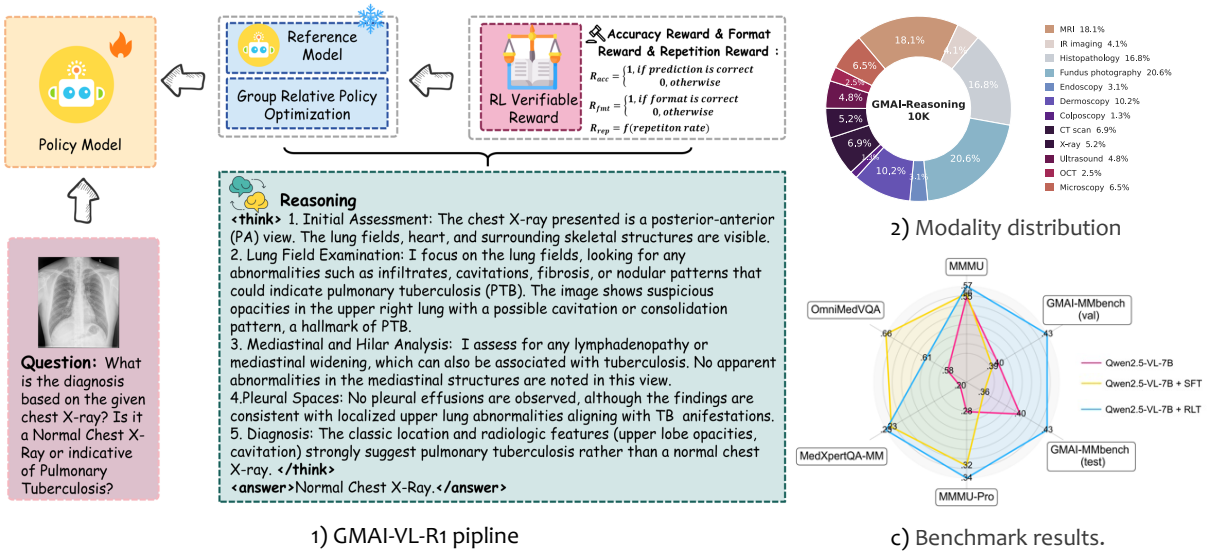

Figure 1: **Overview of GMAI-VL-R1.** (a) GMAI-VL-R1 pipeline, where group-sampled reasoning responses are optimized with accuracy, format, and repetition rewards. (b) Modality distribution of GMAI-Reasoning10K, a curated 10K medical VQA dataset spanning 12 imaging modalities. (c) Benchmark results showing that the RLT-based model generally outperforms the SFT-based model.

in Fig. 1-b). It contains 10,000 rigorously curated medical VQA pairs from 95 public datasets, spanning 12 imaging modalities such as X-ray, CT, and MRI. Unlike contemporary works (Lai et al., 2025) that rely only on question-answer supervision, we use GPT-4o to generate detailed CoT reasoning instructions and apply a specialized filtering strategy to retain high-quality reasoning steps. This design gives SFT a strong reasoning-supervision baseline, enabling a fair comparison between SFT and RLT.

We evaluate GMAI-VL-R1 across 6 large-scale medical multimodal benchmarks (Hu et al., 2024; Chen et al., 2024b; Yue et al., 2024a; Wang et al., 2024; Zuo et al., 2025), covering medical question answering, disease diagnosis, and recognition tasks. As shown in Fig. 1-c), RLT improves generalization over both the base model and SFT across diverse benchmarks. On MMMU, GMAI-VL-R1 achieves an accuracy of 57.33%, representing a 2% improvement over the baseline model. For MMMU-pro, the model attains an accuracy of 34.03%, yielding a substantial gain of 5.56%. On GMAL-MMbench, GMAI-VL-R1 reaches accuracies of 43.14% on the validation set and 43.84% on the test set, corresponding to improvements of 3.12% and 3.25% over the baseline, and in both cases, it remains 3.48% and 7.72% higher than the SFT model. Furthermore, the method obtains 23.80% (+3.50%) on MedXpertQA-MM and 61.01% (+2.60%) on OmniMedVQA. These findings demonstrate that RLT consistently improves medical multimodal reasoning and serves as a versatile optimization strategy for medical vision-language models.

Overall, our contributions in this work are threefold:

- We develop GMAI-VL-R1, a multimodal medical model that directly applies reinforcement learning tuning to enhance reasoning capabilities, moving beyond pattern memorization in medical AI.
- We construct GMAI-Reasoning10K, a high-quality dataset spanning 12 imaging modalities with detailed Chain-of-Thought annotations, establishing a reliable benchmark for developing and comparing tuning approaches.
- We demonstrate through extensive experiments across multiple benchmarks that our approach achieves competitive performance compared to previous state-of-the-art methods, achieves superior generalization in out-of-distribution scenarios, while requiring significantly less training data.

2 Related Work

Medical Vision-Language Models are typically built on general-purpose large models and adapted to medical applications using specialized

datasets (Hartsock and Rasool, 2024). For example, Med-Flamingo (Moor et al., 2023) enhances OpenFlamingo-9B with 800K medical image-text pairs, focusing on medical image analysis and report generation. RadFM (Wu et al., 2023b) improves PMC-LLaMA (Wu et al., 2023a) with 16 million radiology images and text. Med-PaLM (Tu et al., 2024) adapts PaLM-E (Driess et al., 2023) for medical tasks using one million samples, excelling in diagnostics and Q&A. LLaVA-Med (Li et al., 2024a) uses PubMed Central data to enhance LLaVA (Touvron et al., 2023), improving biomedical image understanding and open-domain conversation. MedTrinity-25M (Xie et al., 2024) creates 25M image-text pairs for fine-tuning, though results remain limited. Qilin-Med-VL (Liu et al., 2023) and BioMedGPT (Zhang et al., 2024) also use large image-text pairs, but their performance is limited by data quality, hindering generalization. Med-Gemini (Saab et al., 2024) enhances Gemini with long-format Q&A datasets, improving performance in complex medical reasoning tasks.

