# OPVD: On-Policy Chain-of-Visual-Thought Distillation for Medical Reasoning

Ningyue Zhang[1] and Qiushi Yang *[2]

[1]Department of Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China
[2]Department of Electrical Engineering, City University of Hong Kong, Hong Kong SAR, China

**Abstract.** Recent advancements in enhancing the medical reasoning capabilities of Vision-Language Models (VLMs) have predominantly relied on rule-based reinforcement learning (RL) paradigms. However, such approaches are often constrained by their dependence on *sparse answer-level rewards*, which fail to provide dense supervisory signals for intermediate reasoning steps, and by the *computational inefficiency inherent* in requiring extensive sampling for policy exploration. To address these limitations, we propose *On-Policy Chain-of-Visual-Thought Distillation* (OPVD) for vision-language medical reasoning, a concise yet efficient framework designed to imbue VLMs with robust medical reasoning abilities. OPVD introduces an interleaved vision-language Chain-of-Visual-Thought (CoVT) mechanism to construct reliable and granular reasoning trajectories. Unlike previous approaches, our OPVD leverages these trajectories to perform token-level dense supervision, effectively distilling multi-modal CoVT knowledge into the model via an on-policy self-distillation strategy where the teacher and student share the same policy model. This eliminates the need for external teachers or costly sampling procedures while maximizing data efficiency. Extensive experiments conducted on the public multi-modal medical reasoning dataset demonstrate that OPVD remarkably outperforms existing models trained via standard supervised fine-tuning and RL methods, establishing a new state-of-the-art in efficient and accurate medical visual reasoning.

**Keywords:** On-Policy Distillation · Medical Reasoning · Visual-Language Model · Chain-of-Thought.

## 1 Introduction

Vision-language medical reasoning [10, 16, 6] represents a cornerstone of modern clinical decision support systems, bridging the gap between complex multi-modal patient data and actionable diagnostic insights. The ability to simultaneously

* Corresponding author: qsyang2-c@my.cityu.edu.hk

interpret medical imagery, such as MRI, CT or X-ray, and correlate these visual findings with textual clinical information is critical for accurate diagnosis, treatment planning, and prognosis. As healthcare increasingly relies on artificial intelligence to augment human expertise, developing Vision-Language Models (VLMs) that possess robust, step-by-step reasoning capabilities akin to expert clinicians has become an urgent priority [10, 16, 12, 2]. Such models must not only identify abnormalities but also articulate the logical deduction process that links visual evidence to medical conclusions, thereby enhancing trustworthiness and interpretability in high-stakes clinical environments.

Recent research in vision-language medical reasoning [10, 16, 6, 15, 9, 19, 7, 8, 14, 3] has largely gravitated towards rule-based Reinforcement Learning (RL) paradigms to unlock the latent reasoning potential of VLMs. Prominent among these approaches is the application of Group Relative Policy Optimization (GRPO) and its variants [4, 17], which have demonstrated remarkable success in stimulating complex reasoning chains. For instance, MedVLM-R1 [16] adapts GRPO to medical Visual Question Answering (VQA) tasks, encouraging the model to explore diverse Chain-of-Thought (CoT) trajectories to derive correct answers. Similarly, Med-R1 [10] systematically analyzes the impact of incorporating explicit CoT processes across various medical tasks, confirming that structured reasoning significantly boosts diagnostic accuracy. ViTAR [3] extends the DeepEyes [23] paradigm by encouraging the model to explore subtle local visual regions for reasoning, optimized via GRPO. While these RL-driven methods have set new benchmarks, they predominantly rely on trial-and-error exploration. Concurrently, although a nascent body of work in general Large Language Models (LLMs) has begun to explore distillation strategies [22, 20] for efficiently enhancing reasoning, such methodologies remain largely unexplored within the specialized domain of medical reasoning, leaving a significant gap in efficient training frameworks for medical VLMs.

Despite the successes of rule-based RL methods, they suffer from two fundamental limitations that hinder their scalability and efficiency. First, these approaches typically optimize models using sparse rewards derived solely from the final answer correctness and format adherence. This lack of dense supervisory signals for intermediate reasoning steps often leads to inefficient optimization, where the model struggles to learn how to reason correctly even when it occasionally arrives at the right answer. Second, RL-based training necessitates extensive multiple sampling per instance to explore the policy space adequately. This requirement incurs prohibitive computational costs and can introduce training instability due to the high variance in reward estimates. Motivated by these challenges, we propose a paradigm shift: instead of relying on sparse feedback and exhaustive exploration, we enable the model to engage in selective self-learning during optimization. By injecting medical diagnosis-specific priors to construct reliable vision-language CoT trajectories and leveraging these trajectories for token-level dense supervision, we aim to harness the model's own capabilities more efficiently to elevate its reasoning proficiency.

