Improving Medical Visual Reinforcement Fine-Tuning via Perception and Reasoning Augmentation

Guangjing Yang[1]*, ZhangYuan Yu[1]*, Ziyuan Qin[2]*, Xinyuan Song[2], Huahui Yi[3]*,
Qingbo Kang[3], Jun Gao[3], Yiyue Li[3], Chenlin Du[4], Qicheng Lao[1]†
[1]Beijing University of Posts and Telecommunications
[2]Emory University [3]Sichuan University [4]Peking University
qicheng.lao@bupt.edu.cn

While recent advances in Reinforcement Fine-Tuning (RFT) have shown that rule-based reward schemes can enable effective post-training for large language models, their extension to cross-modal, vision-centric domains remains largely underexplored. This limitation is especially pronounced in the medical imaging domain, where effective performance requires both robust visual perception and structured reasoning. In this work, we address this gap by proposing *VRFT-Aug*, a visual reinforcement fine-tuning framework tailored for the medical domain. VRFT-Aug introduces a series of training strategies designed to augment both perception and reasoning, including prior knowledge injection, perception-driven policy refinement, medically informed reward shaping, and behavioral imitation. Together, these methods aim to stabilize and improve the RFT process.

Through extensive experiments across multiple medical datasets, we show that our approaches consistently outperform both standard supervised fine-tuning and RFT baselines. Moreover, we provide empirically grounded insights and practical training heuristics that can be generalized to other medical image tasks. We hope this work contributes actionable guidance and fresh inspiration for the ongoing effort to develop reliable, reasoning-capable models for high-stakes medical applications.

1. Introduction

Recently, Reinforcement Learning (RL)-based fine-tuning [1–6] for large language models (LLMs) has shown significant progress in complex reasoning tasks.

The emergence of methods such as DeepSeek-R1 [3] and the GRPO [2] algorithm has demonstrated the feasibility of fine-tuning large models using rule-based rewards [4, 7] instead of learned reward models [8–10], substantially lowering the barrier to applying RL in large-scale model training and introducing a promising new paradigm. While RL-based fine-tuning has been actively explored in LLMs, its application to large vision-language models (LVLMs) [11–13]—referred to as Visual Reinforcement Fine-Tuning (V-RFT) [14–17]—remains largely underexplored.

Despite its promise, the effectiveness of V-RFT remains constrained by fundamental challenges in visual perception and reasoning. First, a pretrained LVLMs may lack the capacity to capture subtle visual cues or localize key regions without explicit supervision [18]. This leads to unreliable or sparse rewards during early-stage exploration, hindering stable policy updates [19–21]. Second, many vision-language tasks require multi-step reasoning [22] or structured decision-making, which cannot be effectively learned through scalar reward signals alone. Without explicit reasoning supervision or prior knowledge, V-RFT models are prone to shortcut learning or shallow pattern memorization [23], rather than developing genuine reasoning ability. These limitations highlight a pressing need to enhance V-RFT with augmented perception and reasoning mechanisms, enabling more robust learning in visually and cognitively demanding tasks.

*These authors contributed equally to this work.
†Corresponding author

Third Conference on Parsimony and Learning (CPAL 2026).

The gap is even more pronounced in the context of medical imaging domain, where it is still unclear how to effectively perform RL post-training on pretrained LVLMs to improve their clinical utility and generalization.

Before delving into the technical details, we highlight a key distinction between medical image recognition and general-domain vision tasks—an insight that forms the cornerstone of our work. **Specifically, we find that successful medical image understanding hinges on the fusion of perception and reasoning, rather than relying on either in isolation.** The former emphasizes how information is received and interpreted, while the latter focuses on how information is organized, abstracted, and logically manipulated. Perceptual tasks are characterized by their reliance on accurate interpretation of sensory input—once the content of an image is clearly perceived, further analysis may require little to no reasoning. This is exemplified by many Visual Question Answering (VQA) benchmarks [24–26] in the general domain, where models are asked about attributes, positions, or colors of objects. As long as the model can correctly parse the visual elements, it can answer such questions without needing to perform complex inference. In contrast, reasoning tasks [27, 28] demand an additional layer of logical composition. They require the model to synthesize multiple pieces of information to arrive at a coherent, logically grounded conclusion. Unlike natural images, medical images are not readily interpretable by untrained individuals. Recognizing subtle patterns such as tumors on a CT scan—and further judging their malignancy—often requires both perceptual decoding of the visual content and the integration of domain-specific knowledge [29]. The task thus involves both visual pattern recognition (perception) and medical reasoning based on those patterns.

This naturally gives rise to a central question: *Can reinforcement learning—originally envisioned as a tool to enhance reasoning capabilities—effectively address tasks that require a hybrid of perception and reasoning, such as medical image understanding?* In this work, we take a step toward answering this question by proposing VRFT-Aug, a visual reinforcement fine-tuning framework tailored for the medical domain. VRFT-Aug introduces a series of improvements aimed at two core challenges:

1. Augment LVLM expertise perception capability by dual-channel knowledge injection.
2. Augment LVLM medical reasoning skill by reward shaping.

To achieve these goals, we systematically investigate how RL techniques—specifically GRPO, used as our baseline V-RFT method—can be adapted and extended to better support perception and reasoning in visually and cognitively demanding medical tasks.

Perception Augmentation via Knowledge Injection. Because medical image recognition requires extensive domain-specific prior knowledge [30–34], we first propose a pipeline that enhances pretrained models by integrating such knowledge through both explicit and implicit mechanisms. Specifically, we utilize prompt engineering to incorporate medical knowledge, improving the model's ability to recognize and distinguish domain-specific entities.

Inspired by [33–35], we introduce the visual attributes–such as color, shape, and location–to the prompts of a medical concept. And the prompt will incentivize the LVLM to recognize objects that share identical visual attributes. Then we propose an implicit method for knowledge injection by exploring cross-task training, leveraging diverse medical vision tasks to encourage transferability and robust generalization. This enables the model to acquire both local (e.g., lesion boundary) and global (e.g., anatomical structure) understanding, crucial for handling the multi-scale nature of clinical reasoning.

Reasoning Augmentation via Reward Shaping. Prior studies suggest that the hallucinated content in the reasoning process on the language side leads to incorrect output for tasks like VQA and image captioning [36–38]. This observation suggests that the reasoning process may influence the perceived content during the text decoding process. **Firstly, drawing inspiration from human cognitive mechanisms, we explore whether enforcing repeated recitation of expressive descriptions of medical concepts, as specified in the prompts, could help mitigate hallucinations and guide the model toward more accurate conclusions.** Interestingly, our empirical observations reveal

a nuanced outcome. While such repetition during the model's internal reasoning (analogous to a human's internal monologue) can indeed accelerate convergence to a sub-optimal plateau, it often fails to achieve optimal performance in the long run. This suggests that, despite some shared linguistic structures, large models do not always benefit from human-inspired heuristics in the same way, and over-reinforcing certain patterns may limit the model's flexibility and generalization. **Secondly, we design a specialized reward function based on multi-grade fuzzy scheme tailored for ordinal classification tasks commonly found in the medical domain, aiming to help the model distinguish subtle inter-class differences and mitigate the sparse reward problem during early-stage exploration.** By providing more nuanced feedback, the designed reward promotes stable learning and supports the development of accurate reasoning patterns in fine-grained classification tasks.

