Visual Reinforcement Fine-Tuning via Bootstrapped Medical Reasoning

Yequan Bie* **Peng Xie*** **Zhixuan Chen** **Yihui Wang**
Andong Tan **Linshan Wu** **Sunan He** **Jianda Mao**
Yangqiu Song **Kani Chen**✉ **Hao Chen**✉
The Hong Kong University of Science and Technology
{ybie, pxieaf}@connect.ust.hk

Abstract

Multimodal large language models (MLLMs) have made significant strides in natural vision-language understanding, yet their potential in specific domains, such as healthcare, remains largely untapped. Since existing medical models often lack the reasoning capabilities needed for complex decision-making, adapting general MLLMs for medical reasoning has attracted increasing interest, with reinforcement fine-tuning (RFT) emerging as a promising approach. However, medical reasoning tasks require both precise thinking processes and generalization of well-justified answers, presenting unique challenges due to the inherent scarcity of annotated data and nuanced visual complexity of medical images. Current visual RFT methods prioritize answer correctness through verifiable rewards while neglecting the reasoning process, leading to limited reasoning capabilities and sub-optimal performance, which are essential in high-stakes scenarios like healthcare. To address these issues, we propose **MedR2FT**, a **Med**ical **R**easoning **R**einforcement **F**ine-**T**uning framework that enhances medical reasoning through rationale bootstrapping and reasoning rewarding. Specifically, MedR2FT leverages distilled rationales from fundamental training phases to supervise subsequent reasoning processes with semantic rewards, alleviating the challenge of scarce reasoning data and enhancing model reasoning. Extensive experiments demonstrate that our method consistently outperforms baselines on visual question answering and lesion detection with a significant margin. MedR2FT explicitly attends to the reasoning process, significantly enhancing MLLM adaptability for medical reasoning. The code is available at https://github.com/Tommy-Bie/MedR2FT.

1 Introduction

Multimodal large language models (MLLMs) have showcased significant progress in reasoning and generation over natural images and languages [1–6]. The powerful multimodal and reasoning capability in different tasks, such as image classification and visual question answering [7, 8], makes these models widely deployed in real-world applications. The achievements are mainly attributed to pre-training and supervised fine-tuning (SFT) on a large amount of high-quality data [9, 10]. However, unlike the natural field, the medical imaging domain faces some critical challenges. For example, the well-annotated data in healthcare is very limited and expensive, as medical data labeling is time-consuming and requires the efforts of human experts. This makes it even harder for medical models to be supervised fine-tuning on large-scale and high-quality Chain-of-Thought (CoT) data [11]. Another challenge is that, as a high-stakes field, healthcare demands high trustworthiness

* Equal contributions.
✉ Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

standards [12, 13], requiring AI models to provide accurate decision-making process, *e.g.*, lesion localization and finding evaluation, before giving the final answer or diagnosis to gain doctors' and patients' trust. This poses another issue since existing medical models are struggling to provide traceable reasoning processes due to the lack of medical reasoning data and the inherently nuanced visual features of medical images. Some models get even lower performance when thinking before answering [14]. Hence, most models can only give black-box predictions without offering reasonable thinking processes, which hinders their adoption in medical applications.

The scarcity of CoT data for SFT in the medical domain has prompted researchers to explore reinforcement learning (RL) as an alternative training paradigm. Conventional reinforcement learning from human feedback (RLHF) [15, 16] relies heavily on preference data, which is also particularly challenging to obtain in medical contexts due to the need for expert-level domain knowledge. Consequently, recent efforts have shifted toward reinforcement learning with verifiable rewards (RLVR), especially leveraging RFT to adapt general-purpose MLLMs with reasoning capabilities to medical tasks, as shown in Figure 1. This approach enables the construction of large medical reasoning models at a relatively low cost compared to training from scratch [17, 18]. Nevertheless, due to the limited availability of reasoning-aware supervision in medicine, most existing RLVR methods [19, 14] primarily focus on rewarding the output format and answer correctness, while overlooking supervision of the reasoning process, which leads to sub-optimal performance and undermines model interpretability.

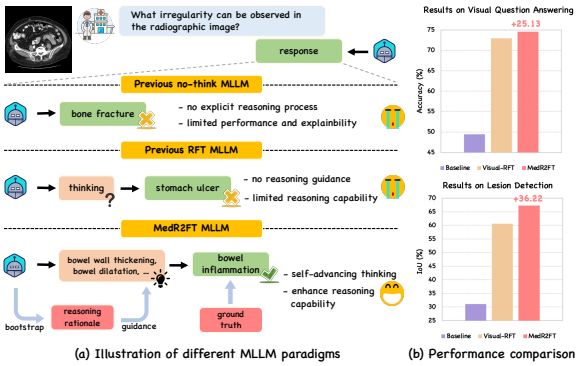

Figure 1: (a) Previous methods prioritize answer correctness while overlooking the reasoning process, while MedR2FT enhances the model capability with explicit reasoning bootstrapping and guidance. (b) MedR2FT significantly improves the performance of the base model on different tasks.

To address the above issues, in this paper, we propose MedR2FT, a reinforcement fine-tuning method designed to enhance medical reasoning through rationale bootstrapping and reasoning rewarding. Specifically, our framework mitigates the challenge of medical reasoning data scarcity by bootstrapping rationales from the training dynamics of MLLMs. The model is first trained with verifiable rewards, while the output reasoning processes are filtered and extracted to form a rationale-augmented dataset. This distilled dataset is then used to further fine-tune the model, reinforcing its reasoning capabilities. To robustly measure and guide the reasoning process, we introduce a semantic-aware reward that integrates logical coherence and medical relevance using an expert reward model. We demonstrate that this self-advancing training approach improves both the reasoning quality and overall performance of MLLMs across diverse medical datasets and multimodal tasks.

The main contributions of our work are as follows: (i) We propose **MedR2FT**, a reinforcement fine-tuning framework for medical image reasoning. It leverages rational bootstrapping and iteratively optimizes the model reasoning, alleviating the scarcity of reasoning data and enabling reasoning supervision to facilitate model domain adaptation in the medical field. (ii) We introduce and employ various rewards, including a semantic-aware reasoning reward, to guide the model optimization, which evaluates both logical coherence and medical relevance of the reasoning process. This enables more precise and reliable supervision beyond conventional answer correctness. (iii) Extensive quantitative and qualitative experiments on multiple medical datasets, modalities, and tasks demonstrate the effectiveness of our framework, showing significant improvements in model reasoning and adaptation, highlighting the importance of reasoning-aware RFT method for model development.

2 Related Work

2.1 General and Medical MLLMs

Recent years have witnessed remarkable progress in MLLMs, which refer to the models that use powerful large language models (LLMs) as a brain with the ability to receive, reason, and output with

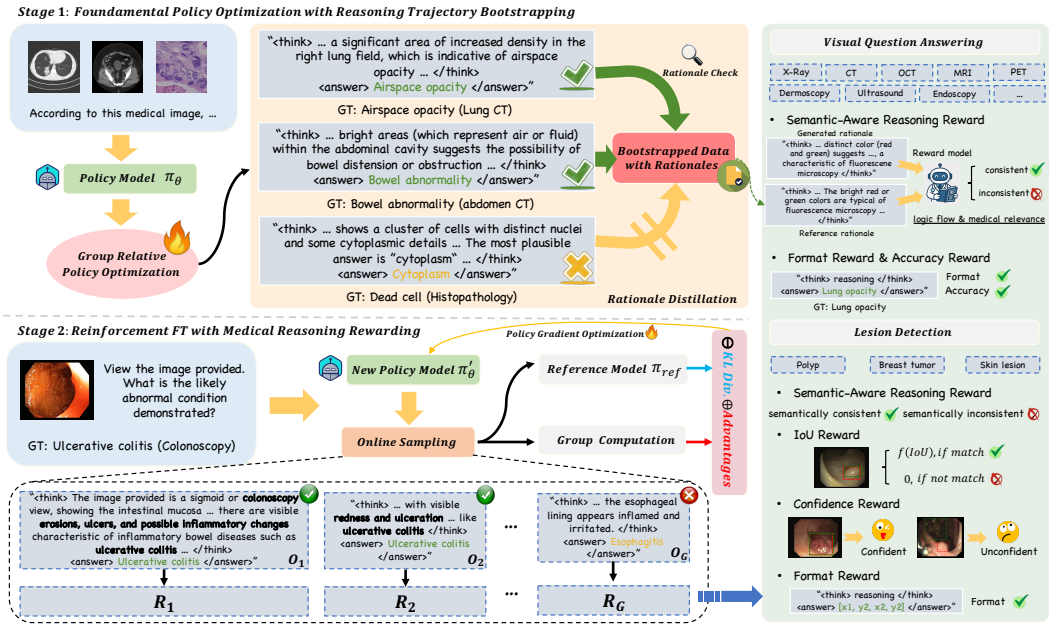

Figure 2: The pipeline of the MedR2FT framework. The main idea of the framework is to bootstrap the rationale data to guide the reasoning process of MLLMs through reinforcement fine-tuning, achieving effective performance improvement and task adaptation.

multimodal (usually images, texts, and videos) information [20]. Models like GPT-4 [4], DeepSeek [21] have been deployed in real-world applications as commercial models. InternVL [22, 6] adopts the widely-used "ViT-MLP-LLM" paradigm, which combines a pre-trained ViT [23] with LLMs [24] through an MLP projector. Qwen-VL [5], as one of the most representative open-source MLLMs, consists of a ViT-based vision encoder and an LLM decoder. Following the success of MLLMs in natural domains, their development has also begun to gain momentum in healthcare. For example, LLaVA-Med [25] improves on LLaVA [2] by fine-tuning on a combination of medical images and dialogue. PMC-LLaMA [26] is based on LLaMA [1] with medical instruction tuning. MedDr [27] incorporates diagnosis-guided data generation with InternVL [28] to achieve a medical generalist model. Lingshu [29] adopts medical alignment and instruction tuning based on the Qwen-VL [5]. However, existing MLLMs fall short in offering reliable medical reasoning processes, leading to sub-optimal performance and reasoning capability. Therefore, we leverage model adaptation to alleviate this issue through reasoning-aware visual reinforcement fine-tuning.

