Breaking Failure Cascades: Step-Aware Reinforcement Learning for Medical Multimodal Reasoning

Anonymous ACL submission

Abstract

Recent multimodal large language models have shown great promise in clinical image reasoning, but existing post-training pipelines remain predominantly outcome-centric, relying on final answer correctness or sequence-level preferences. This approach suffers from sparse credit assignment, making it difficult to optimize the internal reasoning process essential for clinical applications. Our analysis reveals that cascading errors from early-stage reasoning failures are a leading cause of incorrect predictions in medical visual question answering (VQA) benchmarks. Motivated by this, we propose **M**edical **R**easoning-aware **P**olicy **O**ptimization **(MRPO)**, an RL algorithm that incorporates step-wise process rewards. When the final answer is incorrect, MRPO assigns exponentially larger penalties to tokens in earlier invalid reasoning steps, breaking failure cascades without compromising successful paths. Across three multimodal LLM backbones, MRPO consistently outperforms standard GRPO and a recent RL baseline, and on Qwen3-VL-8B-Instruct even outperforms substantially larger medical MLLMs such as HuatuoGPT-Vision-34B by 2.79 points. Moreover, MRPO reduces early-stage reasoning failures from 64.0% to 13.0%, demonstrating that targeted mitigation of cascading failures improves both reasoning quality and final answer accuracy.

1 Introduction

Recent advances in multimodal large language models (MLLMs) have extended their capabilities to clinical image reasoning (Chen et al., 2024; Team et al., 2025a), where prior work has explored chain-of-thought fine-tuning on curated reasoning traces (Sun et al., 2025; Kim et al., 2025) as well as reinforcement learning (RL) post-training (Fan et al., 2025; Lai et al., 2025). However, most post-training techniques for medical MLLMs remain outcome-centric, supervising primarily through final answer correctness or sequence-level preferences (Huang et al., 2025; Liu et al., 2025). This inherently suffers from sparse credit assignment problem, when a reasoning trajectory fails, the learning signal cannot identify which intermediate steps caused the failure (Mu et al., 2025; Xie et al., 2025a). This is particularly severe in free-form generation, where rewards are sparse and delayed, typically observed only after the entire response is generated (Chaudhari et al., 2024). Consequently, learning signals are distributed uniformly across all tokens, making it difficult for models to learn how to reason correctly step by step. Since real clinical environments predominantly involve open-ended queries, addressing this limitation is essential for practical deployment of medical MLLMs.

Beyond sparse credit assignment, we identify a structural failure mode in medical multimodal reasoning: early-stage reasoning failure tends to propagate and accumulate, forming failure cascades that dominate final prediction errors. Through systematic analysis of sentence-level reasoning traces generated by existing MLLMs on open-ended medical visual question answering (VQA) benchmarks, we find a strong correlation between the position of the first invalid reasoning step and the probability of producing an incorrect final answer. Once an early reasoning step becomes invalid, subsequent steps are significantly more likely to fail, even when later reasoning capabilities would otherwise suffice.

To mitigate this, we propose **M**edical **R**easoning-aware **P**olicy **O**ptimization **(MRPO)**, an RL algorithm that reshapes the GRPO (Shao et al., 2024)-based advantage to assign larger penalties to tokens in earlier failed reasoning steps. MRPO leverages an answer reward and a step-wise reasoning process reward computed by an external judge model that evaluates each step as valid or invalid. When the answer reward indicates an incorrect prediction, we reshape the advantage to assign exponentially larger penalties to earlier invalid steps, preserving

successful reasoning trajectories while increasing the probability of generating correct tokens at the first invalid step in failed traces. As a result, MRPO corrects early-stage reasoning failures and encourages valid reasoning in later steps, improving both answer accuracy and reasoning quality.

To demonstrate its effectiveness, we apply MRPO to three multimodal LLM backbones, Qwen2.5-VL-7B-Instruct (Bai et al., 2025b), Qwen3-VL-8B-Instruct (Bai et al., 2025a), and InternVL3-8B-Instruct (Zhu et al., 2025), on diverse open-ended medical VQA benchmarks. With only 13K training samples, MRPO consistently achieves the highest average performance across all three backbones, outperforming GRPO and the recent RL baseline GDPO (Liu et al., 2026). On Qwen3-VL-8B-Instruct, MRPO outperforms larger medical MLLMs including HuatuoGPT-Vision-34B by 2.79 points, suggesting that targeted reasoning supervision can be a competitive alternative to large-scale medical instruction tuning. Moreover, reasoning failure analysis shows that MRPO substantially improves the reasoning failure pattern, reducing early-stage failures from 64.0% to 13.0% and mitigating downstream failure accumulation.

Our contributions are summarized as follows:

- Through systematic analysis of sentence-level reasoning traces on open-ended medical VQA, we identify cascading failures as a dominant cause of incorrect predictions, where early-stage reasoning failures propagate and accumulate to derail the final answer.
- We propose MRPO, a GRPO-based RL algorithm that incorporates step-wise process rewards and reshapes advantages with exponentially larger penalties on earlier failed steps, directly targeting cascading failures.
- Across three backbones, MRPO consistently achieves the highest average performance over standard GRPO and GDPO. On Qwen3-VL-8B-Instruct, it outperforms larger medical MLLMs such as HuatuoGPT-Vision-34B by 2.79 points. Moreover, MRPO improves the reasoning failure pattern, reducing early-stage failures from 64.0% to 13.0% and mitigating downstream failure accumulation.

2 Related Works

Multimodal large language models (Team et al., 2025b; Zhu et al., 2025) have been adapted to medical vision-language tasks (Li et al., 2023; Sellergren et al., 2025), evolving from supervised fine-tuning (Sun et al., 2025; Kim et al., 2025) to GRPO-based RL (Lai et al., 2025; Pan et al., 2025; Su et al., 2025) following DeepSeek-R1 (Guo et al., 2025). To move beyond final-answer supervision, recent work evaluates individual reasoning steps and uses them as process rewards in RL (Fan et al., 2025; Zhi et al., 2025), as inference-time verifiers (Yun et al., 2025), or as offline preferences (Yang et al., 2026). However, none redistributes the learning signal by where a failure occurs, and thus cannot directly correct the failed step. This is studied more directly in general-domain reasoning, where token-level credit assignment such as FSPO (Li and Ng, 2025) and CAPO (Xie et al., 2025a) allocates learning signals selectively across the trajectory. MRPO combines this token-level credit assignment with the medical step-level evaluation. By assigning exponentially stronger penalties to earlier invalid steps, MRPO corrects early reasoning failures before they cascade and steers the model toward valid reasoning. A detailed overview is in Appendix A.

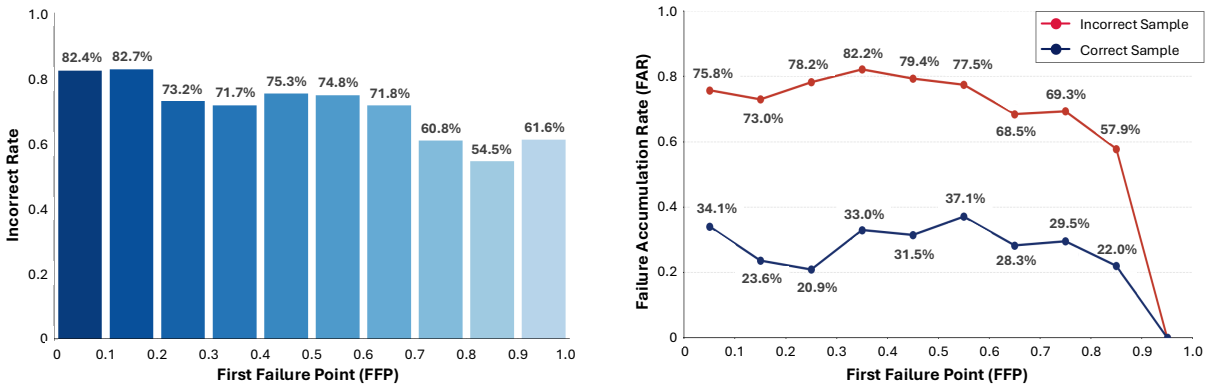
3 Preliminary Experiments

To analyze structural failures in reasoning patterns for open-ended medical VQA, we conduct preliminary experiments across a diverse set of multimodal LLMs and medical VQA benchmarks.

Experiment Purpose. We first analyze the relationship between the point of first reasoning failure and the final answer incorrect rate. To this end, we define the *First Failure Point* (FFP), which represents the relative position in a reasoning trace where the first invalid step occurs. Given K total reasoning steps and the index k of the first invalid step, we compute $\mathrm{FFP} = k/K$. A lower FFP indicates that reasoning fails earlier in the trace. We further examine how reasoning behaves after the first failure by defining the *Failure Accumulation Rate* (FAR), which measures the proportion of failed steps among the remaining steps after the first failed step. We compute FAR as follows:

$$\mathrm{FAR} = \frac{\#\text{ failed steps after the first failure}}{\#\text{ total steps after the first failure}} \quad (1)$$

Models and Benchmarks. We evaluate four MLLMs spanning general-purpose and medical-specialized domains: (i) General MLLMs: Qwen3-VL-8B-Instruct (Bai et al., 2025a), InternVL3-8B-Instruct (Zhu et al., 2025); (ii) Medical



(A) Incorrect rate across First Failure Point (FFP) bins.

(B) Failure Accumulation Rate (FAR) across FFP bins.

Figure 1: **Step-wise medical multimodal reasoning analysis.** (A) Incorrect rate across FFP bins. Earlier first failures are associated with substantially higher incorrect rates. (B) FAR across FFP bins for correct and incorrect instances. Incorrect predictions show greater failure accumulation, particularly when the first failure occurs early.

MLLMs: HuatuoGPT-Vision-7B (Chen et al., 2024), Lingshu-7B (Team et al., 2025a). We use three medical VQA benchmarks, VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), and PathVQA (He et al., 2020). To obtain gold rationales for each medical VQA instance, we use MedThink (Gai et al., 2024), which provides medical decision-making rationales for these three benchmarks. We align each MedThink gold reasoning one-to-one with its corresponding test instance. We then exclude binary and multiple-choice questions to focus our experiment on open-ended instances.

Evaluation Metrics. We assess the correctness of answers to open-ended VQA using an LLM-as-judge approach (Zheng et al., 2023) with GPT-5-mini[1], following the evaluation prompt design in PeFoMed (He et al., 2025). The prompt used for this evaluation is provided in Appendix H.1.

Evaluating the validity of each reasoning step in medical VQA is inherently challenging, as it requires identifying domain-specific key findings, anatomical landmarks, and pathological features relevant to the question. Without explicit guidance on what constitutes valid observations or inferences, an LLM judge lacks sufficient context to assess whether a step contributes meaningfully to the diagnostic process. We thus introduce two complementary metrics that together define step validity. **(i) Gold Alignment** measures whether a reasoning step is consistent with the gold reasoning trajectory. By providing expert-annotated reasoning as a reference, we anchor the evaluation to clinically appropriate observations and diagnostic directions. **(ii) Answer Contribution:** measures whether a reasoning step directly contributes to deriving the ground-truth answer. This metric captures cases where a step may diverge from the gold reasoning path yet still validly supports the correct conclusion, acknowledging that multiple reasoning trajectories can lead to the same diagnosis. Each metric is scored binarily at the sentence level using GPT-5-mini, and a step is *valid* if it scores 1 on either metric, reflecting that multiple valid diagnostic pathways can lead to the same conclusion in medical practice. The prompt is in Appendix H.2.

To validate the reliability of our two LLM-as-judge protocols, we conduct human evaluation on answer correctness and step-wise reasoning quality. Both protocols achieve substantial agreement with human judgments, as detailed in Appendix B.1.

Results and Analysis. Figure 1A groups instances into bins of width 0.1 based on FFP and visualizes the incorrect rate within each bin. Aggregating all four models, we observe a clear trend in which earlier FFPs correspond to earlier failures in step-wise reasoning and are associated with a higher probability of incorrect answers. Figure 1B visualizes FAR across FFP bins separately for correct and incorrect instances, revealing two observations. First, across all bins, incorrect instances exhibit higher FAR than correct ones, suggesting a strong link between failure accumulation and answer correctness. Second, for incorrect instances, earlier FFPs show even higher FAR, indicating stronger downstream failure accumulation when the first failure occurs early in the trajectory.

In summary, our analysis demonstrates that initial reasoning failures trigger an error cascade that

[1]We use `gpt-5-mini-2025-08-07`

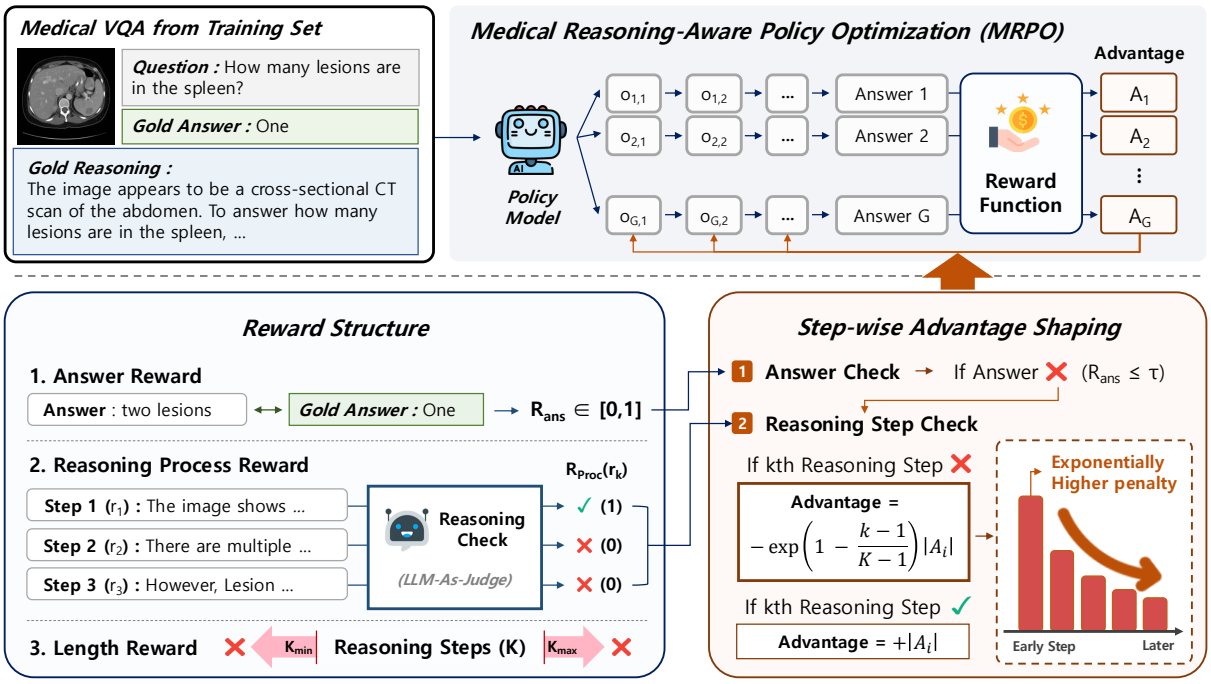


Figure 2: **Overview of the MRPO algorithm.** The policy model generates multiple reasoning paths, each evaluated by answer, step-wise reasoning process reward, and length reward. When the answer is judged incorrect, MRPO assigns larger penalties to earlier failed steps to correct early-stage reasoning failures.

leads to incorrect medical VQA outcomes. Standard GRPO-based methods struggle to address this, as they distribute learning signals uniformly across all tokens, failing to pinpoint early errors. We thus propose a medical reasoning-aware RL algorithm that overcomes this limitation by evaluating reasoning step-wise and assigning larger penalties to tokens in the initial stages of a failed trajectory, targeting the root causes of reasoning failure.

4 Approach

We propose **M**edical **R**easoning-aware **P**olicy **O**ptimization (**MRPO**), which improves medical VQA reasoning by correcting failures in the early stages of the reasoning process. As shown in Figure 2, MRPO incorporates both an answer reward and a step-wise reasoning reward during policy optimization. Based on these signals, MRPO shapes advantages to assign larger penalties to tokens in reasoning steps that fail earlier when the final answer is judged incorrect. This discourages early-stage reasoning failures and encourages valid reasoning in subsequent steps, thereby improving both reasoning quality and answer accuracy.

4.1 Reward structure

We define three reward components, (1) Answer Reward, (2) Reasoning Process Reward, and (3) Length Reward. The final reward is computed as a weighted combination of these components.