To realize this vision, we present On-Policy Chain-of-Visual-Thought Distillation (OPVD), a novel framework that operationalizes dense supervision through an interleaved vision-language reasoning process. Our approach proceeds in a parallel branches within a shared-parameter policy model. First, acting as a teacher, the policy model is provided with the ground-truth answer alongside the input query and image. Guided by this prior, it generates a preliminary reasoning path and identifies critical local visual regions (e.g., subtle lesions). We then employ automated "crop-and-zoom" tools to magnify these regions, feeding them back as enhanced visual cues to generate a comprehensive, high-quality text-CoT trajectory and the final solution. Second, acting as a student, the same policy model processes the original input without answer priors. We impose an on-policy Chain-of-Visual-Thought (CoVT) distillation constraint that forces the student's generated reasoning tokens to align with the teacher's dense, step-by-step output at the token level. This strategy requires only two rounds of interleaved vision-text sampling per instance, yet effectively transfers reasoning knowledge via dense supervision, bypassing the need for costly multi-group exploration.

In summary, our main contributions are threefold:

- We propose OPVD, a novel on-policy self-distillation framework tailored for medical VLMs, which replaces sparse reward-based RL with token-level dense supervision derived from interleaved CoVT trajectories.
- We design a lightweight "crop-and-zoom" iterative reasoning mechanism that leverages ground-truth priors to construct reliable teacher trajectories, enabling efficient knowledge transfer with minimal sampling overhead compared to traditional RL methods.
- Extensive experiments on the public dataset demonstrate that OPVD outperforms both standard supervised fine-tuning and advanced RL-based competitors, establishing a new benchmark for efficiency and accuracy in multimodal medical reasoning.

## 2 Approach

### 2.1 Overview

As illustrated in Figure 1, we develop On-Policy chain-of-Visual-thought Distillation (OPVD) framework, which consists of two parallel branches instantiated from the same Vision-Language Models (VLMs): a *teacher policy model* $\pi_T$ and a *student policy model* $\pi_S$. Crucially, both policies share the same model parameters $\theta$, differing only in their conditioning contexts during training. The teacher branch acts as a guide with access to privileged information. Its inputs include the query $Q$, the original image $I$, and the ground-truth answer $A$. For cases requiring fine-grained visual understanding, a cropped and zoomed-in local visual region $L$ is also generated and fed into the model. The teacher generates a comprehensive reasoning trajectory with the final answer prediction

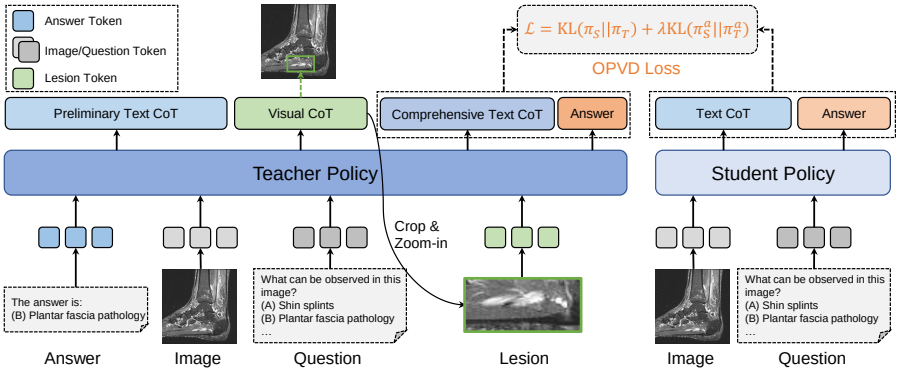


Fig. 1: Overview of OPVD, which employs a shared-parameter model acting as both teacher and student policy models. The teacher utilizes privileged answers and iterative visual zooming to construct reliable Chain-of-Visual-Thought (CoVT) trajectories. The student learns via on-policy distillation to match these trajectories without requiring external sampling or sparse rewards.