2.2 MLLM Post-Training

While pretraining on large-scale datasets has laid the foundation for large AI models, the research community is increasingly investigating post-training techniques to alleviate existing issues, *e.g.*, limited reasoning capacity and domain-specific performance, as well as to achieve further breakthroughs [30, 31]. Fine-tuning is a representative post-training strategy that adapts pre-trained models to specific tasks by leveraging specific labeled datasets. Recent research reveals that the traditional SFT paradigm relies on memorizing correct answers, therefore needs a large amount of data to avoid over-fitting [19], which exhibits limited potential in data-scarce domains like healthcare. To address this issue, recent researchers have introduced RFT to the vision and medical domains [19, 17, 18, 14, 32, 33], where they improve the performance by enabling the model to explore multiple reasoning paths. For example, MedVLM-R1 [17] introduces group relative policy optimization (GRPO) [21] RL methods to radiology images. MedGround-R1 [34] fine-tunes Qwen-VL [5] through verifiable rewards for the chest X-ray image grounding task. Wang et al. [35] propose a hierarchical reward learning framework for radiology report generation. However, existing visual RFT methods in the medical field prioritize answer correctness while overlooking the reasoning process due to the lack of medical reasoning data. In this work, we alleviate this issue by bootstrapping rationales and introducing a semantic-aware training strategy for reasoning supervision and enhancement.

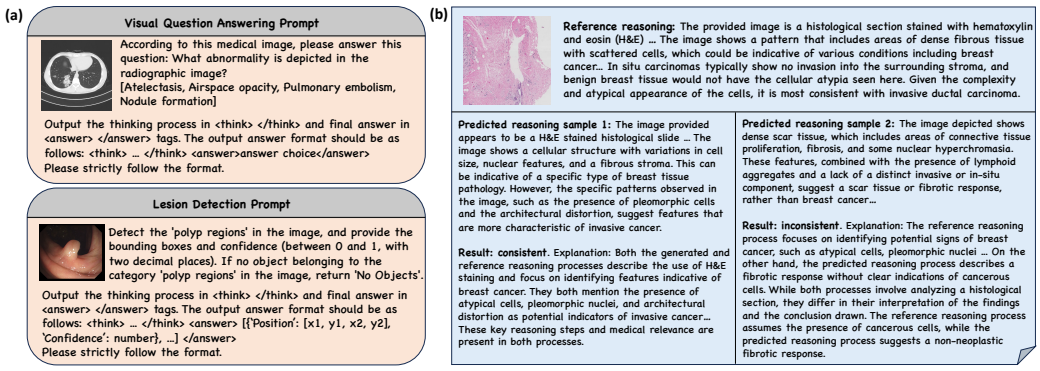

Figure 3: (a) The prompt examples used for MLLM training and inference, including visual question answering and lesion detection. (b) Example results of semantic consistency measurement.

3 Method

In this section, we present the details of our proposed MedR2FT. The overall framework is shown in Figure 2, which mainly comprises two stages. The first stage is the foundational policy optimization of MLLM and reasoning trajectory bootstrapping (Sec. 3.1), where we train the model using verifiable rewards and distill the thinking processes with rationale checking to form a reasoning-augmented database. The second stage is visual reinforcement fine-tuning with medical reasoning reward (Sec. 3.2), where we further utilize the bootstrapped database from the first stage to guide the reasoning optimization of the MLLM via semantic-aware rewards for tasks of visual question answering and lesion detection across various medical imaging modalities.

3.1 Fundamental Policy Optimization based Reasoning Trajectory Bootstrapping

Fundamental Policy Optimization. In the first stage, MedR2FT employs group relative policy optimization [36, 21, 19] with verifiable rewards to train a fundamental medical policy model that incorporates domain-specific priors. Unlike previous RL methods such as PPO [37] and DPO [38], GRPO directly compares and calculates the advantages of groups of candidate responses. Specifically, given the input medical image-question pair q from the training set distribution $P(Q)$, the old policy model $\pi_{\theta_{\text{old}}}$ samples a group of responses $\{o_1, o_2, ..., o_G\}$, where G is the number of generated responses. The rewards, denoted by $\{r_1, r_2, ..., r_G\}$, will be computed for each response based on pre-defined evaluation functions or reward models (the reward design will be introduced in detail in Sec. 3.2). The advantage A_i (*i.e.*, the relative quality) of each generated response is calculated by normalizing the reward using the mean and standard deviation of the response group:

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

Then the policy optimization encourages the model to favor better answers with high advantages within the sampling group by maximizing the following objective:

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

where ϵ and β are hyperparameters, and clip$(\cdot)$ is the clipping function. $\mathbb{D}_{KL}(\pi_\theta||\pi_{\text{ref}})$ is the KL divergence that measures the distribution difference between the current updated policy model π_θ and the reference policy model π_{ref}:

$$\mathbb{D}_{KL}(\pi_\theta||\pi_{\text{ref}}) = \frac{\pi_{\text{ref}}(o_i|q)}{\pi_\theta(o_i|q)} - \log\frac{\pi_{\text{ref}}(o_i|q)}{\pi_\theta(o_i|q)} - 1. \tag{3}$$

The KL divergence serves to constrain the updated policy from deviating excessively from the reference policy, thereby mitigating the risk of catastrophic forgetting or model output collapse.

Table 1: Quantitative comparison on visual question answering with the state-of-the-art MLLMs (accuracy (%)). Our method is highlighted in a grey background. The best results are shown in **bold**. †Note that Lingshu is fine-tuned on the training set of PMC-VQA and PathVQA.

Method	OMedVQA	GMAIMM	PMCVQA	PathVQA	Avg.
General MLLM					
Qwen2.5-VL-3B [5]	52.22	34.76	37.77	57.85	45.65
Qwen2.5-VL-7B [5]	57.97	38.09	40.02	61.63	49.43
Qwen2.5-VL-32B [5]	64.62	44.94	45.45	65.70	55.18
InternVL3-8B [6]	66.57	47.82	45.09	67.91	56.85
Medical MLLM					
HuatuoGPT-V-7B [16]	59.18	39.28	40.43	68.86	51.94
MedDr-40B [27]	55.94	41.45	42.05	71.36	52.70
Lingshu-7B†[29]	83.00	44.12	55.37	84.68	66.79
RFT MLLM					
Baseline (Qwen2.5-VL-7B)	57.97	38.09	40.02	61.63	49.43
+ Visual-RFT [19]	95.67	51.01	60.43	83.85	72.74
+ MedR2FT (Ours)	**97.03**	**53.55**	**61.59**	**86.08**	**74.56**
Δ	**↑1.36**	**↑2.54**	**↑1.16**	**↑2.23**	**↑1.82**

Reasoning Trajectory Bootstrapping. In order to alleviate the issue of medical reasoning data scarcity and supervise the reasoning process of the MLLM, we bootstrap the rationales from the fundamentally tuned model. Specifically, given the trained policy model π_θ in the first stage, we prompt the model to think and answer the questions on the training set through the prompts presented in Figure 3(a). We filter and extract the high-quality rationales from the responses, together with the corresponding input image-question pair and predicted answer, to form the reasoning-enhanced datasets. Specifically, given the training set Q of image-question pairs, the new dataset is

$$\mathcal{D}_{\text{reasoning}} = \{(q, o) \mid q \in Q,\ o \sim \pi_\theta(O|q),\ f_1(\text{ans}(o)) = 1 \wedge f_2(o) = 1\}, \tag{4}$$

where $\text{ans}(o)$ means extracting the answer from the model response o according to the <answer> tags. $f_1(\cdot)$ and $f_2(\cdot)$ are discriminator functions that determine whether a response satisfies pre-defined criteria. Specifically, we first ensure the output answer correctness of the response, *i.e.*, $f_1(x) = 1$ if the predicted answer x matches the ground truth in the VQA task, or if the Intersection-over-Union (IoU) between the predicted and ground truth bounding boxes exceeds 0.5 in the lesion detection task. Otherwise, $f_1(x) = 0$, and the corresponding failed sample is filtered out. Moreover, we conduct further checks on the responses filtered from the first discriminator function to ensure the reasoning consistency (logical coherence between the reasoning process and the final answer). On the one hand, if the reasoning process in <think> tags states the conclusion, we guarantee that the conclusion should be matched with the output answer indicated in <answer> tags, and we check whether the reasoning logic matches the final answer according to the clinical factuality based on retrieval augmented generation. On the other hand, if there is no conclusion mentioned in the reasoning process, then we directly check the clinical fact based on the retrieved knowledge. $f_2(x) = 1$ only when the above conditions are satisfied; otherwise, responses with failed rationales are filtered out (more details are in Appendix B.3). These answer and reasoning verifications provide a double-check rationale filtering mechanism, aiming to create accurate reasoning-augmented data.