Answer Reward. To assess how well the generated answer aligns with the reference answer (Jeong et al., 2024), we adopt a weighted combination of lexical overlap ($\lambda_1 = 0.25$) and semantic similarity ($\lambda_2 = 0.5$):

$$R_{\text{ans}} = \lambda_1 \cdot (\text{ROUGE-1} + \text{BLEU-1}) + \lambda_2 \cdot \text{BERTScore} \tag{2}$$

ROUGE-1 (Lin, 2004) and BLEU-1 (Papineni et al., 2002) capture lexical overlap, but can be sparse for short open-ended medical answers. We therefore include BERTScore (Zhang et al., 2019) to provide a denser semantic signal, computed using BiomedBERT (Chakraborty et al., 2020).

Reasoning Process Reward. We segment the reasoning text r into K sentence-level steps $\{r_k\}_{k=1}^{K}$, and evaluate each step r_k given the ground-truth answer $a^\star$ and gold reasoning $r^\star$ using two criteria. **(i) Gold Alignment** assesses whether step r_k is consistent with $r^\star$. A step is aligned if it identifies key findings in $r^\star$, maintains correct anatomical localization, and follows the diagnostic pathway. A step is misaligned if it contradicts $r^\star$, identifies incorrect anatomical location or laterality, or gives only generic instead of specific findings. **(ii) Answer Contribution** assesses

whether step r_k directly contributes to deriving $a^\star$. A step is contributive if it explicitly mentions $a^\star$ or identifies findings required to derive it. This captures cases where a step may diverge from the gold reasoning path yet still supports the correct conclusion, acknowledging that multiple reasoning trajectories can lead to the same diagnosis.

These criteria match the reasoning evaluation metrics in Section 3. Using GPT-5-mini as an LLM judge, we assign a binary score $\text{score}(r_k, c) \in \{0, 1\}$ for each criterion c, and set $R_{\text{proc}}(r_k) = 1$ if either criterion scores 1, and $R_{\text{proc}}(r_k) = 0$ otherwise. The prompt is provided in Appendix H.2.

$$R_{\text{proc}}(r_k) = \begin{cases} 1, & \text{if score}(r_k, c) = 1,\ \exists c \in \mathcal{C} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Length Reward. Reasoning traces with insufficient length may omit essential diagnostic steps, while excessively lengthy outputs introduce redundant information. To prevent such behaviors, we introduce a length-based reward R_{len} that regularizes the number of reasoning steps. Given a trace of K sentence-level steps, we define:

$$R_{\text{len}}(r) = \begin{cases} -\frac{K_{\min} - K}{K_{\min}}, & \text{if } K < K_{\min} \\ -\frac{K - K_{\max}}{K_{\max}}, & \text{if } K > K_{\max} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $K_{\min}$ and $K_{\max}$ denote the minimum and maximum acceptable step counts, set to 4 and 10 in our experiments. This reward imposes a linear penalty proportional to the deviation from the acceptable range, encouraging the model to generate reasoning traces of appropriate length.

The total reward combines the answer reward, the step-averaged reasoning process reward, and the length reward as follows:

$$R_{\text{tot}} = R_{\text{ans}} + \frac{1}{K} \sum_{k=1}^{K} R_{\text{proc}}(r_k) + R_{\text{len}} \tag{5}$$

4.2 Medical Reasoning-aware Advantage Shaping

We adopt GRPO (Shao et al., 2024) as our base policy optimization algorithm. Given an input x, GRPO generates a set of G candidate outputs $\{y_1, \ldots, y_G\}$, with rewards $\{R_1, \ldots, R_G\}$, and computes a group-normalized advantage for each candidate output y_i as

$$A_i = \frac{R_i - \text{mean}(\{R_1, \ldots, R_G\})}{\text{std}(\{R_1, \ldots, R_G\})} \tag{6}$$

Most GRPO-based methods compute the advantage at the sequence level and apply the same learning signal uniformly to all tokens, which fails to provide differentiated supervision across reasoning steps. Consequently, it is inadequate for mitigating the early-stage reasoning failures and ensuing error accumulation observed in Section 3. To address this, MRPO reshapes advantages step-wise to assign larger penalties to earlier failed reasoning steps when the final answer is judged incorrect.

Step-wise Advantage Shaping. We adjust token-level advantages using the answer reward R_{ans} and the step-wise reasoning process reward $R_{\text{proc}}(r_k)$ for each step r_k. We judge the final answer as correct if $R_{\text{ans}} > \tau$ and as incorrect otherwise, where τ is set to the midpoint between the mean R_{ans} of samples labeled correct and incorrect by the LLM-as-judge in Section 3, yielding $\tau = 0.6$.

When the final answer is incorrect and a reasoning step is evaluated as invalid ($R_{\text{proc}}(r_k) = 0$), we reshape the advantage to assign exponentially larger penalties to tokens in earlier failed steps. Concretely, for an output y_i, we modify the advantage of each token $o_{i,t} \in r_k$ as follows:

$$\hat{A}_{i,t} = \begin{cases} -\exp\left(1 - \frac{k-1}{K-1}\right)|A_i| & \text{if } \substack{R_{\text{ans}} \le \tau \\ \text{and } R_{\text{proc}}(r_k)=0} \\ +|A_i| & \text{if } \substack{R_{\text{ans}} \le \tau \\ \text{and } R_{\text{proc}}(r_k)=1} \\ A_i & \text{otherwise} \end{cases} \tag{7}$$

This reshaping more strongly decreases the probability of inaccurate tokens in earlier failed reasoning steps, substantially increasing the probability of generating correct tokens at the first failure point and facilitating valid reasoning in subsequent steps. In contrast, when the final answer is judged correct, we do not reweight the reasoning tokens, reflecting the intuition that minor imperfections did not materially hinder arriving at the correct answer. With this design, MRPO selectively corrects reasoning only for incorrect predictions, improving failed reasoning traces without disrupting successful trajectories. Based on the adjusted advantages, we optimize the policy using the following GRPO-based training objective for each input prompt x:

Model	Benchmarks					AVG
	PMC-VQA	VQA-Med	Quilt-VQA	Rad-VQA	MIMIC-VQA	
General MLLMs						
Qwen2.5-VL-7B-Instruct	28.35	5.18	19.06	27.20	15.08	22.28
Qwen3-VL-8B-Instruct	31.00	9.41	23.90	33.15	15.31	25.61
Qwen3-VL-8B-Thinking	30.75	10.82	22.92	31.20	**22.39**	26.73
InternVL3-8B-Instruct	30.00	7.59	22.24	37.60	15.20	26.29
LLaVA-v1.6-7B	11.45	3.29	9.94	21.60	7.19	12.67
LLaVA-v1.6-34B	15.90	4.71	13.12	24.90	9.80	16.00
Medical MLLMs						
LLaVA-Med-v1.5-7B	20.65	3.53	14.92	23.60	9.57	17.07
HuatuoGPT-Vision-7B	27.90	8.94	19.61	32.70	16.06	24.28
HuatuoGPT-Vision-34B	31.30	9.65	21.27	33.60	17.63	26.15
Chiron-o1-8B	29.30	6.82	20.86	29.85	16.82	24.05
QoQ-Med-7B	28.90	7.76	21.82	26.60	14.56	22.58
Medical Reasoning MLLMs (GRPO)						
MedVLM-R1-2B	23.35	5.41	17.54	25.10	12.88	19.51
MedVLThinker-7B	28.90	6.82	20.30	28.85	15.66	23.29
Our Method(MRPO)						
Qwen2.5-VL-7B-Instruct(MRPO)	30.85	5.65	22.79	29.65	18.65	25.04
Qwen3-VL-8B-Instruct(MRPO)	**33.00**	**12.00**	23.90	**40.20**	17.46	**28.94**
InternVL3-8B-Instruct(MRPO)	31.60	7.59	**24.31**	39.15	20.42	28.75

Table 1: **Performance comparison of MRPO against existing MLLMs.** All models are evaluated on five out-of-distribution benchmarks, where VQA-Med denotes VQA-Med-2021, Rad-VQA denotes RadImageNet-VQA, and MIMIC-VQA denotes MIMIC-Ext-MIMIC-CXR-VQA. The best result in each column is in **bold** and the second-best is underlined. AVG denotes the average across all five benchmarks.

$$\mathcal{J}_{\text{MRPO}}(\theta) = E_{\{y_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(x)} \frac{1}{G} \sum_{i=1}^{G} \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \Big[\min \Big(\rho_{i,t}(\theta)\, \hat{A}_{i,t},\ \text{clip}(\rho_{i,t}(\theta),\, 1-\epsilon,\, 1+\epsilon)\, \hat{A}_{i,t} \Big) - \beta\, D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}) \Big] \tag{8}$$

$$\rho_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid x, y_{i,<t})} \tag{9}$$

5 Experiments

5.1 Experimental Setup

Dataset. Our training set is derived from the training splits of three medical VQA benchmarks, VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), and PathVQA (He et al., 2020), filtered to open-ended instances, and augmented with MedThink (Gai et al., 2024), which provides gold reasoning annotations. For evaluation, test splits of these three benchmarks serve as our in-distribution test sets. We further adopt five out-of-distribution benchmarks spanning diverse imaging modalities unseen during training, including PMC-VQA (Zhang et al., 2024), VQA-Med-2021 (Ben Abacha et al., 2021), Quilt-VQA (Seyfioglu et al., 2025), RadImageNet-VQA (Butsanets et al., 2026), and MIMIC-Ext-MIMIC-CXR-VQA (Bae et al., 2024). Details are provided in Appendix C.1.

Backbones. We apply MRPO to three open-source multimodal LLM backbones of comparable scale, namely Qwen2.5-VL-7B-Instruct (Bai et al., 2025b), Qwen3-VL-8B-Instruct (Bai et al., 2025a), and InternVL3-8B-Instruct (Zhu et al., 2025).

Evaluation Metrics. To evaluate **answer accuracy**, we adopt an LLM-as-judge approach using GPT-5-mini to assess answer correctness with a binary label. Accuracy is reported as the proportion of examples judged correct. The evaluation prompt is in Appendix H.1, and its reliability is validated through a human alignment study in Appendix B.1.

Baselines. We compare against three categories of baselines. **(i) General MLLMs,** general-purpose open-source MLLMs of varying sizes, such as Qwen2.5-VL (Bai et al., 2025b), Qwen3-VL (Bai et al., 2025a), InternVL3 (Zhu et al., 2025), and LLaVA-v1.6 (Liu et al., 2023); **(ii) Medical MLLMs,** medical-domain MLLMs trained on specialized biomedical datasets, including LLaVA-Med (Li et al., 2023), HuatuoGPT-Vision (Chen et al., 2024), Chiron-o1 (Sun et al., 2025), and QoQ-Med (Dai et al., 2025); **(iii) Medical Reasoning MLLMs,** medical reasoning MLLMs trained with

Method	In-Distribution			Out-of-Distribution					AVG
	VQA-RAD	SLAKE	PathVQA	PMC-VQA	VQA-Med	Quilt-VQA	Rad-VQA	MIMIC-VQA	
Qwen2.5-VL-7B-Instruct									
Base Model	42.00	58.29	17.10	28.35	5.18	19.06	27.20	15.08	23.36
SFT	38.50	63.88	18.83	28.50	4.71	18.37	26.20	15.02	23.59
GRPO	**44.00**	65.27	20.52	**31.05**	3.76	20.99	29.30	16.71	26.06
GDPO	42.50	**66.29**	19.60	28.60	**7.76**	20.17	**32.00**	16.47	25.92
MRPO	41.50	65.89	**21.30**	30.85	5.65	**22.79**	29.65	**18.65**	**26.79**
Qwen3-VL-8B-Instruct									
Base Model	43.00	59.69	**21.48**	31.00	9.41	23.90	33.15	15.31	26.83
SFT	**45.00**	66.29	21.27	30.35	7.76	19.48	33.40	15.20	26.79
GRPO	**45.00**	67.14	20.11	30.25	9.65	**24.03**	**40.60**	18.79	28.69
GDPO	42.50	67.28	19.09	32.10	**12.47**	22.51	40.10	**21.46**	29.01
MRPO	41.50	**68.27**	20.43	**33.00**	12.00	23.90	40.20	17.46	**29.09**
InternVL3-8B-Instruct									
Base Model	**45.50**	64.73	22.97	30.00	7.59	22.24	37.60	15.20	28.07
SFT	44.00	68.83	26.57	29.50	4.71	17.12	35.20	17.11	27.94
GRPO	42.00	68.27	28.15	30.50	7.76	23.48	38.75	19.49	30.84
GDPO	39.50	**71.95**	26.18	30.90	**10.12**	23.20	37.85	17.46	30.11
MRPO	41.00	71.78	**29.54**	**31.60**	7.59	**24.31**	**39.15**	**20.42**	**31.94**

Table 2: **Cross-backbone ablation of training methods.** We compare MRPO against the base model, SFT, GRPO, and GDPO on three backbones, namely Qwen2.5-VL-7B-Instruct, Qwen3-VL-8B-Instruct, and InternVL3-8B-Instruct. All methods are evaluated on three in-distribution and five out-of-distribution benchmarks. The best result in each column is in **bold** and the second-best is underlined. AVG denotes the average across all benchmarks.

GRPO, including MedVLM-R1 (Pan et al., 2025) and MedVLThinker (Huang et al., 2025).

5.2 Main Results

Comparison against MLLMs. Table 1 compares MRPO against existing MLLMs on diverse medical VQA benchmarks. MRPO consistently improves over the baseline across all three backbones, most notably on Qwen3-VL-8B-Instruct, where the average rises from 25.61 to 28.94, an improvement of 3.33 points. Built on this backbone, MRPO attains the highest average among all evaluated models. It obtains the top score on three of the five benchmarks, with a notable 7.05 point gain over the baseline on RadImageNet-VQA. Compared to general-purpose MLLMs, it outperforms all baselines regardless of scale, including the strongest reasoning variant Qwen3-VL-8B-Thinking (26.73) by 2.21 points. Against medical MLLMs, it outperforms the largest baseline HuatuoGPT-Vision-34B (26.15) by 2.79 points despite its substantially smaller 8B backbone and only 13K training samples. Together, these results show that targeted reasoning supervision can be a competitive alternative to large-scale medical instruction tuning, inducing transferable reasoning that generalizes across diverse out-of-distribution benchmarks rather than overfitting to narrow clinical contexts.

Cross-backbone Ablation. To verify that the gains of MRPO are not specific to a particular backbone, we compare MRPO against the base model, SFT, GRPO, and GDPO (Liu et al., 2026) across three backbones in Table 2. GDPO is a recent variant of GRPO that decouples reward normalization across groups. Training details are provided in Appendix C.2. MRPO achieves the highest average score on all three backbones, reaching 26.79 on Qwen2.5-VL-7B-Instruct, 29.09 on Qwen3-VL-8B-Instruct, and 31.94 on InternVL3-8B-Instruct. While SFT improves in-distribution performance, it fails to transfer to out-of-distribution benchmarks and even degrades several of them. In contrast, RL-based methods including MRPO yield consistent improvements on both in-distribution and out-of-distribution benchmarks, indicating that RL induces transferable reasoning capability rather than overfitting to the training distribution. Among the RL methods, MRPO consistently outperforms both GRPO and GDPO, with average improvements of 0.73, 0.40, and 1.10 points over GRPO on the three backbones. The improvement over GRPO is largest on InternVL3-8B-Instruct, where MRPO achieves the best score on five of the eight benchmarks. This suggests that step-wise advantage reshaping provides finer-grained credit assignment than the sequence-level optimization of GRPO and GDPO.

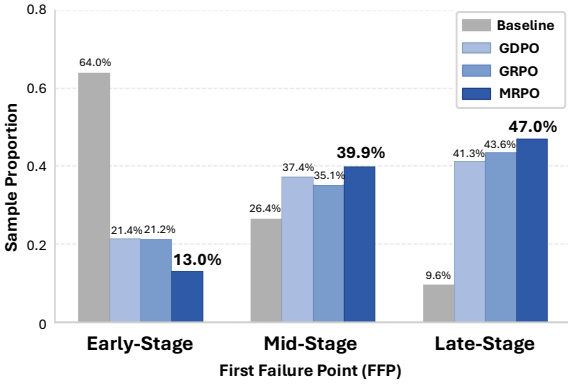


Figure 3: **Sample distribution across First Failure Point (FFP) stages.** Grouped into Early-Stage (0.0–0.4), Mid-Stage (0.4–0.7), and Late-Stage (0.7–1.0).

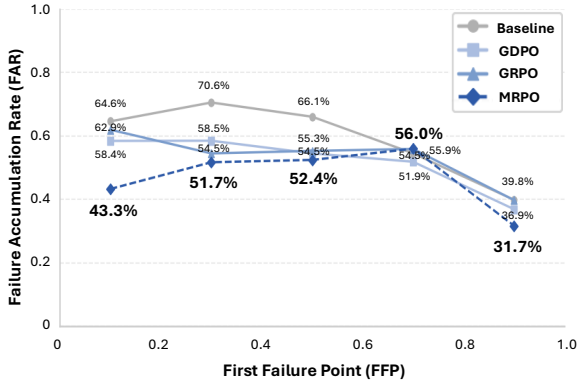


Figure 4: **Failure Accumulation Rate (FAR) across FFP bins.** MRPO shows the lowest FAR across all methods, indicating reduced failure accumulation.

