

Toward Effective Reinforcement Learning Fine-Tuning for Medical VQA in Vision-Language Models

Anonymous ACL submission

Abstract

Recently, reinforcement learning (RL)-based tuning has shifted the trajectory of Multimodal Large Language Models (MLLMs), particularly following the introduction of Group Relative Policy Optimization (GRPO). However, directly applying it to medical tasks remains challenging for achieving clinically grounded model behavior. Motivated by the need to align model response with clinical expectations, we investigate four critical dimensions that affect the effectiveness of RL-based tuning in medical visual question answering (VQA): base model initialization strategy, the role of medical semantic alignment, the impact of length-based rewards on long-chain reasoning, and the influence of bias. We conduct extensive experiments to analyze these factors for medical MLLMs, providing new insights into how models are domain-specifically fine-tuned. Additionally, our results also demonstrate that GRPO-based RL tuning consistently outperforms standard supervised fine-tuning (SFT) in both accuracy and reasoning quality.

1 Introduction

Encouraged by its success in Large Language Models (LLMs), researchers have extended Group Relative Policy Optimization (GRPO) (Shao et al., 2024) to multimodal domains such as image understanding, audio processing, UI design, and physical world interaction (Chen et al., 2025; Shen et al., 2025a; Lai et al., 2025; Lu et al., 2025; Xia and Luo, 2025; Zhao et al., 2025; Azzolini et al., 2025). Specifically, studies like (Huang et al., 2025; Zhou et al., 2025a) report promising results, including emergent reasoning in compact models. However, applying GRPO-based RL to medical visual question answering (VQA), which demands clinically accurate outputs, remains underexplored.

In this work, we analyze GRPO-based RL for medical Multimodal Large Language Model (MLLMs) across five critical aspects:

(1) Training from Scratch vs. Fine-Tuning.

Training from scratch allows for unconstrained reasoning exploration, but initializing from an instructionally fine-tuned model stabilizes training and accelerates convergence (Zhang et al., 2023a; Chung et al., 2024). To investigate this trade-off, we conduct experiments comparing both approaches. Our results show that prior instruction supervised fine-tuning (SFT) improves both answer accuracy and domain alignment.

(2) Rewarding Medical Semantic Accuracy.

Generic rewards for format or output correctness are insufficient for clinical tasks. We introduce a medical semantic reward using LLM-generated evaluations, guided by prompt engineering. This significantly boosts both response quality and clinical alignment.

(3) Does Longer Reasoning alone Help?

Although many studies have shown that deep reasoning can be beneficial (Cheng et al., 2024; Kumar et al., 2025), we found that relying solely on length-based rewards (i.e., Extended Chain Reward (ECR) and Correctness-Weighted Length Reward (CWR)) often leads to verbose and less accurate answers. This observation calls into question the effectiveness of using length-based rewards along as a mechanism to promote meaningful long-form reasoning in medical VQA.

(4) Does bias exist in medical MLLMs?

Normalization is commonly used to stabilize training. However, recent study (Liu et al., 2025) suggest that the question-level normalization may inadvertently bias model behavior, encouraging the generation of longer but incorrect responses by amplifying the per-token gradient signal. To further evaluate its impact in medical VQA, we implement Dr.GRPO (Liu et al., 2025). Our results demonstrate its effectiveness in improving both answer accuracy and token efficiency.

(5) SFT vs. GRPO-based RL tuning.

SFT is a widely adopted strategy to enhance the reasoning

ability of MLLMs. To compare its effectiveness against GRPO-based RL tuning, we implemented three different SFT approaches and additionally evaluated two publicly available MLLMs. Our results show that the GRPO-based RL tuning consistently outperforms SFT methods, yielding higher answer accuracy and more clinically grounded responses.

Our main contributions can be summarized as follows:

- We present a systematic analysis of GRPO-based RL in medical MLLMs, focusing on initialization strategies, medical semantic alignment, the impact of length-based rewards and bias-related behaviors.
- We validate our findings through large-scale experiments on medical VQA benchmarks, offering practical insights into aligning RL with clinically meaningful behavior.
- We find that GRPO-based RL tuning outperforms traditional fine-tuning methods (e.g., SFT), highlighting its potential for developing more capable and aligned medical MLLMs.

2 Preliminary

Instead of relying on an extra reward and value model as in PPO (Schulman et al., 2017), GRPO simplifies the process by using the average reward from the policy model’s sampled responses as a baseline for advantage estimation. Specifically, given an input question q , we first sample a group of responses $\{o_1, o_2, \dots, o_G\}$ from the policy model $\pi_{\theta_{\text{old}}}$ and compute corresponding rewards $\mathbf{r} := \{r_1, r_2, \dots, r_G\}$. Then GRPO seeks to optimize the following objective and update the model π_{θ} , denoted as:

$$\begin{aligned} \mathcal{J}_{GRPO}(\theta) := & \mathbb{E}_{q \sim p_Q, \{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(O|q)} \\ & \left\{ \frac{1}{G} \sum_{i=1}^G \min \left[\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)} \hat{A}_i, \right. \right. \\ & \left. \left. \text{clip} \left(\frac{\pi_{\theta}(o_i|q)}{\pi_{\theta_{\text{old}}}(o_i|q)}, 1 - \epsilon, 1 + \epsilon \right) \hat{A}_i \right] \right. \\ & \left. - \beta \cdot D_{KL}(\pi_{\theta} || \pi_{\text{ref}}) \right\}, \end{aligned} \quad (1)$$

where the advantage is denoted as:

$$\hat{A}_i := \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})} \quad (2)$$