3.2 Visual Reinforcement Fine-Tuning with Reasoning Rewarding

To further refine the model and enhance its reasoning capability, we leverage the reasoning-augmented data distilled from the first-stage training to reinforce fine-tuning the model in the second stage, where we design and adopt various rewards for different tasks.

Semantic-Aware Reasoning Reward. To enhance the reasoning capability of the policy model, thereby improving both the reasoning capability and answer correctness of the policy model, we design a semantic-aware reasoning reward to fully leverage the bootstrapped reasoning-augmented data distilled from the first stage. Specifically, during the second-stage fine-tuning, we prompt a medical expert foundation model to measure the semantic consistency between the current predicted thinking rationale t_{pred} and the reference rationale t_{ref} in the training data $D_{\text{reasoning}}$ for tasks of visual question answering and lesion detection. The semantic-aware reasoning reward r_s is computed as:

$$r_s = J(t_{\text{pred}}, t_{\text{ref}}, s_k) \wedge J(t_{\text{pred}}, t_{\text{ref}}, s_m), \tag{5}$$

Table 2: Quantitative comparison on lesion detection with the state-of-the-art MLLMs (IoU (%)). Our method is highlighted in a grey background. The best results are shown in **bold**. Note that we do not include InternVL and MedDr for lesion detection evaluation here due to their different image resize strategies for a fair comparison.

Method	Kvasir	ClinicDB	ColonDB	EndoScene	BUSI	ISBI2017	Avg.
General MLLM							
Qwen2.5-VL-7B [5]	43.88	27.29	25.37	29.55	16.32	44.14	31.09
Qwen2.5-VL-32B [5]	61.38	45.51	39.62	57.08	42.77	50.63	49.50
Medical MLLM							
HuatuoGPT-V-7B [16]	20.50	5.82	0.00	4.89	5.72	10.58	7.92
Lingshu-7B [29]	10.30	1.57	2.85	1.46	11.61	13.82	6.94
RFT MLLM							
Baseline (Qwen2.5-VL-7B)	43.88	27.29	25.37	29.55	16.32	44.14	31.09
+ Visual-RFT [19]	71.56	54.14	55.16	67.63	40.22	74.76	60.58
+ MedR2FT (Ours)	**73.80**	**64.80**	**62.12**	**70.62**	**56.21**	**76.30**	**67.31**
Δ	**↑2.24**	**↑10.66**	**↑6.96**	**↑2.99**	**↑15.99**	**↑1.54**	**↑6.73**

where s denotes the conditions that are used to measure the semantic and logical consistency. Specifically, s_k denotes the key reasoning steps, while s_m includes *medical logic flow* and *medical relevance* (*e.g.*, causal links and diagnostic reasoning, as shown in Figure 3(b)). $J(\cdot)$ refers to the consistency score given by the expert model judgment, where the reward is set to 1 if the two rationales satisfy the conditions (semantically consistent), otherwise the reward is 0. More details about the reward are in Appendix B.2.

Format Reward & Confidence Reward. As described in Figure 3(a), we prompt the model to explicitly output its thinking process in the "<think>... </think>" tags and the final answer in "<answer>...</answer>" tags. The given format reward is 1 if the response format matches the pre-defined format. For the lesion detection task, we adopt an additional confidence reward to measure the model's certainty in its detection results. Specifically, for a predicted bounding box $bbox_{pred}$, if the IoU between $bbox_{pred}$ and the ground truth bounding box $bbox_{gt}$ is greater than zero, then we regard it as a successful match and reward the model with the value of the non-zero IoU. To summarize, the detection confidence reward r_c can be obtained by:

$$r_c = \begin{cases} c, & \text{if IoU}(bbox_{pred}, bbox_{gt}) > 0, \\ 1 - c, & \text{if IoU}(bbox_{pred}, bbox_{gt}) = 0, \end{cases} \tag{6}$$

where c is the detection confidence output by the model. This implies that for successfully matched boxes, higher confidence scores are preferred. Conversely, a lower confidence reward is given when the IoU is zero, which indicates a failed match.

Accuracy Reward & IoU Reward. To evaluate the answer correctness during the training process of the MLLM, accuracy reward and IoU reward are computed for visual question answering and lesion detection tasks, respectively. Specifically, the accuracy reward is set to 1 if the predicted answer enclosed within the "<answer>...</answer>" tags matches the ground truth, otherwise the reward is 0. The IoU reward corresponds to the IoU between $bbox_{pred}$ and $bbox_{gt}$ when a successful match occurs, and is set to 0 if no lesion is detected.

Aggregate Rewards. The rewards for VQA tasks include format reward, accuracy reward, and semantic-aware reasoning reward (stage 2), *i.e.,* $r_{\text{VQA}} = r_{\text{format}} + r_{\text{acc}} + \alpha r_s$, where α is a reward-balancing hyperparameter. The rewards for lesion detection tasks contain format reward, confidence reward, IoU reward, and semantic-aware reasoning reward, *i.e.,* $r_{\text{DET}} = r_{\text{format}} + r_c + r_{\text{IoU}} + \alpha r_s$.

Table 3: Ablation study of MedR2FT. FPO means the first-stage fundamental policy optimization, RB denotes rationale bootstrapping, and SAR reward refers to our semantic-aware reasoning reward used in the second-stage.

Method	PathVQA	EndoScene	ClinicDB
Baseline	61.63	29.55	27.29
+ FPO (Visual-RFT)	83.85	67.63	54.14
+ RB w/ Rouge-L reward	85.54	68.33	63.13
+ RB w/ SAR reward	**86.08**	**70.62**	**64.80**

4 Experiments

4.1 Experimental Setup

Datasets. Our model is fine-tuned and evaluated separately on each dataset within a comprehensive benchmark of 10 datasets spanning diverse domains and tasks, including multiple image modalities,

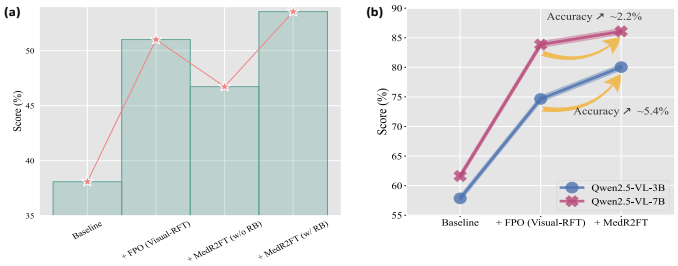

Figure 4: (a) Ablation study of the rationale bootstrapping (RB). (b) Ablation study of different base model sizes.

Table 4: Ablation study of using different reward models (IoU performance). Lingshu-7B is adopted in the proposed MedR2FT.

Reward model	Performance
HuatuoGPT-7B [16]	63.25%
InternVL-8B [22]	64.09%
Lingshu-7B [29]	**64.80%**

e.g., X-Ray, CT, OCT, dermoscopy, ultrasound, endoscopy, pathology, *etc.* for the VQA task [39–42], and polyp, breast tumor, skin lesion for the detection task [43–48]. More details about the datasets and data pre-processing are in Appendix A.

Baselines. We compare our method with representative general-purpose MLLMs, including Qwen2.5-VL [5] and InternVL3 [6], and medical MLLMs including HuatuoGPT-Vision [16], MedDr [27], and Lingshu [29]. The state-of-the-art visual reinforcement fine-tuning method [19] is also included as the compared method. More compared MLLMs and RFT baselines can be found in Appendix C.1.

Implementation Details. Following previous works, we adopt the representative Qwen2.5-VL-7B [5] as the base model for RFT. The GRPO algorithm is implemented based on Virft [19] and Open-R1[1]. The max prompt length is set to 1024, gradient accumulation steps is set to 2, and the number of generations G is set to 4. We employ Lingshu-7B [29] as the semantic reward model. Through grid search, α is set to 0.5 and 1 for VQA and lesion detection, respectively. Experiments are conducted on 10 RTX A6000 GPUs (48G memory). More implementation details are in Appendix B.