Additional ablations are provided in Appendix D.

5.3 Reasoning Analysis

First Failure Point Analysis. To verify that MRPO addresses the cascading failure problem identified in Section 3, we analyze how the distribution of reasoning failures changes under MRPO. We include GRPO and GDPO trained under identical conditions to isolate the effect of step-wise advantage reshaping. Following the evaluation protocol in Section 3, we assess step validity and compare the proportion of samples falling into each of three FFP stages: Early, Mid, and Late-Stage.

Figure 3 shows the FFP distribution averaged across the three backbones. The baseline concentrates 64.0% of failures in the early-stage range, indicating frequent failures at the beginning of the trajectory. All RL methods reduce this, but MRPO lowers early-stage failures the most, to 13.0% versus 21.2% for GRPO and 21.4% for GDPO. Correspondingly, MRPO shifts the largest share of failures to the late-stage at 47.0%, compared to 9.6% in the baseline. These results demonstrate that MRPO's exponential penalty on early invalid steps effectively prevents early-stage failures and shifts the failure distribution toward later stages.

Failure Accumulation Analysis. Figure 4 compares the FAR across FFP bins at 0.2 intervals, averaged over the three backbones, where a lower FAR indicates more effective recovery after the first failure. In the Early-to-Mid-Stage range from FFP 0.0 to 0.6, the baseline exhibits consistently high FAR, from 64.6% to 70.6%. In contrast, MRPO records the lowest FAR across this range, with a particularly pronounced reduction in the 0.0–0.2 bin, where its FAR of 43.3% falls markedly below the baseline's 64.6% as well as GRPO (62.9%) and GDPO (58.4%). This indicates that MRPO reduces downstream failure propagation, with the strongest effect when failures begin early.

Together with the FFP results in Figure 3, MRPO mitigates cascading failures along two complementary axes: delaying failure onset and reducing subsequent accumulation. Detailed per-backbone results, an instance-level paired comparison between GRPO and MRPO, and qualitative analyses are provided in Appendix E and Appendix F.

6 Conclusion

In this paper, we show that in open-ended medical VQA, early-stage reasoning failures systematically propagate and dominate final prediction errors, exposing a key limitation of outcome-centric approaches. To address this, we propose MRPO, an RL algorithm that targets the first point of reasoning failure by integrating step-wise reasoning rewards into policy optimization. When the final answer is incorrect, MRPO assigns exponentially larger penalties to earlier invalid steps, correcting root causes of failure while preserving successful trajectories. Across three multimodal LLM backbones, MRPO consistently outperforms standard GRPO and GDPO, achieving competitive performance with substantially larger medical MLLMs using only 13K training samples. Failure analysis confirms that MRPO reduces early-stage reasoning failures and mitigates downstream failure accumulation, validating that our approach addresses the cascading failure problem. These findings suggest that explicitly addressing early reasoning failures offers a promising direction for developing more reliable medical multimodal reasoning.

Limitations

Our work has several limitations.

First, MRPO relies on an external LLM judge, GPT-5-mini, to compute the step-wise reasoning process reward during RL training. While this design enables fine-grained sentence-level supervision without training a separate process reward model, it incurs additional API cost and a dependency on the judge; a detailed cost breakdown is in Appendix C.3. As shown in Appendix D.4, GPT-5-mini provides the highest absolute performance ceiling among the three process reward models we evaluate, which is why we adopt it as the default. Conversely, the weakest judge yields a notable degradation, indicating that a sufficiently strong judge is required for the step-wise reward signal to be effective. A strong open-source multimodal judge such as MedGemma-27B nonetheless yields competitive results, making it a viable alternative when API access is constrained, though it does not reach the ceiling of GPT-5-mini. This motivates a promising future direction of training or distilling dedicated medical VQA process reward models that match frontier-LLM judgment quality while eliminating the dependency on external APIs.

Second, the construction of step-wise reasoning rewards in our framework relies on gold reasoning annotations from MedThink (Gai et al., 2024). Extending MRPO to settings without such gold rationales would require annotation-free reward signals or weaker forms of step-level supervision, both promising directions for future investigation.

Third, our evaluation is currently confined to medical VQA. The cascading failure problem and the step-wise advantage reshaping mechanism may generalize to other domains requiring multi-step reasoning, such as scientific question answering or legal reasoning, but we leave the empirical verification of this generalization to future work.

References

Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei JI, Eric Chang, Tackeun Kim, and Edward Choi. 2024. MIMIC-Ext-MIMIC-CXR-VQA: A Complex, Diverse, And Large-Scale Visual Question Answering Dataset for Chest X-ray Images. *PhysioNet*. Version 1.0.0.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

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

Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A. Hasan, and Henning Müller. 2021. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *CLEF 2021 Working Notes*, CEUR Workshop Proceedings, Bucharest, Romania. CEUR-WS.org.

Léo Butsanets, Charles Corbière, Julien Khlaut, Pierre Manceron, and Corentin Dancette. 2026. Radimagenet-vqa: A large-scale ct and mri dataset for radiologic visual question answering. *Preprint*, arXiv:2512.17396.

Souradip Chakraborty, Ekaba Bisong, Shweta Bhatt, Thomas Wagner, Riley Elliott, and Francesco Mosconi. 2020. BioMedBERT: A pre-trained biomedical language model for QA and IR. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 669–679, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shreyas Chaudhari, Pranjal Aggarwal, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, Karthik Narasimhan, Ameet Deshpande, and Bruno Castro da Silva. 2024. Rlhf deciphered: A critical analysis of reinforcement learning from human feedback for llms. *Preprint*, arXiv:2404.08555.

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

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

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *Preprint*, arXiv:2307.08691.

Ziqing Fan, Cheng Liang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Chestx-reasoner: Advancing radiology foundation models with reasoning through step-by-step verification. *Preprint*, arXiv:2504.20930.

Xiaotang Gai, Chenyi Zhou, Jiaxiang Liu, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. Medthink: Explaining medical visual question answering via multimodal decision-making rationale. *Preprint*, arXiv:2404.12372.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Jinlong He, Pengfei Li, Gang Liu, Genrong He, Zhaolin Chen, and Shenjun Zhong. 2025. Pefomed: Parameter efficient fine-tuning of multimodal large language models for medical imaging. *Preprint*, arXiv:2401.02797.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. 2025. Medvlthinker: Simple baselines for multimodal medical reasoning. *Preprint*, arXiv:2508.02669.

Minbyul Jeong, Hyeon Hwang, Chanwoong Yoon, Taewhoo Lee, and Jaewoo Kang. 2024. Olaph: Improving factuality in biomedical long-form question answering. *arXiv preprint arXiv:2405.12701*.

Hyunjae Kim, Hyeon Hwang, Jiwoo Lee, Sihyeon Park, Dain Kim, Taewhoo Lee, Chanwoong Yoon, Jiwoong Sohn, Jungwoo Park, Olga Reykhart, Thomas Fetherston, Donghee Choi, Soo Heon Kwak, Qingyu Chen, and Jaewoo Kang. 2025. Small language models learn enhanced reasoning skills from medical textbooks. *npj Digital Medicine*, 8(1):240.

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

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):180251.

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

Junyi Li and Hwee Tou Ng. 2025. Reasoning models hallucinate more: Factuality-aware reinforcement learning for large reasoning models. *Preprint*, arXiv:2505.24630.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. *Preprint*, arXiv:2102.09542.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. 2026. Gdpo: Group reward-decoupled normalization policy optimization for multi-reward rl optimization. *Preprint*, arXiv:2601.05242.

Yizhou Liu, Jingwei Wei, Zizhi Chen, Minghao Han, Xukun Zhang, Keliang Liu, and Lihua Zhang. 2025. Breaking reward collapse: Adaptive reinforcement for open-ended medical reasoning with enhanced semantic discrimination. *Preprint*, arXiv:2508.12957.

Linjie Mu, Yannian Gu, Zhongzhen Huang, Yakun Zhu, Shaoting Zhang, and Xiaofan Zhang. 2025. Medceg: Reinforcing verifiable medical reasoning with critical evidence graph. *Preprint*, arXiv:2512.13510.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Prasanna Parthasarathi, Mathieu Reymond, Boxing Chen, Yufei Cui, and Sarath Chandar. 2025. Grpo-λ: Credit assignment improves llm reasoning. *Preprint*, arXiv:2510.00194.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward

Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Preprint*, arXiv:1912.01703.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2025. Medgemma technical report. *Preprint*, arXiv:2507.05201.

Mehmet Saygin Seyfioglu, Wisdom O. Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. 2025. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. *Preprint*, arXiv:2312.04746.

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

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. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *Preprint*, arXiv:2504.07615.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, and 11 others. 2022. Large language models encode clinical knowledge. *Preprint*, arXiv:2212.13138.

Yanzhou Su, Tianbin Li, Jiyao Liu, Chenglong Ma, Junzhi Ning, Cheng Tang, Sibo Ju, Jin Ye, Pengcheng Chen, Ming Hu, Shixiang Tang, Lihao Liu, Bin Fu, Wenqi Shao, Xiaowei Hu, Xiangwen Liao, Yuanfeng Ji, and Junjun He. 2025. Gmai-vl-r1: Harnessing reinforcement learning for multimodal medical reasoning. *Preprint*, arXiv:2504.01886.

Haoran Sun, Yankai Jiang, Wenjie Lou, Yujie Zhang, Wenjie Li, Lilong Wang, Mianxin Liu, Lei Liu, and Xiaosong Wang. 2025. Chiron-o1: Igniting multimodal large language models towards generalizable medical reasoning via mentor-intern collaborative search. *Preprint*, arXiv:2506.16962.

Hongze Tan, Jianfei Pan, Jinghao Lin, Tao Chen, Zhihang Zheng, Zhihao Tang, and Haihua Yang. 2025. Gtpo and grpo-s: Token and sequence-level reward shaping with policy entropy. *Preprint*, arXiv:2508.04349.

LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. 2025a. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *Preprint*, arXiv:2506.07044.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 69 others. 2025b. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Guofu Xie, Yunsheng Shi, Hongtao Tian, Ting Yao, and Xiao Zhang. 2025a. Capo: Towards enhancing llm reasoning through generative credit assignment. *Preprint*, arXiv:2508.02298.

Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, and Yuyin Zhou. 2025b. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. *Preprint*, arXiv:2408.02900.

Qi Yang, Bolin Ni, Shiming Xiang, Han Hu, Houwen Peng, and Jie Jiang. 2025. R-4b: Incentivizing general-purpose auto-thinking capability in mllms via bi-mode annealing and reinforce learning. *Preprint*, arXiv:2508.21113.

Zongxian Yang, Jiayu Qian, Zegao Peng, Haoyu Zhang, Yu-An Huang, KC Tan, and Zhi-An Huang. 2026. Med-refl: Medical reasoning enhancement via self-corrected fine-grained reflection. *Preprint*, arXiv:2506.13793.

Jaehoon Yun, Jiwoong Sohn, Jungwoo Park, Hyunjae Kim, Xiangru Tang, Yanjun Shao, Yonghoe Koo, Minhyeok Ko, Qingyu Chen, Mark Gerstein, Michael Moor, and Jaewoo Kang. 2025. Med-prm: Medical reasoning models with stepwise, guideline-verified process rewards. *Preprint*, arXiv:2506.11474.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Wenchuan Zhang, Penghao Zhang, Jingru Guo, Tao Cheng, Jie Chen, Shuwan Zhang, Zhang Zhang, Yuhao Yi, and Hong Bu. 2025a. Patho-r1: A multimodal reinforcement learning-based pathology expert reasoner. *Preprint*, arXiv:2505.11404.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Pmc-vqa: Visual instruction tuning for medical visual question answering. *Preprint*, arXiv:2305.10415.

Yuting Zhang, Kaishen Yuan, Hao Lu, Yutao Yue, Jintai Chen, and Kaishun Wu. 2025b. Medtvt-r1: A multimodal llm empowering medical reasoning and diagnosis. *Preprint*, arXiv:2506.18512.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*.

Weihai Zhi, Jiayan Guo, and Shangyang Li. 2025. Medgr^2: Breaking the data barrier for medical reasoning via generative reward learning. *Preprint*, arXiv:2508.20549.

Shuang Zhou, Wenya Xie, Jiaxi Li, Zaifu Zhan, Meijia Song, Han Yang, Cheyenna Espinoza, Lindsay Welton, Xinnie Mai, Yanwei Jin, Zidu Xu, Yuen-Hei Chung, Yiyun Xing, Meng-Han Tsai, Emma Schaffer, Yucheng Shi, Ninghao Liu, Zirui Liu, and Rui Zhang. 2025. Automating expert-level medical reasoning evaluation of large language models. *Preprint*, arXiv:2507.07988.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

A Related Work

A.1 Reasoning in Medical Multimodal Large Language Models

Multimodal large language models (Team et al., 2025b; Yang et al., 2025; Zhu et al., 2025) have significantly advanced visual understanding through instruction-based learning, and this paradigm has been widely adapted to medical vision–language tasks (Singhal et al., 2022; Sellergren et al., 2025). Early works on medical MLLMs primarily relied on supervised fine-tuning (SFT) with carefully curated or synthetic medical data (Sun et al., 2025; Kim et al., 2025; Xie et al., 2025b). LLaVA-Med (Li et al., 2023) fine-tuned LLaVA (Liu et al., 2023) on PubMed Central data for biomedical image understanding.

Following the success of DeepSeek-R1 (Guo et al., 2025) in enhancing reasoning capabilities through GRPO (Shao et al., 2024), subsequent work has increasingly adopted this paradigm to adapt RL to the medical domain (Zhang et al., 2025a,b). Med-R1 (Lai et al., 2025), MedVLM-R1 (Pan et al., 2025), and GMAI-VL-R1 (Su et al., 2025) leverage GRPO-based training to enhance medical multimodal reasoning. However, most of these approaches compute advantages at the sequence level (Liu et al., 2025; Su et al., 2025), applying a uniform learning signal across all tokens that cannot distinguish which reasoning step led to an incorrect answer.

A.2 Process Supervision for Medical Reasoning

To address this, a growing body of work has moved toward evaluating individual reasoning steps in the medical domain, employing the resulting signal in various ways. Some approaches integrate step-level process rewards directly into RL training. ChestX-Reasoner (Fan et al., 2025) mines step-by-step reasoning from clinical reports and uses the resulting process rewards to guide a two-stage SFT-then-RL pipeline. MedGR^2 (Zhi et al., 2025) trains a generative reward model whose composite reward, combining reasoning quality and answer correctness, supervises GRPO. Others convert step-level assessments into offline preferences, as in Med-REFL (Yang et al., 2026), which distills tree-of-thoughts reflection values into the policy through direct preference optimization (DPO). Still others train a process reward model and apply it at inference time, as in Med-PRM (Yun et al., 2025), which verifies each step against retrieved clinical guidelines and selects traces.

The importance of step-level errors is thus widely recognized, and prior work (Yun et al., 2025; Zhou et al., 2025) has emphasized that identifying and correcting errors at specific reasoning steps is essential for reliable clinical decision making. However, none of these methods redistributes the learning signal according to where a failure occurs. In all these cases the resulting signal is summed into the reward and applied at the sequence level (Fan et al., 2025; Zhi et al., 2025), or distilled into trajectory-level preferences (Yang et al., 2026), and thus does not directly correct the step at which the failure occurs during training.

A.3 Step-wise Credit Assignment

The limitation above, namely that step-level signals are not redistributed according to where a failure occurs, has been studied more directly in general-domain reasoning. There, recent work explores token-level credit assignment (Tan et al., 2025; Parthasarathi et al., 2025; Xie et al., 2025a) to allocate learning signals more selectively across

Dataset	N	Human–LLM Alignment	
		Cohen's κ	Agreement Rate
VQA-RAD	100	0.745	88.0%
SLAKE	100	0.740	87.0%
PathVQA	100	0.617	85.0%
Total	300	0.717	86.7%

(a) Answer correctness alignment.

Criteria	N	Human–LLM Alignment	
		Cohen's κ	Agreement Rate
Gold Alignment	178	0.742	87.1%
Answer Contribution	178	0.680	83.9%
Reasoning Process Reward	178	0.712	85.5%

(b) Step-wise reasoning quality alignment.