$\{\mathrm{CoT}_{\mathrm{com}}, \pi_T^a\} = \pi_T(Q, I, A, L)$. Simultaneously, the student branch operates under standard inference conditions, taking only the query $Q$ and image $I$ as input to produce its own trajectory with the predicted answer $\{\mathrm{CoT}_{\mathrm{S}}, \pi_S^a\} = \pi_S(Q, I)$. The core of our method lies in aligning the student's output with the high-quality trajectories generated by the teacher through token-level dense supervision, thereby achieving efficient knowledge transfer.

### 2.2 Chain-of-Visual-Thought (CoVT) Construction

To obtain reliable reasoning trajectories for distillation, we design a mechanism where the teacher model leverages both textual and visual priors to construct high-quality, interleaved vision-language CoT information. This process involves one or two forward passes depending on the complexity of the visual evidence. Specifically, in the first pass, the teacher receives the query $Q$, image $I$, and ground-truth answer $A$. The prompt includes a system instruction guiding the model to determine if specific local regions require re-examination. Based on this, the model generates a preliminary text CoT. If the model identifies difficult-to-discriminate regions (e.g., small lesions), it outputs the bounding box coordinates $\mathcal{B}$ for these areas. We then invoke an external tool to crop and zoom in on these regions from the original image $I$, resizing them to match the original input dimensions to form the local image $L$.

In the second pass (if triggered), $L$ is concatenated with $Q, I$, and $A$ and fed back into the teacher. This allows the model to refine its reasoning based on enhanced visual details, ultimately producing a comprehensive text CoT trajectory $\mathrm{CoT}_{\mathrm{com}}$ and the predicted answer $\pi_T^a$. By interleaving text reasoning with adaptive visual refinement, the teacher constructs a robust trajectory that serves as the dense supervision target for the student.

### 2.3 On-Policy CoVT Distillation

To enable the model to continuously self-correct and improve its reasoning capabilities, we adopt an on-policy self-distillation strategy. Unlike traditional methods that rely on static datasets or multiple stochastic samples, our approach dynamically generates targets based on the current model state.

Let the student generate a response trajectory $\mathrm{CoT_S} \sim \pi_S(\cdot \mid Q, I)$. Both the teacher and student policies then evaluate this student-generated trajectory. At each token position $n$, they induce next-token distributions conditioned on the same prefix $\mathrm{CoT_S^{<n}} = (\mathrm{CoT_S^1}, \ldots, \mathrm{CoT_S^{n-1}})$. However, their conditioning contexts differ:

$$\pi_S(\mathrm{CoT_S^n} \mid Q, I, \mathrm{CoT_S^{<n}}) \quad \text{and} \quad \pi_T(\mathrm{CoT_S^n} \mid Q, I, A, L, \mathrm{CoT_S^{<n}}), \tag{1}$$

where $\pi_S(\cdot|\cdot)$ and $\pi_T(\cdot|\cdot)$ refers to the probability distributions over the vocabulary $\mathcal{V}$ for the student and teacher, respectively. Note that while the teacher has access to the privileged answer $A$ and local image $L$ to guide the generation, it evaluates the *student's* partial generation $\mathrm{CoT_S^{<n}}$ rather than its own optimal path. This setup encourages the teacher to provide informed guidance on how to correct or continue the student's current reasoning step, effectively creating a dynamic curriculum. The learning objective minimizes the divergence between these two distributions along the student's rollout, allowing gradients to backpropagate only through the student's logits.

### 2.4 Optimization and Inference

During the optimization phase, we calculate the full-vocabulary divergence distillation loss for both the reasoning trajectory and the final answer prediction. To ensure the model prioritizes diagnostic accuracy, we apply a weighting coefficient to the loss term associated with the final answer. The total loss function $\mathcal{L}_{total}$ is defined as:

$$\mathcal{L}_{\mathrm{OPVD}} = \mathrm{KL}(\pi_S(\cdot \mid Q, I) \parallel \pi_T(\cdot \mid Q, I, A, L)) + \lambda \mathrm{KL}(\pi_S^a \parallel \pi_T^a), \tag{2}$$

where $\mathrm{KL}(\cdot||\cdot)$ denotes the Kullback-Leibler divergence, $\pi_S^a$ and $\pi_T^a$ represent the distributions over the final answer tokens, and $\lambda$ is a hyperparameter controlling the weight of the answer alignment. Note that gradients are backpropagated only to the student policy for update, while the teacher policy shares parameters with the student initially but remains frozen throughout training.