4.2 Results on Visual Question Answering

In Table 1, we show the visual question answering comparison results of our method on four medical datasets, including three multi-modality datasets OmniMedVQA [39], GMAIMMBench [40], PMCVQA [41], and a pathology dataset PathVQA [42]. The results of Qwen baseline [5] are shown in grey, while both Visual-RFT [19] and our MedR2FT are fine-tuned based on the Qwen baseline. It can be observed that our method consistently outperforms other baseline models of comparable or even larger scale (*e.g.*, Qwen2.5-VL-32B [5] and MedDr-40B [27]) with a significant margin, including general MLLMs, medical MLLMs, and the RFT baseline. Specifically, our MedR2FT significantly improves the performance of the base model, gaining an average accuracy improvement of 25.1% over the original base model. In addition, our method achieves a 2.5% relative performance boost over the second-best method Visual-RFT [19] on average, which demonstrates that MedR2FT can effectively improve the model performance by enhancing the reasoning capability of MLLM. More comparison and experimental results are in Appendix C.

4.3 Results on Lesion Detection

In Table 2, we report the comparison results on six lesion detection datasets, including polyp detection for endoscopy images (Kvasir-SEG [43], CVC-ClinicDB [44], CVC-ColonDB [45], EndoScene [46]), tumor detection for breast ultrasound images (BUSI [47]), and skin lesion detection for dermatological images (ISBI2017 [48]). Our method consistently outperforms the baseline methods. Specifically, MedR2FT achieves an average improvement of 36.2% compared to the base model. Moreover, MedR2FT respectively gains 2.24%, 10.66%, 6.96%, 2.99%, 15.99%, 1.54% IoU performance boost over Visual-RFT [19] on each lesion detection dataset. It is noteworthy that existing medical MLLMs such as HuatuoGPT [16], often fall short in lesion detection tasks, while our method can significantly improve the performance of MLLM on lesion detection across various image modalities.

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

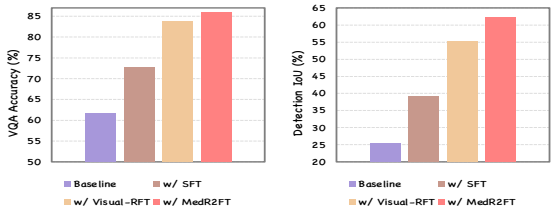

Figure 5: Results of comparison with SFT.

Table 5: Reasoning quality evaluation. The full mark of the evaluation is 10.

Method	Scorer1	Scorer2	Overall
Baseline	5.453	5.049	5.251
+ FPO (Vis-RFT)	7.051	6.675	6.863
+ MedR2FT	**7.363**	**6.863**	**7.113**
Δ (baseline)	**↑1.910**	**↑1.814**	**↑1.862**

4.4 Ablation Study & Additional Analysis

In this section, we adopt a *research-oriented question answering* approach to analyze and validate the effectiveness of our method, including ablation studies, reasoning capability, and other analyses.

Do both stages of MedR2FT individually contribute to improving MLLM performance? We conduct ablation experiments, as shown in Table 3. It can be observed that the first-stage fundamental policy optimization can establish a solid medical foundation for the original baseline MLLM. Furthermore, the second-stage reasoning reinforcement fine-tuning significantly improves model performance with the semantic-aware reasoning reward based on bootstrapped reasoning-augmented data from the first stage. In addition, we further replace the reasoning reward with the Rouge-L reward (in gray) to measure the difference between the predicted and the reference reasoning rationale, where we can also see performance improvement compared to the baseline. However, the MedR2FT model with our semantic-aware reasoning reward achieves better performance. For example, the proposed reasoning reward gains a 2.29% improvement over the method with the Rouge-L reward on the EndoScene dataset. More details about the ablation study and other rewards are in Appendix C.2.

Is rationale bootstrapping (RB) essential? Are the filtered rationales correct/effective, or how does performance change when rationale bootstrapping and checking are removed and all unfiltered stage-1 data are used for second-stage training? In Figure 4(a), we show the ablation results of RB on GMAIMMBench. Specifically, the performance degrades when the RB is removed (*i.e.,* not filter the high-quality rationales but use all the reasoning samples from the first stage for two-stage training instead). This is consistent with the fact that using wrong rationales to guide the reasoning optimization is harmful for model reasoning, leading to limited performance (around 7% performance degradation compared to the full method). The results demonstrate the reliability and effectiveness of the distilled rationales and that RB is essential for model reasoning enhancement in our framework. More details and discussions about the rationale correctness are in Appendix B.3.

Do both stages of our method consistently enhance performance across different base model sizes? Figure 4(b) shows the results of using different base model sizes on PathVQA. It can be observed that larger base model achieves higher performance, and our method can consistently improve model performance under different model sizes with a significant margin.

How to choose the reward model for reasoning guidance? Does the choice of reward model significantly impact performance? We conduct experiments using extra different LLM-based reward models, including a medical and a general expert model (HuatuoGPT-7B [16] and InternVL2.5-8B [22]), as shown in Table 4. Although Lingshu achieves relatively higher performance, the differences are not substantial. Moreover, we find that using state-of-the-art closed-source models (*e.g.*, GPT-5.4 [49]) as the reward model also yields comparable results. This is mainly because evaluating logical coherence and relevance between two texts is a relatively simple task for current LLMs, indicating that reward judgments are robust and reliable. Considering the trade-off between performance and cost, we select the open-source model Lingshu as the final reward model. More discussions about the reward model judgment reliability are in the Appendix C.5.

How does MedR2FT compare with traditional SFT? We conduct comparative experiments against SFT, as shown in Figure 5. Specifically, we observe that the base model trained with MedR2FT achieves around 13% improvements over SFT on the PathVQA dataset, while for the ColonDB polyp detection task, MedR2FT yields more than 20% improvements compared to SFT. More details and results of SFT can be found in the Appendix B.4 and C.3.

What improvements does MedR2FT introduce over Visual-RFT, and what problems is it designed to address? MedR2FT introduces a two-stage reasoning-aware RFT framework that explicitly strengthens model reasoning through rationale bootstrapping and reasoning supervision.

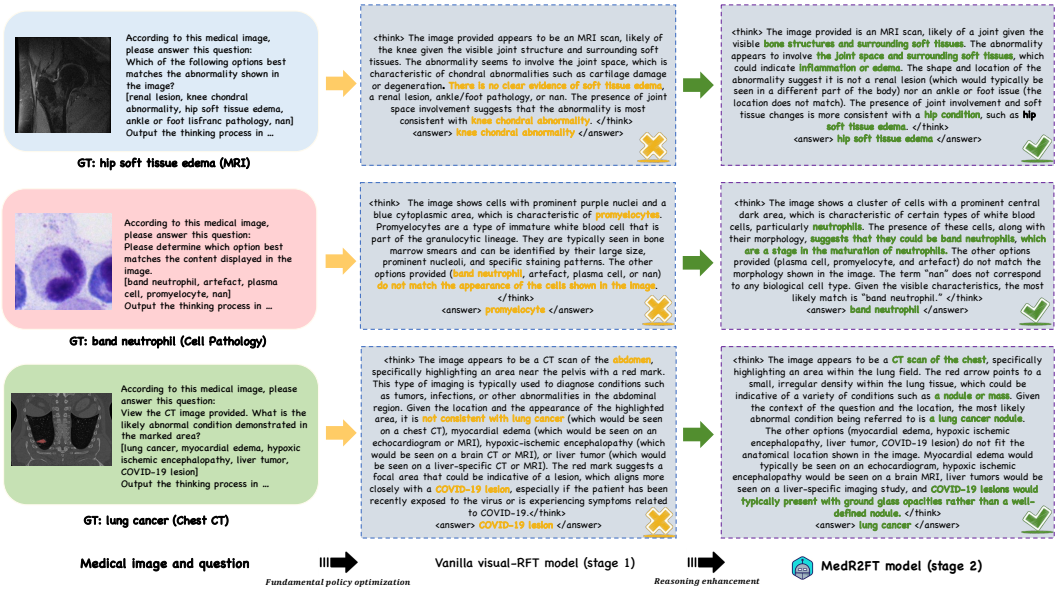

Figure 6: Qualitative results of the reasoning process on visual question answering. We show the reasoning process of the first-stage model (which is a vanilla visual-RFT model) and the second-stage MedR2FT model. The important texts are highlighted in yellow and green for convenient comparison, where green indicates more accurate and complete reasoning or findings.

The goal is to address the issue that existing RFT methods focus on answer correctness rather than the reasoning process, while also mitigating the challenge of scarce medical reasoning data.