Table 3: **Human–LLM alignment on reasoning quality.** We report alignment between human judgments and LLM-based evaluation for answer correctness and step-wise reasoning quality. For answer correctness, the overall Cohen's κ is 0.717 across 300 samples. For step-wise reasoning quality, both criteria achieve $\kappa > 0.68$, and the reasoning process reward achieves $\kappa = 0.712$, indicating substantial agreement with human judgments.

the reasoning trajectory. For example, FSPO (Li and Ng, 2025) uses step-wise factual verification to reward supported reasoning steps and penalize hallucinated ones. CAPO (Xie et al., 2025a) leverages an LLM to generate step-wise critiques for token-level rewards. MRPO combines this token-level credit assignment with the medical step-level evaluation. By assigning exponentially stronger penalties to earlier invalid steps, MRPO corrects faulty reasoning more effectively than prior medical process supervision and steers the model toward valid reasoning.

B Reliability of LLM-as-Judge Evaluation

B.1 Human–LLM Evaluator Alignment

To validate the reliability of our LLM-as-judge evaluation for **answer correctness** (Section 3 and 5.1) and **step-wise reasoning quality evaluation** (Sections 3, 4.1, and 5.3), we conduct a comparison study with a medical-student human evaluator.

We randomly sample 100 instances from the test sets of VQA-RAD, SLAKE, and PathVQA, and generate responses using Qwen2.5-VL-7B-Instruct. Both the LLM evaluator and the human evaluator independently assess the generated responses. Answer correctness is evaluated using the prompt in Appendix H.1 and step-wise reasoning quality using the prompt in Appendix H.2, all with GPT-5-mini as the LLM evaluator. The human evaluator is provided with the same criteria specified in each prompt.

Metrics. We measure inter-rater agreement between the two evaluators using **agreement rate** and **Cohen's** κ. Agreement rate represents the proportion of samples on which both evaluators made identical judgments. However, this metric can be misleading under class imbalance, as high agreement may occur by chance. To address this, we additionally report Cohen's κ, which adjusts for chance agreement. Following the guidelines of Landis and Koch (1977), a Cohen's κ above 0.61 indicates substantial agreement.

Answer Correctness Table 3a reports the alignment between the LLM evaluator and the human evaluator on the 300 selected instances, broken down by benchmark. While the agreement rate approaches 90% across all three benchmarks, we additionally report Cohen's κ to account for class imbalance. For VQA-RAD and SLAKE, Cohen's κ exceeds 0.7, indicating substantial alignment. For PathVQA, Cohen's κ is lower due to more severe class imbalance, where the proportion of incorrect answers is higher. Under such imbalance, Cohen's κ can be reduced even when agreement rate remains high. Additionally, PathVQA involves longer answers and higher question difficulty, making correctness judgments more challenging and resulting in lower inter-evaluator agreement. Finally, when aggregating results across all three benchmarks, we obtain Cohen's $\kappa = 0.717$, indicating strong alignment between the human evaluator and the LLM evaluator.

Step-wise Reasoning Quality For step-wise reasoning quality evaluation, we first segment each generated rationale into sentence-level steps and then score each step based on two metrics: Gold Alignment and Answer Contribution, which are derived from the process reward criteria introduced in Section 3, 4.1, and Section 5.3. Since each instance contains an average of 3.9 reasoning steps, evaluating all steps for the full set of 300 instances would be prohibitively costly. We therefore randomly sample 50 instances from the 300 and evaluate all

Method	Judge Model		
	GPT-5-mini	GPT-5.4	Claude-4.5-haiku
Qwen2.5-VL-7B-Instruct			
GRPO	26.06	24.37	24.96
MRPO	26.79	25.01	26.17
Qwen3-VL-8B-Instruct			
GRPO	28.69	26.91	28.05
MRPO	29.09	27.57	28.73
InternVL3-8B-Instruct			
GRPO	30.84	29.32	28.28
MRPO	31.94	30.69	29.15

Table 4: **Performance comparison under different judge models.** Average accuracy of GRPO and MRPO across three backbones, evaluated by GPT-5-mini, GPT-5.4, and Claude-4.5-haiku.

Method	Judge Model		
	GPT-5-mini	GPT-5.4	Claude-4.5-haiku
Early-Stage			
Baseline	64.0	52.3	54.8
GRPO	21.2	24.1	23.3
MRPO	13.0	19.4	16.7
Mid-Stage			
Baseline	26.4	33.5	30.4
GRPO	35.1	36.6	35.9
MRPO	39.9	37.7	37.0
Late-Stage			
Baseline	9.6	14.2	14.8
GRPO	43.6	39.3	40.8
MRPO	47.0	42.9	46.3

Table 5: **Sample distribution across First Failure Point (FFP) stages under different judge models.** Proportions of early, mid, and late-stage failures for the baseline, GRPO, and MRPO under different judge models.

reasoning steps within these instances, resulting in a total of 178 reasoning steps.

Table 3b reports the alignment between the LLM evaluator and the human evaluator on the 178 reasoning steps. Agreement rates exceed 80% for both metrics, and Cohen's κ exceeds 0.68 for both Gold Alignment and Answer Contribution. We further measure alignment on the **reasoning process reward** derived from these two metrics. As described in Section 4.1, this reward is computed binarily, assigning 1 when either metric receives a score of 1 and 0 otherwise. Both evaluators independently score the two metrics and compute the corresponding process reward. We obtain an agreement rate of 85.5% and Cohen's $\kappa = 0.712$, indicating substantial agreement between the two evaluators.

B.2 Cross-Judge Evaluation

Since MRPO employs GPT-5-mini both as the process reward judge during training and as the evaluator for answer correctness, a concern arises as to whether the observed gains in both answer accuracy and reasoning quality stem from evaluator-aligned overfitting to a single judge rather than genuine reasoning improvement. To address this, we repeat the evaluations from Section 5.2 and Section 5.3, replacing the judge with GPT-5.4 and Claude-4.5-haiku to re-assess answer accuracy and reasoning step evaluation across all three backbones and benchmarks.

Answer Accuracy. Table 4 reports answer accuracy under the three judges. Using each alternative judge, we re-evaluate answer accuracy on all benchmarks from Section 5.2, and report the average score across benchmarks in the table. While the absolute scores vary across judges, reflecting differences in their scoring strictness, the relative ordering between GRPO and MRPO is preserved in every case. Across all three backbones and all three judges, MRPO consistently outperforms GRPO without a single exception. This consistency indicates that the advantage of MRPO is not an artifact of the specific judge used during training, but rather a substantive improvement in answer quality that holds independently of the evaluating judge.

First Failure Point (FFP) Analysis. Under each judge, we compute the proportion of samples across FFP stages (Early, Mid, and Late-Stage), following Section 5.3.

Table 5 reports the resulting proportion of samples falling into each FFP stage for the baseline, GRPO, and MRPO under each judge. Although the exact magnitudes differ across judges, the same monotonic trend emerges consistently: moving from the baseline to GRPO and then to MRPO progressively reduces the share of early-stage failures while shifting failures toward later stages. Under all three judges, MRPO records the lowest early-stage failure proportion and the highest late-stage proportion, confirming that its effect of delaying the first failure point is preserved regardless of the judge. Taken together, these results demonstrate that MRPO's mitigation of cascading failures is a robust property of the method rather than a judge-specific phenomenon.

C Experimental Setup

C.1 Dataset

Training Dataset. Our training set is derived from the training splits of three medical VQA benchmarks, VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), and PathVQA (He et al., 2020).

- **VQA-RAD.** VQA-RAD is the first manually constructed dataset for radiology VQA, containing 315 radiology images from the MedPix database paired with 2,247 clinician-generated question-answer pairs. The images cover head CT and MRI, chest X-ray, and abdominal CT, with both open-ended and closed-ended questions.

- **SLAKE.** SLAKE is a bilingual English-Chinese medical VQA dataset comprising 642 radiology images including CT, MRI, and X-ray, paired with 14,028 question-answer pairs annotated by experienced physicians. It covers five body regions. We use only the English subset in our experiments.

- **PathVQA.** PathVQA is the first pathology VQA dataset, containing 4,289 pathology images and 32,632 question-answer pairs across eight categories. The dataset was constructed using a semi-automated pipeline from pathology textbooks and digital libraries, with all pairs manually verified. The majority of questions are open-ended.

From these sources, we exclude binary and multiple-choice questions to focus on open-ended instances, where answers are provided as free-form text. Gold reasoning annotations are obtained from MedThink (Gai et al., 2024), a rationale-augmented resource built on top of the same benchmarks. To ensure that every training example is paired with a gold reasoning annotation, we align our training instances to MedThink through exact matching on the image, question, and answer triple. This procedure yields a one-to-one mapping between the final training set and the gold reasoning annotations. After alignment, the resulting training set contains 13,375 open-ended QA instances over 3,556 unique images.

Evaluation Dataset. For evaluation, we use two types of benchmarks: **in-distribution benchmarks**, consisting of the test splits of our training datasets, and five **out-of-distribution benchmarks**. For the in-distribution benchmarks, we filter the test splits of the training datasets, VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), and PathVQA (He et al., 2020), to open-ended QA only, yielding 200, 706, and 3,357 samples respectively, for a total of 4,263 samples. For the out-of-distribution evaluation, we additionally adopt five benchmarks, PMC-VQA (Zhang et al., 2024), VQA-Med-2021 (Ben Abacha et al., 2021), Quilt-VQA (Seyfioglu et al., 2025), RadImageNet-VQA (Butsanets et al., 2026), and MIMIC-Ext-MIMIC-CXR-VQA (Bae et al., 2024), that span diverse imaging modalities unseen during training.

- **PMC-VQA.** PMC-VQA is a large-scale medical VQA dataset comprising 227K question–answer pairs over 149K images, automatically constructed from figure–caption pairs in PubMed Central articles via a scalable pipeline. The dataset spans diverse imaging modalities and diseases, and although it provides multiple-choice options for every question, we evaluate in an open-ended setting in our experiments. Specifically, we use the 2,000-sample manually-verified clean test set.

- **VQA-Med-2021.** VQA-Med-2021 is a radiology VQA benchmark released as part of the ImageCLEF 2021 challenge, with a pronounced focus on questions about abnormalities in radiology images. Its test set consists of 500 radiology images paired with 500 abnormality questions, and the reference answers of the test set were manually validated by a medical doctor to ensure data quality. Among these, we use only the 425 QA pairs that have a single ground-truth answer.

- **Quilt-VQA.** Quilt-VQA is a histopathology VQA benchmark comprising 985 images paired with 1,283 human-generated question–answer pairs, extracted from educational histopathology videos on YouTube. The benchmark covers a wide range of diagnostic concepts and supports both open-ended and closed-ended questions for evaluating pathology-focused multimodal models. As Quilt-VQA is released as an evaluation set, in our experiments we use only the 724 open-ended QA pairs.

- **RadImageNet-VQA.** RadImageNet-VQA is a large-scale CT and MRI VQA dataset built

from expert-curated annotations, providing 750K images paired with 7.5M question–answer samples. It covers three diagnostic tasks of abnormality detection, anatomy recognition, and pathology identification across 8 anatomical regions and 97 pathology categories, and supports open-ended, closed-ended, and multiple-choice questions. For evaluation, the authors provide a stratified benchmark of 1,000 images with 9,000 QA pairs. From this benchmark, we use only the 2,000 open-ended questions in our experiments.

- **MIMIC-Ext-MIMIC-CXR-VQA.** MIMIC-Ext-MIMIC-CXR-VQA is a complex and large-scale chest radiograph VQA dataset comprising approximately 377K question-answer pairs derived from the MIMIC-CXR database. The dataset is constructed using 48 question templates that involve set and logical operations, designed to evaluate diverse aspects of chest X-ray interpretation. From the test set, we filter to open-ended questions of the *query* semantic type, and we further restrict these to the 1,724 test samples that have a single ground-truth answer.

C.2 Implementation Detail

All experiments are conducted with PyTorch (Paszke et al., 2019) on 8×NVIDIA A100 GPUs, and we adopt FlashAttention-2 (Dao, 2023) to improve training efficiency. Our implementation builds on VLM-R1 (Shen et al., 2025), an open-source GRPO framework for VLMs.

We trained MRPO on three backbones, Qwen3-VL-8B-Instruct, Qwen2.5-VL-7B-Instruct, and InternVL3-8B-Instruct. To quantify the effectiveness of MRPO and compare it with other methods, all RL methods including MRPO, GRPO, and GDPO are trained on the same training dataset under identical settings. Specifically, each method is trained for 1 epoch with a batch size of 64 and a learning rate of 10^{-6}, sampling 8 rollouts per prompt. Following the standard GRPO configuration (Shao et al., 2024), we set the KL coefficient $\beta = 0.04$ and the clipping range $\epsilon = 0.2$ for all RL methods.

For SFT, we train the model on the same training dataset augmented with gold reasoning annotations, so that the model jointly learns to produce the reasoning trace and the final answer. We employ LoRA (Hu et al., 2021) with rank 8, alpha 32, and dropout 0.05, using a learning rate of 2×10^{-5} for 3 epochs.

Metric	GRPO	GDPO	MRPO
Training Time	110h 25min	113h 39min	120h 54min
Input Tokens	234.3M	245.0M	273.3M
Output Tokens	73.0M	72.8M	78.7M
Total Cost	$192.96	$201.74	$215.48

Table 6: **Training resource comparison across RL methods.** We compare training time, token usage, and total cost for GRPO, GDPO, and MRPO.

C.3 Training Cost and Efficiency

We report a detailed comparison of training time, token usage, and total cost across RL methods on Qwen2.5-VL-7B-Instruct in Table 6. MRPO, GRPO, and GDPO all issue only a single API call per rollout to jointly evaluate all reasoning sentences. With 13K samples and 8 rollouts, this amounts to roughly 107K calls per epoch across all configurations.

As shown in Table 6, the total cost ranges from $192.96 (GRPO) to $215.48 (MRPO), with GDPO falling in between at a comparable level. MRPO incurs only a marginal increase over GRPO (≈12%), which arises not from additional judge queries but from MRPO generating longer, more structured reasoning traces that consume more tokens. This indicates that step-wise process supervision is practically affordable at this scale. For larger-scale training, the results in Appendix D.4 show that while MRPO maintains consistent gains over GRPO even with open-source judges, their absolute performance still falls short of API-based judges, as the quality of step-level evaluation governs the magnitude of RL gains. This motivates the development of dedicated local process reward models that match frontier-judge quality. We leave this as future work, which would enable both stronger RL-driven gains and the complete elimination of API dependency.

D Ablation Study

To better understand the contribution of each design choice in MRPO, we conduct four ablation studies on two backbones, Qwen2.5-VL-7B-Instruct and Qwen3-VL-8B-Instruct. Specifically, we examine the SFT cold-start initialization prior to RL training, the token-level shaping function for advantage reshaping, the advantage reweighting strategy, and

Method	In-Distribution				Out-of-Distribution					
	VQA-RAD	SLAKE	PathVQA	AVG	PMC-VQA	VQA-Med	Quilt-VQA	Rad-VQA	MIMIC-VQA	AVG
Qwen2.5-VL-7B-Instruct										
Base	42.00	58.29	17.10	25.10	28.35	5.18	19.06	27.20	15.08	22.28
SFT	38.50	63.88	18.83	26.30	28.50	4.71	18.37	26.20	15.02	21.91
GRPO	**44.00**	65.27	20.52	29.06	31.05	3.76	20.99	29.30	16.71	24.20
MRPO	41.50	65.89	21.30	29.63	30.85	5.65	**22.79**	**29.65**	**18.65**	**25.04**
SFT + GRPO	42.00	66.82	22.34	29.67	30.40	**6.82**	21.13	26.65	13.81	22.71
SFT + MRPO	42.50	**68.68**	**23.44**	**30.85**	**31.15**	6.59	22.24	26.90	14.79	23.35
Qwen3-VL-8B-Instruct										
Base	43.00	59.69	21.48	28.81	31.00	9.41	23.90	33.15	15.31	25.61
SFT	**45.00**	66.29	21.27	29.84	30.35	7.76	19.48	33.40	15.20	24.89
GRPO	**45.00**	67.14	20.11	29.06	30.25	9.65	**24.03**	**40.60**	**18.79**	28.46
MRPO	41.50	68.27	20.43	29.35	**33.00**	**12.00**	23.90	40.20	17.46	**28.94**
SFT + GRPO	40.00	67.42	25.95	33.47	28.25	6.59	19.75	35.20	15.43	24.82
SFT + MRPO	40.50	**68.41**	**26.57**	**34.15**	29.20	7.53	19.48	35.75	14.50	25.05

Table 7: **Ablation on SFT cold-start initialization.** We compare six configurations across two backbones, Qwen2.5-VL-7B-Instruct and Qwen3-VL-8B-Instruct, including the base model, SFT, GRPO, MRPO, SFT followed by GRPO, and SFT followed by MRPO. VQA-Med denotes VQA-Med-2021, Rad-VQA denotes RadImageNet-VQA, and MIMIC-VQA denotes MIMIC-Ext-MIMIC-CXR-VQA.

the choice of process reward model. All ablations are conducted on the same training dataset and evaluated on three in-distribution and five out-of-distribution medical VQA benchmarks.