Despite significant progress, existing medical multimodal models still face limitations in both data and training methods. For instance, these models typically rely on limited medical datasets, which hinder their ability to generalize effectively, especially in complex medical scenarios that require reasoning and decision-making. To address these challenges, we leverage high-quality medical reasoning data and apply reinforcement learning (RL) to enhance the reasoning and decision-making capabilities of medical multimodal models.

Reasoning Models Reasoning models are essential in multimodal learning, particularly for tasks that require integrating modalities like images and text for decision-making. As multimodal reasoning demand grows, frameworks and techniques have evolved. The introduction of OpenAI-O1 (Jaech et al., 2024) marked a key development, though it still struggles with complex vision-language tasks. Chain-of-Thought (CoT)(Wei et al., 2022) breaks reasoning into steps, improving capability, while Progressive Reasoning Models (PRM)(Lightman et al., 2023) optimize outcomes through intermediate supervision. Sky-T1 (Team, 2025) enhances high-level reasoning with supervised fine-tuning (SFT), and LIMA (Zhou et al., 2023) proves that limited SFT data can still achieve strong reasoning performance. Similarly, (Huang et al., 2025) shows minimal SFT data can boost reasoning.

However, while SFT primarily optimizes the structure and logic of reasoning, its capacity for more complex tasks remains limited. In response, recent advancements in reinforcement learning (RL)-based reasoning models have shown significant progress (Ziegler et al., 2019; Ouyang et al., 2022; Zhou et al., 2024; DeepSeek-AI, 2025). For example, DeepSeek-R1 (DeepSeek-AI, 2025) leverages RL to refine the reasoning process, improving efficiency and accuracy in complex tasks. The Qwen-QwQ (Team, 2024) model expands reasoning capabilities, particularly in long-context reasoning, by utilizing large-scale context processing and RL to deepen the model's reasoning abilities. Moreover, Kimi-R1.5 (Team et al., 2025) introduced an innovative framework combining long-chain-of-thought (long-CoT) with RL, significantly enhancing performance in multimodal reasoning tasks and demonstrating potential for more complex tasks.

Thus, the combination of SFT and RL not only strengthens the fundamental logic and structure of reasoning but also enhances knowledge generalization and flexibility through RL's strategy adjustments (Chu et al., 2025). This combined approach allows the model to perform effective reasoning and decision-making with limited data, expanding its capability in more complex tasks.

3 Methodology

3.1 GMAI-Reasoning10K

To explore reinforcement learning tuning in medical multimodal analysis, we first curate a high-quality visual question-answering dataset tailored for medical image analysis.

Visual Question Answering. Our dataset construction process began with the collection of multimodal medical data from 95 datasets from reliable sources such as Kaggle, GrandChallenge, and Open-Release, covering 12 imaging modalities (e.g., X-ray, CT, MRI). The data preprocessing pipeline was based on methods from SAMed-20M (Ye et al., 2023), which included augmentation and standardization; for instance, 3D data (CT/MRI) had individual slices extracted and pixel values normalized to 0–255, while video data was processed by extracting frames at 2 frames per second. Key metadata, such as background information, modality, and labels, was extracted from each dataset and subsequently used to construct informative prompts via GPT for generating multiple-

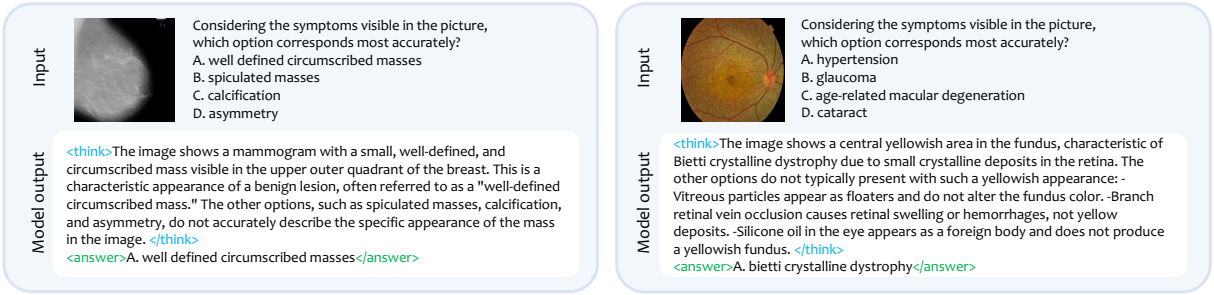

Figure 2: **Case study illustrating the model's reasoning ability under Reinforcement Learning Tuning (RLT)**. Given medical images, the model identifies the most accurate diagnosis based on visible symptoms. RLT encourages the model to engage in reasoning and select the correct answer from multiple choices.

choice questions (with a single correct answer). To ensure quality, a reject sampling strategy was applied to discard samples that did not meet our predefined criteria, such as relevant annotations, or correct labels, resulting in a high-quality dataset comprising 10K samples. This data distillation approach has also been adopted by numerous studies (DeepSeek-AI, 2025; Team, 2025; Liu et al., 2025). It should be noted that our data selection avoids any overlap with the test data used in commonly-used benchmarks. The distribution of this dataset is illustrated in Fig. 1-b), and more curation details can be found in the supplementary material.