Does MedR2FT indeed enhance the model reasoning capability? (i) Visualization & qualitative analysis. To demonstrate that visual reinforcement fine-tuning with reasoning rewarding (Stage 2) can effectively enhance the reasoning capability of the MLLM, we conduct qualitative analyses on both visual question answering and lesion detection tasks. For the VQA task, as shown in Figure 6, the top case asks the model to identify the most closely matched abnormality in a magnetic resonance image (MRI), where the ground-truth answer is "hip soft tissue edema". The model trained with fundamental Visual-RFT (*i.e.*, the Stage-1 model) states that "there is no clear evidence of soft tissue edema" and reasons that "the abnormality is most consistent with knee chondral abnormality". In contrast, MedR2FT after Stage-2 training correctly identifies that the abnormality "involve the joint space and surrounding soft tissues" and "indicate inflammation or edema", concluding that it is more consistent with "hip soft tissue edema". These results demonstrate that our method can effectively enhance the reasoning capability of the base model through rationale bootstrapping and reasoning guidance. More qualitative VQA results are provided in Appendix C.4. For the lesion detection task, Figure 7 shows that MedR2FT achieves more fine-grained visual recognition and reasoning. In the top case, the Visual-RFT model predicts a region substantially larger than the ground-truth polyp, resulting in an IoU of 0.26, whereas MedR2FT localizes the lesion more precisely and achieves an IoU of 0.90. The reasoning process reveals that MedR2FT identifies the polyp's "raised, rounded appearance and coloration that contrasts with the surrounding tissue", while the Visual-RFT model fails to distinguish the polyp from the surrounding tissue. In the bottom case, Visual-RFT treats only the "red spot" as the skin lesion, whereas MedR2FT correctly reasons that the bounding box should cover the *"red spot and the surrounding reddish"* region, producing a localization closer to the ground truth. These results demonstrate more comprehensive reasoning from both holistic and fine-grained perspectives. (ii) Additional LLM evaluations. To further measure the rationale quality, we adopt several LLM experts to evaluate the reasoning process. In addition to the answer correctness, the experts give their scores based on the following perspectives: (1) Relevance: whether the content of reasoning process is indeed the steps required to answer the question. (2) Factuality: whether the content of the reasoning is consistent with medical facts or clinical guidelines. (3) Logical coherence: whether the reasoning process can logically lead to the final conclusion. Specifically, after checking subsets of the evaluation results by users with medical background, two state-of-the-art expert scorers are adopted. The first selected expert scorer is MedGemma-27B [50], and the second expert scorer

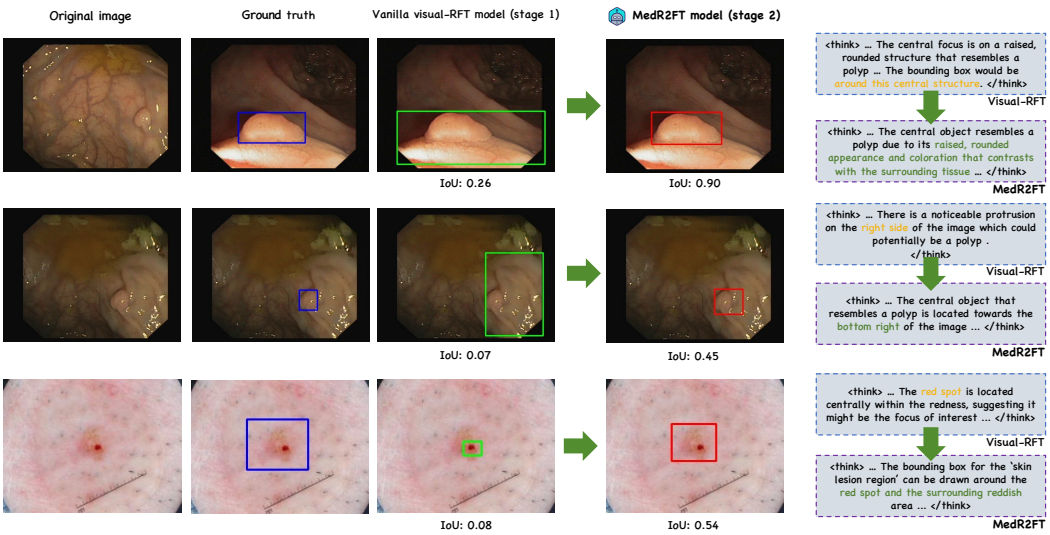

Figure 7: Qualitative results of the reasoning process on lesion detection. The important texts are highlighted in yellow and green for convenient comparison, where green indicates more accurate and complete reasoning or localization.

is Lingshu-32B [29]. The reasoning quality evaluation results of PathVQA are shown in Table 5, where the full mark for the evaluation is 10. It can be observed that MedR2FT has more accurate and coherent reasoning process than the baselines, with an overall score of 7.1 on PathVQA. This demonstrates that our method can effectively enhance the model reasoning capability. More details, results, prompt templates are in Appendix C.4.

5 Conclusion

In this paper, we propose MedR2FT, a visual reinforcement fine-tuning framework for medical reasoning. To alleviate the challenges of data scarcity in medical reasoning and the current practice of existing medical MLLMs prioritizing answer correctness while overlooking the reasoning process, MedR2FT leverages the semantic-aware optimization strategy to guide the reasoning trajectory of MLLMs based on bootstrapped rationale data distilled during model training. Extensive quantitative experimental results, ablation studies, and qualitative analyses across multiple medical image datasets, modalities, and tasks demonstrate that MedR2FT substantially enhances the reasoning capability of the base MLLM, thereby achieving promising performance. We hope that this reinforcement fine-tuning approach holds the potential to benefit doctors and patients in healthcare with a more reliable and precise decision-making process and diagnostic performance.

Acknowledgments and Disclosure of Funding

This work was supported by Innovation and Technology Commission (Project No. PRP/086/24FX and ITC-SKLNSD26SC01) and the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. T45-401/22-N and JLFS/M-604/24).

References

[1] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

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

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

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

[7] Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. Natural language understanding and inference with mllm in visual question answering: A survey. *ACM Computing Surveys*, 57(8):1–36, 2025.

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

[9] Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*, 2025.

[10] Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2087–2098, 2025.

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

[12] Tim Hulsen. Explainable artificial intelligence (xai): concepts and challenges in healthcare. *AI*, 4(3):652–666, 2023.

[13] Junlin Hou, Sicen Liu, Yequan Bie, Hongmei Wang, Andong Tan, Luyang Luo, and Hao Chen. Self-explainable ai for medical image analysis: A survey and new outlooks. *arXiv preprint arXiv:2410.02331*, 2024.

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

[15] Songhua Yang, Hanjie Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 19368–19376, 2024.

[16] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, et al. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*, 2024.

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

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

[19] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2034–2044, 2025.

[20] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

[21] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

[22] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[23] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[24] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*, 2024.

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

[26] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine. *Journal of the American Medical Informatics Association*, 31(9):1833–1843, 2024.

[27] Sunan He, Yuxiang Nie, Hongmei Wang, Shu Yang, Yihui Wang, Zhiyuan Cai, Zhixuan Chen, Yingxue Xu, Luyang Luo, Huiling Xiang, et al. Gsco: Towards generalizable ai in medicine via generalist-specialist collaboration. *arXiv preprint arXiv:2404.15127*, 2024.

[28] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

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

[30] Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip HS Torr, Fahad Shahbaz Khan, and Salman Khan. Llm post-training: A deep dive into reasoning large language models. *arXiv preprint arXiv:2502.21321*, 2025.

[31] Guiyao Tie, Zeli Zhao, Dingjie Song, Fuyang Wei, Rong Zhou, Yurou Dai, Wen Yin, Zhejian Yang, Jiangyue Yan, Yao Su, et al. A survey on post-training of large language models. *arXiv e-prints*, pages arXiv–2503, 2025.

[32] Weihai Zhi, Jiayan Guo, and Shangyang Li. MedGR^2: Breaking the data barrier for medical reasoning via generative reward learning. *arXiv preprint arXiv:2508.20549*, 2025.

[33] Wei Dai, Peilin Chen, Chanakya Ekbote, and Paul Pu Liang. Qoq-med: Building multimodal clinical foundation models with domain-aware grpo training. *arXiv preprint arXiv:2506.00711*, 2025.

[34] Huihui Xu, Yuanpeng Nie, Hualiang Wang, Ying Chen, Wei Li, Junzhi Ning, Lihao Liu, Hongqiu Wang, Lei Zhu, Jiyao Liu, et al. Medground-r1: Advancing medical image grounding via spatial-semantic rewarded group relative policy optimization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 391–401. Springer, 2025.

[35] Yuan Wang, Shujian Gao, Jiaxiang Liu, Songtao Jiang, Haoxiang Xia, Xiaotian Zhang, Zhaolu Kang, Yemin Wang, and Zuozhu Liu. Beyond n-grams: A hierarchical reward learning framework for clinically-aware medical report generation. *arXiv preprint arXiv:2512.02710*, 2025.

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

[37] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

[39] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 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, 2024.

[40] Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37:94327–94427, 2024.

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

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

[43] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*, pages 451–462. Springer, 2019.

[44] Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015.

[45] Jorge Bernal, Javier Sánchez, and Fernando Vilarino. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012.

[46] David Vázquez, Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Antonio M López, Adriana Romero, Michal Drozdzal, and Aaron Courville. A benchmark for endoluminal scene segmentation of colonoscopy images. *Journal of healthcare engineering*, 2017(1):4037190, 2017.

[47] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.

[48] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[49] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[50] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

[51] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*, 2024.

[52] Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, et al. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *arXiv preprint arXiv:2502.09838*, 2025.

[53] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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

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

A Dataset Details

In this section, we provide details on each dataset and our data preprocessing. There are in total of ten datasets used for training and evaluation, including the tasks of visual question answering and lesion detection. Note that each image in the processed detection dataset is assumed to contain a single lesion region with corresponding ground truth (the images without lesion regions are filtered out), and we directly use IoU as the evaluation metric.