D.1 SFT Cold-Start Initialization

Table 7 reports the effect of initializing the model with supervised fine-tuning prior to RL training. We compare six configurations against the base model, including SFT alone, GRPO alone, MRPO alone, SFT followed by GRPO, and SFT followed by MRPO. The cold-start SFT is trained under the same setting as SFT alone, with training details provided in Appendix C.2.

A consistent pattern emerges across both backbones. SFT cold-start substantially improves in-distribution performance but degrades out-of-distribution performance. For Qwen3-VL-8B-Instruct, combining SFT with MRPO further improves PathVQA from 20.43 to 26.57 and SLAKE from 68.27 to 68.41, achieving the highest in-distribution average among all configurations. However, the same configuration drops the out-of-distribution average from 28.94 to 25.05, falling below MRPO without SFT cold-start. The same pattern holds for Qwen2.5-VL-7B-Instruct, where SFT+MRPO yields gains on PathVQA from 21.30 to 23.44 and SLAKE from 65.89 to 68.68 in the in-distribution setting, but underperforms MRPO alone on out-of-distribution benchmarks.

This trade-off suggests that SFT on gold reasoning annotations overfits the model to the specific reasoning patterns of the training distribution, which improves performance on the matched in-distribution test sets but weakens the transferable reasoning capability that RL alone is able to induce. The effect is particularly pronounced on PathVQA, where SFT cold-start brings the largest gains, since the gold reasoning annotations from MedThink most directly align with the PathVQA test distribution. In contrast, the out-of-distribution benchmarks such as PMC-VQA, VQA-Med, and Quilt-VQA show consistent degradation, indicating that the reasoning patterns learned through SFT do not generalize to unseen imaging modalities and question styles. Based on this finding, we adopt direct RL training without SFT cold-start in our main experiments, prioritizing the generalization capability that better reflects the demands of real-world clinical deployment, where models routinely encounter queries beyond the training distribution.

D.2 Token-Level Shaping Function

Table 8 reports the effect of different token-level shaping functions in MRPO's advantage reshaping. We compare four functions for assigning penalties to tokens in failed reasoning steps: uniform, linear, quadratic, and exponential. For a failed reasoning step at relative position $\frac{k-1}{K-1}$, where k denotes the step index and K denotes the total number of steps, the penalty multipliers are defined as -1 for uni-

Shaping Function	In-Distribution			Out-of-Distribution					AVG
	VQA-RAD	SLAKE	PathVQA	PMC-VQA	VQA-Med	Quilt-VQA	Rad-VQA	MIMIC-VQA	
Qwen2.5-VL-7B-Instruct									
Uniform	**43.00**	61.47	18.17	29.20	4.47	20.72	27.10	15.43	24.16
Linear	42.50	62.46	19.66	29.95	5.41	19.89	28.70	16.13	25.18
Quadratic	40.50	64.45	18.59	29.25	**5.88**	20.17	28.35	17.81	25.05
Exponential (MRPO)	41.50	**65.89**	**21.30**	**30.85**	5.65	**22.79**	**29.65**	**18.65**	**26.79**
Qwen3-VL-8B-Instruct									
Uniform	36.00	61.33	19.24	28.60	8.00	20.17	33.35	13.63	25.19
Linear	**41.50**	62.75	19.72	30.30	7.53	21.55	**40.55**	16.94	27.70
Quadratic	40.50	62.61	19.48	31.50	7.76	22.10	39.80	14.50	27.35
Exponential (MRPO)	**41.50**	**68.27**	**20.43**	**33.00**	**12.00**	**23.90**	40.20	**17.46**	**29.09**

Table 8: **Ablation study on token shaping strategies.** We compare four token-level shaping functions, namely uniform, linear, quadratic, and exponential, on two backbones, Qwen2.5-VL-7B-Instruct and Qwen3-VL-8B-Instruct. VQA-Med denotes VQA-Med-2021, Rad-VQA denotes RadImageNet-VQA, and MIMIC-VQA denotes MIMIC-Ext-MIMIC-CXR-VQA. The exponential function corresponds to the proposed shaping strategy in MRPO.

Reweighting Strategy	In-Distribution			Out-of-Distribution					AVG
	VQA-RAD	SLAKE	PathVQA	PMC-VQA	VQA-Med	Quilt-VQA	Rad-VQA	MIMIC-VQA	
Qwen2.5-VL-7B-Instruct									
Soft Reweighting	**45.00**	**66.71**	19.63	28.50	**6.35**	21.13	28.55	14.10	25.00
Full Reweighting	35.50	64.02	20.67	29.65	4.71	19.20	**30.40**	15.55	25.55
Selective Reweighting (MRPO)	41.50	65.89	**21.30**	**30.85**	5.65	**22.79**	29.65	**18.65**	**26.79**
Qwen3-VL-8B-Instruct									
Soft Reweighting	**45.00**	64.45	**20.88**	30.20	6.59	22.38	36.95	14.62	27.29
Full Reweighting	41.50	64.45	20.52	30.15	8.00	**25.83**	38.80	11.95	27.34
Selective Reweighting (MRPO)	41.50	**68.27**	20.43	**33.00**	**12.00**	23.90	**40.20**	**17.46**	**29.09**

Table 9: **Ablation study on advantage reweighting strategies.** We compare three strategies, namely Soft Reweighting, Full Reweighting, and Selective Reweighting, on two backbones, Qwen2.5-VL-7B-Instruct and Qwen3-VL-8B-Instruct. VQA-Med denotes VQA-Med-2021, Rad-VQA denotes RadImageNet-VQA, and MIMIC-VQA denotes MIMIC-Ext-MIMIC-CXR-VQA. The selective reweighting strategy corresponds to the proposed method in MRPO, which applies advantage reshaping only to instances with incorrect final answers.

form, $-\left(1-\frac{k-1}{K-1}\right)$ for linear, $-\left(1-\frac{k-1}{K-1}\right)^2$ for quadratic, and $-\exp\left(1-\frac{k-1}{K-1}\right)$ for exponential. Each multiplier is applied to $|A_i|$.

Uniform applies the same penalty regardless of position, while the other three apply progressively stronger penalties to tokens in earlier failed steps, with exponential providing the steepest decay from early to late positions. A clear pattern emerges across both backbones. All position-aware shaping functions substantially outperform the uniform baseline, confirming that penalizing early failures more strongly than later ones is beneficial. Among them, exponential shaping yields the largest improvement, gaining 2.63 points on Qwen2.5-VL-7B-Instruct and 3.90 points on Qwen3-VL-8B-Instruct over the uniform baseline, substantially exceeding the gains achieved by linear shaping at 1.02 and 2.51 points and quadratic shaping at 0.89 and 2.16 points. Exponential shaping also achieves the highest average score on both backbones, reaching 26.79 on Qwen2.5-VL-7B-Instruct and 29.09 on Qwen3-VL-8B-Instruct. Based on this finding, we adopt exponential shaping as the default token-level function in MRPO, as it provides the strongest signal for correcting the root causes of reasoning failures.

D.3 Advantage Reweighting Strategy

Table 9 reports the effect of different advantage reweighting strategies. We compare three strategies, namely Soft Reweighting, Full Reweighting, and the proposed selective strategy in MRPO. Full Reweighting applies the step-wise advantage reshaping to all training instances regardless of answer correctness, Soft Reweighting applies the same reshaping to all instances but scaled by a factor of 0.5, and MRPO applies the reshaping only

RL method	PRM	In-Distribution			Out-of-Distribution					AVG
		VQA-RAD	SLAKE	PathVQA	PMC-VQA	VQA-Med	Quilt-VQA	Rad-VQA	MIMIC-VQA	
Qwen2.5-VL-7B-Instruct										
- (Base Model)	-	42.00	58.29	17.10	28.35	5.18	19.06	27.20	15.08	23.26
GRPO	MedGemma	41.00	64.73	18.41	30.30	**6.35**	20.30	28.85	17.34	25.26
GRPO	Med-PRM	40.50	65.01	17.27	29.80	6.12	16.99	26.15	14.21	23.64
GRPO	GPT-5-mini	**44.00**	65.27	20.52	**31.05**	3.76	20.99	29.30	16.71	26.06
MRPO	MedGemma	42.00	**66.29**	19.06	30.60	4.94	20.99	28.50	16.94	25.49
MRPO	Med-PRM	39.50	62.32	16.98	28.60	**6.35**	18.23	25.85	14.04	23.16
MRPO	GPT-5-mini	41.50	65.89	**21.30**	30.85	5.65	**22.79**	**29.65**	**18.65**	**26.79**
Qwen3-VL-8B-Instruct										
- (Base Model)	-	43.00	59.69	**21.48**	31.00	9.41	23.90	33.15	15.31	26.83
GRPO	MedGemma	**45.00**	65.86	19.42	31.20	11.76	21.69	38.45	**18.97**	28.14
GRPO	Med-PRM	43.00	64.87	12.98	27.25	7.53	15.19	32.00	18.62	23.59
GRPO	GPT-5-mini	**45.00**	67.14	20.11	30.25	9.65	**24.03**	**40.60**	18.79	28.69
MRPO	MedGemma	44.00	66.71	19.54	31.05	11.76	22.38	38.65	18.91	28.26
MRPO	Med-PRM	40.00	66.29	13.40	27.00	5.18	16.99	33.25	21.00	24.34
MRPO	GPT-5-mini	41.50	**68.27**	20.43	**33.00**	**12.00**	23.90	40.20	17.46	**29.09**

Table 10: **Ablation study on process reward models.** We compare three process reward models, namely MedGemma-27B, Med-PRM, and GPT-5-mini, under both GRPO and MRPO training on two backbones, Qwen2.5-VL-7B-Instruct and Qwen3-VL-8B-Instruct. VQA-Med denotes VQA-Med-2021, Rad-VQA denotes RadImageNet-VQA, and MIMIC-VQA denotes MIMIC-Ext-MIMIC-CXR-VQA. GPT-5-mini corresponds to the default process reward model used in MRPO.

when the answer reward falls below the threshold τ, leaving the advantages of correctly answered trajectories unchanged.

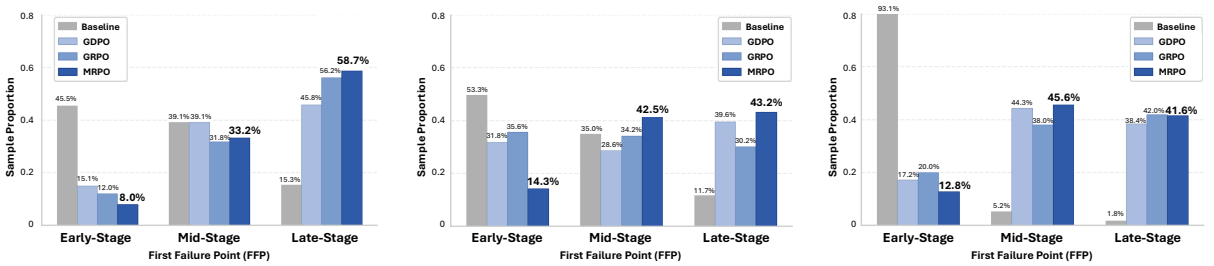
On both backbones, MRPO consistently outperforms both Full and Soft Reweighting. On Qwen3-VL-8B-Instruct, MRPO achieves an average of 29.09, surpassing Full Reweighting at 27.34 and Soft Reweighting at 27.29 by 1.75 and 1.80 points respectively. The same pattern holds for Qwen2.5-VL-7B-Instruct, where MRPO reaches 26.79 compared to 25.55 and 25.00 for the two alternatives. These results indicate that applying advantage reshaping indiscriminately to all training instances harms performance by disrupting reasoning trajectories that already lead to correct answers. Restricting reshaping to failed predictions allows the model to retain the reasoning patterns that already work while concentrating the learning signal on the root causes of failure. Based on this finding, we adopt selective reweighting in MRPO, applying advantage reshaping only to instances where the final answer is judged incorrect.

D.4 Process Reward Model

Table 10 reports the effect of different process reward models used for computing the step-wise reasoning reward R_{proc}. We compare three alternatives under both GRPO and MRPO training. MedGemma-27B is a strong open-source medical multimodal LLM repurposed as a local judge, Med-PRM is a process reward model trained specifically for medical reasoning evaluation that takes only the question and reasoning text as input without the image, and GPT-5-mini is a frontier general-purpose LLM accessed via API.

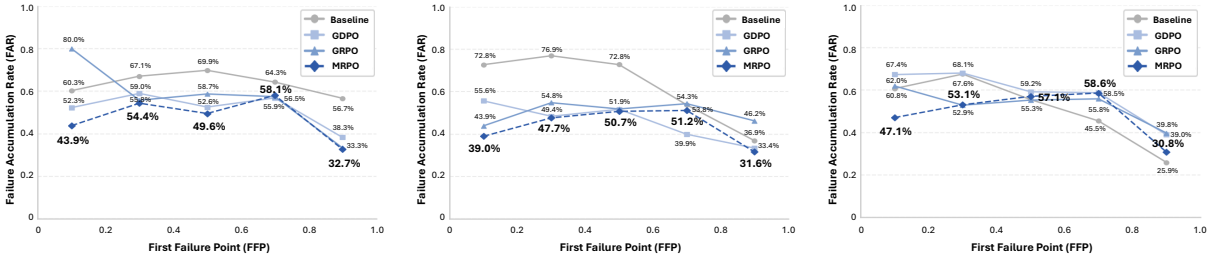
Two clear patterns emerge from the results. First, MRPO outperforms GRPO in most configurations, indicating that step-wise advantage reshaping benefits across a wide range of process reward models. On Qwen3-VL-8B-Instruct, MRPO outperforms GRPO with all three PRMs, achieving improvements of 0.12, 0.75, and 0.40 points with MedGemma, Med-PRM, and GPT-5-mini, respectively. The same trend holds on Qwen2.5-VL-7B-Instruct with MedGemma and GPT-5-mini, while only with the weakest judge, Med-PRM, does MRPO underperform GRPO by a small margin. This pattern suggests that the core benefit of MRPO, the step-wise reshaping mechanism, is largely robust to the choice of judge model, as long as the judge provides sufficiently reliable step-level signals.

Second, GPT-5-mini achieves the highest average across both backbones and both training paradigms, which is why we adopt it as the de-



(A) Qwen2.5-VL-7B-Instruct (B) Qwen3-VL-8B-Instruct (C) InternVL3-8B-Instruct

Figure 5: **Sample distribution across First Failure Point (FFP) stages for each backbone.** Proportions of samples falling into Early, Mid, and Late-Stage FFP ranges for the baseline, GRPO, GDPO, and MRPO on Qwen2.5-VL-7B-Instruct, Qwen3-VL-8B-Instruct, and InternVL3-8B-Instruct.



(A) Qwen2.5-VL-7B-Instruct (B) Qwen3-VL-8B-Instruct (C) InternVL3-8B-Instruct

Figure 6: **Failure Accumulation Rate (FAR) across FFP bins for each backbone.** FAR across First Failure Point (FFP) bins for the baseline, GRPO, GDPO, and MRPO on Qwen2.5-VL-7B-Instruct, Qwen3-VL-8B-Instruct, and InternVL3-8B-Instruct.

fault process reward model in MRPO: it provides the highest absolute performance ceiling, reaching the best average on every backbone. On Qwen2.5-VL-7B-Instruct, MRPO with GPT-5-mini reaches 26.79, surpassing MedGemma at 25.49 and Med-PRM at 23.16. On Qwen3-VL-8B-Instruct, MRPO with GPT-5-mini reaches 29.09, surpassing MedGemma at 28.26 and Med-PRM at 24.34. We attribute this to the difference in step-level evaluation quality. GPT-5-mini is a stronger evaluator of reasoning steps and aligns well with human judgments, achieving substantial agreement on step-wise reasoning quality with Cohen's κ above 0.7, as shown in our human-LLM alignment study in Appendix B.1. In contrast, MedGemma-27B is fundamentally weaker than GPT-5-mini as a judge, and Med-PRM receives only the question and reasoning text without the image, so it cannot properly assess visual reasoning and consequently fails to accurately identify which reasoning steps are invalid. Since the effectiveness of MRPO's step-wise advantage reshaping fundamentally depends on accurately identifying which reasoning steps are invalid, a judge with stronger reasoning-evaluation capability translates directly into stronger downstream RL performance.