Chain-of-Thought Instruction. In contrast to concurrent RFT approaches (Liu et al., 2025) in general vision tasks that directly use questions and answers as instructions for SFT, we constructed detailed chain-of-thought (COT) explanations for each VQA pair to ensure a fair comparison between RFT and SFT. Specifically, for each VQA pair, we prompted GPT-4o to generate a comprehensive COT for each VQA pair, see the Appendix for detailed prompt information. Similarly, we used a special curation strategy to filter and refine high-quality COTs. This process ultimately resulted in the creation of the GMAI-Reasoning10K dataset, which comprises 10K VQA pairs along with their corresponding reasoning COTs, serving as a resource for developing and comparing instruction-tuning-based methods.

3.2 Reinforcement Learning Tuning.

Our Reinforcement Learning Tuning (RLT) pipeline is shown in Fig. 1-a), which is based on Group Relative Policy Optimization (GRPO) framework. We are the first to apply GRPO to the multimodal medical domain, with extensive validation conducted at scale. We apply this simple yet effective framework to the base model (e.g., Qwen2.5-VL-7b (Bai et al., 2025)), which enables the model to develop advanced medical reasoning capabilities on its own without relying on supervised data.

Algorithm 1 Reinforcement Learning Tuning

Input: Medical image I, question q, and answer x
Define policy model π_θ^{RL} and reference model π_{ref}
Define reward: $r(x, q, y) = r_{\text{acc}} + r_{\text{fmt}} + r_{\text{rep}}$
for each iteration t **do**
  for each input (x_i, q_i) **do**
    Sample outputs: $\{y_i^j\}_{j=1}^k \sim \pi_{\theta_{\text{RL}}^t}(x_i, q_i)$
    Compute advantage:
$$A(x_i, y_i^j) = r(x_i, y_i^j) - \frac{1}{k}\sum_{l=1}^{k} r(x_i, y_i^l)$$
  end for
  GRPO loss:
$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}\Big[A(x,q,y)\cdot\min\Big(\frac{\pi_\theta(y|x,q)}{\pi_{\theta_{\text{RL}}^t}(y|x,q)}, 1+\epsilon\Big)\Big]$$
  Total loss: $\mathcal{L}_{\text{RL}} = \mathcal{L}_{\text{GRPO}} + \beta \cdot D_{KL}(\pi_\theta \parallel \pi_{\text{ref}})$
  Update: $\theta_{\text{RL}}^{t+1} \leftarrow \theta_{\text{RL}}^t - \alpha \nabla_\theta \mathcal{L}_{\text{RL}}$
end for
return Final model $\pi_{\theta_{\text{RL}}}$

Preliminaries. Let I denote a medical image and (q, x) the associated VQA pair. We initialize a reference model π_{ref} and a reinforcement learning model π_θ^{RL} (both obtained from base model), which will be optimized via our RLT pipeline. The RLT pipeline begins with the base Q&A pair (q, x), where q represents the question and x denotes the ground truth answer. The baseline policy, π_{ref}, guides the learning of the reinforcement learning policy, π_θ^{RL}, through a carefully designed reward mechanism.

Group Sampling. During each iteration t, for each input q_i, we sample a group of k outputs $\{y_i^j\}_{j=1}^k$ from the current policy $\pi_{\theta^t}^{\text{RL}}(q_i)$. Group sampling enables the model to explore diverse re-

sponses for the same input, which is essential for robust policy evaluation.

Reward Function. Each output is evaluated using a reward function $r(x, y)$ that encapsulates three key components: $r(x, y) = r_{\text{acc}}(x, y) + r_{\text{fmt}}(x, y) + r_{\text{rep}}(x, y)$, where:

- Accuracy Reward (r_{acc}): Assesses whether the response is correct (e.g., matching a standard answer for multiple-choice questions).
- Format Reward (r_{fmt}): Ensures that the model encloses its reasoning within designated tags (e.g., `<think>` and `</think>`) and wraps the final answer within `<answer>` and `</answer>` tags.
- Repetition Penalty Reward (r_{rep}): Penalizes repetitive or redundant outputs to encourage concise and diverse reasoning.

Advantage Computation. For each sampled output y_i^j, we compute the advantage function:

$$A(x_i, y_i^j) = r(x_i, y_i^j) - \frac{1}{k} \sum_{l=1}^{k} r(x_i, y_i^l), \quad (1)$$

which quantifies the relative performance of each output compared to the average reward of the group.

Policy Optimization. The core of our approach is the GRPO loss, defined as:

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}\left[A(x, y) \cdot \min\left(\frac{\pi_\theta(y|x)}{\pi_{\theta_{\text{RL}}^t}(y|x)}, 1 + \epsilon\right)\right], \quad (2)$$

where the clipping operation ensures stability in the updates. To regularize the learning process and prevent drastic deviations from the supervised baseline, we include a KL divergence term:

$$\mathcal{L}_{\text{RL}} = \mathcal{L}_{\text{GRPO}} + \beta \cdot D_{KL}(\pi_\theta \parallel \pi_{\text{ref}}). \quad (3)$$

The hyperparameter β (initialized to 0.04) acts as a weighting factor to balance the KL divergence regularization term and the main GRPO loss, ensuring a controlled trade-off between the policy model's stability and creativity. The model parameters are updated using gradient descent:

$$\theta_{\text{RL}}^{t+1} \leftarrow \theta_{\text{RL}}^t - \alpha \nabla_\theta \mathcal{L}_{\text{RL}}, \quad (4)$$

This iterative process continues until convergence, thereby progressively self-improving the model's medical reasoning capabilities.