Here, ϵ and β denote the PPO clipping hyperparameter and weight of KL-divergence regularization,

respectively. We omit the token-level average per response in Eq. 1 for simplicity. Additionally, we strictly follow DeepSeek-R1 (Guo et al., 2025), using rule-based reward (i.e., format and response reward) as our base reward design.

3 Empirical Studies on RL tuning

We selected Qwen2-VL-2B (Wang et al., 2024) as our base model. All experiments were conducted on the subset of the PMC-VQA (Zhang et al., 2023b) benchmark, including 10K training samples and 7K testing samples. In this section, we analyze and answer four key questions that influence the effectiveness of GRPO-based RL tuning in medical VQA. The training parameters follow the settings used in previous work (Zhou et al., 2025b), and additional details are provided in Appendix A. Here, response accuracy, similarity score, perplexity (Chung, 2025), thinking reward (Jiang et al., 2023), and thinking token length are considered as metrics wherever such measurements are available. We outline the details of metrics in Appendix B.

3.1 From Scratch vs. Fine-Tuning.

Recent studies on GRPO-based RL tuning in MLLMs have typically relied on base models that were already instruction fine-tuned (Chen et al., 2025; Zheng et al., 2025; Shen et al., 2025b; Wang and Peng, 2025). However, these models often fail to exhibit the "aha moment" in their learning curves, suggesting that instruction tuning may hinder the reasoning exploration. Indeed, (Zhou et al., 2025b) shows that cold-start GRPO-based RL without SFT can more effectively promote reasoning behavior in the MLLMs setting. To examine the role of domain knowledge, we compare GRPO-based RL on Qwen2-VL-2B (trained from scratch) and Qwen2-VL-2B-Instruct (instruction-tuned). As shown in Tab. 1, the scratch-trained model has a higher similarity score and a +1.61 gain in thinking reward, meaning its reasoning is more aligned and useful. However, it also shows lower accuracy and higher perplexity Score, which suggests less correct and less fluent answers.

This means that while training from scratch encourages more reasoning, it lacks the medical knowledge and language fluency that instruction tuning provides (see Fig. S2 in Appendix C for examples). These results show that how a model is initialized affects its performance. Instruction-tuned models give more accurate and fluent an-

Template for Semantic LLM Judgment

Prompt: Evaluate the following medical statement for semantic correctness and clinical coherence:

Given Statement: Reasoning Text (e.g., text between <think></think>)

Answer 'Yes' if the statement is medically coherent and accurate, or 'No' otherwise.

Answer:

Figure 1: Illustration of the prompt template used to evaluate the effectiveness of medical semantic alignment. See more details in Sec. 3.2.

swers. For medical VQA, utilizing domain-specific pretraining methods (e.g., cold-start, pretraining) helps strike a balance between answer quality and reasoning.

3.2 Medical Semantic Alignment.

Aligning the model’s reasoning path with the target task can enhance the effectiveness of GRPO-based RL tuning. To further investigate the impact of medical semantic alignment in the medical VQA setting, we introduce a semantic alignment reward that encourages model responses to match the judgments of predefined expert LLMs. Specifically, we use Qwen2-VL-2B-Instruct as the base model and employ BioGPT (Luo et al., 2022) and BioMistral (Labrak et al., 2024) as the reference LLMs.

As illustrated in Fig. 1, we design a prompt template in which the reference LLM is asked to assess whether the reasoning enclosed within the <think></think> tags is clinically grounded during training. If the reasoning is valid, the LLM responds with "Yes," and a reward of 1 is assigned; otherwise, it responds with "No," and a reward of 0 is given. Results in Tab. 1 show that adding semantic alignment improves both performance and reasoning quality. Accuracy increases by 1.82%, and the Similarity Score improves by 0.25, indicating that the model’s reasoning becomes more semantically aligned with reference answers. Overall, the integration of medical semantic rewards leads to notable improvements in both accuracy and reasoning depth. Example visualizations are provided in Fig.S3 in Appendix C.

3.3 The Influence of Long-Chain Reasoning.

Chain-of-thought (CoT) reasoning has been shown to improve performance in large language mod-

els (Team et al., 2025; Guo et al., 2025; LLMS), but its role in medical VQA is less understood (Zhou et al., 2025b; Zhang et al., 2024; Dong et al., 2024). To explore this, we incorporated an Extended Chain Reward (ECR) during GRPO-based RL tuning of Qwen2-VL-2B-Instruct, alongside a medical semantic alignment reward. ECR incentivizes longer reasoning chains by assigning an additional reward based on output length.