2. Related Works

Large Vision Language Models Large Vision-Language Models (LVLMs) are an evolution of traditional Vision-Language Models (VLMs) [39–41], integrating powerful LLMs [42–45] with advanced visual perception backbones. This fusion enhances multimodal understanding and complex reasoning across text and visual data, making LVLMs a key step toward Artificial General Intelligence (AGI). LVLMs are categorized into two main types: commercial closed-source models accessible via APIs (e.g., GPT-4o [46], Gemini [47, 48]) and open-source models available for local deployment (e.g., LLaVA [49], InterVL [13], Qwen VL [11, 12]). The rapid growth of open-source communities has accelerated progress in medical LVLMs, with notable examples like LLaVA-Med [50], developed from LLaVA, and MedRegA [18], built on InterVL with continued medical pre-training. Our experiments are based on the advanced Qwen 2.5 VL model.

Reinforcement Learning OpenAI's o1 [1] pioneered using reinforcement learning (RL) to enhance model reasoning, introducing the test-time scaling law. DeepSeek R1 [3] extended this with GRPO [2] and rule-based rewards, becoming the first open-source model to replicate o1's complex reasoning, sparking interest in LLM reasoning research [51, 52]. In the LVLM domain, R1-V achieved superior performance with GRPO, while VisualThinker-R1-Zero [53] showed that applying R1 to base VLMs led to "visual aha moments". MM-Eureka [54] observed similar effects using RLOO [55], and Vision-R1 [56] introduced a multimodal CoT dataset for enhanced training. Curr-ReFT [57] proposed a three-stage RL framework. Visual-RFT [14] uniquely focused on RL for visual perception, while VLM-R1 [15] validated R1-style RL across diverse visual tasks. MedVLM-R1 [58], Med-RLVR [59], and Med-R1 [60] extended RL to the medical domain. Building on these advances, we optimized RL for medical vision with enhanced perception and reward mechanisms.

3. Methods

3.1. Preliminary

Visual Reinforcement Fine-Tuning (V-RFT) fine-tunes pretrained LVLMs using reinforcement learning techniques such as PPO [61] or GRPO [2, 3], enhancing their decision-making capabilities through task-specific, rule-based reward functions (e.g., classification accuracy or IoU). For a downstream task dataset D consisting of N samples, each sample is defined by an input prompt P and its corresponding image I_i, where i represents the index of the current sample. The policy model π_θ generates a response O_i, which is then evaluated using a rule-based reward function R with respect to the task ground truth G_i. Formally, V-RFT aims to optimize the following objective:

$$\begin{aligned}&\max_{\pi_\theta} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{O\sim\pi_\theta(P,I)} R_{\text{V-RFT}}(P, I_i, G_i)\\ =&\max_{\pi_\theta} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{O\sim\pi_\theta(P,I)} \left[R(\pi_\theta(O_i \mid P, I_i), G_i) - \beta\,\mathrm{KL}\left[\pi_\theta(O_i \mid P, I_i) \| \pi_{\text{ref}}(O_i \mid P, I_i)\right]\right],\end{aligned} \tag{1}$$

where π_{ref} is the reference model before optimization, and β is a hyperparameter controlling the impact of KL-divergence. The rule-based reward function R is defined as:

$$R(\pi_\theta(O_i \mid P, I_i), G_i) = \begin{cases} 1.0 & \text{if } O_i == G_i \text{ or } \mathrm{IoU}(O_i, G_i) \geq threshold, \\ 0.0 & \text{otherwise.} \end{cases} \tag{2}$$

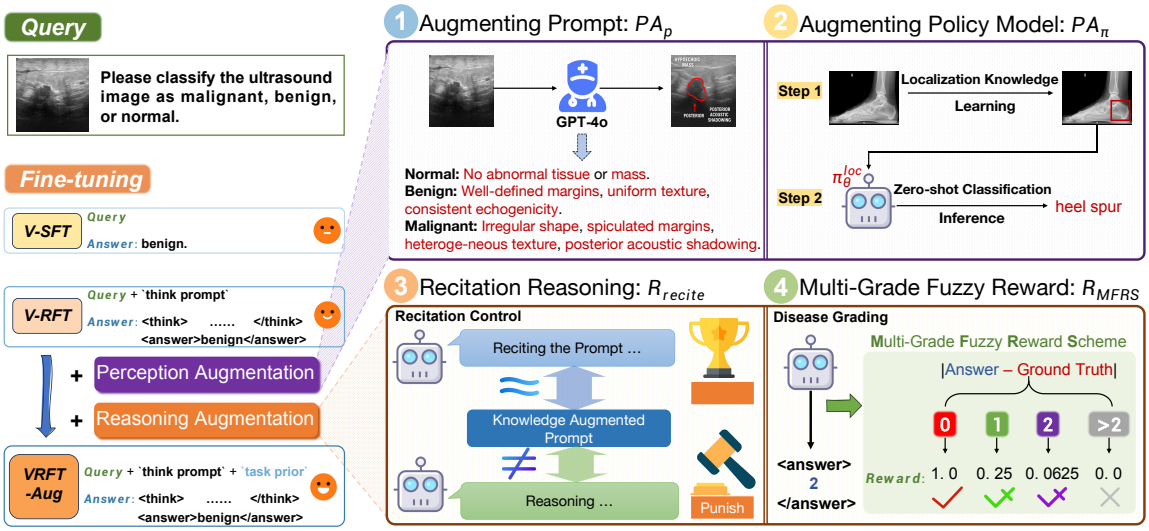

Figure 1: **Overview of VRFT-Aug**. VRFT-Aug incorporates enhancements from both Perception and Reasoning perspectives, introducing four improvement strategies for medical vision tasks: Augmenting Prompt (PA_p), Augmenting Policy Model (PA_π), Recitation Reasoning (R_{recite}), and Multi-Grade Fuzzy Reward (R_{MFRS}).

where IoU is the intersection over union metric, and the $threshold$ is typically set to 0.5 as default.

To improve V-RFT with perception and reasoning capabilities in the medical domain, we propose optimizing Eq. (1) by augmenting its three key components: the prompt P, the policy model π_θ, and the reward function R. For **perception augmentation**, we apply contextual augmentation through a structured prompt $\hat{P}$ (Section 3.2), and implicit knowledge injection by refining the policy $\hat{\pi_\theta}$ (Section 3.3). For **reasoning augmentation**, we adopt reward shaping to guide the learning process: R_{recite} is designed to capture the model's recitation pattern (Section 3.4), while R_{MFRS}, a task-specific reward based on multi-grade fuzzy reward scheme, is proposed to address the sparse reward problem in medical-grade classification and improve learning effectiveness (Section 3.5).