Benchmark	# Samples	Out-of-distribution?
MMMU(H&M) (Yue et al., 2024a)	150	✓
MMMU-pro (Wang et al., 2024)	288	✓
GMAI-MMBench (*val*) (Chen et al., 2024b)	4,550	✓
GMAI-MMBench (*test*) (Chen et al., 2024b)	21,281	✓
MedXpertQA-MM (Zuo et al., 2025)	2,000	✓
OmniMedVQA (Hu et al., 2024)	127,995	✗

Table 1: **Evaluation benchmarks used in our experiments**. Note that OmniMedVQA and GMAI-Reasoning10K share the same curated dataset distribution, despite being derived from non-overlapping test and training splits.

4 Experiments

4.1 Setup

Benchmarks. We utilized various widely used benchmarks to conduct a large-scale and comprehensive evaluation of RTL's effectiveness in multimodal medical image analysis. Our evaluation benchmarks include OmniMedVQA (Hu et al., 2024), GMAI-MMBench (Chen et al., 2024b), MMMU (H&M) (Yue et al., 2024a), MMMU-pro (Wang et al., 2024), and MedXpertQA (Zuo et al., 2025), each designed to assess different facets of medical image understanding and question answering. As summarized in Table 1, the data counts and distributions are presented. Notably, OmniMedVQA is treated as an *i.i.d.* test set since it shares the same curated dataset distribution as GMAI-Reasoning10K (albeit derived from non-overlapping test and training splits), while the remaining benchmarks are considered out-of-distribution (OOD). Further details on these benchmarks are provided in Sec. B in the Appendix.

Setting. In our experiments, we compare several methods: (i) the base model; (ii) +SFT, which is fine-tuned using the VQA pairs and corresponding CoT instructions from GMAI-Reasoning10K; and (iii) +RLT, where only the VQA pairs from GMAI-Reasoning10K are used for reinforcement learning tuning. For all experiments, we default to the Qwen-VL-7B model. Specifically, we apply LLaMAFactory (Zheng et al., 2024) for the supervised fine-tuning stage. During training, we use AdamW as the optimizer with a learning rate of $1e^{-5}$ following a cosine decay schedule and a batch size of 32. The SFT training is run for 2 epochs. For the RLT stage, we utilize the repository[1] to perform RL training with GRPO, setting the number of generations per group to 7, and the

[1] https://github.com/EvolvingLMMs-Lab/open-r1-multimodal

Method	MMMU	MMMU-pro	GMAI-MMbench(*val*)	GMAI-MMbench(*test*)	MedXpertQA-MM	OmniMedVQA
Med-Flamingo (7B) (Moor et al., 2023)	28.40	_	12.74	11.64	_	23.82
RadFM (13B) (Wu et al., 2023b)	27.90	_	22.95	22.93	_	23.48
LLaVA-Med (7B) (Li et al., 2024a)	38.60	_	20.54	19.60	_	27.82
HuatuoVision (34B) (Chen et al., 2024a)	50.30	_	_	_	_	**73.23**
MedDR (40B) (He et al., 2024)	_	_	41.95	43.69	_	68.27
Base (Qwen2.5-vl-7B)	55.33	28.47	40.02	40.59	20.30	58.41
Base+SFT	56.00	32.99	39.65	36.12	23.55	66.34
Base+RLT	**57.33**(+2.00)	**34.03**(+5.56)	**43.14**(+3.12)	**43.84**(+3.25)	**23.80**(+3.50)	61.01(+2.60)

Table 2: **Comparison of different tuning strategies across multiple medical multimodal benchmarks, along with state-of-the-art methods**. Numbers in parentheses indicate the performance gain over the base model. The best results are bolded.

RL training process runs for one epoch. After training, we leverage VLMEvalKit (Duan et al., 2024) to evaluate model performance on the benchmarks. We refer readers to Table 5 in the Appendix for more details of the network training.

4.2 Main Results

RLT Generalizes Better. We first explored supervised fine-tuning SFT and reinforcement learning tuning RLT on the base model. As shown in Table 2, our ablation study reveals that while SFT serves as a strong baseline for in-distribution tasks, it surpasses the base model by a large margin on OmniMedVQA 66.34% versus 58.41%. However, SFT appears to rely more on in-distribution pattern learning, which limits its robustness on out-of-distribution tasks. For instance, on benchmarks such as GMAI-MMbench, SFT leads to performance degradation compared to the base model, dropping from 40.02% to 39.65% on GMAI-MMbench (*val*) and from 40.59% to 36.12% on GMAI-MMbench (*test*).

In contrast, the reinforcement learning tuning approach significantly improves performance across both in-distribution and out-of-distribution tasks. Despite the in-distribution benchmarks such as OmniMedVQA, it improves on the base model by 2.60% percent, reaching 61.01% percent, although it remains lower than SFT. However, on out-of-distribution benchmarks, RLT demonstrates a clear advantage. It surpasses both the base and SFT models on GMAI-MMbench val and test by +3.12% and +3.25% respectively, and achieves notable gains on MMMU and MMMU-pro with improvements of +2.00% and +5.56%. Similarly, on MedXpertQA-MM, RLT improves performance by +3.50% over the base model. These results indicate that reinforcement learning tuning encourages more robust and transferable reasoning abilities across diverse and complex tasks. While SFT is effective for tasks that rely on in-distribution pattern recognition, RLT cultivates deeper and more generalizable reasoning capabilities, making it better suited for handling out-of-distribution challenges. This finding is consistent with some recent observations that RLT generalizes while SFT memorizes (Chu et al., 2025; Liu et al., 2025).