OmniMedVQA. OmniMedVQA [39] is a comprehensive medical visual question answering dataset, collected from 73 different medical datasets, including 12 different modalities and covering more than 20 distinct anatomical regions. The dataset contains 118,010 images and 127,995 QA items. In this paper, we adopt the open-source subset of OmniMedVQA with 79,585 VQA samples. The dataset is randomly split into 70% for training and 30% for testing.

GMAIMMBench. GMAIMMBench [40] is constructed from 284 datasets across 38 medical image modalities, 18 clinical-related tasks, 18 departments, and 4 perceptual granularities in a VQA format. There are 25,831 QA pairs in total, and the dataset is randomly split into 70% for training and 30% for testing.

PMCVQA. PMCVQA [41] contains 227k VQA pairs of 149k images that cover various modalities or diseases. We adopt the updated version (version 2) with 186,033 VQA samples. The official data split is used, with 33,430 test samples.

PathVQA. PathVQA [42] is a dataset of question-answer pairs on pathology images. We adopt the official split and use the closed-end part of the dataset, and the numbers of the training set and test set are 9,751 and 3,362, respectively.

Kvasir-SEG. Kvasir-SEG [43] contains 1000 polyp images and their corresponding bounding box ground truth. The dataset is randomly split into 70% for training and 30% for testing.

CVC-ClinicDB. CVC-ClinicDB [44] is a database of frames extracted from colonoscopy videos. This dataset contains 612 polyp images and their corresponding lesion masks. We create the bounding box annotation based on the provided masks for lesion detection. The dataset is randomly split into 70% for training and 30% for testing.

CVC-ColonDB. CVC-ColonDB [45] is a database of annotated video sequences of colonoscopy videos. It contains 15 short colonoscopy sequences, coming from 15 different studies. This dataset provides an annotation of the region of interest (mask) for 380 frames selected from all the sequences. We create the bounding box annotation based on the provided masks for lesion detection. The dataset is randomly split into 70% for training and 30% for testing.

EndoScene. EndoScene [46] is a polyp dataset containing 912 images and corresponding masks. We create the bounding box annotation based on the provided masks for lesion detection, and we adopt the official data split.

BUSI. BUSI [47] is a breast ultrasound image dataset with 780 images and the corresponding mask annotations. We create the bounding box annotation based on the provided masks for lesion detection. We adopt the "benign" and "malignant" breast tumor images, and the dataset is randomly split into 70% for training and 30% for testing.

ISBI2017. ISBI2017 [48] is a skin image dataset containing 2,750 images with masks for skin lesions. We create the bounding box annotation based on the provided masks for lesion detection. We adopt the official data split where the number of test images is 600.

Algorithm 1: MedR2FT

The implementation of our used GRPO algorithm
Given π_θ: the policy model, D: the task prompts, R: the set of reward functions.
Our GRPO algorithm definition: GRPO(π_θ, D, R):
while $iteration \leq num_iteration$ **do**
  The reference model $\pi_{\text{ref}} \leftarrow \pi_\theta$
  while $step \leq num_step$ **do**
    Sample a batch of data Q_b from Q
    Update the old policy model $\pi_{\theta_{\text{old}}} \leftarrow \pi_\theta$
    Sample G outputs $\{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(O|q)$ for each image-question pair $q \in Q_b$
    Compute the reward $\{r_i\}_{i=1}^{G}$ for each sampled output response o_i according to the reward function in R:
    Compute format, confidence, accuracy/IoU rewards according to the reward design introduced in Sec. 3.2 in the main paper (for both two stages)
    if *semantic-aware reasoning reward r_s is in R* **then**
      Compute the reward r_s according to equation (5) in the main paper (for stage two)
    Compute advantage A_i for o_i through group relative advantage estimation described in equation (1) in the main paper
    Update π_θ by maximizing the GRPO objective $\mathcal{J}(\theta)$ described in equation (2) in the main paper
    $step \leftarrow step + 1$
  $iteration \leftarrow iteration + 1$

The main process of MedR2FT
Input: The original task prompts (i.e., data samples from the training dataset) Q, initial policy model $\pi_{\theta_{\text{init}}}$, VQA reward function r_{VQA} and detection reward function r_{DET}.
Output: The new policy model π'_θ.
Initialization: policy model $\pi_\theta \leftarrow \pi_{\theta_{\text{init}}}$.
The fundamental policy optimization
$\pi_\theta \leftarrow \text{GRPO}(\pi_\theta, D \leftarrow Q, R \leftarrow \{r_{\text{VQA}} \leftarrow \{r_{\text{format}}, r_{\text{acc}}\}, r_{\text{DET}} \leftarrow \{r_{\text{format}}, r_c, r_{\text{IoU}}\}\})$ # r_c: confidence reward
Reasoning trajectory bootstrapping
Bootstrap the data and create the reasoning-augmented database through equation (4) in the main paper: $\mathcal{D}_{\text{reasoning}} \leftarrow \{(q, o) \mid q \in Q,\ o \sim \pi_\theta(O|q),\ f_1(\text{ans}(o)) = 1 \wedge f_2(o) = 1\}$
Visual reinforcement fine-tuning with reasoning rewarding
$\pi'_\theta \leftarrow \text{GRPO}(\pi_\theta, D \leftarrow \mathcal{D}_{\text{reasoning}}, R \leftarrow \{r_{\text{VQA}} \leftarrow \{r_{\text{format}}, r_{\text{acc}}, r_s\}, r_{\text{DET}} \leftarrow \{r_{\text{format}}, r_c, r_{\text{IoU}}, r_s\}\})$

B Implementation Details

B.1 MedR2FT Algorithm

In addition to the method illustration in the main paper, the algorithmic format of the proposed MedR2FT framework along with our used GRPO [36] algorithm details is shown in Algorithm 1.

B.2 Details about the Reasoning Reward

To measure the semantic consistency between the generated and the reference rationale, we propose a semantic-aware reasoning reward using the reward model based on Lingshu-7B [29], as described in Sec. 3.2 in the main paper. The prompt for the reward model is shown in Figure 8, which measures the differences between the predicted rationales and the bootstrapped rationales from several aspects: key reasoning steps, medical logical flow & medical relevance.

B.3 Details about the Rationale Checking and Filtering

In Sec. 3.1 of the main paper, we state that we adopt a double-check rationale filtering mechanism during the rationale bootstrapping process. The first step is to ensure the output answer correctness of the rationale, and the second step is to ensure the reasoning consistency and clinical factuality.

Prompt for Resoning Reward Measurement

You are a medical reasoning evaluator. Your task is to evaluate the semantic relevance between the generated reasoning process and the reference reasoning process for a medical visual question-answering task.
The generated reasoning process: {generated_reasoning_process}
The reference reasoning process: {reference_reasoning_process}
Use the following criteria to assess the relevance of the reasoning: Scoring Criteria:
- 1 (Semantically Consistent): The generated reasoning process is logically and semantically consistent with the reference reasoning process. This means it includes the same key reasoning steps, causal relationships, and medical logic, even if the wording or phrasing differs.
- 0 (Semantically Inconsistent): The generated reasoning process significantly deviates from the reference reasoning process. This includes the omission of critical steps, contradictions in logic, or irrelevant reasoning that does not align with the reference.
Guidelines for Evaluation:
1. Analyze the **key reasoning steps** in both the generated reasoning and the reference reasoning.
2. Focus on the **logical flow** and **medical relevance** (e.g., causal links, diagnostic reasoning, or procedural steps).
Provide a brief explanation of your evaluation and include the final score.
The output format is as follows:
Explanation: [Explanation]
Final Score: [0 or 1]

Figure 8: The prompt used for semantic-aware reasoning reward measurement.

The prompt we used for the second-step verification (Lingshu) is shown in Figure 9. To demonstrate the correctness of the filtered rationales, we recruit two users with medical background to verify the rationales. They are asked to "*check whether the provided rationales match the given answer, and assess whether the rationales are reasonable for the specified medical tasks*". Specifically, for a random sample of 50 filtered out rationales (which do not pass our rationale checking process), human raters consider all the samples to be wrong. For the bootstrapped rationales passing the checking process (randomly sampled 100 samples), human raters find that more than 90% of them are reasonable, demonstrating the effectiveness of our rationale bootstrapping and checking mechanism.

Prompt for Rationale Verification and Filtering (2nd Step)

You are tasked with verifying the consistency & clinical factuality of a model's response output. The model response contains two parts: a reasoning trace enclosed in <think>...</think> and a final answer enclosed in <answer>...</answer>.
Your goal:
1. Read the reasoning in <think></think>.
2. Read the answer in <answer></answer>.
3. Identify the conclusion or final result stated in the reasoning and compare it with the answer in <answer></answer>.
4. Determine whether they match.
5. Determine whether the reasoning matches the answer based on your inherent knowledge and the retrieved knowledge.

If they are consistent, output: "Consistent".
If they are inconsistent (e.g., reasoning concludes Answer1 but <answer></answer> contains Answer2, or the reasoning does not match the clinical facts), output: "Inconsistent".
If the reasoning does not clearly state a conclusion, then directly use your inherent knowledge and the retrieved knowledge to check whether the reasoning matches the output answer (also output "Consistent" or "Inconsistent").

Here are the model response output and the retrieved knowledge: xxx

Figure 9: The prompt used for rationale checking.