3.2. Augmenting Prompt P with Task-Relevant Context

Pretrained LVLMs often struggle with medical tasks due to the lack of understanding of domain-specific concepts, which are essential for accurate recognition and reasoning. To address this, we first seek to enhance the model's comprehension of medical tasks by expanding the prompt with task-relevant contextual information.

Inspired by prior works in prompt engineering [33, 62, 63], we enrich prompts with visual attributes—such as color, shape, and spatial location—associated with specific medical concepts, therby encouraging the LVLM to focus on relevant objects and strengthening its task-specific perception. To achieve this, we leverage advanced foundation models, such as GPT-4o, to generate relevant visual attributes and create a structured prompt template enriched with task-specific contextual information.

Specifically, for each task, we query GPT-4o with detailed task information—including data source, imaging modality, sample size, and categories—and provide representative images I_c for each category C. We then extract comprehensive visual attribute descriptions that capture key aspects essential for solving the task, which we define as explicit contextual knowledge K_c. To overcome hallucinations, we manually refine the outputs by consulting medical literature and validating them with medical professionals to ensure clinical accuracy. This contextual knowledge is then used to augment the original input prompt, forming an enhanced prompt $\hat{P}$:

$$\hat{P} = [P, \sum_{C} K_c] = [P, \sum_{C} M_{\text{GPT}}(C, I_c)], \tag{3}$$

where C denotes the category of the task to which the image belongs.

Expanding the prompt with task-specific contextual information enhances model performance by providing richer, more relevant cues.This augmented context serves as perceptual guidance, enabling the model to make more accurate predictions. From a theoretical perspective, since the policy $\pi_\theta(a \mid I, p)$ is conditioned on the prompt p, choosing a more informative prompt p_{rich} results in an initial policy that is closer to the optimal policy π^*:

$$KL(\pi^* \parallel \pi(\cdot \mid I, p_{\text{rich}})) < KL(\pi^* \parallel \pi(\cdot \mid I, p_{\text{naive}})) \tag{4}$$

This alignment reduces the exploration burden and improves sample efficiency.

3.3. Augmenting Policy Model π_θ with Task Relevant Knowledge

Beyond enhancing LVLMs with contextual information through prompts (P), we further explore whether the policy model (π_θ) can transfer knowledge from other relevant tasks through RL, thereby enhancing its perception capability accumulated from cross-task learning.

Inspired by the cognitive workflow of radiologists—"localize first, diagnose later" [64, 65]—we employ the RFT framework to train the model to localize specific regions, lesions, or organs in medical images. These localization priors allow the model to focus its attention on anatomically relevant areas, and thus enhance the perception capability by ruling out irrelevant areas.

Concretely, for medical image classification tasks, we first train the model with a reinforcement learning objective to localize potential regions of abnormality using a small number of samples $(M < N)$. During this stage, only a coarse anatomical region is provided as the grounding reference. The model is tasked with predicting a bounding box coordinate $[x_1, y_1, x_2, y_2]$, without receiving any classification-related information. We denote the model that acquired the localization knowledge as π_θ^{loc} and this implicit knowledge injection process is formulated as:

$$\hat{\pi_\theta} = \pi_\theta^{\text{loc}} = \max_{\pi_\theta} \frac{1}{M} \sum_{i=1}^{M} \mathbb{E}_{O \sim \pi_\theta(P,I)} R_{\text{V-RFT}}(P^{\text{loc}}, I_i, G_i). \tag{5}$$

where P^{loc} is the prompt designed for localization (more details in the Appendix A.5). Then we use the $\hat{\pi_\theta}$ as the base model to perform zero-shot inference for predicting classification labels $\hat{y}_i^{\text{cls}}$:

$$\hat{y}_i^{\text{cls}} = \hat{\pi_\theta}(O_i^{\text{cls}} \mid P^{\text{cls}}, I_i). \tag{6}$$

3.4. Augmenting Reward R with Recitation Reasoning

During our experiments on contextual augmentation (Section 3.2), we notice that the model's generated reasoning outputs often appear to recite the medical prior knowledge we implanted in the prompts, a phenomenon we refer to as "Recitation Reasoning". **This observation closely resembles a stereotypical human behavior: when attempting to recognize an unfamiliar concept, humans often reinforce their understanding by mentally or verbally repeating its defining characteristics.** We hypothesize that mimicking this repetitive pattern—by recite medical descriptors throughout the model's internal reasoning steps—can help stabilize attention and output consistency.

To investigate the impact of recitation reasoning, we augment the reward function R with a recitation reward component R_{recite}, enabling us to study this behavior during training by encouraging or discouraging it through reward shaping. Specifically, we hire the Bilingual Evaluation Understudy (BLEU) [66] score—a widely adopted metric in natural language generation—to measure the similarity between the model's reasoning outputs and the prior medical knowledge provided in the prompt $\hat{P}$. A higher BLEU score indicates greater repetition of prior knowledge in the output, resulting in a higher recitation reward R_{recite}. The formulation of R_{recite} is defined as follows:

$$R_{\text{recite}} = \delta \times \text{BLEU}(O_i, \hat{P}), \quad R_{\text{recite}} \in (-1, 1). \tag{7}$$

Following previous work [3, 67, 68], we also include accuracy reward R_{accuracy} and format reward R_{format}. Aggregating these, we obtain the following formula for calculating the overall reward:

$$\hat{R} = \lambda \times R_{\text{accuracy}} + (1 - \lambda) \times R_{\text{format}} + R_{\text{recite}}, \quad \lambda \in (0, 1). \tag{8}$$

where λ is a weighting parameter. We control the influence of the recitation reward by adjusting the sign of δ: a positive δ encourages repetition by rewarding it, while a negative δ penalizes repetition, which we hypothesize enhances reasoning by stabilizing attention and promoting more independent reasoning.

3.5. Augmenting Reward R with Multi-Grade Fuzzy Approach

In clinical diagnosis, lesions often differ subtly between adjacent disease grades, with progression marked by gradual changes in quantity, distribution, or extent rather than abrupt shifts. These subtle visual cues make learning difficult and data-intensive. For instance, mild to moderate retinal lesions may differ only slightly in features like microaneurysm count or hemorrhage extent, making them challenging to distinguish [69]. In early-stage exploration, RL algorithms may suffer from training collapse in particularly challenging tasks where rewards are infrequent—a well-known issue referred to as the sparse reward problem [70, 71]. Similarly, when a model fails to detect subtle visual differences, it may make near-correct predictions in grade classification without receiving any reward, further exacerbating learning difficulty and hindering the development of accurate reasoning patterns.