As shown in Tab. 1, adding ECR increases token Length by 273.72 and improves Similarity Score by 0.19 and thinking reward by 1.35. However, these gains come at the expense of a 7.87% drop in accuracy, indicating that the model begins to favor more verbose and elaborate reasoning at the cost of factual correctness. To mitigate this, we introduced a Correctness-Weighted Length Reward (CWR), which incentivizes long responses only when the final answer is correct. This is achieved through symbolic and string-based correctness checks. As shown in Tab. 1, when combined with semantic alignment, CWR enhances fluency and structure coherence, as evidenced by improvements in perplexity and similarity Score. However, it remains 3.36% lower than the baseline GRPO model. Although token length still increases considerably, the model continues to exhibit tendencies toward exploiting length-based incentives. These findings suggest that relying on length-focused reward can lead to verbosity. Striking an appropriate balance between factual accuracy and high-quality reasoning remains a key challenge in medical VQA. Additional illustrative examples are provided in Appendix C, Fig. S4.

3.4 Unbiased GRPO.

Normalization can distort model behavior. Token-level normalization may weaken the effect of negative advantages, leading the model to generate longer but incorrect answers. Question-level normalization can exacerbate this issue, particularly with overconfident questions (i.e., those that are too easy or too hard). These effects raise concerns about the reliability of reasoning in medical VQA. To investigate this, we apply Dr. GRPO (Liu et al., 2025), which removes standard deviation normalization and token-level averaging. It computes advantage as a simple difference from the group mean reward.

$$\hat{A}_i := r_i - \text{mean}(\mathbf{r}) \quad (3)$$

Table 1: An illustration of all our experimental results. Specifically, **Qwen2-VL-2B-Instruct**, **LLaVA-7B-v1.5** and **Llama-3.2-11B-Vision-Instruct** are used directly without additional fine-tuning. For experiments involving fine-tuning, **Qwen2-VL-2B-Instruct** is used as the default base model unless otherwise specified (e.g., training from scratch with **Qwen2-VL-2B**). Additional experimental details are provided in Sec.3.

Model	Fine-Tuning Data Scale	Accuracy \uparrow	Similarity Scores \uparrow	Perplexity Scores \downarrow	Thinking Rewards \uparrow	Thinking Token Length
Qwen2-VL-2B-Instruct	-	47.29	-	-	-	-
SFT-based Training						
— LoRA	10K	45.98	-	-	-	-
— Full Fine-Tuning	10K	52.00	-	-	-	-
— DPO Fine-Tuning	10K	46.97	-	-	-	-
Training from Scratch						
Qwen2-VL-2B	-	-	-	-	-	-
— GRPO	10K	51.56	0.49 (± 0.19)	14.16 (± 2.80)	9.27 (± 2.09)	141.46 (± 69.10)
RL-based Training						
Qwen2-VL-2B-Instruct	-	-	-	-	-	-
— GRPO	10K	58.04	0.21 (± 0.24)	13.28 (± 19.64)	7.66 (± 2.49)	66.41 (± 74.34)
— GRPO + Semantic Alignment	10K	59.86	0.46 (± 0.19)	36.54 (± 10.97)	8.07 (± 2.63)	64.42 (± 24.02)
— GRPO + Semantic Alignment + ECR	10K	50.17	0.65 (± 0.07)	20.54 (± 2.92)	9.42 (± 1.91)	338.14 (± 95.69)
— GRPO + Semantic Alignment + CWR	10K	54.68	0.61 (± 0.11)	18.45 (± 2.40)	7.82 (± 3.52)	224.06 (± 73.72)
— Dr.GRPO	10K	61.09	0.24 (± 0.25)	11.10 (± 12.63)	4.17 (± 3.76)	76.11 (± 80.27)
Other Models						
LLaVA-7B-v1.5	-	11.8	0.19 (± 0.24)*	-	-	-
Llama-3.2-11B-Vision-Instruct	-	22.92	0.33 (± 0.27)*	-	-	-

This formulation provides a more stable and interpretable reward signal by reducing the influence of response length and preventing overly sharp gradients. We integrate Dr.GRPO into the GRPO-based RL tuning of Qwen2-VL-2B-Instruct, using the same training setup. As shown in Tab. 1, Dr.GRPO achieves the highest accuracy among all configurations, surpassing the standard GRPO baseline by 3.05%. It also yields improvements in both Perplexity and Similarity Score, indicating enhanced fluency and semantic alignment. These results suggest that removing normalization mechanisms facilitates more stable training dynamics and improves the alignment between reasoning quality and answer correctness. Overall, Dr. GRPO provides a more reliable and interpretable optimization signal, supporting the generation of clinically relevant and efficient responses in medical VQA.