At the inference stage, only the student policy is utilized. Given an input image and query, the model performs a single forward pass to generate the chain-of-thought reasoning and the final diagnosis. Since the teacher branch and the iterative zooming mechanism are only active during training, our method introduces no additional computational overhead or latency during deployment compared to standard VLM inference.

Table 1: Results of VQA-VLMs on MRI (in-domain), and CT and X-Ray (out-of-domain) modalities.

| Methods | Sampling Num. | In-Domain / Out-of-Domain | | | Average |
|---|---|---|---|---|---|
| | | MRI | CT | X-ray | |
| *Zero-Shot VLM* | | | | | |
| Qwen2-VL-2B [18] | - | 61.67 | 50.67 | 53.00 | 55.11 |
| Qwen3-VL-2B [1] | - | 67.28 | 55.80 | 58.62 | 60.57 |
| Qwen2-VL-7B [18] | - | 72.33 | 68.67 | 66.63 | 69.21 |
| Qwen2-VL-72B [18] | - | 68.67 | 60.67 | 72.33 | 67.22 |
| Huatuo-GPT-vision-7B [2] | - | 71.00 | 63.00 | 73.66 | 69.22 |
| *MRI RL trained VLM* | | | | | |
| MedVLM-R1-2B [16] | 6 | 96.83 | 72.03 | 70.30 | 79.72 |
| Med-R1-2B [10] | 4 | 96.11 | 71.26 | 70.09 | 79.15 |
| ViTAR [3] | 16 | 97.30 | 72.48 | 70.78 | 80.19 |
| *MRI Distillation trained VLM* | | | | | |
| Qwen2-VL-2B-SFT | - | 95.05 | 55.26 | 35.10 | 61.80 |
| **Qwen2-VL-2B-OPVD** | 2 | **97.84** | **73.61** | **71.44** | **80.96** |
| w/o CoVT | 1 | 97.25 | 72.33 | 70.72 | 80.10 |
| Qwen3-VL-2B-SFT | - | 96.80 | 56.76 | 36.74 | 63.43 |
| **Qwen3-VL-2B-OPVD** | 2 | **98.54** | **73.77** | **72.13** | **81.48** |
| w/o CoVT | 1 | 98.32 | 73.10 | 71.56 | 80.99 |

## 3 Experiments

### 3.1 Datasets and Setup

***Dataset.*** Following the experimental protocol established by MedVLM-R1 [16], we conduct our evaluations on the publicly available HuatuoGPT-Vision dataset [2]. This benchmark aggregates data from several prominent medical VQA sources, including VQA-RAD [11], SLAKE [13], PathVQA [5], OmniMedVQA [6], and PMC-VQA [21], comprising over 17,300 multiple-choice questions across diverse imaging modalities. Since MedVLM-R1 did not release its specific training splits, we construct our own subsets focusing on radiology modalities (CT, MRI, and X-ray). Specifically, we selected 2,000 MRI image-question pairs for training. For evaluation, we reserved a total of 900 pairs: 300 MRI, 300 CT, and 300 X-ray samples. In this setup, the MRI test set serves as the in-domain evaluation benchmark, while the CT and X-ray sets are utilized to assess out-of-distribution (OOD) generalization capabilities.

***Implementation details.*** We employ Qwen2-VL-2B [1] and Qwen3-VL-2B as the VLM backbones for all experiments. The models are trained using the PyTorch framework on NVIDIA A100 GPUs for 3,000 iterations. Key hyper-parameters include an initial learning rate of $1e-6$, a global batch size of 64, and a distillation weight $\lambda$ set to 5.0. To optimize computational efficiency during the calculation of KL divergence, we restrict the computation to the top 128 vocabulary tokens with the highest probabilities predicted by the student model.