Inspired by multi-objective reward design in reinforcement learning [72], we introduce a **M**ulti-grade **F**uzzy **R**eward **S**cheme (MFRS) tailored for overcoming the sparse reward problem in medical grading tasks. Specifically, we calculate the difference between the predicted output O^{cls} and the ground truth G^{cls}, where both O^{cls} and G^{cls} are integers labels, and design a "fuzzy" reward mechanism that allows for a relaxed reward even when the predicted value is incorrect. The fuzzy reward weights are selected based on extensive early-stage experiments, as shown in the following formula:

$$R_{\text{MFRS}} = \begin{cases} 1.0 & \text{if } O^{\text{cls}} == G^{\text{cls}}, \\ \frac{1}{4} & \text{if abs}(O^{\text{cls}} - G^{\text{cls}}) = 1, \\ \frac{1}{16} & \text{if abs}(O^{\text{cls}} - G^{\text{cls}}) = 2, \\ 0.0 & \text{otherwise.} \end{cases} \tag{9}$$

Therefore, the overall reward is calculated through a weighted average of the updated accuracy reward R_{MFRS} and the format reward R_{format}. The specific formula is as follows:

$$\hat{R} = \alpha \times R_{\text{MFRS}} + \gamma \times R_{\text{format}}, \tag{10}$$

where α and γ are weighting parameters, set to $\alpha = 0.9$ and $\gamma = 0.1$ in this work. As a reward shaping strategy, MFRS works well for medical grade classification tasks and significantly increases the reasoning performance of the model, compared with the Vanilla RFT methods.

The four components are integrated based on task types and training stages: PA_P (**P**erception **A**ugmentation through **P**rompt) is used in all training pipelines, PA_π is optimized with GRPO for tasks involving object-level alignment, R_{recite} mitigates over-repetition in reasoning tasks, and R_{MFRS} is applied for ordinal classification with soft thresholds.

4. Experiments

4.1. Setup

Datasets. To evaluate the effectiveness of our proposed VRFT-Aug in the medical vision domain, we curate datasets from public sources across three representative task types: 1. **Medical Image Classification**, which involves distinguishing anatomical structures or lesions; 2. **Fine-Grained Regional Classification**, targeting the recognition of lesion subtypes within specific anatomical regions; and 3. **Disease Grading**, which assesses both the presence and progression of the disease. We utilize eight datasets from MedMNIST [73], covering diverse imaging modalities such as X-ray, ultrasound, and CT, to comprehensively evaluate medical image classification tasks. For fine-grained regional classification, we adopt the HAM10000 [74] and Heel [75] datasets. To assess disease progression, we use RetinaMNIST from MedMNIST and the processed COVID-19 dataset [76]. Detailed information of the datasets can be found in the Appendix A.1.

Implementation Details. For both medical image classification and localization tasks, we employ Qwen2.5-VL-3B-Instruct [12] as our base reasoning model. Following previous work [68, 77–79], we

Table 1: Comparison of methods. The best results are highlighted in bold, while the second-best results are underlined. Note that, with the exception of RetinaMNIST adopting MFRS reward, all other datasets utilize accuracy reward.

Shot	Method	Breast	Pneumonia	OCT	Retina*	Derma	Tissue	Blood	OrganA	Average
0-shot	Qwen2.5VL-3B	26.92	52.88	25.39	14.75	19.60	8.40	12.30	9.80	21.25
	Qwen2.5VL-7B	8.33	39.90	33.59	21.00	21.84	10.54	7.42	8.81	18.93
10-shot	V-SFT	46.15	55.12	31.25	18.50	26.89	11.71	33.39	33.23	32.03
	V-RFT	60.89	51.28	29.68	27.00	27.17	12.50	43.35	29.40	35.16
	V-RFT + PA$_P$	**61.53**	**64.42**	**51.17**	**31.75**	**31.92**	**17.18**	**48.04**	**30.39**	**42.05**
	Δ	↑ 0.64	↑ 13.14	↑ 21.49	↑ 4.75	↑ 4.75	↑ 4.68	↑ 4.69	↑ 0.99	↑ 6.89
20-shot	V-SFT	58.33	52.24	33.59	26.75	35.01	13.08	45.31	36.22	37.57
	V-RFT	57.69	68.42	45.31	25.00	39.49	12.50	44.33	30.82	40.45
	V-RFT + PA$_P$	**67.94**	**72.91**	**55.46**	**37.25**	**40.05**	**17.77**	**54.68**	**38.63**	**48.09**
	Δ	↑ 10.25	↑ 4.49	↑ 10.15	↑ 12.25	↑ 0.56	↑ 5.27	↑ 10.35	↑ 7.81	↑ 7.64
256-shot	V-SFT	40.38	71.15	44.14	50.00	45.93	19.33	59.96	37.92	46.10
	V-RFT	**73.07**	81.89	70.31	59.50	45.93	15.82	58.59	52.13	57.16
	V-RFT + PA$_P$	**73.07**	**82.69**	**73.43**	**60.25**	**52.66**	**19.92**	**70.70**	**56.25**	**60.93**
	Δ	↑ 0.00	↑ 0.80	↑ 1.56	↑ 0.75	↑ 6.73	↑ 4.10	↑ 12.11	↑ 4.12	↑ 3.77
	↑ 0.00	↑ 0.80	↑ 1.56	↑ 0.75	↑ 6.73	↑ 4.10	↑ 12.11	↑ 4.12	↑ 3.77	

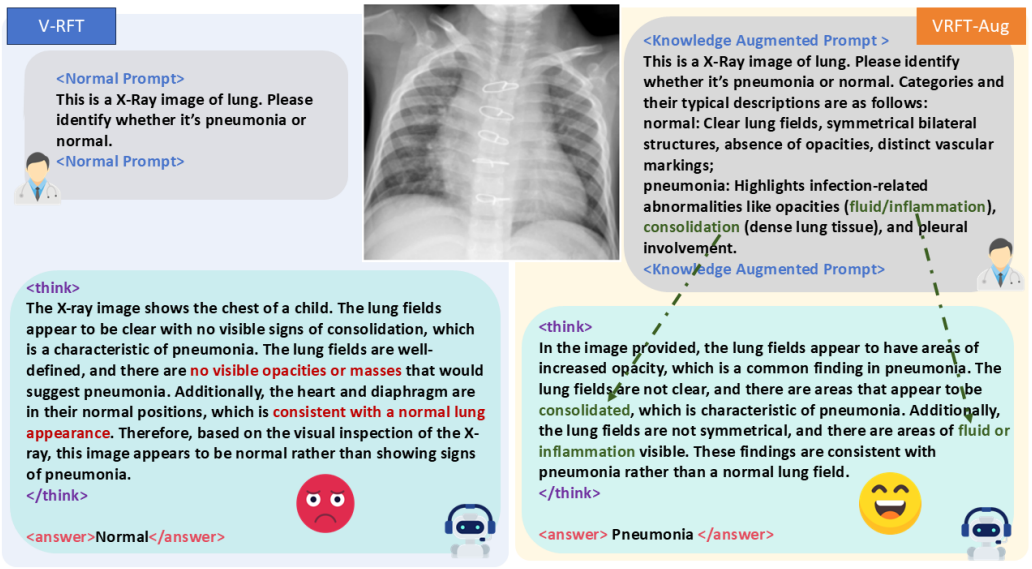

Figure 2: The effectiveness of our proposed perception augmentation on the prompt.

implement the code in Pytorch using 2 NVIDIA A800 80G GPUs. During the RL training, we adopt default GRPO settings, with N set to 8, temperature to 0.9, and KL divergence ratio β to 0.04. For the classification task, the model is fully fine-tuned for 120 steps, using a batch size of 256 and the AdamW optimizer with an initial learning rate of 1e-6 for both SFT and RL. For the localization task, the model is fully fine-tuned for up to 2 epochs, with an initial learning rate of 1e-6 for both SFT and RL. The batch size is set to 1 per device, with 2-step gradient accumulation. Comprehensive details on the experimental settings and evaluation schemes are provided in Appendix A.2.