4 SFT vs. GRPO-based RL tuning

Given that SFT is another widely used approach to endow MLLMs with reasoning ability, an important question arises in the context of clinically meaningful RL-tuned MLLMs: *Which performs better in medical VQA — GRPO-based RL or SFT-based training?* To explore this comparison, we follow prior SFT work (Lee, 2024), evaluating three SFT strategies (i.e., full fine-tuning, LoRA (Hu et al., 2022) and DPO fine-tuning (Rafailov et al., 2023)) against the GRPO-based RL tuning, along with the aforementioned modification. In all above experiments, Qwen2-VL-Instruct serves as the base model. In addition to our fine-tuned models, we include two publicly available MLLMs: LLaVA-7B-v1.5 (Liu et al., 2023) and Llama-3.2-11B-Vision-

Instruct (Grattafiori et al., 2024), both of which are already trained using supervised fine-tuning.

In a medical VQA setting, GRPO-based RL tuning consistently outperforms SFT-based approaches. As shown in Tab. 1, the two public MLLMs exhibit a clear gap in accuracy and semantic alignment compared to our tuned models. Among the SFT methods, full fine-tuning achieves the highest accuracy of 52.00. However, applying GRPO-based RL tuning to the same base model results in a significant improvement in performance. These results suggest that while SFT enables the model to imitate reasoning patterns observed in data, it remains limited in its capacity to induce genuine reasoning ability. Moreover, we also observe that models lose CoT reasoning ability after undergoing SFT (examples refer to Appendix C, Fig. S5). In contrast, GRPO-based RL training allows the MLLMs to automatically explore and exploit medically meaningful reasoning behavior.

5 Conclusion

In this work, we investigate the gap between GRPO-based RL tuning and clinically grounded MLLMs. We first examine the effectiveness of instructional fine-tuning, medical semantic alignment, and unbiased GRPO in improving answer accuracy, while also critically assessing the limitations of length-based rewards in promoting long-chain reasoning. Additionally, our findings show that GRPO-based RL consistently enhances the reasoning ability of medical MLLMs. We believe that this study offers valuable insights for advancing the development of clinically meaningful MLLMs and can inform future research in the medical AI community.

6 Limitations

Despite the promising results of applying GRPO-based reinforcement learning to medical MLLMs, our study has several limitations that open avenues for future research:

1. **Scalability to Larger Datasets:** Our experiments are conducted on a subset of the medical dataset (PMC-VQA). While this setting allows for focused evaluation, it may limit the generalizability of our findings. Future work should consider expanding the study to larger and more diverse medical datasets, which may expose new challenges in model robustness and alignment with real-world clinical variability.
2. **Model Size and Capacity:** All experiments in this work are conducted using Qwen2-VL-2B, a relatively small multimodal model. While this choice ensures training efficiency and interpretability of RL dynamics, it may not fully reflect the behavior of larger foundation models. Extending GRPO-based tuning to larger-scale MLLMs (e.g., >7B parameters) could provide insights into the scalability and generalization capabilities of our empirical study.
3. **Incorporation of Expert-Labeled Chain-of-Thought (CoT) Data:** Currently, our approach does not leverage any expert-labeled reasoning traces or CoT annotations dataset. An interesting direction for future work is to explore the integration of CoT data, either through pretraining or cold-start initialization, to enhance models' reasoning capabilities before RL tuning. This could help bridge the gap between language alignment and step-wise clinical logic.
4. **Reasoning Limitations SFT:** As noted in our findings, models fine-tuned via SFT alone often fail to acquire robust reasoning abilities, especially in complex medical scenarios. This highlights a fundamental limitation of supervised fine-tuning when reasoning is not explicitly annotated. Investigating alternative strategies—such as integrating reasoning-aware objectives or hybrid SFT-RL pipelines—may offer promising solutions to enhance multi-step inference in medical MLLMs.

References

- Alisson Azzolini, Hannah Brandon, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, Francesco Ferroni, Rama Govindaraju, and 1 others. 2025. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*.
- Liang Chen, Lei Li, Haozhe Zhao, Yifan Song, and Vinci. 2025. R1-v: Reinforcing super generalization ability in vision-language models with less than \$3. <https://github.com/Deep-Agent/R1-V>. Accessed: 2025-02-02.
- Kanzhi Cheng, Yantao Li, Fangzhi Xu, Jianbing Zhang, Hao Zhou, and Yang Liu. 2024. Vision-language models can self-improve reasoning via reflection. *arXiv preprint arXiv:2411.00855*.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.
- Matthew Chung. 2025. Training a qwen 2.5 model for medical reasoning with grpo: A tutorial and “aha!” moment. Accessed: 2025-05-16.
- Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkan Yang, Winston Hu, Yongming Rao, and Ziwei Liu. 2024. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. *arXiv preprint arXiv:2411.14432*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shiron Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao,