Based on these findings, we adopt GPT-5-mini as the default process reward model in MRPO. We note that MedGemma-27B yields competitive results that approach those of GPT-5-mini, making it a viable alternative when API access is constrained, and we leave the development of dedicated medical VQA process reward models that match frontier-LLM judgment quality without external API dependency as a promising direction for future work.

E Reasoning Analysis

E.1 Reasoning Analysis Across Backbones

To examine the results of Section 5.3 in greater detail, we decompose the First Failure Point (FFP) analysis by backbone. As shown in Figure 5, across all three backbones, the most consistent finding is the suppression of early-stage, cascade-inducing failures. The baseline concentrates failures in the early stage, and while all RL methods reduce this, MRPO lowers it the most on every backbone, reaching 8.0% on Qwen2.5-VL-7B-Instruct, 14.3% on Qwen3-VL-8B-Instruct, and 12.8% on InternVL3-8B-Instruct, the lowest among all methods and below both GRPO and GDPO in each case. The removed early failures are correspondingly redistributed to later stages, most prominently as a clean early-to-late shift on Qwen2.5-VL-7B-

Instruct, where the late stage rises from 15.3% to 58.7%. Overall, the directional pattern of fewer early-stage failures and more late-stage failures holds consistently across three backbones from different model families, demonstrating that MRPO's distinctive contribution, the suppression of early cascade failures, generalizes across architectures.

Figure 6 reports the Failure Accumulation Rate (FAR) across FFP bins for each backbone, where MRPO attains the lowest FAR in the earliest bin, FFP 0.0–0.2, on all three backbones, reaching 43.9%, 39.0%, and 47.1% for Qwen2.5-VL-7B-Instruct, Qwen3-VL-8B-Instruct, and InternVL3-8B-Instruct, below every competing method and well under the baseline. In other words, even when an early failure does occur, MRPO is the most effective at preventing it from cascading through the remaining steps, indicating improved recovery from early errors. This advantage is concentrated in the early bins, consistent with MRPO's design: because the exponential penalty targets the earliest failed steps, its strongest effect on failure accumulation emerges precisely where the penalty is applied. Together with the suppression of early-stage failures above, this shows that MRPO addresses cascading failures along two complementary and architecture-agnostic axes, delaying the onset of failures and improving recovery once they begin.

E.2 Paired Comparison of GRPO and MRPO

Since MRPO is derived directly from GRPO by reshaping step-wise advantages, we conduct a paired, instance-level comparison between the two methods to isolate the effect of this reshaping on individual predictions. We pool the reasoning traces generated on the test splits of the three in-distribution benchmarks (VQA-RAD, SLAKE, and PathVQA) across all three backbones (Qwen2.5-VL-7B-Instruct, Qwen3-VL-8B-Instruct, and InternVL3-8B-Instruct), and evaluate each instance under both methods. Table 11 summarizes the resulting prediction consistency. The two methods agree on the large majority of instances, with 25.2% jointly correct and 62.0% jointly incorrect. The discordant cases, where exactly one method succeeds, are of primary interest: MRPO is uniquely correct on 6.8% of instances while GRPO is uniquely correct on 6.0%. We analyze these two disjoint groups separately. For the MRPO-only-correct group, we ask how MRPO recovers instances that GRPO fails, by examining where reasoning failures occurred under GRPO and how they are redistributed under MRPO. For the GRPO-only-correct group, we conversely characterize the nature of the failures that MRPO newly introduces. Together, these two analyses reveal not only the net change in accuracy but the underlying shift in reasoning behavior that GRPO and MRPO induce on the same inputs.

GRPO \ MRPO	Correct	Wrong
Correct	3220 (25.2%)	768 (6.0%)
Wrong	875 (6.8%)	7926 (62.0%)

Table 11: **Prediction consistency between GRPO and MRPO.** Rows indicate GRPO predictions and columns indicate MRPO predictions.

GRPO \ MRPO	No FFP	Early	Mid	Late	SUM
No FFP	209	8	17	46	280
Early	47 (30.5%)	11 (7.1%)	27 (17.5%)	69 (44.8%)	154
Mid	91	24	33	35	183
Late	138	17	42	61	258
SUM	485	60	119	211	875

Table 12: **Transition matrix of First Failure Point (FFP) stages between GRPO and MRPO for MRPO-only-correct instances.** Rows indicate the GRPO FFP stage and columns the MRPO FFP stage. The GRPO Early-stage row is highlighted, with percentages computed within that row.

To characterize how MRPO recovers instances that GRPO answers incorrectly, we track, for each of the MRPO-only-correct instances, the First Failure Point (FFP) stage of its reasoning trace under both methods. Table 12 reports the resulting transition matrix, where rows index the GRPO FFP stage and columns index the MRPO FFP stage. The clearest signal is the reduction of early-stage and mid-stage failures. Under GRPO, these instances exhibit a substantial concentration of failures in the early and mid stages, consistent with the cascading-failure pattern that drives incorrect predictions. Under MRPO, both are markedly reduced, with early-stage failures falling from 154 to 60 and mid-stage failures from 183 to 119, indicating that step-wise advantage reshaping effectively corrects the early reasoning errors that GRPO fails to resolve. The redistribution is most pronounced for the early-stage GRPO row, highlighted in Table 12: of the 154 instances that fail early under GRPO, only 7.1% remain in the early stage under

GRPO \ MRPO	No FFP	Early	Mid	Late	SUM
No FFP	252	25	66	105	448
Early	11	13	6	31	61
Mid	35	18	57	13	123
Late	66	11	16	43	136
SUM	364 (47.4%)	67 (8.7%)	145 (18.9%)	192 (25.0%)	768

Table 13: **Transition matrix of First Failure Point (FFP) stages between GRPO and MRPO for GRPO-only-correct instances.** Rows indicate the GRPO FFP stage and columns the MRPO FFP stage. The MRPO SUM row is highlighted, with percentages computed over the total number of samples.

Type	Transition (GRPO → MRPO)	Count
Removed (from MRPO-only-correct)	Early → No-FFP	47
	Early → Mid	27
	Early → Late	69
	Subtotal (early failures removed)	**143**
Introduced (from GRPO-only-correct)	No-FFP → Early	25
	Mid → Early	18
	Late → Early	11
	Subtotal (early failures introduced)	**54**
Net reduction in early-stage failures		**89**

Table 14: **Net change in early-stage cascade failures induced by MRPO.** MRPO removes 143 early-stage failures while introducing only 54, a net reduction of 89.

MRPO, while the remaining 92.9% move out of it. Specifically, 30.5% are resolved without a detectable failure point, 17.5% shift to the mid stage, and 44.8% shift to the late stage. This indicates that MRPO rarely leaves an early failure in place; instead, it either eliminates the initial failure or pushes its onset to a later stage, where the residual error is far less likely to cascade. We further note that a portion of the MRPO-only-correct instances corresponds to GRPO traces with no detected failure point, the No-FFP row. Our analysis suggests that these are cases where the reasoning trajectory is essentially valid but the answer is marked incorrect due to surface-level expression mismatches rather than a genuine reasoning error, and thus fall outside the step-level reasoning behavior analyzed here.

Table 13 reports the symmetric analysis on the GRPO-only-correct group, characterizing the failures MRPO newly introduces. Since GRPO answers these instances correctly, a failure point necessarily emerges under MRPO, so the relevant question is not whether MRPO fails but at which stage. As shown in the highlighted SUM row reporting MRPO's stage-wise distribution, MRPO's failures concentrate in the late stage at 25.0%, exceeding both the mid stage at 18.9% and the early stage at 8.7%, while a large fraction of traces remain free of any detected failure point at 47.4%. In other words, when MRPO answers an instance incorrectly, the failure tends to occur late in the trajectory rather than an early failure that triggers a cascade.

Beyond being rare, these newly introduced early-stage failures must also be weighed against those MRPO removes. Table 14 makes this comparison explicit using the two transition matrices above. The *Removed* column counts, among MRPO-only-correct instances, cases where MRPO turns a GRPO early-stage failure into a failure at another stage or a trace with no detected failure point. The *Introduced* column conversely counts, among GRPO-only-correct instances, cases where MRPO turns a failure at another stage or a No-FFP trace into an early-stage failure. MRPO removes 143 early-stage failures while introducing only 54, a net reduction of 89, roughly $2.6\times$ more removed than introduced. This confirms that MRPO suppresses far more early-stage failures than it creates.

F Qualitative Analysis

F.1 Stage-wise failure taxonomy

We conduct reasoning evaluation on the test sets of VQA-RAD, SLAKE, and PathVQA across three backbones (Qwen2.5-VL-7B-Instruct, Qwen3-VL-8B-Instruct, and InternVL3-8B-Instruct). We randomly sample 30 failed traces from each First Failure Point (FFP) bin of width 0.1, yielding 300 traces in total, and the authors manually examine the first invalid reasoning step of each trace. Analyzing these samples, we find that traces with similar First Failure Point positions exhibit similar failure patterns and error types, allowing the failures to be grouped into three contiguous ranges, 0.0-0.4, 0.4-0.7, and 0.7-1.0, within which the observed error types are largely homogeneous. We refer to these three ranges as the early, mid, and late stages, respectively. The resulting taxonomy reveals that

failure types are clearly stratified by stage, and that their position determines how severely they affect the final answer: early-stage failures correspond to errors in establishing the visual premise and mostly trigger a cascade that corrupts the entire downstream trajectory and leads to an incorrect answer; mid-stage failures arise during the interpretation of correctly perceived content and propagate only partially, often still permitting recovery in later steps; and late-stage failures are confined to terminological expression and rarely affect the validity of the preceding reasoning. We describe the failure types and failure patterns observed at each stage below.

Early-stage errors. Early-stage failures occur when the model fails to establish a correct visual premise at the outset of reasoning, corrupting every subsequent step that builds upon it. We identify two types. **(1) Default Staining/Modality Assumption (Figure 7):** rather than reading the actual image, the model begins reasoning by defaulting to the most common modality (e.g., H&E staining), grounding the entire trace in an unverified assumption. **(2) Wrong Organ/Structure Identification (Figure 8):** the model correctly recognizes the image type but, in its first sentence, confidently specifies an incorrect organ or anatomical structure.

Mid-stage errors. Mid-stage failures arise after the visual premise is established, when the model engages with the image but makes an error in interpretation or diagnostic judgment. Because the initial premise is correct, these errors typically corrupt only the steps that depend on the specific misjudgment rather than the entire trajectory, so later steps may still recover and reach the correct answer. We identify two types. **(1) Structural Misidentification (Figure 9):** the model misrecognizes a microstructure as a different organ or structure. **(2) Pathology Omission (Figure 10):** despite the presence of a lesion in the image, the model describes only normal findings.

Late-stage errors. Late-stage failures occur after the model has correctly perceived the image and carried out largely valid reasoning, with the error confined to the final step. Because the preceding reasoning remains intact, these failures do not propagate and rarely induce a cascade; they instead reflect a local breakdown at the point of producing the final answer. We identify two types. **(1) Non-committal Terminal Conclusion (Figure 11):** although the reasoning proceeds correctly until the late steps, the model fails to converge on a specific answer and instead hedges with vague expressions such as "consistent with" or "possibly", avoiding a definitive conclusion. **(2) Terminal Label/Term Mismatch (Figure 12):** having correctly identified the relevant structure or finding, the model mismaps it at the final step to an incorrect name, laterality (left/right), or specific term.

F.2 Qualitative Case Studies

To complement the quantitative analyses, we present qualitative comparisons of GRPO and MRPO reasoning traces on identical inputs. Each case pairs the two methods on the same question with every step annotated as valid or invalid, so that the divergence point is visible. We organize the cases around three mechanisms: **Case 1.** correcting early failures before they cascade, **Case 2.** recovering from early failures when they occur, and **Case 3.** the nature of the residual failures MRPO introduces.

Case 1: Cascade Correction. The most direct effect of MRPO is correcting the early-stage error that GRPO commits at the outset, before it propagates. Here GRPO establishes an incorrect visual premise in its first step, such as a Wrong Organ/Structure Identification, and every subsequent step inherits this misperception, locking the trace onto a wrong answer. Given the same input, MRPO instead grounds its opening steps in the actual visual evidence, so that the remaining reasoning builds on a correct premise and converges to the right answer. Figure 13 shows a representative gross pathology specimen, where GRPO misreads the liver's reddish-brown granular cut surface as spongy lung parenchyma and cascades into an incorrect organ identification, ultimately answering Lung, while MRPO anchors on the correct premise of a solid visceral organ and recovers the right answer, Liver. This reflects the core mechanism behind MRPO's reduction of early-stage failures (Section 5.3): suppressing the first misperception is sufficient to redirect the entire trajectory, since downstream reasoning is conditioned on the corrected premise.

Case 2: Early Recovery. Beyond preventing early failures, MRPO can also recover from a failure once it occurs, which GRPO rarely does. In these cases both methods reach the same wrong answer early in the trace, but for different reasons

and with diverging subsequent behavior. GRPO commits to the mistaken premise and every following step inherits it, whereas MRPO re-examines the image and reverses the error before it reaches the answer. Figure 14 illustrates this on an MRI-weighting question. Both traces initially settle on T2 but for different reasons. GRPO defaults to the assumption that abdominal MRI is typically T2-weighted without inspecting the image, while MRPO engages the image but misreads the dark peri-hepatic regions as fluid. GRPO never returns to the image and locks in the wrong answer, whereas MRPO re-examines the scan, anchors on the clearly bright subcutaneous fat that signals T1-weighting, and reverses its drift to recover the correct answer. This behavior is the case-level counterpart of MRPO's lower Failure Accumulation Rate in the early bins (Section 5.3). Even when an early error arises, the step-wise penalty discourages the model from compounding it, allowing later steps to correct course rather than propagate the mistake.

Case 3: MRPO Loss. For completeness, we also examine the instances that GRPO answers correctly but MRPO does not. Consistent with the paired analysis in Appendix E.2, where MRPO's losses concentrate in the late stage, these failures are typically not cascading errors but local breakdowns at the final step, after the reasoning has otherwise proceeded correctly. In such cases MRPO perceives the image and carries out valid reasoning, but commits a Terminal Label/Term Mismatch when producing the answer, mapping a correctly described finding to the wrong term. Figure 15 shows a representative instance, where MRPO correctly identifies the spleen and describes its smooth, curved contour, yet labels the shape as "lobulated" at the final step, contradicting its own description and yielding the wrong answer where GRPO succeeds. Such losses indicate that MRPO's residual errors stem from terminal terminology rather than corrupted reasoning, leaving the preceding trajectory intact and far less likely to cascade. This points to refining terminal answer grounding, rather than the reasoning process itself, as a direction for further improvement.

G RL Training Plots

We provide training-dynamics plots for all three backbones, Qwen2.5-VL-7B-Instruct (Figure 16), Qwen3-VL-8B-Instruct (Figure 17), and InternVL3-8B-Instruct (Figure 18), covering answer reward, reasoning process reward, KL divergence, and completion length.

Across the three backbones, MRPO holds a slight edge over GRPO on answer reward and a clearer margin on reasoning process reward. KL divergence runs somewhat higher for MRPO, a natural consequence of its stronger step-wise advantage reshaping, but in all cases it rises early and then settles rather than diverging, so training remains stable. Completion length varies by backbone and shows no consistent ordering between the two methods, except on Qwen3-VL-8B-Instruct, where MRPO produces clearly longer completions.

H Prompts

H.1 Answer Correctness Check Prompt

Answer Correctness Check Prompt

Given a question about a medical image, there is a correct answer to the question and an answer to be determined.

If the answer to be determined matches the correct answer or is a good enough answer to the question, output 'O'; otherwise output 'X'. Evaluate the answer to be determined ('O' or 'X').

Question
- question about the medical image: {problem}
- Image: {image}

Answers
- correct answer (ground truth): {solution}
- answer to be determined: {generated answer}

Your response must be a single character: 'O' (correct) or 'X' (incorrect).

H.2 Step-wise Reasoning Evaluation Prompt

Answer Correctness Check Prompt

You are a medical reasoning verification module.

Task

You will be given:

- Image: A medical image (provided separately)
- Problem: A medical VQA question about

the image
- Ground_Truth: The correct answer to the problem
- Gold_Reasoning: Gold reasoning steps (reference standard)
- Reasoning_Sentences: Generated reasoning sentences to evaluate

For EACH generated sentence, output:

- Gold Alignment: 0/1
- Answer Contribution: 0/1

General Rules (Apply First)

Automatic 0 for BOTH Gold Alignment and Answer Contribution:
- Meta-Commentary: "I will identify...", "Consider the possibilities..."
- Empty or meaningless statements
- Pure repetition of previous steps without new content

Check these rules FIRST before evaluating Gold Alignment and Answer Contribution.