4.2. Experimental Results

Results on Contextual Augmentation. For contextual augmentation, we compare V-SFT and V-RFT baselines on various few-shot settings with our V-RFT+PA$_P$ approach. As is shown in Table 1, both V-SFT and V-RFT can improve the model's performance under the few-shot settings, while our approach consistently outperforms all baselines and maintains a significant lead. With just 10

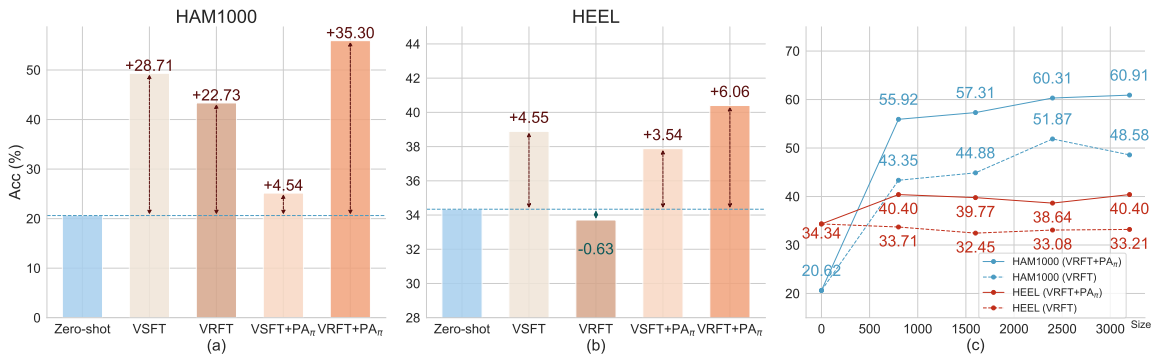

Figure 3: Performance comparison of different methods on the HAM10000 and HEEL. (a) and (b) show that VRFT + PAπ achieves the highest accuracy, with a +35.30% improvement on HAM10000. (c) demonstrates that performance of VRFT + PAπ improves with increasing training samples, reflecting enhanced perception capabilities. VSFT + PAπ and VRFT + PAπ are trained on bounding box prediction tasks (using SFT and GRPO, respectively) and evaluated on classification in a zero-shot manner, while V-SFT and V-RFT are directly trained for classification without localization.

Table 2: Comparison of the best results are highlighted in bold, while the second-best results are underlined. Note that with the exception of RetinaMNIST adopting MFRS reward, all other datasets utilize accuracy reward.

Method	Breast	Pneumonia	OCT	Retina*	Derma	Tissue	Blood	OrganA	Average
Qwen2.5-VL-3B	26.92	52.88	25.39	14.75	19.60	8.40	12.30	9.80	21.25
V-SFT+ PA$_P$	<u>58.33</u>	76.12	55.46	52.75	49.57	17.96	70.50	42.18	52.86
V-RFT + PA$_P$	**73.07**	82.69	<u>71.87</u>	<u>60.25</u>	<u>52.66</u>	**19.92**	<u>70.70</u>	**56.25**	<u>60.93</u>
V-RFT + PA$_P$ + $\delta^+ R_{\text{recite}}$	**73.07**	**83.49**	66.79	49.00	**56.02**	12.50	70.31	51.70	57.86
V-RFT + PA$_P$ + $\delta^- R_{\text{recite}}$	**73.07**	<u>83.01</u>	**75.78**	**63.50**	51.54	<u>17.96</u>	**81.25**	<u>53.40</u>	**62.44**

shots of data, our approach already delivers a boost by **+6.89%** compared with the V-RFT baseline. As the data amount increases, our approach achieves an average performance of **60.93%** in the 256-shot setting, **14.83%**/**3.77%** higher than V-SFT/V-RFT baselines. During the experiment, we have also noticed that contextual augmentation accelerates the training process. The phenomenon indicates that the model is incentivized to focus on feature-distinctive objects, thus enhancing its domain-specific perception and reducing the time required to learn task-relevant patterns.

Results on Implicit Knowledge Injection. We evaluate the classification performance of five methods—zero-shot, V-SFT, V-RFT, V-SFT+PA$_\pi$ and V-RFT+PA$_\pi$ —on the HAM10000 and HEEL test sets. The zero-shot method refers to the Qwen2.5-VL-3B-Instruct model performing disease classification without any fine-tuning. As shown in Fig. 3 (a) and (b), it achieves 20.62% accuracy on HAM10000 and 34.34% on HEEL, indicating limited diagnostic performance and underscoring the need for downstream fine-tuning.

We then apply SFT and vanilla GRPO-based RFT, denoted as V-SFT and V-RFT, respectively. Both methods outperform zero-shot, validating the benefit of fine-tuning. Notably, V-RFT improves accuracy on HAM10000 by **+22.7%**, but slightly underperformed on HEEL (-0.63%). We find that the HEEL dataset suffers from data imbalance, and the less frequent classes have relatively more complex image features. We suspect that under complex or imbalanced data distributions, and in the absence of advanced techniques, the RFT may converge to suboptimal local patterns, overfitting to high-frequency and low-complexity features. In such cases, the simpler SFT may offer greater stability despite lacking reasoning capabilities. Next, we introduce V-SFT+PA$_\pi$ and V-RFT+PA$_\pi$, which incorporate a perception augmentation strategy via task-relevant training to inject implicit spatial knowledge. The model is first trained on localization tasks via SFT or RFT, followed by a zero-shot disease classification on the corresponding test sets. Notably, V-RFT+PA$_\pi$ demonstrates the most significant performance improvement across both datasets, with an impressive increase of

Table 3: Performance variation between MFRS Reward and accuracy reward.

Method	Retina	COVID-19	Average
Qwen2.5-VL-3B	14.75	17.64	16.20
Qwen2.5-VL-7B	21.00	20.26	20.63
V-SFT	50.00	19.60	34.80
V-RFT	59.50	20.26	39.88
V-RFT+PA$_P$+R_{acc}	43.50	24.18	33.84
V-RFT+PA$_P$+R_{MFRS}	**60.25**	**30.06**	**45.16**

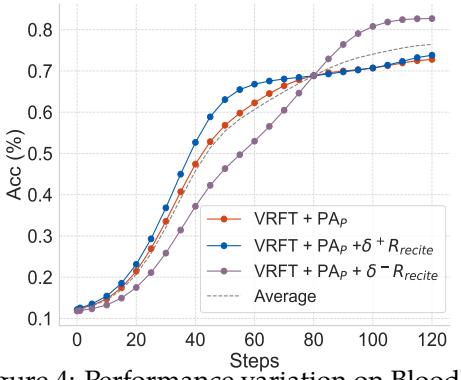

Figure 4: Performance variation on BloodMNIST of different Recitation Reward settings.