RLT Works Efficiently. We compared our RLT approach with current state-of-the-art models that employ extensive supervised fine-tuning data. For example, while models such as Med-Flamingo (7B) and LLaVA-Med (7B) rely on millions of annotated QA pairs, and larger models like HuatuoVision (34B) and MedDR (40B) benefit from vast amounts of SFT data (e.g., 1.3 million samples for HuatuoVision and 2 million for MedDR), our Base+RL model is built on a 7B base model using only 10K unannotated QA pairs. For instance, our Base+RL model achieves an MMMU score of 57.33%, significantly outperforming HuatuoVision's 50.30% on the same benchmark. In addition, on GMAI-MMbench, Base+RL obtains 43.14% on the validation set and 43.84% on the test set, surpassing MedDR's scores of 41.95% and 43.69%, respectively. These results highlight that SFT not only enhances reasoning capabilities but does so efficiently, requiring far less computational and data-intensive resources compared to existing state-of-the-art models, which could be of significant benefit in resource-constrained scenarios such as medical applications.

Fine-grained Analysis of RLT. We further conduct fine-grained analysis on the GMAI-MMBench benchmark, which is a comprehensive medical multimodal benchmark designed to evaluate models on a range of clinical VQA tasks, as detailed in Table 3. The results reveal that RLT significantly improves performance on more generalizable tasks, such as Disease Diagnosis (**DD**), Attribute Recognition

Model Name	Overall (*val*)	Overall (*test*)	AR	BVR	B	CR	C	DD	IQG	MR	M	NT	OR-A	OR-HN	OR-P	OR-T	SG	SAR	SIR	SWR
Random	25.70	25.94	38.20	22.73	22.92	22.72	24.06	26.66	27.13	27.00	20.00	24.75	21.37	22.93	22.33	21.18	32.43	24.23	21.39	23.71
Claude3-Opus (Anthropic, 2024)	32.37	32.44	1.61	39.51	34.31	31.66	12.63	39.26	28.74	30.86	22.40	37.37	25.79	41.07	29.33	33.18	31.31	21.35	23.87	4.00
Qwen-VL-Max (Bai et al., 2023)	41.34	42.16	32.68	44.58	31.38	40.79	10.68	50.53	32.79	44.36	29.20	51.52	41.37	58.00	30.67	41.65	26.95	25.00	24.64	39.14
GPT-4V (Achiam et al., 2023)	42.50	44.08	29.92	48.95	44.00	37.39	12.93	52.88	32.79	44.21	32.80	63.64	39.89	54.13	37.00	50.59	27.55	23.08	25.75	37.43
Gemini 1.0 (Team et al., 2023)	44.38	44.93	42.12	45.10	46.46	37.57	20.45	53.29	35.22	36.94	25.20	51.01	34.74	59.60	34.00	50.00	36.64	23.65	23.87	35.43
Gemini 1.5 (Reid et al., 2024)	47.42	48.36	43.50	56.12	51.23	47.58	2.26	55.33	38.87	48.07	30.00	76.26	51.05	75.87	46.33	62.24	20.57	27.69	30.54	40.57
GPT-4o (Achiam et al., 2023)	53.53	53.96	38.32	61.01	57.08	49.02	46.62	61.45	46.56	56.38	34.00	75.25	53.79	69.47	48.67	65.88	33.93	22.88	29.51	39.43
Med-Flamingo (Moor et al., 2023)	12.74	11.64	6.67	10.14	9.23	11.27	6.62	13.43	12.15	6.38	8.00	18.18	9.26	18.27	11.00	11.53	12.16	5.19	8.47	11.43
LLaVA-Med (Li et al., 2024a)	20.54	19.60	24.51	17.83	17.08	19.86	15.04	19.81	20.24	21.51	13.20	15.15	20.42	23.73	17.67	19.65	21.70	19.81	14.11	20.86
Qilin-Med-VL-Chat (Liu et al., 2023)	22.34	22.06	29.57	19.41	16.46	23.79	15.79	24.19	21.86	16.62	7.20	13.64	24.00	14.67	12.67	15.53	26.13	24.42	17.37	25.71
RadFM (Wu et al., 2023b)	22.95	22.93	27.16	20.63	13.23	19.14	20.45	24.51	23.48	22.85	15.60	16.16	14.32	24.93	17.33	21.53	29.73	17.12	19.59	31.14
MedDr (He et al., 2024)	41.95	43.69	41.20	50.70	37.85	29.87	28.27	52.53	36.03	31.45	29.60	47.47	33.37	51.33	32.67	44.47	35.14	25.19	25.58	32.29
Base(Qwen2.5-VL-7B)	**40.02**	**40.55**	**44.89**	**46.21**	**34.37**	**37.79**	**9.47**	**47.42**	**35.60**	**36.35**	**23.20**	**44.00**	**38.35**	**45.55**	**35.20**	**37.53**	**32.85**	**24.23**	**28.38**	**37.71**
Base+SFT	**39.65**	**36.12**	**41.33**	**38.49**	**31.84**	**39.57**	**1.95**	**41.60**	**31.20**	**22.11**	**22.80**	**47.50**	**39.21**	**41.42**	**31.20**	**33.76**	**31.28**	**24.62**	**25.91**	**30.29**
Base+RLT	**43.14**(+3.12)	**43.84**(+3.29)	**51.41**(+6.52)	**49.00**(+2.79)	**35.63**(+1.26)	**41.35**(+3.56)	**26.47**(+17.00)	**50.64**(+3.22)	**37.20**(+1.60)	**37.54**(+1.19)	**25.60**(+2.40)	**49.00**(+5.00)	**39.55**(+1.20)	**49.94**(+4.39)	**40.00**(+4.80)	**42.24**(+4.71)	**33.81**(+0.96)	**25.58**(+1.33)	**28.29**(-0.09)	**40.57**(+2.86)

Table 3: **Results on the *val* and *test* sets of GMAI-MMBench for clinical VQA tasks**.The sub-items behind are evaluated on the test set. Numbers in parentheses indicate the performance gain over the base model. The full names of the evaluated tasks can be found in Table 5 in literature (Chen et al., 2024b).