Table 6: More quantitative comparison experiments on visual question answering with the state-of-the-art medical and RFT MLLMs (accuracy (%)). Our method is highlighted in a grey background. The best results are shown in **bold**.

Method	Ref.	OMedVQA	GMAIMM	PMCVQA	PathVQA	Average
Medical MLLM						
HealthGPT-M3 [52]	ICML'25	53.04	33.31	27.97	59.93	43.56
RFT MLLM						
MedVLM-R1 [17]	MICCAI'25	73.11	42.87	53.04	75.94	61.24
MedR2FT	-	**97.03**	**53.55**	**61.59**	**86.08**	**74.56**

B.4 Details about Supervised Fine-Tuning

In Sec. 4.4 of the main paper, we compare our method with the traditional supervised fine-tuning (SFT). Specifically, the SFT method is implemented based on LLaMA-Factory [51]. We fully fine-tune the Qwen2.5-VL-7B [5] on the target datasets and tasks with a learning rate of 0.0001. We adopt the same input prompt as in our RFT method, except that the MLLM is instructed to directly generate the final answer without first generating the reasoning process.

C More Experimental Results

C.1 More Results of the Medical and RFT MLLMs

To further demonstrate the effectiveness of our method, we conduct more comparison experiments with medical and RFT MLLMs, as shown in Table 6. Specifically, HealthGPT [52] is a state-of-the-art medical MLLM that performs visual instruction fine-tuning on multimodal medical data. MedVLM-R1 [17] is a RFT model based on Qwen [53] that performs GRPO [36] on sampled data from VQA-RAD [54], SLAKE [55], PathVQA [42], OmniMedVQA [39], and PMCVQA [41]. It can be observed that our MedR2FT consistently outperforms these two methods with a significant margin.

In addition, we report the comparison results with more RFT methods in Figure 10. Specifically, we include MedR1 [14] and MedVLM-R1 [17] for comparison on PathVQA. It can be observed that our MedR2FT achieves the best performance, with more than 10% performance improvement compared to other baselines.

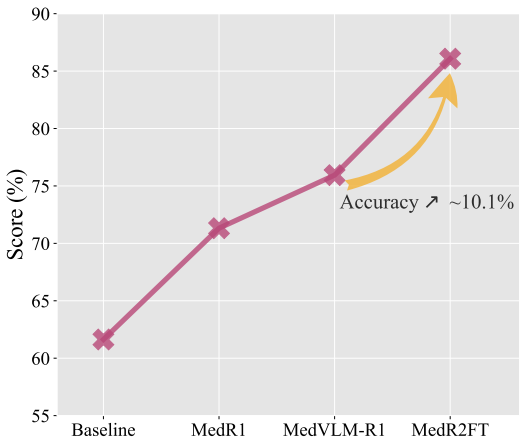

Figure 10: Comparison results with other RFT methods (accuracy).

C.2 More Results of the Ablation Study

Reasoning rewards. We conduct more ablation studies to demonstrate the effectiveness of the proposed method. Specifically, as shown in Figure 11, we show the results of our MedR2FT with different reasoning rewards on CVC-ClinicDB. It can be observed that the reasoning rewards

can improve the model performance compared to the baseline model and the Visual-RFT method. Furthermore, our method with the semantic-aware reasoning (SAR) reward outperforms other rewards such as Rouge-L and BLEU score, validating the effectiveness of the reasoning supervision signal.

Reward-balancing hyperparameter. In Table 7, we report the ablation results of the reward-balancing hyperparameter α (*i.e.,* weight of the reasoning reward) on GMAIMMBench. Specifically, we find that for VQA tasks, $\alpha = 0.5$ is a relatively promising choice (also for other datasets) during the reasoning rewarding. The adopted hyperparameter values have been stated in implementation detail (Sec. 4.1) in the main paper.

Table 7: Ablation of reward-balancing hyperparameter α (accuracy (%)).

Hyperparameter α	$\alpha = 0.3$	$\alpha = 0.5$	$\alpha = 0.7$	$\alpha = 1.0$
Performance	53.09	53.55	53.29	52.80

Extended Stage-1 training. To examine whether the improvements of MedR2FT can be attributed to a larger number of optimization steps, we continue training the converged Stage-1 checkpoint using the Stage-1 Visual-RFT objective. The number of additional optimization steps is set equal to that used in Stage 2, ensuring that the extended Stage-1 baseline and the full two-stage pipeline have the same total training budget. As shown in Table 8, extending Stage-1 training yields nearly unchanged performance on the VQA datasets and degrades performance on several lesion detection datasets. In contrast, the full MedR2FT model consistently improves upon the original Stage-1 checkpoint across all datasets. These results indicate that the gains of MedR2FT arise from the reasoning-aware supervision introduced in Stage 2 rather than from additional training alone. Since the Stage-1 rewards evaluate answer correctness without directly supervising the reasoning process, simply extending this stage cannot effectively improve the reasoning trajectories and may lead to overfitting to task-specific answer signals.

Table 8: Comparison with extended Stage-1 training under matched total training steps. Accuracy (%) is reported for VQA datasets, and IoU (%) is reported for lesion detection datasets.

Method	OMVQA	PMCVQA	PathVQA	Kvasir	BUSI	ISBI
Stage 1 (original)	95.67	60.43	83.85	71.56	40.22	74.76
Stage 1 (extended, matched steps)	95.64	60.46	83.79	71.49	38.25	74.44
MedR2FT (Ours, w/ Stage 2)	**97.03**	**61.59**	**86.08**	**73.80**	**56.21**	**76.30**

C.3 More Results of the Supervised Fine-Tuning

We report additional comparison results for our method versus SFT, as shown in Table 9. Specifically, our method consistently outperforms the traditional SFT training paradigm, demonstrating the effectiveness of our reasoning-aware RFT training strategy.

C.4 More Results of the Reasoning Capability and Quality

We show more results and analysis of the model's reasoning capability. As illustrated in Figure 12, we can observe that in the top case, the question is to identify the most closely matched abnormality shown in the highlight area for an endoscopy, where the ground truth answer is "barrett's esophagus". The model trained with fundamental visual-RFT reasons that "the most likely abnormality is esophageal cancer", while the MedR2FT model correctly concludes that the abnormality would be "barrett's esophagus". The bottom example shows an abdominal MRI, and the question is to identify the abnormality in the highlighted area. Specifically, the fundamental visual-RFT model claims it is a "colon cancer primaries", while the full MedR2FT model correctly reasons that "The highlighted area shows a mass with enhancement after gadolinium administration, which is characteristic of neuroblastoma. Neuroblastoma is a type of cancer that can present with such findings on MRI", which is consistent with the ground truth. These additional results and analyses further demonstrate that MedR2FT can effectively enhance the reasoning capability of the base model through reasoning-aware fine-tuning.

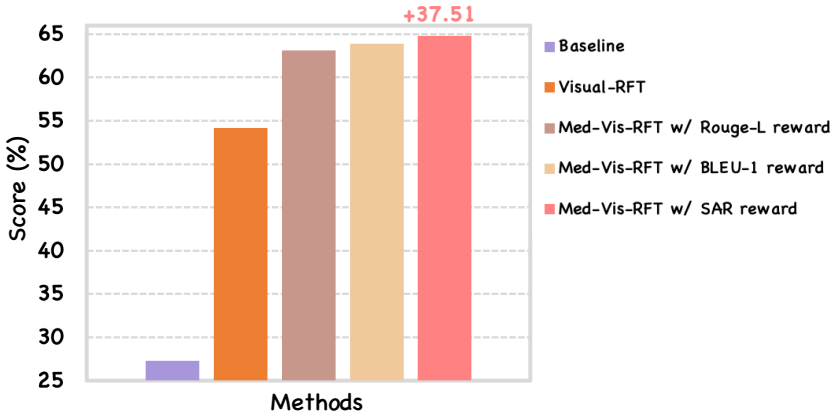

Figure 11: Ablation study on different reasoning rewards.

Table 9: More quantitative comparisons with SFT on five lesion detection datasets (IoU (%)). Our method is highlighted in a grey background. The best results are shown in **bold**.

Method	ClinicDB	ColonDB	EndoScene	BUSI	ISBI2017	Average
Baseline (Qwen2.5-VL-7B)	27.29	25.37	29.55	16.32	44.14	28.53
+ SFT	30.52	39.04	16.24	37.22	60.38	36.68
+ MedR2FT (Ours)	**64.80**	**62.12**	**70.62**	**56.21**	**76.30**	**66.01**

Moreover, in order to demonstrate the reasoning quality and reliability, as mentioned in Sec. 4.4 in the main paper, we adopt two LLM experts (MedGemma-27B [50] and Lingshu-32B [29]) to evaluate the reasoning process. The prompt we used for the reasoning quality evaluation based on LLM experts is shown in Figure 13. In addition, for human study, we recruit two users with medical background to *verify the rationale correctness*. Specifically, in a random sample of 200 cases with correct predicted answers from both VQA and detection data, human raters find that 91.5% of the rationales are clear and clinically correct, which are causally grounded in the image evidence. In future work, we plan to explore how to get more reliable rationale reasoning (*e.g.*, incorporating more strict rationale filtering and human-in-the-loop) and further improve the model reasoning.