428	Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b . <i>Preprint</i> , arXiv:2310.06825.	483
429		484
430		485
431	Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip HS Torr, Salman Khan, and Fahad Shahbaz Khan. 2025. Llm post-training: A deep dive into reasoning large language models. <i>arXiv preprint arXiv:2502.21321</i> .	486
432		487
433		488
434		
435		489
436		490
437	Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. <i>arXiv preprint arXiv:2402.10373</i> .	491
438		492
439		493
440		
441		494
442	Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, and Xiaofeng Yang. 2025. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. <i>arXiv preprint arXiv:2503.13939</i> .	495
443		496
444		497
445		498
446	Yuwon Lee. 2024. Qwen2-vl-finetune .	499
447	Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. <i>Advances in neural information processing systems</i> , 36:34892–34916.	500
448		501
449		502
450		503
451	Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. Understanding r1-zero-like training: A critical perspective. <i>arXiv preprint arXiv:2503.20783</i> .	504
452		505
453		506
454		
455	INGIN LLMS. Demystifying long chain-of-thought reason.	507
456		508
457	Zhengxi Lu, Yuxiang Chai, Yaxuan Guo, Xi Yin, Liang Liu, Hao Wang, Han Xiao, Shuai Ren, Guanqing Xiong, and Hongsheng Li. 2025. Ui-r1: Enhancing action prediction of gui agents by reinforcement learning. <i>arXiv preprint arXiv:2503.21620</i> .	509
458		
459		510
460		511
461		512
462	Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. <i>Briefings in bioinformatics</i> , 23(6):bbac409.	513
463		514
464		515
465		516
466		517
467	Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. <i>Advances in Neural Information Processing Systems</i> , 36:53728–53741.	518
468		519
469		520
470		521
471		522
472		
473	John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms . <i>Preprint</i> , arXiv:1707.06347.	523
474		524
475		525
476		526
477	Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseek-math: Pushing the limits of mathematical reasoning in open language models. <i>arXiv preprint arXiv:2402.03300</i> .	527
478		528
479		529
480		530
481		531
482		
		532
		533
		534
		535
		536
		537
		538
		539
		540
		541
		542
		543
		544
		545
		546
		547
		548
		549
		550
		551
		552
		553
		554
		555
		556
		557
		558
		559
		560
		561
		562
		563
		564
		565
		566
		567
		568
		569
		570
		571
		572
		573
		574
		575
		576
		577
		578
		579
		580
		581
		582
		583
		584
		585
		586
		587
		588
		589
		590
		591
		592
		593
		594
		595
		596
		597
		598
		599
		600
		601
		602
		603
		604
		605
		606
		607
		608
		609
		610
		611
		612
		613
		614
		615
		616
		617
		618
		619
		620
		621
		622
		623
		624
		625
		626
		627
		628
		629
		630
		631
		632
		633
		634
		635
		636
		637
		638
		639
		640
		641
		642
		643
		644
		645
		646
		647
		648
		649
		650
		651
		652
		653
		654
		655
		656
		657
		658
		659
		660
		661
		662
		663
		664
		665
		666
		667
		668
		669
		670
		671
		672
		673
		674
		675
		676
		677
		678
		679
		680
		681
		682
		683
		684
		685
		686
		687
		688
		689
		690
		691
		692
		693
		694
		695
		696
		697
		698
		699
		700
		701
		702
		703
		704
		705
		706
		707
		708
		709
		710
		711
		712
		713
		714
		715
		716
		717
		718
		719
		720
		721
		722
		723
		724
		725
		726
		727
		728
		729
		730
		731
		732
		733
		734
		735
		736
		737
		738
		739
		740
		741
		742
		743
		744
		745
		746
		747
		748
		749
		750
		751
		752
		753
		754
		755
		756
		757
		758
		759
		760
		761
		762
		763
		764
		765
		766
		767
		768
		769
		770
		771
		772
		773
		774
		775
		776
		777
		778
		779
		780
		781
		782
		783
		784
		785
		786
		787
		788
		789
		790
		791
		792
		793
		794
		795
		796
		797
		798
		799
		800
		801
		802
		803
		804
		805
		806
		807
		808
		809
		810
		811
		812
		813
		814
		815
		816
		817
		818
		819
		820
		821
		822
		823
		824
		825
		826
		827
		828
		829
		830
		831
		832
		833
		834
		835
		836
		837
		838
		839
		840
		841
		842
		843
		844
		845
		846
		847
		848
		849
		850
		851
		852
		853
		854
		855
		856
		857
		858
		859
		860
		861
		862
		863
		864
		865
		866
		867
		868
		869
		870
		871
		872
		873
		874
		875
		876
		877
		878
		879
		880
		881
		882
		883
		884
		885
		886
		887
		888
		889
		890
		891
		892
		893
		894
		895
		896
		897
		898
		899
		900
		901
		902
		903
		904
		905
		906
		907

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

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

A Implementation Details

For distributed training, we employed DeepSpeed with ZeRO Stage 2 and bfloat16 mixed precision. All experiments were conducted on a single machine equipped with four NVIDIA A100 GPUs, each with 80GB of memory. We launched four training processes using the standard DeepSpeed multinode launcher and disabled offloading for both optimizer and model parameters. The environment was configured for local execution with static rendezvous, no CPU fallback, and no TPU usage. The main training function was set to main, and communication was established via port 44326.

