OphVLM-R1: Efficient Ophthalmic Reasoning via Curriculum RL

Zishi Qi[1][0009−0003−6437−5426], Xiaoya Hu[1][0000−0002−0394−3931]*, Huilin Pan[1][0000−0002−3997−4246], Ang Gao[1], Jiaxin Hou[1][0009−0009−7277−0059], Jiankun Li[1][0009−0006−0480−3574], and Yongao Qian[1][0009−0006−1232−3685]

School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan, China
{qizishi, huxy, huilinpan, hust511, hou_jiaxin, lijiankun, qianya}@hust.edu.cn
*Corresponding author: huxy@hust.edu.cn

Abstract. Ophthalmic multimodal large language models face three challenges: training data lacking structured reasoning chains, single-stage training that fails to cultivate deep clinical reasoning, and large model sizes that limit deployment in resource-constrained settings. We present OphVLM-R1, a 2B-parameter model that addresses all three through a complete data–model–optimization pipeline centered on curriculum reinforcement learning. We construct OphReason-Vision, a corpus of 15,418 Chain-of-Thought reasoning trajectories from 100K+ clinical cases and 30+ public datasets, verified through automated quality control and expert-collaborative review achieving inter-rater agreement Cohen's $\kappa = 0.82$. Training proceeds in two stages: LoRA-based supervised fine-tuning followed by progressive curriculum reinforcement learning using Group Sequence-level Policy Optimization (GSPO) across four diagnostic tasks of increasing complexity. The curriculum follows the clinical diagnostic pathway from perceptual lesion localization through multi-image comparison to long-form clinical reasoning, augmented by hard-sample dynamic backtracking for long-tail difficult cases. Methodologically, we demonstrate that curriculum RL grounded in clinical pathways yields substantial improvements: within the 2B parameter class, OphVLM-R1 outperforms InternVL3.5-4B across all benchmarks with an average accuracy of 56.41% versus 51.95%, and exceeds the domain-adapted OphthaReason-Qwen-3B at 54.11%. On out-of-domain OmniMedVQA-Eye, our model achieves 88.24%, comparable to the off-the-shelf 7B model Lingshu-7B at 87.42% , and scores 42.58% on Fundus-MMBench. Ablation studies confirm that curriculum reinforcement learning yields +26.21 percentage points over supervised fine-tuning alone, with progressive ordering and sequence-level optimization each contributing independently. Our project can be found at https://qizishi.github.io/OphVLM-R1.

Keywords: Ophthalmic Multimodal Large Language Model · Curriculum Reinforcement Learning · Chain-of-Thought Reasoning

1 Introduction

Ophthalmic diseases are among the leading causes of visual impairment worldwide, with approximately 2.2 billion people affected and at least 1 billion cases preventable through timely diagnosis [1]. Accurate diagnosis requires specialist interpretation of multimodal imaging such as Color Fundus Photography and Optical Coherence Tomography, yet the global shortage of ophthalmologists results in high misdiagnosis rates in primary care.

Multimodal large language models have demonstrated strong visual reasoning in general [2,3] and medical [4–6] domains. In ophthalmology, models such as OphGLM [7], EyecareGPT [8], FundusExpert [9], Lingshu-7B [10], and Fundus-R1 [11] have advanced from classifiers to multimodal reasoning systems. Despite this progress, three challenges remain. First, existing datasets lack structured reasoning chains, and automated synthesis risks medical hallucinations [12]. Second, single-stage training fails to cultivate deep reasoning; when reinforcement learning is used, GRPO suffers from token-level variance in long chains [13]. Third, the best models require 7B+ parameters [9, 10], limiting deployment in resource-constrained settings.

We propose OphVLM-R1, a 2B-parameter model that addresses these challenges through curriculum reinforcement learning for medical multimodal reasoning. While curriculum learning [14] is established in supervised settings and RL has been applied to medical vision-language models, no prior work designs a curriculum grounded in clinical diagnostic pathways for reinforcement learning in ophthalmic reasoning. Our key insight is that clinical competence develops progressively—from visual perception to integrated reasoning—and training should mirror this progression. Our contributions are:

- **OphReason-Vision Dataset.** We construct a three-stage closed-loop pipeline integrating 100K+ clinical cases with 30+ public datasets, producing 15,418 structured reasoning trajectories with expert-collaborative review achieving inter-rater agreement Cohen's $\kappa = 0.82$.
- **Clinical-Pathway Curriculum RL.** We develop a two-stage training pipeline: LoRA-based cold-start fine-tuning followed by progressive curriculum reinforcement learning using Group Sequence-level Policy Optimization, which operates at the sequence level to reduce policy-ratio variance for long reasoning chains. The four-stage curriculum progresses from perceptual lesion localization through multi-image comparison to long-form clinical reasoning, following the diagnostic workflow from visual identification to knowledge-intensive decision-making.
- **Hard Sample Dynamic Backtracking.** We introduce an on-policy resampling mechanism that tracks persistently failing training samples and reintroduces them with elevated sampling probability, functioning as a prioritized experience replay tailored to long-tail clinical cases.

Within the 2B parameter class, our model outperforms InternVL3.5-4B on all three benchmarks with an average of 56.41% versus 51.95%, and exceeds the domain-adapted OphthaReason-Qwen-3B at 54.11%. On out-of-domain

OmniMedVQA-Eye, it achieves 88.24%, surpassing the off-the-shelf 7B Lingshu-7B at 87.42%.