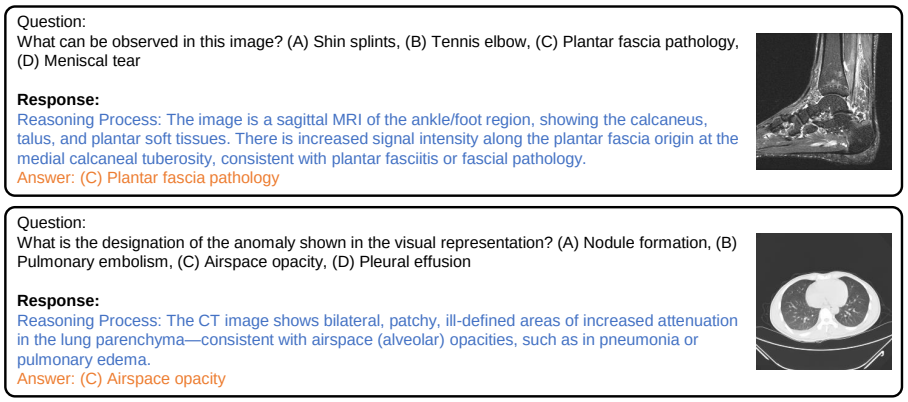


Fig. 2: Example of medical VQA generated by OPVD, highlighting its ability to generate reliable, interpretable reasoning and accurate answers.

***Prompt template and evaluation metrics.*** In the CoVT construction phase, the teacher policy model receives a structured text prompt (illustrated in Fig. 3) containing the input image, the query, and the ground-truth answer as a prior. For instance, given an image and the query "What can be observed in this image? (A) Shin splints, (B) Tennis elbow, (C) Plantar fascia pathology, (D) Meniscal tear" along with the ground truth "The answer is: (C) Plantar fascia pathology" the model is guided to generate the intermediate reasoning trajectory. For evaluation, we use standard multiple-choice accuracy: 1 point for the correct option letter, 0 otherwise; results are reported as correct ratio.

### 3.2 Comparison Results

We evaluate the in-domain and out-of-domain (OOD) generalization performance by training on MRI data and testing on MRI, CT, and X-ray datasets. As shown in Table 1, our OPVD framework outperforms larger zero-shot models. Specifically, when fine-tuning Qwen2-VL-2B, our method achieves accuracies of 97.84%, 73.61%, and 71.44% on the MRI, CT, and X-ray test sets, respectively. This surpasses the second-best comparison, ViTAR, by margins of 0.54%, 1.13%, and 0.66%. Notably, while ViTAR requires 16 sampling iterations per step, incurring eight times our computational cost, OPVD maintains superior efficiency with minimal sampling. Furthermore, consistent improvements over standard SFT are observed across both Qwen2-VL-2B and Qwen3-VL-2B backbones, validating the model-agnostic nature of our approach. These results confirm that our on-policy strategy for structuring visual-language thought processes effectively enhances reasoning capabilities with high data efficiency. From Fig. 2, we show two examples of medical VQA generated by our OPVD, which demonstrates that our method can produce reliable, interpretable reasoning processes and yield correct answers.

**Text Prompt Template**

Question:
What can be observed in this image? (A) Shin splints, (B) Tennis elbow, (C) Plantar fascia pathology, (D) Meniscal tear

The ground truth answer is: (C) Plantar fascia pathology

Analyze the provided image, question, and ground truth answer. First, deduce the logical reasoning steps required to bridge the visual evidence with the answer.

If specific target regions in the image are ambiguous, difficult to identify, or critical for understanding the reasoning, output the bounding boxes in JSON format.

Output Format:
1. Reasoning Process: [Your step-by-step deduction]
2. Critical Regions (if any): [JSON list of bboxes or "None"]

Fig. 3: Text prompt template of the teacher policy model at the first round. Similarly, the template at the second round incorporates additional cropped and zoomed-in local regions to generate final reasoning process and the answer.

Table 2: Ablations on distillation part re-weighting term $\lambda$.

| $\lambda$ | In-Domain / Out-of-Domain | | | Average |
|---|---|---|---|---|
| | MRI | CT | X-ray | |
| 1.0 | 98.06 | 73.25 | 71.54 | 80.95 |
| 2.0 | 98.21 | 73.37 | 71.66 | 81.08 |
| 5.0 | 98.54 | 73.77 | 72.13 | 81.48 |
| 10.0 | 98.09 | 73.27 | 71.57 | 80.98 |

### 3.3 Ablation Studies

***Effectiveness of chain-of-visual-thought.*** To investigate the contribution of local visual cues within the Chain-of-Visual-Thought (CoVT), we conduct an ablation study by removing the visual zoom-in mechanism. In this variant, the teacher policy generates a preliminary text-only CoT based solely on the answer context for distillation. As detailed in Table 1 (w/o CoVT), upon Qwen2-VL-2B, eliminating the visual component resulted in performance drops of 0.59%, 1.28%, and 0.72% on the MRI, CT, and X-ray test sets, respectively. This degradation underscores the critical role of interleaved visual information in constructing reliable reasoning trajectories, thereby boosting overall inference accuracy.