We used either the Qwen2-VL-2B or Qwen2-VL-2B-Instruct model as the backbone, trained on a multimodal subset of the PMC-VQA dataset. The input resolution was limited to 401,408 pixels, with a maximum prompt length of 1024 tokens. Training was conducted for two epochs using a per-device batch size of 1 and gradient accumulation over 2 steps. Mixed-precision training with bfloat16 was enabled, while gradient checkpointing was disabled. Flash Attention 2 was used for efficient attention computation. Logging was performed at every step, and checkpoints were saved every 100 steps.

The model was trained for a total of 1500 steps with a learning rate of 1×10^{-6} and a temperature of 1.0. To facilitate Guided Response Preference Optimization (GRPO), we set the maximum response length to 700 tokens. At each optimization step, 8 responses were sampled, and a KL divergence coefficient of 0.04 was applied to regularize training.

Table S2: Hyper-parameters

Setting	Value
Batch Size per Device	1
Gradient Accumulation Steps	2
Training Steps	1500
Learning Rate	1×10^{-6}
Temperature	1.0
Maximum Response Length	700
Number of Responses per GRPO Step	8
KL Coefficient	0.04

B Evaluation Metrics

To comprehensively assess the quality of the model’s intermediate reasoning, we employ a di-

verse set of metrics, including similarity score, perplexity, thinking reward, and reasoning token length.

B.1 Similarity Score

Similarity Score measures how well the model’s reasoning aligns with the reference answer. We compute the semantic similarity between the reasoning and the ground-truth answer using a pre-trained cross-encoder model (cross-encoder/stsb-roberta-base). This captures whether the model’s internal reasoning is semantically consistent with the correct final answer.

B.2 Perplexity Score

Perplexity evaluates the fluency and linguistic quality of the model’s reasoning. We compute perplexity over the reasoning using a pretrained biomedical language model (microsoft/biogpt), following a standard left-to-right likelihood estimation. This metric captures how coherent and well-formed the reasoning appears from a language modeling perspective. Lower perplexity indicates more fluent, consistent, and syntactically stable reasoning.

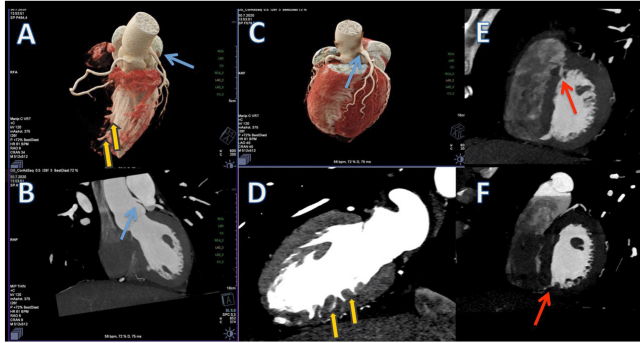
B.3 Thinking Reward

Thinking Reward assesses the usefulness and relevance of the model’s reasoning content. To compute this score, we prompt a pretrained language model (mistral-7B-instruct) with the question, reference answer, and the reasoning generated from model, and ask it to assign a score from 1 to 10 based on the quality of the reasoning. This provides an external evaluation of whether the reasoning meaningfully contributes to answering the question.

B.4 Thinking Token Length

Thinking Token Length quantifies the length of the model’s reasoning by computing the total number of tokens. This metric provides a basic measure of reasoning verbosity and helps analyze the relationship between reasoning length and quality.

C Visualization



Question: What does the Cardiac CT 3D rendering in image A demonstrate? Choose between the following options: A: High take-off LMCA and multiple crypts in the interventricular septum, B: High take-off LMCA, C: High take-off LMCA from another view, D: Multiple crypts in the inferior wall of the left ventricle

GRPO Training From scratch Qwen2-VL-2B

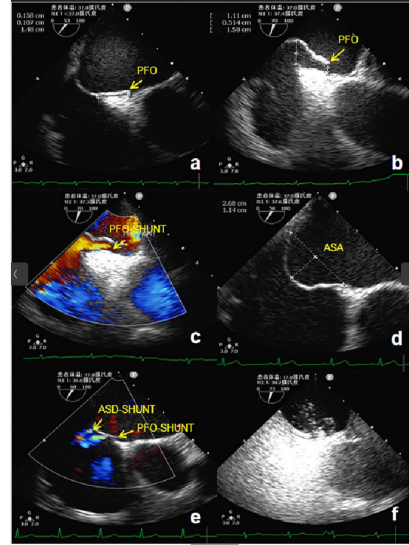
"<think>\nTo determine the correct answer, let's analyze the Cardiac CT 3D rendering in image A step by step:\n\n1. The image shows a 3D rendering of the heart.\n2. The blue arrows point to the left main coronary artery (LMCA).\n3. The yellow arrows point to the interventricular septum.\n4. The red arrow points to the inferior wall of the left ventricle.\n\nBased on this analysis, the correct answer is:\n\nC: High take-off LMCA from another view\n</think>\n\n"