Evaluation Criteria

Gold Alignment (Gold_Reasoning Consistency)

First, extract Key Elements from Gold_Reasoning:
- Modality, Context, Key Findings, Anatomical Location (including laterality), Diagnostic Direction

"Is this step consistent with Gold_Reasoning?"

Gold Alignment ONLY checks whether the step matches Gold_Reasoning, regardless of contribution to answer.

- **Gold Alignment = 1:**

- Early (First 1-2 steps) : Correctly identifies modality (e.g., "X-ray", "CT", "MRI", "electron microscopy"), tissue type (e.g., "histological section", "gross specimen"), or staining method (e.g., "H&E", "immunostain") that matches Gold_Reasoning
- Middle (Middle steps) : Identifies specific findings mentioned in Gold_Reasoning, including: abnormalities, pathological features, key structures, AND correct anatomical location/laterality. Must match Gold's level of specificity.
- Later (Last 1-2 steps) : Reaches or clearly approaches the same diagnostic conclusion as Gold_Reasoning. Must demonstrate diagnostic reasoning toward Ground_Truth, not just restate observations.
- If Gold_Reasoning contains diagnostic conclusions (e.g., "tuberculosis", "adenocarcinoma", "hemorrhage", "fracture"), generated reasoning MUST progress toward that diagnosis to get Alignment = 1

- **Gold Alignment = 0:**

- Wrong Location/Laterality: Gold says "LEFT" but generated says "right"
- Contradiction: Directly contradicts Gold_Reasoning
- Misdirection: Different diagnostic direction than Gold
- Content not mentioned or supported by Gold_Reasoning
- Missing Critical Findings: Gold identifies pathological/abnormal findings but generated only describes normal or generic features
- Specificity Mismatch: Gold is specific (e.g., "lymphoma", "abnormal features") but generated is generic (e.g., "cellular structure", "tissue")
- If Gold identifies specific pathology (e.g., "granuloma", "infarction", "metastasis", "fibrosis") but generated only describes generic features (e.g., "tissue changes", "some abnormality", "lesion")
- Describing only normal-appearing structures when Gold identifies abnormalities

Answer Contribution (Ground_Truth Derivation)

"Does this step directly help reach Ground_Truth: '{solution}'?"

Answer Contribution = 1:

- Directly mentions Ground_Truth or semantically equivalent terms

- Identifies a finding that is explicitly required to derive Ground_Truth
- States the specific diagnosis, location, or structure that matches Ground_Truth

Answer Contribution = 0:

- No direct relevance to Ground_Truth
- Generic observation that applies to any image of this type
- Describes features not connected to Ground_Truth
- Evasion: "Unknown", "cannot determine", "None"
- Context-only: states modality/setting without advancing toward Ground_Truth

Output Format (JSON ONLY):

Return JSON only, with this exact structure:

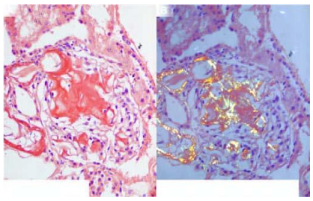
```
{
  "Reasoning_Check": {
    "step1": {
      "Gold Alignment": 1,
      "Answer Contribution": 1
    },
    "step2": { ... }
  }
}
```

Rules:

- Apply General Rules FIRST (meta-commentary → both 0)
- Evaluate Gold Alignment and Answer Contribution INDEPENDENTLY
- Both values must be 0 or 1
- Do NOT output explanations outside JSON

Inputs:

- "Problem": {problem}
- "Image": {image}
- Ground_Truth: {ground_truth}
- Gold_Reasoning: {gold_reasoning}
- "Reasoning_Sentences": {reasoning_sentences}



Question | What show apple-green birefringence?

Gold Answer | congophilic areas

Gold Reasoning |

The apple-green birefringence is characteristic of amyloid deposits when stained with Congo red dye and viewed under polarized light. The molecular structure of the amyloid fibrils allows them to bind to Congo red, and when they are aligned in a certain way, they exhibit this distinctive apple-green color when viewed under the polarized light. This is a diagnostic feature used in pathology to confirm the presence of amyloid in tissue samples. The image likely shows a comparison between tissue sections viewed under regular light microscopy (on the left, without the birefringence) and under polarized light microscopy (on the right, exhibiting the apple-green birefringence).

Early-Stage Failure Case 1 : Default Staining/Modality Assumption

Step1 | ✗ The image appears to be a histological slide, likely **stained for lipid or cholesterol**, given the context of birefringence.

The model defaults to a "lipid/cholesterol stain" assumption instead of reading the actual Congo-red stain, whose apple-green birefringence is the pathognomonic sign of amyloid. This premise leads it to reinterpret the apple-green signal as lipids rather than amyloid, propagating through every step to the wrong answer ("cholesterol esters" instead of congophilic/amyloid deposits

Step2 | ✗ Apple-green birefringence is a characteristic feature seen under polarized light microscopy when certain lipids, such as cholesterol esters, are present.

Step3 | ✗ The staining method used in the image suggests it could be a lipid stain, which would explain the birefringence observed.

Step4 | ✗ The presence of this birefringence indicates the presence of these lipids in the tissue sections.

Answer | ✗ Cholesterol esters

Figure 7: **Early-stage failure: Default Staining/Modality Assumption.** The model defaults to the most common modality (e.g., H&E staining) instead of reading the actual image, grounding the entire trace in an unverified premise that corrupts all subsequent steps.

Question | What is the largest organ in the picture?

Gold Answer | Liver

Gold Reasoning |

The explanation for identifying the liver as the largest organ in the picture is based on the recognition of the typical shape and location as seen on a transverse (axial) CT scan of the abdomen. The liver typically appears as a large, dark area in the upper right quadrant of the abdomen on such scans, with a characteristic shape that can be distinguished from other abdominal organs. Given that it's one of the largest solid organs in the human body and what we can see from the CT scan's cross-section, the liver is visible and takes up the most space compared to other organs identified in the same slice.

Early –Stage Failure Case 2 : Wrong Organ/Structure Identification

Step1 | ❌ The image provided is a CT scan of the **chest**.

The model correctly identifies the CT modality but misreads the anatomical level as "chest" rather than upper abdomen, constraining the answer space to thoracic organs. This premise leads it to fixate on the basal lung air pockets while overlooking the liver dominating the frame, yielding the wrong answer

Step2 | ❌ In such scans, the largest organ typically visible is the lung.

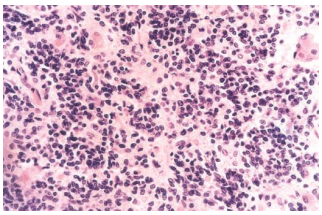
Step3 | ❌ The lung occupies a significant portion of the thoracic cavity and is usually the most prominent structure in a chest CT scan.

Step4 | ❌ The lung appears as the large, dark areas on either side of the heart

Step5 | ❌ Given the context of the image, the lung is the largest organ present.

Answer | ❌ Lung

Figure 8: **Early-stage failure: Wrong Organ/Structure Identification.** The model recognizes the image type correctly but, in its first sentence, confidently specifies an incorrect organ or anatomical structure, leading the downstream reasoning to cascade into a wrong answer.



Question | What did the microscopic appearance of medulloblastoma show?

Gold Answer | small

Gold Reasoning |

Medulloblastoma is a type of brain tumor that predominantly affects children. Under the microscope, medulloblastoma cells typically appear small, round, and densely packed together with hyperchromatic nuclei. The cells exhibit high nuclear-to-cytoplasmic ratios, meaning the nuclei take up most of the cell volume. The provided image likely shows these features, revealing the small and densely arranged cells characteristic of this malignancy. The "small" in the answer refers to the appearance of the cells when viewed under the microscope, not the overall size of the tumor.

Mid-Stage Failure Case 1 : Structural Misidentification

Step1 | ✓ The image provided is a microscopic view of a tissue sample, likely stained to highlight cellular and structural details.

Step2 | ✓ Medulloblastoma is a type of brain tumor that originates in the cerebellum.

Step3 | ✓ Microscopically, medulloblastoma typically shows a dense population of small, round cells with uniform nuclei and scant cytoplasm.

Step4 | ✗ The cells are often arranged in a pattern that can resemble a **"fried-egg" appearance**, with a central nucleus surrounded by a halo of cytoplasm.

The "fried-egg" appearance (perinuclear halo) is the hallmark of oligodendroglioma, not medulloblastoma. This is Structural Misidentification: it imposes a different tumor's cytologic pattern absent from the image.

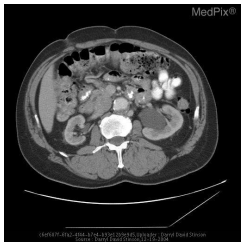
Step5 | ✗ The background may show a pinkish or purplish hue due to the staining method used, which helps in differentiating the cellular components.

Step6 | ✓ The image shows a dense collection of **small, round cells with darkly stained nuclei**, consistent with the microscopic appearance of medulloblastoma.

In the final step, the model returns to the actual image and re-derives the correct features (small round cells, hyperchromatic nuclei), recovering to the right answer

Answer | ✓ dense population of small, round cells with uniform nuclei and scant cytoplasm

Figure 9: **Mid-stage failure: Structural Misidentification.** After establishing a correct premise, the model misrecognizes a microstructure as a different organ or structure, corrupting only the steps that depend on this misjudgment while later steps may still recover.



Question | What are the hyperintensities surrounding the aorta?

Gold Answer | Calcified atherosclerosis

Gold Reasoning |

The hyperintensities noted on the CT scan surrounding the aorta represent areas where the tissues have a higher radiodensity compared to surrounding structures. In this context, the term \"calcified atherosclerosis\" refers to the deposition of calcium within the atherosclerotic plaques. Atherosclerosis is a condition in which the arteries become narrowed and hardened due to a buildup of plaque around the artery wall. The plaque consists of fat, cholesterol, calcium, and other substances found in the blood. When this plaque calcifies, it can appear as hyperintense on a CT scan due to the high radiodensity of calcium, which blocks X-rays more than the soft tissue or blood, resulting in the brighter area seen on the image. This phenomenon is indicative of atherosclerotic disease, which can be a risk factor for various cardiovascular events, such as heart attack or stroke.

Mid-Stage Failure Case 2 : Pathology Omission

Step1 | ✓ The image is a CT scan showing a cross-sectional view of the abdomen.

Step2 | ✓ The aorta is visible in the center, and surrounding it are areas of hyperintensity.

Step3 | ✓ Hyperintensities in CT imaging typically indicate regions with higher density or contrast uptake.

Step4 | ✗ In this context, the hyperintensities surrounding the aorta could be due to several reasons such as **inflammation, fat deposition, or contrast material.**

This step omits calcification, the real finding (calcified plaque) along the aortic wall, steering the differential toward benign causes. The omission cascades: later steps commit to "fat" and never reconsider, locking in the wrong answer

Step5 | ✗ Given the location and appearance, these hyperintensities likely represent fat surrounding the aorta, a normal finding in abdominal CT scans.

Step6 | ✗ Fat appears hyperintense on CT due to its lower density than surrounding tissues.

Answer | ✗ Fat

Figure 10: **Mid-stage failure: Pathology Omission.** Despite the presence of a lesion in the image, the model describes only normal findings, an interpretation error that arises after the visual premise is correctly established.

Question | Which sign do you see in the aortopulmonary window in this image?

Gold Answer | middle mogul

Gold Reasoning |

The aortopulmonary window is an anatomical space between the aortic arch and the pulmonary artery. On a chest radiograph, like the one provided, the middle mogul sign refers to an abnormal prominent bulge or mass in this area. It is typically seen as a silhouette or contour abnormality on the left border of the mediastinum at the level of the aortic arch. The sign suggests an enlargement or mass in the aortopulmonary window, which may be indicative of certain pathological conditions such as lymphadenopathy, aortic aneurysm, or other masses. The provided answer of "middle mogul" is based on the recognition of this radiographic feature in the image.

Late-Stage Failure Case 1 : Non-committal Terminal Conclusion

Step1 | ✓ The image provided is a chest X-ray.

Step2 | ✓ In this X-ray, the aortopulmonary window is a region where the aorta and pulmonary artery are visible.

Step3 | ✓ The aortopulmonary window is typically seen as a clear area between the aorta and the pulmonary artery.

Step4 | ✓ In this image, the aorta and pulmonary artery are visible, and there is a clear space between them, which is the aortopulmonary window.

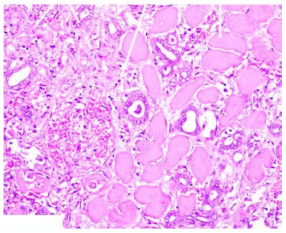
Step5 | ✓ Such a prominent contour can correspond to a mass, lymphadenopathy, or an aneurysm in this region.

Step6 | ✗ This prominent contour points to a mass or lymphadenopathy, **though further imaging would be needed to confirm it.**

The reasoning correctly identifies the abnormal contour through Step 5 and even points toward the right region, but at the final step the model hedges with "would be needed to confirm" rather than committing to the specific sign (middle mogul). The error is confined to the terminal step and leaves the earlier reasoning intact.

Answer | ✗ Mass

Figure 11: **Late-stage failure: Non-committal Terminal Conclusion.** The reasoning proceeds correctly until the late steps, but the model fails to converge on a specific answer and hedges with vague expressions such as “consistent with” or “possibly,” avoiding a definitive conclusion.



Question | What do renal tubules and glomeruli show?

Gold Answer | typical coagulative necrosis

Gold Reasoning |

Coagulative necrosis is a type of tissue necrosis where the architecture of dead tissues is preserved for a couple of days. The necrosis is typically caused by ischemia or infarction, and in the kidneys, it often affects the renal tubules and glomeruli due to their high sensitivity to oxygen deprivation. In the given image, which is likely a microscopic image of kidney tissue, the typical coagulative necrosis can be identified by the presence of pale, eosinophilic (pink) staining cells with loss of nuclear staining, indicating that these cells are dead but still maintain their basic structural outlines. Even without the cellular detail that would be provided with higher magnification or additional stains, the alteration of normal tissue architecture and the presence of pale cells suggest necrosis, which in the context of the kidneys often manifests as coagulative necrosis.

Late-Stage Failure Case 2 : Terminal Label/Term Mismatch

Step1 | ✓ The image is a histological section stained with hematoxylin and eosin (H&E), commonly used in pathology to visualize tissue architecture.

Step2 | ✓ The structures visible include tubular formations (renal tubules) and vascular structures with a central lumen and surrounding capillary loops (glomeruli).

Step3 | ✓ These are hallmark features of kidney tissue.

Step4 | ✓ The presence of both tubules and glomeruli indicates the image is showing a kidney biopsy or section, and the question implies a pathological state.

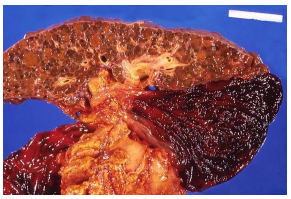
Step5 | ✗ The image shows **tubules with pale, shrunken cytoplasm and loss of nuclear staining**, indicating widespread cellular degeneration consistent with severe pathology.

Step 5 fits Terminal Label/Term Mismatch: the model correctly describes coagulative necrosis (pale, shrunken cells with loss of nuclear staining) but maps it at the final step to the wrong term, "atrophy." The error stays local rather than cascading, as only the terminal naming slips, collapsing an otherwise correct trace into a wrong answer.

Answer | ✗ Atrophy

Figure 12: **Late-stage failure: Terminal Label/Term Mismatch.** Having correctly identified the relevant structure or finding, the model mismaps it at the final step to an incorrect name, laterality, or specific term, while the preceding reasoning remains intact.

Qualitative Case Study: Case 1) Cascade Correction



Question | What does this image show?

Gold Answer | Liver

Gold Reasoning |

The image displays a human liver. This organ has been identified by its distinctive lobular shape and the dark reddish-brown color which is typical of the liver due to its rich blood supply and iron content. The visceral surface is shown, and typical liver morphology, such as the larger right lobe and smaller left lobe, can be seen. The liver's texture and the vessels that are visible on its surface also contribute to this identification.

1. GRPO

Step1 | ✗ The specimen has a reddish-brown, **spongy surface resembling aerated lung parenchyma.**

Early-stage failure(Wrong Organ/Structure Identification) : GRPO misreads the liver's reddish-brown granular cut surface as "spongy lung parenchyma," setting a wrong-organ premise at the outset. This false premise cascades: the nodular areas become alveolar spaces and the central area a bronchial lumen, locking the trace onto "lung."

Step2 | ✗ The pale nodular areas look like air-filled alveolar spaces.

Step3 | ✗ A central cavity suggests a bronchial lumen running through the lobe.

Step4 | ✗ Together, the spongy texture and cavity indicate a lung specimen.