***Influence of answer re-weighting term $\lambda$.*** We examine the impact of the hyperparameter $\lambda$, which balances the distillation loss between the reasoning process and the final answer. Experiments were conducted with $\lambda \in \{1.0, 2.0, 5.0, 10.0\}$, as summarized in Table 2. Optimal performance is achieved at $\lambda = 5.0$. Deviating from this value proved detrimental: insufficient weighting ($\lambda < 5.0$) causes objective misalignment, while excessive emphasis on the answer ( $\lambda$>5.0) impairs the model's ability to learn robust intermediate reasoning steps. This suggests that a balanced focus on both process and outcome is essential for effective knowledge distillation.

## 4 Conclusion

In this work, we propose OPVD, an efficient framework that addresses sparse rewards and high computation in RL-based medical VLMs. By using interleaved vision-language trajectories for token-level dense supervision via on-policy self-distillation, OPVD enables robust reasoning—without external teachers or heavy sampling. Experiments on public benchmarks show OPVD significantly surpasses SFT and RL baselines in efficient, accurate medical visual reasoning.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
2. Chen, J., Ouyang, R., Gao, A., Chen, S., Chen, G.H., Wang, X., Zhang, R., Cai, Z., Ji, K., Yu, G., Wan, X., Wang, B.: Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale (2024)
3. Chen, K., Rui, S., Jiang, Y., Wu, J., Zheng, Q., Song, C., Wang, X., Zhou, M., Liu, M.: Think twice to see more: Iterative visual reasoning in medical vlms. arXiv preprint arXiv:2510.10052 (2025)
4. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
5. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
6. Hu, Y., Li, T., Lu, Q., Shao, W., He, J., Qiao, Y., Luo, P.: Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In: CVPR. pp. 22170–22183 (2024)
7. Huang, X., Wu, J., Liu, H., Tang, X., Zhou, Y.: Medvlthinker: Simple baselines for multimodal medical reasoning. arXiv preprint arXiv:2508.02669 (2025)
8. Huang, Z., Mu, L., Zhu, Y., Zhao, X., Zhang, S., Zhang, X.: Elicit and enhance: Advancing multimodal reasoning in medical scenarios. arXiv preprint arXiv:2505.23118 (2025)
9. Kumar, A., Zhuang, V., Agarwal, R., Su, Y., Co-Reyes, J.D., Singh, A., Baumli, K., Iqbal, S., Bishop, C., Roelofs, R., et al.: Training language models to self-correct via reinforcement learning. arXiv preprint arXiv:2409.12917 (2024)
10. Lai, Y., Zhong, J., Li, M., Zhao, S., Li, Y., Psounis, K., Yang, X.: Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. In: IEEE TMI. IEEE (2026)
11. Lau, J.J., Gayen, S., Ben Abacha, A., et al.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 1–10 (2018)
12. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)

13. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: ISBI. pp. 1650–1654 (2021)
14. Liu, C., Wang, H., Pan, J., Wan, Z., Dai, Y., Lin, F., Bai, W., Rueckert, D., Arcucci, R.: Beyond distillation: Pushing the limits of medical llm reasoning with minimalist rule-based rl. arXiv preprint arXiv:2505.17952 (2025)
15. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. NeurIPS **35**, 27730–27744 (2022)
16. Pan, J., Liu, C., Wu, J., Liu, F., Zhu, J., Li, H.B., Chen, C., Ouyang, C., Rueckert, D.: Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In: MICCAI. pp. 337–347. Springer (2025)
17. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
18. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
19. Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)
20. Ye, T., Dong, L., Wu, X., Huang, S., Wei, F.: On-policy context distillation for language models. In: arXiv preprint arXiv:2602.12275 (2026)
21. Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W.: Pmc-vqa: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415 (2023)
22. Zhao, S., Xie, Z., Liu, M., Huang, J., Pang, G., Chen, F., Grover, A.: Self-distilled reasoner: On-policy self-distillation for large language models. In: arXiv preprint arXiv:2601.18734 (2026)
23. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing" thinking with images" via reinforcement learning. In: arXiv preprint arXiv:2505.14362 (2025)