<answer>
High take-off LMCA from another view
</answer>

GRPO Fine-tuning From Qwen2-VL-2B-Instruct

"<think>\nTo determine the correct answer, let's analyze the cardiac CT 3D rendering in image A. The image shows a 3D rendering of the heart with a blue arrow pointing to a high take-off left main coronary artery (LMCA) and a yellow arrow pointing to multiple crypts in the interventricular septum.\n\n</think>\n\n"

<answer>
High take-off LMCA and multiple crypts in the interventricular septum
</answer>



Question: What is the image (d) showing? Choose between the following options: A: Normal heart valves, B: PFO during Valsalva maneuver, C: ASA, D: ASD "

GRPO Training From scratch Qwen2-VL-2B

"<think>\nTo determine what image (d) is showing, let's analyze the image step by step:\n\n1. The image shows a transesophageal echocardiogram (TEE) view.\n2. The image shows a cross-sectional view of the heart.\n3. The image shows a color Doppler flow pattern, which is typically used to visualize blood flow.\n4. The image shows a small opening, which is characteristic of a patent foramen ovale (PFO).\n5. The image shows a small area of increased color flow, which is characteristic of a patent foramen ovale shunt (PFO-shunt).\n\nBased on these observations, the image (d) is showing a patent foramen ovale during a Valsalva maneuver.\n</think>\n\n"

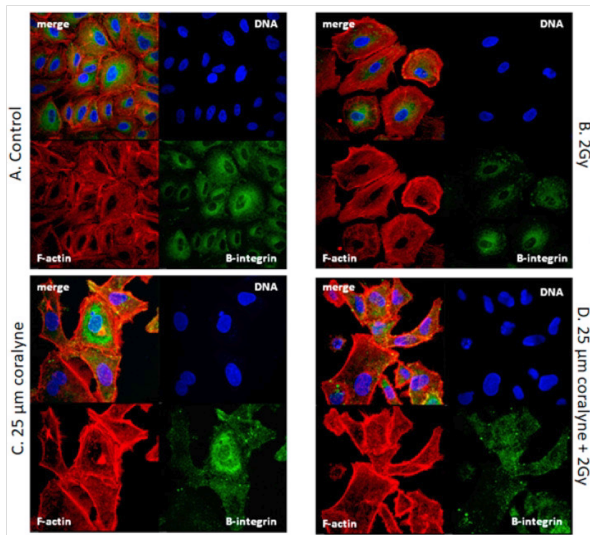
<answer>
B: PFO during Valsalva maneuver
</answer>

GRPO Fine-tuning From Qwen2-VL-2B-Instruct

"<think>\nTo determine what image (d) is showing, let's analyze the information provided in the image. The image shows a transesophageal echocardiogram (TEE) view, and the highlighted areas indicate the presence of a patent foramen ovale (PFO) and an atrial septal defect (ASD).\n\n</think>\n\n"

<answer> ASA **</answer>**

Figure S2: Visual comparison of reasoning outputs on two medical imaging questions. Red highlights indicate incorrect answers, while green highlights indicate correct answers. Although training Qwen2-VL-2B from scratch with GRPO-based RL tuning model generates longer sequences, its reasoning is often redundant and inaccurate. GRPO-based RL tuning based on Qwen2-VL-2B-Instruct produces more concise and clinically accurate reasoning, leading to correct answers.



Question: Which instrument was used to obtain the images? Choose between the following options:

- A: Fluorescence microscope ,
- B: Transmission electron microscope ,
- C: Scanning electron microscope ,
- D: Confocal microscope

GRPO

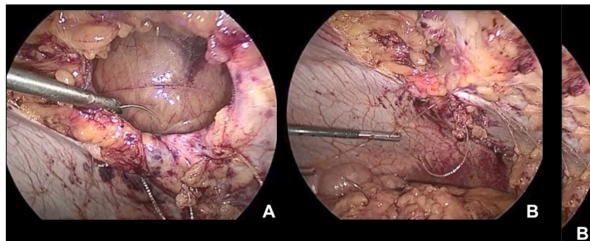
"<think>To determine the type of instrument used to obtain the images, let's analyze the images and the staining methods used.</think>

<answer> Confocal microscope </answer>"

GRPO + Medical Alignment

"<think>The images are fluorescence microscopy images, which are typically obtained using a Confocal microscope.</think>

<answer>Confocal microscope</answer>"



GRPO

"<think>\n\nTo determine the name of the structure being closed in the image, let's analyze the image step by step.\n\n</think>\n\n

<answer> Barbed suture. </answer>"

Question: What is the name of the structure being closed in the image? Choose between the following options:

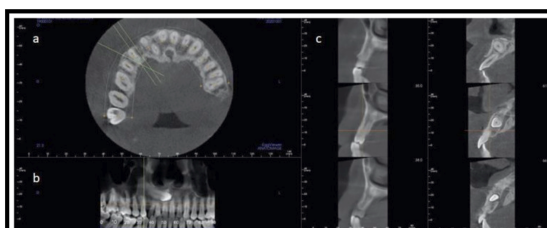
- A: Incision site. ,
- B: Abdominal wall. ,
- C: Fascial defect. ,
- D: Barbed suture.