+35.30% on the HAM10000 dataset. In contrast, although V-SFT+PA$_\pi$ also shows an improvement, the enhancement is less pronounced. These results indicate that training on the localization task to enhance the model's spatial perception ability is more effective in improving medical image classification performance. Moreover, it highlights that reinforcement learning, integrated during the inference process, further strengthens the model's anatomical localization perception. As observed in Fig. 3 (c), compared to V-RFT, the perception capability of V-RFT+PA$_\pi$ progressively improves as the model encounters more training samples, leading to continuous performance enhancement. In conclusion, we can assert that enhancing the model's anatomical localization perception capability significantly stimulates stronger performance in medical image classification.

Results on Recitation Reward. In this section, we compare our proposed V-RFT+PA$_P$ approach with different Recitation Reward modifications. In addition to quantitative results in Table 2, we also provide a curve graph of performance variation on BloodMNIST in Fig. 4. It can be observed from the figure that although repeated recitation of medical concepts can accelerate convergence to a sub-optimal plateau, it fails to achieve optimal performance in the long term. For other datasets in Table 2, the addition of positive Recitation Reward results in an average performance of 57.86%, 3.07% lower than the original proposed approach. The phenomenon indicates that over-reinforcing certain patterns may limit the model's flexibility and generalization. By contrast, a negative Recitation Reward can reduce the model's dependence on specific patterns. As is shown in Table 2, the average accuracy of negative R_{recite} setting is 62.44%, creating a +1.51% improvement. Compared to the positive R_{recite} setting, although the negative R_{recite} setting causes a slight decline in DermaMNIST, TissueMNIST, and OrganAMNIST, the overall impact is only -0.74%, much smaller than the -3.59% decline observed with the positive R_{recite} setting, highlighting the advantage of the negative setting in terms of model flexibility and generalization.

Results on MFRS Reward. To evaluate the validity of the MFRS Reward, we compare the classification performance of V-RFT+PA$_p$+R$_{\text{MFRS}}$ and V-RFT+PA$_p$+R$_{\text{accuracy}}$ in Table 3. It can be concluded that when replace R_{MFRS} in Eq. (10) with R_{accuracy}, the average performance shows a noticeable decline from 45.16% to 33.84%, which even lags behind V-SFT/V-RFT by 0.96%/6.04%. These experimental results indicate that Vanilla RFT methods tend to suffer from the sparse reward problem [70, 71] due to the slight difference between categories in medical grade classification tasks. By allowing a "fuzzy" reward mechanism, the model can learn partial patterns by making near-correct predictions in the early stage instead of being trapped in invalid strategies.

5. Conclusion

We studied why existing RL-based vision reward fine-tuning (V-RFT) struggles in medical visual recognition and showed that improving large vision–language models requires advances in both **perception** and **reasoning**. Our experiments reveal that GRPO-based V-RFT leaves notable gaps in medical settings. To address this, we proposed **VRFT-Aug**, which injects domain knowledge through prompt manipulation and cross-task training, and introduces medical-specific reward functions

such as the recitation reward and a multi-grade fuzzy reward. As the first RL framework targeting complex medical recognition, VRFT-Aug offers a foundation for future medical reasoning models, and its prompt-based knowledge injection strategy may extend to other visually complex domains.

References

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

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

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

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

[5] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2023.

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

[7] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. T\" ulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

[8] Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*, 2024.

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

[10] Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, et al. Internlm-xcomposer2. 5-reward: A simple yet effective multi-modal reward model. *arXiv preprint arXiv:2501.12368*, 2025.

[11] 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.

[12] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[13] 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.

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

[15] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

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

[17] Ming Li, Jike Zhong, Shitian Zhao, Yuxiang Lai, and Kaipeng Zhang. Think or not think: A study of explicit thinking in rule-based visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.16188*, 2025.

[18] Lehan Wang, Haonan Wang, Honglong Yang, Jiaji Mao, Zehong Yang, Jun Shen, and Xiaomeng Li. Interpretable bilingual multimodal large language model for diverse biomedical tasks. *arXiv preprint arXiv:2410.18387*, 2024.

[19] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay, 2018. URL `https://arxiv.org/abs/1707.01495`.

[20] Coline Devin, Abhishek Gupta, Trevor Darrell, Pieter Abbeel, and Sergey Levine. Learning modular neural network policies for multi-task and multi-robot transfer, 2016. URL `https://arxiv.org/abs/1609.07088`.

[21] Lerrel Pinto and Abhinav Gupta. Learning to push by grasping: Using multiple tasks for effective learning, 2016. URL `https://arxiv.org/abs/1609.09025`.

[22] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning, 2024. URL `https://arxiv.org/abs/2309.07915`.

[23] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety, 2016. URL `https://arxiv.org/abs/1606.06565`.

[24] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people, 2018. URL `https://arxiv.org/abs/1802.08218`.

[25] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017. URL `https://arxiv.org/abs/1612.00837`.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL `https://arxiv.org/abs/1405.0312`.

[27] Leonard Salewski, A. Sophia Koepke, Hendrik P. A. Lensch, and Zeynep Akata. *CLEVR-X: A Visual Reasoning Dataset for Natural Language Explanations*, page 69–88. Springer International Publishing, 2022. ISBN 9783031040832. doi: 10.1007/978-3-031-04083-2_5. URL `http://dx.doi.org/10.1007/978-3-031-04083-2_5`.

[28] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning, 2019. URL https://arxiv.org/abs/1903.02741.

[29] Bjoern H Menze, András Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin S. Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, Levente Lanczi, Elizabeth R. Gerstner, Marc-André Weber, Tal Arbel, Brian B. Avants, Nicholas Ayache, Patricia Buendia, D. Louis Collins, Nicolas Cordier, Jason J. Corso, Antonio Criminisi, Tilak Das, Hervé Delingette, Çağatay Demiralp, Christopher R. Durst, Michel Dojat, Senan Doyle, Joana Festa, Florence Forbes, Ezequiel Geremia, Ben Glocker, Polina Golland, Xiaotao Guo, Andaç Hamamci, Khan M. Iftekharuddin, Rajesh Jena, Nigel M. John, Ender Konukoglu, Danial Lashkari, José Antonio Mariz, Raphael Meier, Sérgio Pereira, Doina Precup, Stephen John Price, Tammy Riklin-Raviv, Syed M. S. Reza, Michael T. Ryan, Duygu Sarikaya, Lawrence H. Schwartz, Hoo-Chang Shin, Jamie Shotton, Carlos Alberto Silva, Nuno J. Sousa, Nagesh K. Subbanna, Gábor Székely, Thomas J. Taylor, Owen M. Thomas, N. Tustison, Gözde B. Ünal, Flor Vasseur, Max Wintermark, Dong Hye Ye, Liang Zhao, Binsheng Zhao, Darko Zikic, Marcel Prastawa, Mauricio Reyes, and Koenraad Van Leemput. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34:1993–2024, 2015. URL https://api.semanticscholar.org/CorpusID:1739295.