Model	Type	MMMU(H&M)	GMAI-MM(*val*)
Base(7B)	directly	55.33	37.47
	cot	55.33	40.02
Base(7B)+RLT	directly	54.67	45.07
	cot	57.73	43.14

(a) **Inference strategy**.

Type	MMMU(H&M)	GMAI-MM(*val*)
Base(3B)	53.33	39.05
Base(3B)+RLT	51.00(-2.33)	40.45(+1.40)
Base(7B)	55.33	40.02
Base(7B)+RLT	57.73(+2.40)	43.14(+3.12)

(b) **Model size**.

Step	MMMU(H&M)	GMAI-MM(*val*)
0	55.33	40.02
10	56.00(+0.67)	41.08(+1.06)
20	54.00(-1.33)	42.29(+2.27)
100	56.60(+1.27)	42.97(+2.95)
165	57.73(+2.40)	43.14(+3.12)

(c) **Training steps (RLT)**.

Table 4: Ablation experiments on the MMMU H&M and the GMAI-MMBench *val* set.

(**AR**), and Organ Recognition (**OR-A/HN/P/T**). These tasks require nuanced analysis and reasoning. For instance, RLT improves Disease Diagnosis performance from 47.42% (Base+SFT) to 50.64% (Base+RL). In a more extreme example, for counting (**C**) tasks, SFT leads to a 7.52% drop in accuracy, while RL boosts performance by 17%, highlighting its advantage in tasks that require high-level reasoning and deeper understanding of image content. However, tasks like Surgical Instrument Recognition (**SIR**) show no significant performance difference between RLT and SFT, indicating that RL does not provide substantial benefits for more straightforward recognition tasks. This is likely because, unlike organs or attributes with varied representations, surgical instruments have a more uniform appearance, suggesting that for simpler recognition tasks, fine-tuning is sufficient to achieve competitive results and RLT's impact is less pronounced.

No "Aha Moment" Yet. In language models, "aha moments" refer to self-validation behaviors in which the model explicitly corrects or refines its reasoning during generation. Such behaviors are often observed in complex reasoning tasks such as mathematics or coding, where longer chain-of-thought outputs may reveal intermediate corrections and insights. However, in our medical visual question-answering experiments, we did not observe clear "aha moments"; instead, the generated responses show noticeably shorter reasoning chains, as shown in Figure 3. This may be because medical VQA tasks often rely more on direct associations between visual cues and clinical concepts than on extended symbolic reasoning. Therefore, whether and how "aha moments" emerge in medical multimodal reasoning remains an open direction for further exploration.

4.3 More results

Inference Strategy. We compare two inference strategies: direct answer generation versus Chain of Thought (CoT)−based reasoning followed by answer generation. The results, summarized in Table 4 (a), show that while CoT reasoning can enhance the model's performance in some cases, it does not always lead to improvements. In particular, we observe that CoT reasoning slightly improves performance on GMAI-MMBench (*val*), but it does not consistently outperform direct an-

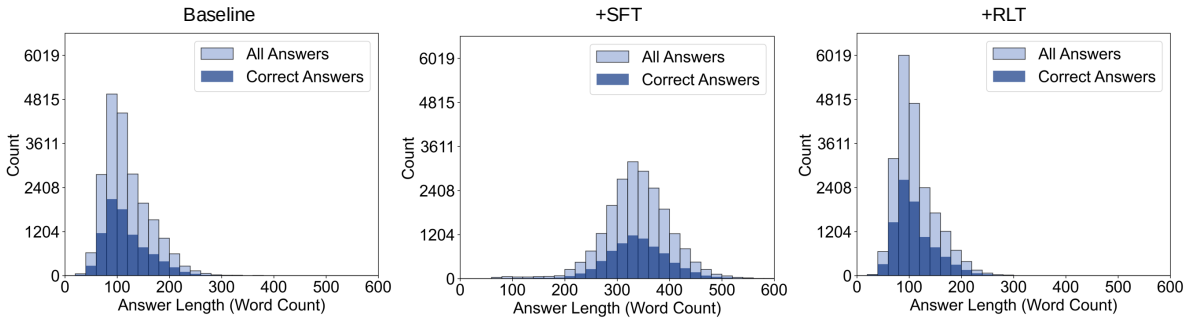

Figure 3: **Distribution of generated answer lengths (in word count) for the Baseline, +SFT, and +RLT models**. Each histogram displays the total number of answers (light bars) and correct answers (dark bars).

swer generation across all settings. After applying RLT, both inference strategies show significant improvements. Direct answer generation achieves scores of 54.67% on MMMU and 45.07% on GMAI-MMBench, whereas CoT reasoning leads to 57.73% on MMMU but a slight drop to 43.14% on GMAI-MMBench. In some cases, however, CoT reasoning hurts performance because overly long or repetitive outputs can obscure the final answer, suggesting that structured reasoning should be controlled to avoid unnecessary verbosity.

Model Size. To evaluate the impact of model size, we compare the performance of two model variants: Qwen2.5-VL-3B and Qwen2.5-VL-7B, both before and after RLT. After RLT, the 7B model shows a notable improvement, while the 3B model improves by 1.40 points on GMAI-MMBench but decreases by 2.33% on MMMU. These results suggest that stronger base models benefit more from RLT, as the 7B model can better leverage reinforcement signals to refine its reasoning. This highlights the importance of model scaling when applying RL to medical multimodal reasoning.