Answer | ✗ Lung

2. MRPO

Step1 | ✓ The image is a gross pathology specimen on a blue background with a scale bar, not a microscopic or radiological image.

MRPO anchors on the actual evidence from the start reading them as a solid visceral organ. Building on this correct premise, it converges to the liver, avoiding the lung misinterpretation that derailed GRPO.

Step2 | ✓ It has a deep reddish-brown color, a lobulated shape, and a granular, nodular cut surface.

Step3 | ✓ The large size and heterogeneous parenchyma point to a solid visceral organ.

Step4 | ✓ This color and nodular cut surface are most consistent with the liver.

Step5 | ✓ Therefore, the image shows a liver.

Answer | ✓ liver

Figure 13: **Case 1: Cascade correction.** GRPO's incorrect premise in the first step cascades into a wrong organ identification, while MRPO anchors on the correct visual evidence and reaches the right answer.

Qualitative Case Study: Case 2) Early Recovery

Question | What is the mr weighting in this image?

Gold Answer | T1

Gold Reasoning |

The image provided is an MRI scan displaying high signal intensity in areas of fatty tissue and low signal intensity in fluid areas. In T1-weighted MR images, fat appears brighter (has a high signal), and water or cerebrospinal fluid appears darker (has a low signal). The image characteristics consistent with T1 weighting are the bright appearance of the subcutaneous fat and other fatty structures, and the darker appearance of the fluid-filled spaces, such as in the ventricles or any potential fluid collections.

1. GRPO

Step1 | ✓ The image is a cross-sectional MRI scan

Step2 | ✗ **Most abdominal MRI scans are T2-weighted by default, so this is most likely T2-weighted.**

Early-stage failure(Default Staining/Modality Assumption) : Rather than reading the image, it defaults to the assumption that abdominal MRI is usually T2-weighted, committing to T2 without checking the signal characteristics. This unverified premise cascades through steps 2–4, which lock in T2 without returning to the image.

Step3 | ✗ T2-weighted images show fluid as bright, the contrast typically sought in abdominal imaging.

Step4 | ✗ The appearance fits a standard T2-weighted study.

Step5 | ✗ This indicates the image is T2-weighted.

Answer | ✗ T2

2. MRPO

Step1 | ✓ The image is an axial abdominal MRI showing the liver.

Step2 | ✗ The dark regions in and around the liver look like fluid, suggesting a fluid-sensitive T2 sequence.

Step3 | ✗ This fluid reading would point toward T2-weighting.

Step4 | ✓ However, the subcutaneous fat is clearly bright, and bright fat with dark fluid is the signature of T1-weighting.

MRPO misreads the dark regions around the liver as fluid, drifting toward a T2 interpretation. At step 4, it re-examines the image and anchors on the clearly bright subcutaneous fat, the signature of T1-weighting, reversing the earlier drift and recovering the correct answer, T1.

Step5 | ✓ Anchoring on this fat-bright, fluid-dark pattern, the image is T1-weighted.

Answer | ✓ T1

Figure 14: **Case 2: Early recovery.** GRPO defaults to a T2 assumption without inspecting the image and MRPO misreads dark regions as fluid; GRPO locks in the error, while MRPO re-anchors on the T1-characteristic bright subcutaneous fat and recovers the correct answer.

Qualitative Case Study: Case 3) MRPO Loss

Question | What is the shape of spleen in the picture?"

Gold Answer | Oval

Gold Reasoning |

The spleen appears oval in the image because of the way it is cut in the particular cross-section provided by the CT scan. The spleen's shape can vary depending on the angle of the image slice, and in this specific image, the plane of the CT cut captures an oval shape of the organ.

1. GRPO

Step1 | ✓ The image provided appears to be an axial CT scan of the abdomen.

Step2 | ✓ The spleen is visible on the left side of the image.

Step3 | ✓ Observing its shape, it appears to be elongated and somewhat oval.

Answer | ✓ Oval

2. MRPO

Step1 | ✓ The image is an axial CT slice of the abdomen.

Step2 | ✓ The spleen is visible on the left side of the body, behind the stomach.

Step3 | ✓ Its outline appears as a smooth, curved soft-tissue structure following the contour of the abdominal wall.

Step4 | ✗ Judging from this contour, **its shape is lobulated rather than a simple rounded form.**

Late-stage failure (Terminal Label/Term Mismatch) : After correctly identifying the spleen and its smooth, curved contour in steps 1–3, the model maps that same contour to the wrong shape term, "lobulated," at the final step. The error is purely terminal, contradicting the smooth outline it just described, collapsing an otherwise correct trace into the wrong answer instead of "oval."

Answer | ✗ Lobulated

Figure 15: **Case 3: MRPO loss.** MRPO correctly identifies and describes the spleen but labels its shape as "lobulated" at the final step, a terminal term mismatch that leaves the preceding reasoning intact.

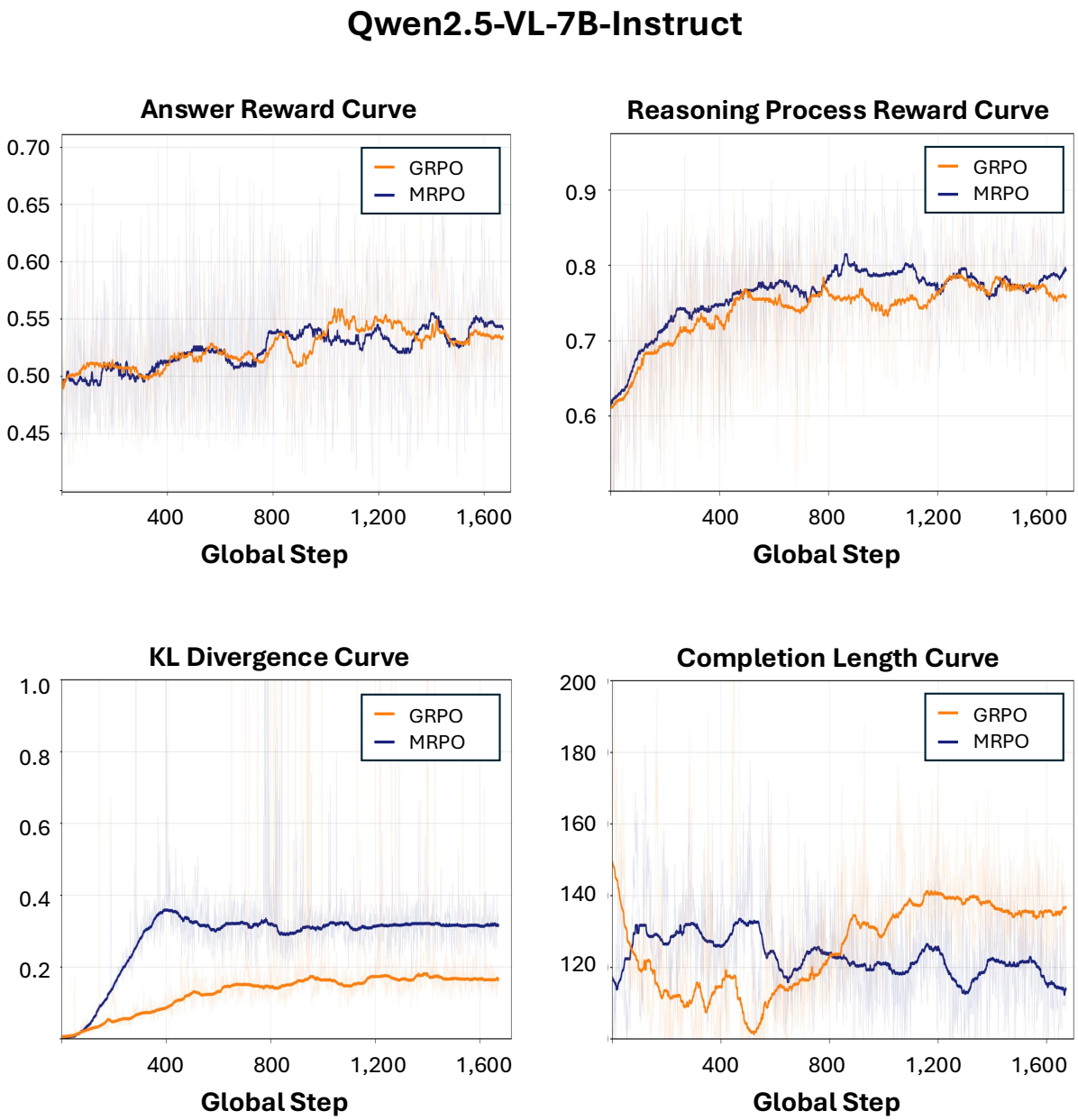


Figure 16: **Training dynamics of GRPO and MRPO on Qwen2.5-VL-7B-Instruct.** Curves show the answer reward, reasoning process reward, KL divergence, and completion length over training.

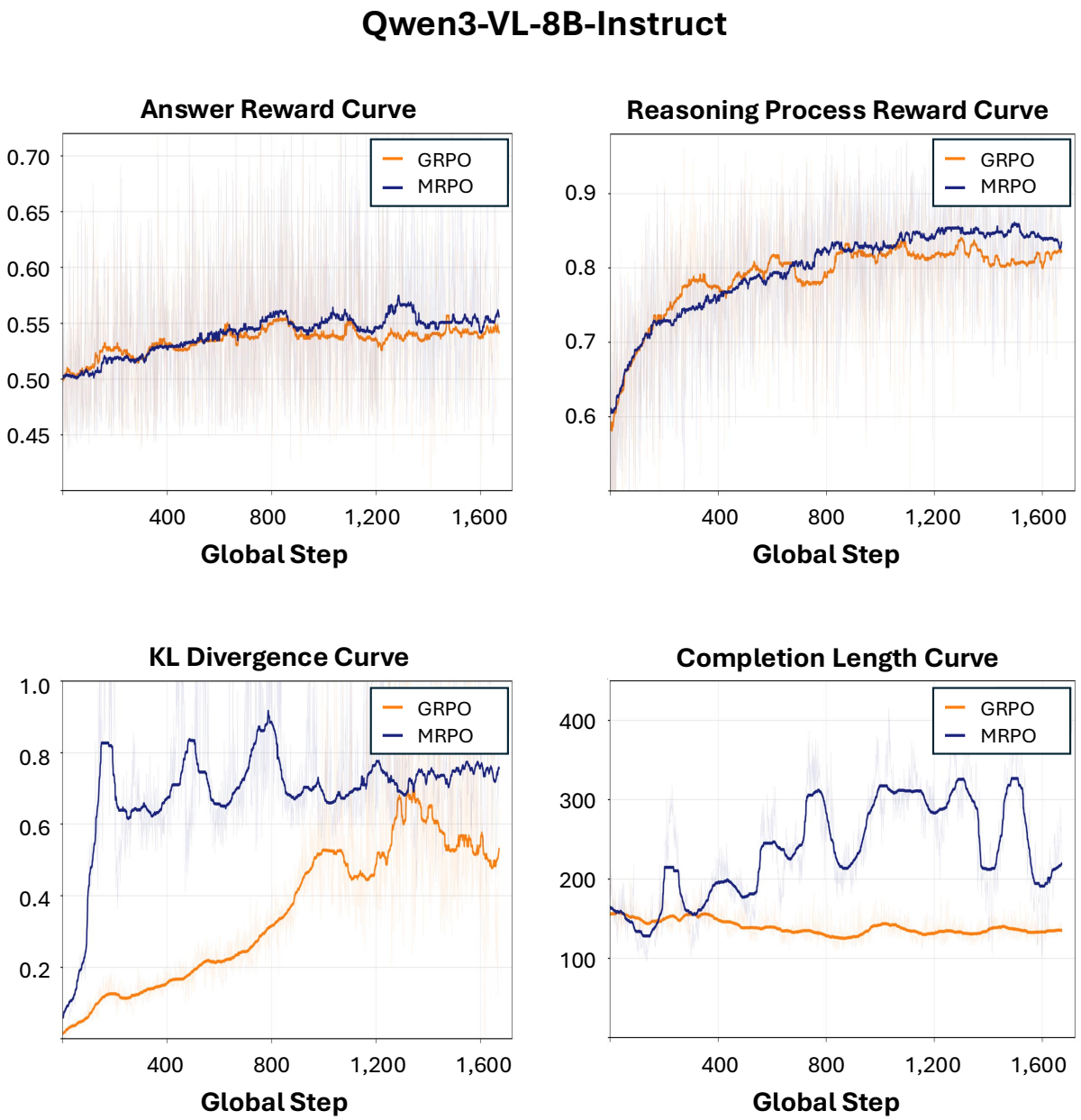


Figure 17: **Training dynamics of GRPO and MRPO on Qwen3-VL-8B-Instruct.** Curves show the answer reward, reasoning process reward, KL divergence, and completion length over training.

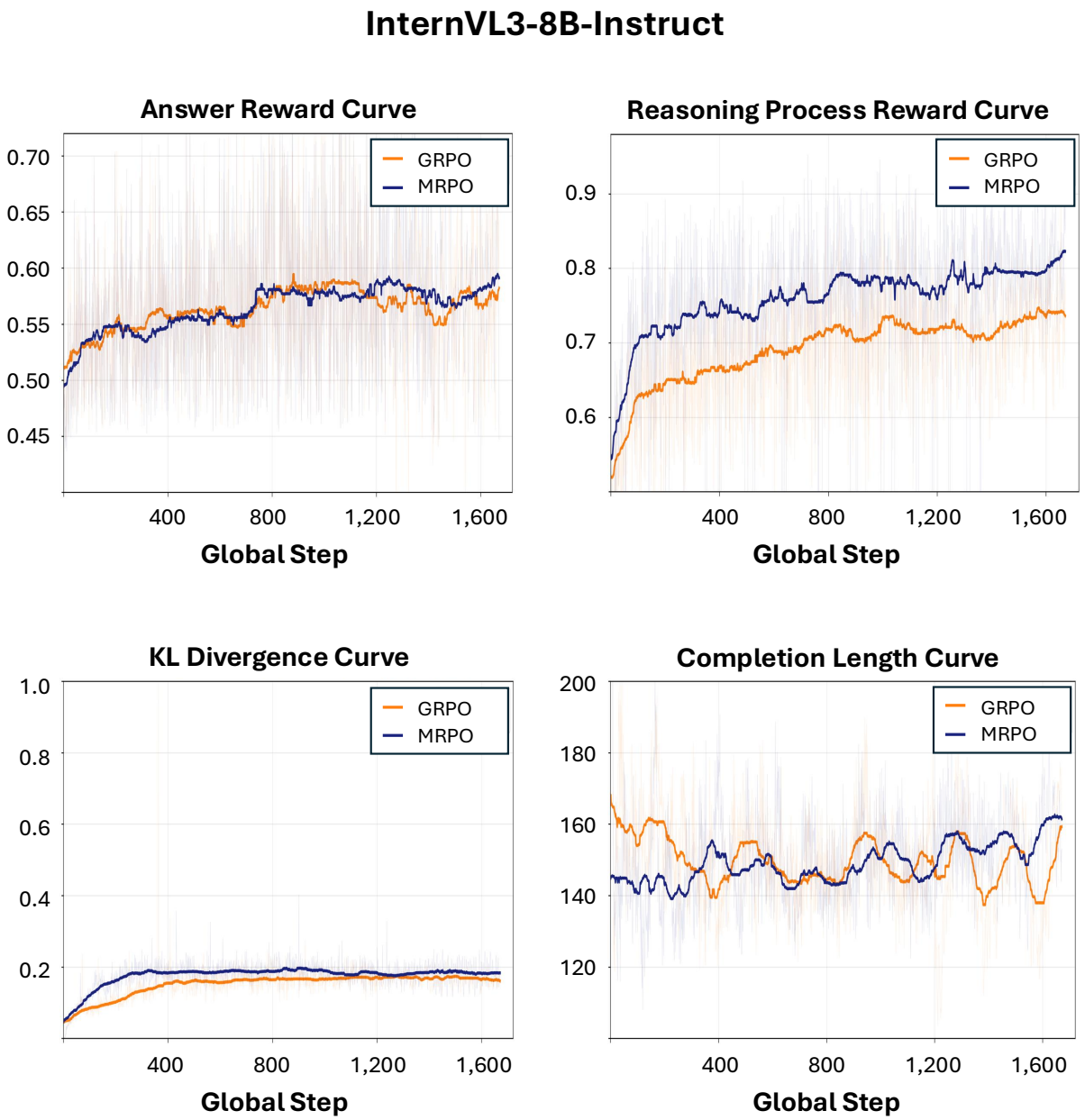


Figure 18: **Training dynamics of GRPO and MRPO on InternVL3-8B-Instruct.** Curves show the answer reward, reasoning process reward, KL divergence, and completion length over training.