[30] Shancheng Jiang, Zehui Wu, Haiqiong Yang, Kun Xiang, Weiping Ding, and Zhen-Song Chen. A prior knowledge-guided distributionally robust optimization-based adversarial training strategy for medical image classification. *Inf. Sci.*, 673:120705, 2024. URL https://api.semanticscholar.org/CorpusID:269696095.

[31] Yunhe Gao, Zhuowei Li, Di Liu, Mu Zhou, Shaoting Zhang, and Dimitris N. Metaxas. Training like a medical resident: Context-prior learning toward universal medical image segmentation, 2024. URL https://arxiv.org/abs/2306.02416.

[32] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training in radiology, 2023. URL https://arxiv.org/abs/2301.02228.

[33] Ziyuan Qin, Huahui Yi, Qicheng Lao, and Kang Li. Medical image understanding with pretrained vision language models: A comprehensive study. *International Conference on Learning Representations*, 2022.

[34] Yuguang Yang, Tongfei Chen, Haoyu Huang, Linlin Yang, Chunyu Xie, Dawei Leng, Xianbin Cao, and Baochang Zhang. Prompt as knowledge bank: Boost vision-language model via structural representation for zero-shot medical detection, 2025. URL https://arxiv.org/abs/2502.16223.

[35] Yongjian Wu, Yang Zhou, Jiya Saiyin, Bingzheng Wei, Maode Lai, Jianzhong Shou, Yubo Fan, and Yan Xu. Zero-shot nuclei detection via visual-language pre-trained models, 2023. URL https://arxiv.org/abs/2306.17659.

[36] Kyungmin Min, Minbeom Kim, Kang il Lee, Dongryeol Lee, and Kyomin Jung. Mitigating hallucinations in large vision-language models via summary-guided decoding, 2025. URL https://arxiv.org/abs/2410.13321.

[37] Yaqi Sun, Kyohei Atarashi, Koh Takeuchi, and Hisashi Kashima. Exploring causes and mitigation of hallucinations in large vision language models, 2025. URL https://arxiv.org/abs/2502.16842.

[38] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models, 2024. URL https://arxiv.org/abs/2310.00754.

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[40] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10975, 2022.

[41] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[42] 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.

[43] 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.

[44] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[45] 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.

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

[47] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[48] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

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

[50] 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.

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

[52] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025.

[53] Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero's" aha moment" in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*, 2025.

[54] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

[55] Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*, 2024.

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

[57] Huilin Deng, Ding Zou, Rui Ma, Hongchen Luo, Yang Cao, and Yu Kang. Boosting the generalization and reasoning of vision language models with curriculum reinforcement learning. *arXiv preprint arXiv:2503.07065*, 2025.

[58] 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. *arXiv preprint arXiv:2502.19634*, 2025.

[59] Sheng Zhang, Qianchu Liu, Guanghui Qin, Tristan Naumann, and Hoifung Poon. Med-rlvr: Emerging medical reasoning from a 3b base model via reinforcement learning. *arXiv preprint arXiv:2502.19655*, 2025.

[60] 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.

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

[62] Jiaqi Wang, Enze Shi, Sigang Yu, Zihao Wu, Chong Ma, Haixing Dai, Qiushi Yang, Yanqing Kang, Jinru Wu, Huawen Hu, et al. Prompt engineering for healthcare: Methodologies and applications. *arXiv preprint arXiv:2304.14670*, 2023.

[63] Stefan Denner, Markus Bujotzek, Dimitrios Bounias, David Zimmerer, Raphael Stock, Paul F Jäger, and Klaus Maier-Hein. Visual prompt engineering for medical vision language models in radiology. *arXiv preprint arXiv:2408.15802*, 2024.

[64] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[65] Weijie Fan, Yi Yang, Jing Qi, Qichuan Zhang, Cuiwei Liao, Li Wen, Shuang Wang, Guangxian Wang, Yu Xia, Qihua Wu, et al. A deep-learning-based framework for identifying and localizing multiple abnormalities and assessing cardiomegaly in chest x-ray. *Nature Communications*, 15 (1):1347, 2024.

[66] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

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

[68] Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm-r1: A stable and generalizable r1-style large vision-language model, 2025. URL https://arxiv.org/abs/2504.07615.

[69] Srinivas R Sadda, Muneeswar G Nittala, Wongsiri Taweebanjongsin, Aditya Verma, Swetha B Velaga, Ahmed Roshdy Alagorie, Connie M Sears, Paolo S Silva, and Lloyd P Aiello. Quantitative assessment of the severity of diabetic retinopathy. *American Journal of Ophthalmology*, 218:342–352, 2020.

[70] Desik Rengarajan, Gargi Vaidya, Akshay Sarvesh, Dileep Kalathil, and Srinivas Shakkottai. Reinforcement learning with sparse rewards using guidance from offline demonstration. *arXiv preprint arXiv:2202.04628*, 2022.

[71] Murad Dawood, Nils Dengler, Jorge de Heuvel, and Maren Bennewitz. Handling sparse rewards in reinforcement learning using model predictive control. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 879–885. IEEE, 2023.

[72] Rui Yang, Xiaoman Pan, Feng Luo, Shuang Qiu, Han Zhong, Dong Yu, and Jianshu Chen. Rewards-in-context: Multi-objective alignment of foundation models with dynamic preference adjustment. In *International Conference on Machine Learning*, pages 56276–56297. PMLR, 2024.

[73] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

[74] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1): 1–9, 2018.

[75] Osamah Taher and Kasım Özacar. Medcapsnet: A modified densenet201 model integrated with capsule network for heel disease detection and classification. *Heliyon*, 10(14), 2024.

[76] VV Danilov, Alex Proutski, Alexander Kirpich, DE Litmanovich, and Yuriy Gankin. Dataset for covid-19 segmentation and severity scoring. *Mendeley Data*, 10:4, 2022.

[77] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. URL http://arxiv.org/abs/2403.13372.

[78] Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. Easyr1: An efficient, scalable, multi-modality rl training framework. https://github.com/hiyouga/EasyR1, 2025. Accessed: 2025-04-25.

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

[80] agchung. Actualmed-covid-chestxray-dataset. https://github.com/agchung/Actualmed-COVID-chestxray-dataset, 2020. Accessed: 2025-05-07.

[81] Amith Khandakar Tawsifur Rahman, Dr. Muhammad Chowdhury. Covid-19 radiography database. https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database, 2022. Accessed: 2025-05-07.

[82] Joseph Paul Cohen, Paul Morrison, and Lan Dao. Covid-19 image data collection. *arXiv 2003.11597*, 2020. URL https://github.com/ieee8023/covid-chestxray-dataset.

[83] agchung. Figure 1 covid-19 chest x-ray dataset initiative. https://github.com/agchung/Figure1-COVID-chestxray-dataset, 2020. Accessed: 2025-05-07.

A. Appendix

A.1. Detailed Information of Datasets Used for Relative Task

Medical Image Classification We use eight datasets from MedMNIST [73], representing various imaging modalities, including X-Ray, Ultrasound, and CT: BreastMNIST, PneumoniaMNIST, OCTMNIST, RetinaMNIST, DermaMNIST, TissueMNIST, BloodMNIST, and OrganAMNIST. Most of these datasets contain over 15,000 images. For training efficiency, we randomly sample up to 256 images per class, except for RetinaMNIST and BreastMNIST, which have fewer than 1,500 images. This setup is treated as a 256-shot setting, with 10-shot and 20-shot settings derived similarly using a consistent test set.