Training Steps. To evaluate the effect of the RLT training steps, we assess the performance of the 7B model at various intervals. The results, summarized in Table 4 (c), demonstrate a generally stable improvement in performance as the number of training steps increases. After 10 RLT steps on the base model, the scores increase by 0.67% on MMMU and 1.06% on GMAI-MMBench, showing that even a small amount of RL feedback can enhance performance. The results in the table also demonstrate the stability of our RLT method, as longer training enables the model to consistently refine its decision-making and reasoning.

5 Conclusion

In this paper, we introduce GMAI-VL-R1, a multimodal medical reasoning model that leverages reinforcement learning tuning to enhance reasoning and reflective capabilities. To address reasoning challenges in complex medical decision-making, we construct GMAI-Reasoning10K and apply RLT to improve the model's reasoning quality and generalization ability. Experimental results demonstrate that GMAI-VL-R1 outperforms current state-of-the-art multimodal medical models across several benchmark tasks, particularly in medical image diagnosis and visual question answering. These findings highlight the value of reinforcement learning for building more generalizable multimodal medical reasoning models. Notably, with only 10K curated VQA pairs, GMAI-VL-R1 rivals models trained on orders-of-magnitude more SFT data, indicating that RLT mainly cultivates transferable reasoning rather than memorizing in-distribution patterns.

6 Limitations

Although GMAI-VL-R1 shows strong performance and generalization across multiple benchmarks, it still has limitations. First, the model is evaluated primarily on curated benchmarks rather than prospective real-world clinical workflows, so deployment performance may vary across institutions, devices, and populations. Second, RLT improves reasoning behavior but does not guarantee factual correctness, especially under ambiguous visual evidence or rare conditions. Third, our setting covers only a limited range of model scales and reward designs, leaving data diversity, reward robustness, and long-horizon clinical reasoning for future work. The model is intended for research use only and is not a substitute for clinician judgment.

References

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

Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. 2022. Multimodal biomedical ai. *Nature medicine*, 28(9):1773–1784.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Junying Chen, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, and 1 others. 2024a. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*.

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, and 1 others. 2024b. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *arXiv preprint arXiv:2408.03361*.

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

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, and 1 others. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Haodong Duan, Junming Yang, and 1 others. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201.

Iryna Hartsock and Ghulam Rasool. 2024. Vision-language models for medical report generation and visual question answering: a review. *Frontiers in Artificial Intelligence*, 7:1430984.

Sunan He, Yuxiang Nie, Zhixuan Chen, Zhiyuan Cai, Hongmei Wang, Shu Yang, and Hao Chen. 2024. MedDr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. *arXiv preprint arXiv:2404.15127*.

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

Zhongzhen Huang, Gui Geng, Shengyi Hua, Zhen Huang, Haoyang Zou, Shaoting Zhang, Pengfei Liu, and Xiaofan Zhang. 2025. O1 replication journey–part 3: Inference-time scaling for medical reasoning. *arXiv preprint arXiv:2501.06458*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. 2022. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1):171.

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

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Jiajia Li, Zhouyu Guan, Jing Wang, Carol Y Cheung, Yingfeng Zheng, Lee-Ling Lim, Cynthia Ciwei Lim, Paisan Ruamviboonsuk, Rajiv Raman, Leonor Corsino, and 1 others. 2024b. Integrated image-based deep learning and language models for primary diabetes care. *Nature Medicine*, pages 1–11.

Tianbin Li, Yanzhou Su, Wei Li, Bin Fu, Zhe Chen, Ziyan Huang, Guoan Wang, Chenglong Ma, Ying Chen, Ming Hu, Yanjun Li, Pengcheng Chen, Xiaowei Hu, Zhongying Deng, Yuanfeng Ji, Jin Ye, Yu Qiao, and Junjun He. 2024c. Gmai-vl & gmai-vl-5.5m: A large vision-language model and a comprehensive multimodal dataset towards general medical ai. *Preprint*, arXiv:2411.14522.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike,

John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Jana Lipkova, Richard J Chen, Bowen Chen, Ming Y Lu, Matteo Barbieri, Daniel Shao, Anurag J Vaidya, Chengkuan Chen, Luoting Zhuang, Drew FK Williamson, and 1 others. 2022. Artificial intelligence for multimodal data integration in oncology. *Cancer cell*, 40(10):1095–1110.

Junling Liu, Ziming Wang, Qichen Ye, Dading Chong, Peilin Zhou, and Yining Hua. 2023. Qilin-med-vl: Towards chinese large vision-language model for general healthcare. *arXiv preprint arXiv:2310.17956*.

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

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, and 1 others. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1.5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

NovaSky Team. 2025. Sky-t1: Train your own o1 preview model for under $450. Available at: https://novasky-ai.github.io/posts/sky-t1. Accessed: 2025-01-09.

Qwen Team. 2024. Qwq: Reflect deeply on the boundaries of the unknown.

Omkar Chakradhar Thawakar, Abdelrahman M Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Khan. 2024. Xraygpt: Chest radiographs summarization using large medical vision-language models. In *Proceedings of the 23rd workshop on biomedical natural language processing*, pages 440–448.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, and 1 others. 2024. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

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

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023a. Pmc-llama: Further fine-tuning llama on medical papers. *arXiv preprint arXiv:2304.14454*, 2(5):6.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023b. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*.

Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, and 1 others. 2024. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*.

Jin Ye, Junlong Cheng, Jianpin Chen, Zhongying Deng, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, and 1 others. 2023. Samed2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks. *arXiv preprint arXiv:2311.11969*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, and 1 others. 2024b. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, and 1 others. 2024. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, pages 1–13.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, and 1 others. 2023. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021.

Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. 2024. Archer: Training language model agents via hierarchical multi-turn rl. *arXiv preprint arXiv:2402.19446*.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *Preprint*, arXiv:2501.18362.