GRPO + Medical Alignment

"<think>The image shows a surgical procedure involving the abdominal cavity, where a suture is being used to close a defect in the abdominal wall. The suture is being used to close a fascial defect, which is a common surgical procedure in abdominal surgery.</think>

<answer>Fascial defect.</answer>"

Figure S3: Examples showing that medical alignment improves visual reasoning. Correct answers are shown in green, incorrect in red, and medical knowledge is highlighted in yellow. With medical alignment, the model produces more accurate and informed responses by grounding its reasoning in domain-specific knowledge.



Question: Which section shows the impacted canine and its relationship with the adjacent teeth?

A: Axial section , B: Image analogous to a panoramic radiography , C: Sagittal sections , D: Both A and C

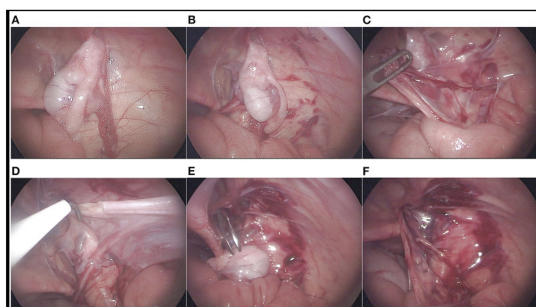
To determine which section shows the impacted canine and its relationship with adjacent teeth, let's analyze the images:

1. Axial section (a): Shows the jaw and teeth in a single view... ..useful for locating teeth in the arch.
2. Panoramic-like image (b): Provides an overview of the jaw... ..but lacks detailed spatial info.
3. Sagittal sections (c): Show side views of the teeth and bone... ..useful for depth and adjacency.
4. Both A and C (d): Combines horizontal position from (a) and depth/detail from (c)... ..

Based on this, the axial section (a) is chosen for identifying the impacted canine and its relation to adjacent teeth.

Answer: Axial section

Correct Answer: Sagittal sections



Question: What is mobilized in step C? Choose between the following options: A: Spermatic vessels proximally , B: Spermatic vessels distally , C: Gubernaculum proximally , D: Gubernaculum distally

To determine what is mobilized in step C, let's review the sequence:

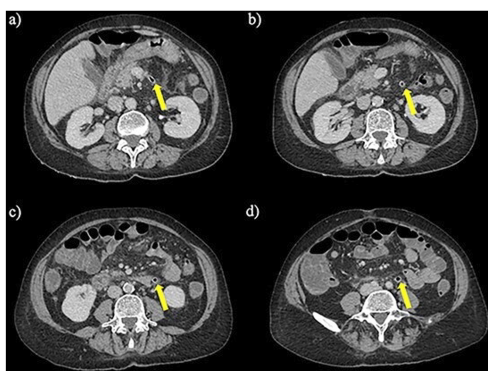
1. Step A: Shows the spermatic cord, including spermatic vessels and gubernaculum.
2. Steps B: All focus on the spermatic cord, particularly the spermatic vessels and gubernaculum.
3. Step C: The image shows a view of the spermatic cord, but it appears to be... ..
4. Step D: The image shows a view of the spermatic cord, but it appears to be more focused on the spermatic
5. Step E: The image shows a view of the spermatic cord, but it appears to be more
6. Step F: The image shows a view of the spermatic cord, but it appears to be more focused on the spermatic

From the analysis, it clear that in step C, the focus is

Answer: Spermatic vessels distally

Correct Answer: Spermatic vessels proximally

Figure S4: Examples of incorrect but verbose reasoning in long-chain answers. Although the model generates extensive intermediate thinking steps, the reasoning is often repetitive, includes irrelevant details, and ultimately leads to an incorrect answer.



Question: What area of the body is the CT scan focusing on? Choose between the following options: A:Thorax , B:Abdomen , C:Head , D:Limbs

Prompt

Output the thinking process in <think> </think> and final answer in <answer> </answer> tags.

Qwen2-VL-2B-Instruct

"<think>To determine the area of the body that the CT scan is focusing on, let's analyze the images step by step.\n\n1. **Image a**: This image shows a cross-sectional view of the abdomen.\n2. **Image b**: This image shows a cross-sectional view of the abdomen.\n3. **Image c**: This image shows a cross-sectional view of the abdomen.\n4. **Image d**: This image shows a cross-sectional view of the abdomen. All the images show a cross-sectional view of the abdomen. Therefore, the CT scan is focusing on the abdomen.</think>\n\n<answer>B: Abdomen</answer>"

SFT with LoRA Training

<answer>Abdomen</answer>

Figure S5: Comparison between the original Qwen2-VL-2B-Instruct and its LoRA fine-tuned variant. While the original model generates step-by-step visual reasoning to support its prediction, the LoRA-SFT version directly outputs the answer without any intermediate explanation.