Fine-Grained Regional Classification. To simulate the clinical workflow of locating and identifying lesions, we use two datasets: HAM10000 [74] and Heel [75]. HAM10000 contains 10,015 dermoscopic images of seven skin lesion types, providing region of interest (ROI) masks without bounding boxes. We derive bounding boxes from the ROI edges. The Heel dataset consists of 3,956 X-ray images of foot lesions, designed for heel bone disease localization and classification.

Severity Grading. In addition to RetinaMNIST from MedMNIST, we also utilized Danilov's preprocessed dataset [76], which consolidates four publicly available datasets for COVID-19 and pneumonia classification. These datasets include Actualmed COVID-19 Chest X-ray [80], COVID-19 Radiography [81], COVID Chest X-Ray [82], and Figure1 COVID Chest X-ray [83]. Danilov's preprocessing standardizes these datasets and provides human-labeled severity scores ranging from 0 to 6, making them suitable for severity grading tasks. We combined these preprocessed datasets for consistent usage in our experiments.

A.2. Comprehensive Details of Experimental Settings and Evaluation Metric

Perception Augmentation Policies. As mentioned earlier, we use two approaches to enhance the model's perceptual capabilities in the medical imaging domain. One is by explicitly injecting medical prior knowledge into the model through prompt engineering to directly perform medical diagnosis, while the other is by transferring inherent knowledge through training on other tasks to improve the model's medical diagnostic ability.

- **Prior Knowledge Augmentation**: We train the model on MedMNIST datasets [73] using two distinct prompt settings. The first setting only provides $\{\boldsymbol{Class\ Names}\}$, while the second setting includes explicit knowledge injection, which provides both $\{\boldsymbol{Class\ Names}\}$ and corresponding $\{\boldsymbol{Visual\ Attributes}\}$. In addition to original dataset settings, we also apply SFT and RL on limited data, adopting 10-shot and 20-shot settings to evaluate the fine-tuned model's generalization ability. Note that for the overall reward R we formulate it as $R = 0.9 \times R_{\text{accuracy}} + 0.1 \times R_{\text{formate}}$ by default. While for the RetinaMNIST dataset we adopt the MFRS reward for better performance, that is, $\hat{R} = 0.9 \times R_{\text{MFRS}} + 0.1 \times R_{\text{format}}$.
- **Visual Perception Augmentation**: We first train the model employing the R1 framework on the training set of the HAM10000 and Heel datasets, respectively, learning to localize specific regions, lesions, or organs. For example, detecting the bounding boxes for skin lesions in HAM10000 images and localizing the heel bone region in Heel images. Subsequently, without any additional training, we directly apply the model to the corresponding test sets for medical disease diagnosis in a zero-shot manner.

Reasoning Augmentation via Reward Design.

- **Recitation Reward**: We conduct two experiments by varying the value of δ in equation 7. When $\delta = 0.2$, the model is rewarded for repeating explicit knowledge during the thinking process. Conversely, when $\delta = -2$, the model is penalized for recitation.
- **MFRS Reward**: We train the model on the RetinaMNIST dataset in MedMNIST and the COVID-19 Dataset[76] using two reward settings to evaluate MFRS reward's validity. The difference is whether to replace R_{MFRS} in euqation 10 with R_{accuracy}.

Comparative Evaluation & Metric. For the **Explicit Knowledge Injection** experiment, we primarily compare the performance of SFT and RL fine-tuning on the test sets, as well as the few-shot experiments. For the **Implicit Knowledge Injection** experiment, we compare the performance of RL and SFT in two approaches: (1) direct classification training, and (2) localization followed by direct classification. We use a metric similar to VQA choice accuracy. Each test sample consists of a medical question and a medical image, and the model must choose a diagnosis from a predefined list of lesion types. A correct diagnosis is made only when the model's prediction matches the ground truth. Finally, we evaluate the model's diagnostic performance by calculating the overall accuracy on the test set.

A.3. Broader Impact

This paper presents work aimed at extending reinforcement learning fine-tuning into the domain of medical imaging. Our goal is to enhance model transparency by enabling visible reasoning processes during medical image interpretation. While this direction may have important implications for clinical AI applications, we believe no specific societal concerns need to be highlighted at this stage.

A.4. Basic Rewards for Reinforcement Fine-Tuning

Format Reward. Following previous work [3, 67, 68], we introduced format rewards to evaluate whether the model's generated output adheres to the expected structured format. Specifically, the model is enforced to enclose its thinking process between the `<think>...</think>` tags, include a bounding box within `<answer>{...[x1, y1, x2, y2]...}</answer>` for the detection task, or place the predicted label into `\boxed{...}` for the classification task, receiving 1 or 0 reward value based on compliance.

Vanilla Accuracy Reward. Detection task requires the model to provide the bounding box for a specific region, lesion, or organ in the medical image. Denote GT^{det} as the ground truth bounding box, O^{det} as the model output content, and f_{det} as the function to extract the bounding box located by the VLM from its output content. The accuracy reward for detection task is defined as follows:

$$R_{acc}^{det} = \begin{cases} 1.0 & \text{if IoU}(GT^{\text{det}}, f_{det}(O^{\text{det}})) > threshold, \\ 0.0 & \text{otherwise.} \end{cases} \tag{11}$$

where IoU is the intersection over union metric, and the $threshold$ is typically set to 0.5 as default.

The vanilla accuracy reward for classification tasks is the most commonly used exact match, i.e., the model receives a reward score of 1 if the final answer exactly matches the ground truth when both are converted to lowercase; otherwise, the score is 0.

A.5. Prompt Template for Localization Task

Prompt Template

This is a $\{data\ modality\}$ image of $\{lession/organ\}$. Please identify the category of the $\{lession/organ\}$ based on the image. Categories and their typical descriptions are as follows: $\{Class\ Names: Visual\ Attributes\}$. You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within <think> </think> tags. The final answer MUST BE put in `\boxed{...}`.

We need to construct training data for the localization task, using the following prompt template:

Prompt Template for Detection Task

Analyze the image and provide the bounding box for the {*target object*}. Ensure the bounding box accurately covers it and does not include too much unrelated areas. Output the bounding box in the format [x1, y1, x2, y2]. Generate your thinking process on how you determined the box. First output the thinking process in <think> </think> tags and then output the final answer in <answer> </answer> tags. Output the final answer in JSON format.

A.6. Limitation

Our work is still limited to medical classification tasks, and has yet to explore fine-grained tasks such as segmentation. In addition, our current approach to knowledge injection lacks certain clinically grounded experiential knowledge. We plan to further investigate these directions in future work.

A.7. Principle of BLEU Metric

BLEU calculates similarity by comparing the overlap of n-grams between the candidate text and the reference text, making it particularly suitable for quantifying the similarity between the model's inference outputs and the prior knowledge.