Appendix

A Training Setting

Table 5: Training settings of GMAI-VL-R1's Stage I (SFT) and Stage II (RL).

Settings	SFT	RL
freeze LLM	False	False
freeze MLP	False	False
freeze Vision Encoder	False	False
learning rate	1e-4	1e-5
optimizer	AdamW	AdamW
optimizer hyper-parameters	$\beta_1 = 0.9, \beta_2 = 0.999$	$\beta_1 = 0.9, \beta_2 = 0.999$
total batch size	8x4x8	8x1x8
drop rate	0.0	0.0
numerical precision	DeepSpeed bf16	DeepSpeed bf16
GPUs for training	8xA100 (80G)	8xA100 (80G)

B Benchmarks

Below is a brief overview of the benchmarks used in our experiments:

- **MMMU Health & Medicine track**: The Health & Medicine track of the MMMU (Yue et al., 2024a) benchmark spans a wide range of medical fields, derived from university exams, quizzes, and textbooks. It evaluates the model's reasoning ability in complex medical scenarios and the specialized knowledge in health and medicine.

- **MMMU-Pro Health & Medicine track**:MMMU-Pro (Yue et al., 2024b) is an enhanced version of the MMMU (Yue et al., 2024a) benchmark, designed to test multimodal models' reasoning by filtering text-only questions, expanding answer options, and adding vision-only input. It better mimics real-world scenarios, requiring integration of visual and textual information, and reveals current models' limitations.

- **OmniMedVQA**: OmniMedVQA (Hu et al., 2024) provides a rich dataset of paired medical images and text, designed to evaluate the model's ability to recognize and understand fundamental medical imaging concepts, focusing on cross-modal reasoning and information integration.

- **GMAI-MMBench**: GMAI-MMBench (Chen et al., 2024b) focuses on assessing the model's ability to identify fine-grained objects in complex clinical scenarios, challenging its capacity to handle long-context tasks and accurately recognize and reason over detailed medical features. In addition, it consists of both validation and test sections.

- **MedXpertQA-MM**: MedXpertQA (Zuo et al., 2025) is a comprehensive and challenging benchmark designed to evaluate expert-level medical knowledge and advanced reasoning, introducing expert-level exam questions with diverse clinical information, including patient records and examination results, distinguishing it from traditional medical multimodal benchmarks.

C System Promot

Table 6: System prompt

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user a concise final answer in a short word. The reasoning process and answer are enclosed within <think> reasoning process here </think>and <answer> answer here </answer>tags, respectively, i.e., <think>reasoning process here </think><answer>answer here </answer>"

D License and Intended Use

All datasets, base models, and toolkits used in this work (including Qwen2.5-VL, LLaMAFactory, open-r1-multimodal, VLMEvalKit, and the 95 source medical VQA datasets) are publicly available under their original licenses and are used solely for non-commercial academic research, consistent with their stated intended use. GMAI-Reasoning10K and GMAI-VL-R1 will be released under a research-only license; they are intended for medical AI research and are not validated or authorized for clinical deployment or patient care.

Table 7: Prompt Template for Medical Imaging Multi-Choice Question Construction

QUESTION: Based on the given medical image and the answer set, propose a question whose answer can be inferred from the image. The correct answer to the question must come from one of the elements in the answer set; no new or extended answers may be added.

CHOICE: Label each option sequentially with letters (A, B, C, D, ...). Each option's content must strictly match one element in the answer set, without repetition or new additions. If the answer set has more than 4 items, randomly select 4 of them; if it has fewer than 4, only create as many options as exist. Ensure that the correct answer is among these options.

ANALYSIS: First, provide an overview of your solution strategy: how you combine image and clinical knowledge to make a judgment. Then, explain the reasoning process step by step: at each step, detail what information you derive from the image and from medical knowledge, and what conclusions you draw. Finally, analyze each option in detail, explaining whether or not it could be the correct answer, and arrive at the final conclusion.

ANSWER: Provide only the letter corresponding to the correct option (e.g., "A" or "B"), without repeating the text of that option.

Sample Output (json format):

```
{
"QUESTION": "Enter the question here, which should be inferable from the image",
"CHOICE": {
"A": "Content of option A",
"B": "Content of option B",
"C": "Content of option C",
"D": "Content of option D"
},
"ANALYSIS": "Enter the detailed analysis here",
"ANSWER": "Letter corresponding to the correct option"
}
```

Table 8: Prompt Template for Medical Imaging Question Construction

As a professional medical imaging expert, your responsibility is to thoroughly explore medical images and provide accurate, precise answers based on your clinical expertise. You will be given an image and a set of possible answers, and your task is to construct a question and provide a reasoned answer.

The reasoning process should be enclosed within <think>tags, and the final answer should be enclosed within <answer>tags. Additionally, the question should be enclosed within <question>.

Steps:

1. **Question Construction:**

Based on the provided image and the possible answers, construct a question that can be answered based on the image. The question must be clear and directly relevant to the information in the image.

2. **Reasoning Process:**

- Analyze the image step by step without referencing the possible answers directly.

- Provide the step-by-step reasoning process, explaining the rationale and supporting each step.

- Consider the clinical knowledge required and how the image provides insight into the condition being examined.

- Clearly explain the steps of the reasoning process that lead to the final conclusion.

3. **Final Answer:**

- After reasoning, provide the final answer derived from the image analysis. This must be one of the options in the answer set. The answer should be concise and logically follow from the reasoning process.

Input Information:

- Imaging Modality: {modality}

- Background: {knowledge}

- Answer Set: {answer_set}

Output Format:

<question>Your question here, based on the image and answer set </question>

<think >Your detailed reasoning process, step by step </think>

<answer >Your final answer, matching one of the options in the answer set </answer>