C.5 More Details and Results of the Model & Judgment Reliability

In addition to the discussion about different reward models in Sec. 4.4 in the main paper, we also conduct extra analysis to demonstrate the robustness and reliability of the reward judge models. Specifically, we demonstrate the reliability of the reward model (Lingshu) judgment in rationale bootstrapping from the following aspects: (1) **Performance**. We conduct experiments using extra different LLM-based reward models, including a medical and a general expert model, and the performance does not vary significantly (Table 4 of the main paper). This is mainly because judging the logical coherence and relevance between two texts is a relatively simple and straightforward task for current LLMs, where the results indicate the model judgment is robust and reliable, as mentioned in the main paper. (2) **Consistency**. We compute a consistency score to measure the judge consistency between different reward models during the model training (*i.e.*, proportion of consistent judgments). The consistency score for Lingshu and HuatuoGPT is 92.5%, and is 90.1% for Lingshu and InternVL, showing high consistency and robustness. (3) **Human study**. We recruit two users with medical background to verify the accuracy of Lingshu judgment. They are asked to "*check whether the model judgement for the predicted rationales and reference rationales is accurate based on key reasoning steps, medical logical flow & medical relevance*". Specifically, in a random sample of 200 cases from both VQA and detection data, human raters consider 189 of the model's judgments to be correct, resulting in an accuracy of 94.5%, further demonstrating the robustness and the quality of our model judgment.

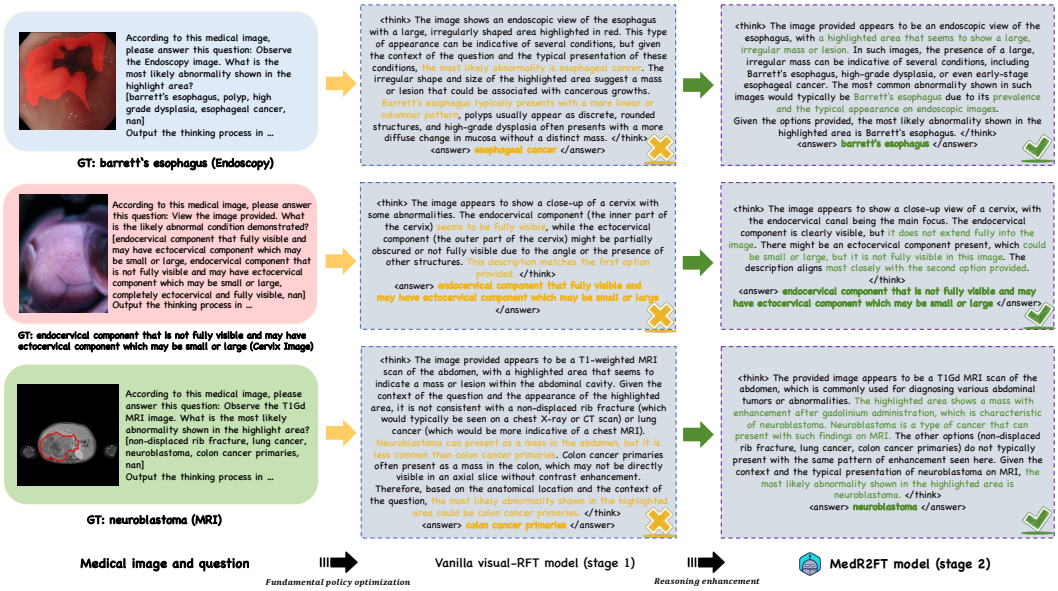

Figure 12: More qualitative results of the reasoning process. We show the reasoning process of the first-stage model (which is a vanilla visual-RFT model) and the second-stage MedR2FT model. The important texts are highlighted in yellow and green for convenient comparison, where green indicates more accurate and complete reasoning or findings.

Prompt for Resoning Quality Evaluation

You are a helpful medical assistant. Your task is to evaluate the quality of the model response to a given medical question. You will be provided with a question, the model's response, the predicted answer extracted from the response, and the groud truth solution. Please evaluate the response based on the following criteria:
1) Answer correctness: You need to check whether the response correctly answer the question. If the predicted answer does not match the ground truth solution, output <score>0</score> directly and ignore the other criteria. Otherwise, the response will be given 4 points for its correctness.
2) **Relevance**: You need to check whether the content of the reasoning process is indeed the steps required to answer the question and give your score (0: not relevant, 1: partially relevant, 2: fully relevant).
3) **Factuality**: You need to check whether the content of the reasoning is consistent with medical facts or clinical guidelines and give your score (0: not factual, 1: partially factual, 2: fully factual).
4) **Logical coherence**: You need to check whether the reasoning process can lead to the final conclusion and give your score (0: not coherent, 1: partially coherent, 2: fully coherent).
You need to calculate the total score (the full mark is 4+3*2=10, and the lowest score is 0, i.e., wrong answer) by summing the individual scores. Only output the final score and your output should be in the format without any additional text: <score>total_score_value</score>.
The question is ... The model response is ... The predicted answer is ... The ground truth solution is ...

Figure 13: The prompt used for reasoning quality evaluation.

C.6 Statistical Significance Analysis

We evaluate statistical significance using paired bootstrap tests between MedR2FT and the second-best method, Visual-RFT. For each considered dataset, we resample the test set with replacement 1,000 times, using identical sampled indices for both methods, and recompute accuracy for VQA or IoU for lesion detection. Table 10 reports the results with the bootstrap standard deviation in parentheses. All improvements are statistically significant ($p < 0.001$).

Table 10: Paired-bootstrap significance test comparing MedR2FT with Visual-RFT. Values in parentheses denote the standard deviation over 1,000 bootstrap replicates. Accuracy (%) is reported for VQA datasets, and IoU (%) is reported for lesion detection datasets.

Method	OMedVQA	GMAIMM	PMCVQA	PathVQA	Kvasir	ClinicDB	ColonDB
Visual-RFT	95.67 (0.13)	51.01 (0.57)	60.43 (0.27)	83.85 (0.61)	71.56 (2.34)	54.14 (2.07)	55.16 (2.91)
MedR2FT (Ours)	**97.03 (0.11)**	**53.55 (0.56)**	**61.59 (0.27)**	**86.08 (0.59)**	**73.80 (2.16)**	**64.80 (2.11)**	**62.12 (2.86)**
Bootstrap p-value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

C.7 Rationale-Level Diversity Analysis

For each test question, we sample four rollouts and compute Self-BLEU, Distinct-1, and Distinct-2 among the generated rationales. Lower Self-BLEU and higher Distinct scores indicate greater diversity. As shown in Table 11, the three metrics remain nearly unchanged on GMAIMMBench, indicating that rationale diversity is preserved, while all three improve on CVC-ClinicDB, indicating more diverse rationales after Stage 2. These results show no reduction in rationale diversity associated with mode collapse.

Table 11: Rationale-level diversity across training stages.

Model	Dataset					
	GMAIMM			ClinicDB		
	Self-BLEU↓	Distinct-1↑	Distinct-2↑	Self-BLEU↓	Distinct-1↑	Distinct-2↑
Stage-1 (Visual-RFT)	0.1247	0.4339	0.8260	0.1339	0.4775	0.8331
MedR2FT (w/ Stage-2)	0.1255	0.4290	0.8303	0.1024	0.5049	0.8630

D Limitation and Future Work

Although we conduct comprehensive experiments to verify the quality of the rationales, more reliable rationales are necessary for real-world medical applications. In the future, we plan to incorporate more strict rationale filtering and human-in-the-loop (or even build a large-scale medical reasoning dataset annotated by doctors) to further enhance the reasoning model development as well as alleviate the challenge of data scarcity. Moreover, our experiments and analysis reveal that failures occur more frequently in fine-grained disease characterization. This finding reflects the persistent bottleneck of fine-grained visual perception in current MLLMs, motivating future work to enhance their capabilities for fine-grained medical reasoning and diagnosis.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: As shown in the abstract and introduction, the main claims accurately reflect the paper's contributions and scopes.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: As shown in Appendix Sec. D, we discuss the limitation and future work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: Not applicable.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide implementation details in the main paper (Sec. 4.1) and Appendix Sec. B. The code will also be made publicly available.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All the data preprocess and experimental code will be made publicly available. All the data used in this work are public datasets, which are clearly cited in the paper (e.g., Appendix Sec. A).

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We provide the training and test details in the main paper (Sec. 4.1) and Appendix Sec. B.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report paired-bootstrap significance tests in Appendix Sec. C.6. Table 10 reports the bootstrap standard deviations and p-values, with all improvements being statistically significant ($p < 0.001$).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the computer resources information in Sec. 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conforms with the Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We mention the societal impacts in Sec. 5 and Appendix Sec. D.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper properly respected all the licence and terms of use and cites each asset clearly.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets. The code will be made publicly available with the proper license.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We include the instructions given to the human users, as shown in Appendix Sec. B.3 and C.5.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No potential risks and not applicable in this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The proposed framework uses LLM for rationale evaluation, and the usage is clearly described.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.