2 Related Work

Vision-language pre-training has produced foundational architectures including CLIP [15], BLIP-2 [16], LLaVA [2,17], and Flamingo [3]. In the medical domain, LLaVA-Med [4], Med-PaLM [5, 18], and HuatuoGPT-Vision [6] have validated multimodal clinical reasoning. In ophthalmology, OphGLM [7], EyecareGPT [8], FundusExpert [9], Lingshu-7B [10], and Fundus-R1 [11] have advanced from classifiers to reasoning systems. Critically, all existing ophthalmic models rely on single-stage SFT or flat GRPO-based RL without curriculum design.

Curriculum learning [14] prescribes training in order of increasing difficulty, proven effective in RL settings [19]. GRPO [13] computes importance ratios at the token level, accumulating variance in long chains; our GSPO formulation elevates the ratio to the sequence level. We design a four-stage curriculum aligned with clinical diagnostic pathways. Evaluation frameworks include OmniMedVQA [21], EH-Benchmark [12], and Fundus-MMBench [9]; we additionally introduce OphReason-Vision as an in-domain evaluation set with structured reasoning annotations.

3 Method

This section presents the OphReason-Vision dataset construction pipeline in Section 3.1, followed by the two-stage training architecture in Section 3.2.

3.1 OphReason-Vision Dataset Construction

Existing ophthalmic multimodal training data suffers from three deficiencies: high heterogeneity across sources with incompatible formats, absence of structured reasoning chains where public datasets provide only image-level labels, and skewed difficulty distribution that overrepresents common conditions while underserving rare diseases. We address these with a three-stage closed-loop pipeline, illustrated in Figure 1, that transforms 100K+ raw clinical cases and 30+ public datasets into 15,418 verified reasoning trajectories.

The first stage, data standardization, integrates 100K+ clinical cases with 30+ public datasets using a dual-stream strategy. The text stream parses unstructured electronic medical records into standardized JSON, resolving inconsistent terminology through a curated ophthalmic synonym table. The visual stream generates detailed textual descriptions for datasets that contain only image-level labels, enriching the visual information available for reasoning synthesis.

The second stage, structured reasoning synthesis, uses Intern-S1 to generate multi-dimensional instructions covering lesion localization, multimodal diagnosis, and knowledge question answering. For each instruction, it produces

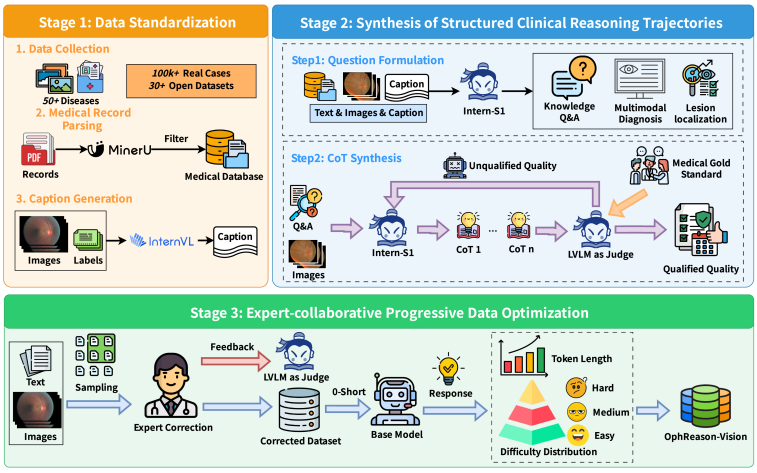

Fig. 1. Three-stage closed-loop pipeline of OphReason-Vision dataset construction.

a Chain-of-Thought reasoning chain following the clinical diagnostic workflow: visual sign identification, knowledge retrieval, pathological analysis, and clinical decision. Given input x, reasoning chain $z = (z_1, \ldots, z_n)$, and answer y, the chain probability is $P(z|x) = \prod_{t=1}^{|z|} P(z_t|x, z_{<t})$. Quality control employs an LVLM-as-a-Judge mechanism using Intern-S1 with threshold $\tau = 0.7$, determined via a pilot study on 500 expert-reviewed samples to maximize F1. The judge evaluates four dimensions: medical correctness, reasoning consistency, step completeness, and clarity, flagging key error categories including hallucinated findings, incorrect disease classification, and logically inconsistent steps. The scoring function is $\mathcal{Q} : (x, z, y) \to [0, 1]$, retaining only samples above threshold τ.

The third stage, expert-collaborative optimization, enlists three board-certified ophthalmologists to review the 18% of samples flagged as difficult, achieving inter-rater agreement Cohen's $\kappa = 0.82$. Disagreements are resolved through discussion until consensus. Difficulty grading partitions the dataset by the base model's perplexity:

$$d(x, y) = 1 - P_\theta(y|x), \quad \text{Level}(x, y) = \begin{cases} \text{Easy}, & d < \tau_1 \\ \text{Medium}, & \tau_1 \le d < \tau_2 \\ \text{Hard}, & d \ge \tau_2 \end{cases} \tag{1}$$

The final dataset contains 15,418 records summarized in Table 1: 13,418 for training and 2,000 for evaluation. The evaluation set uses held-out clinical cases with no shared patient IDs. To mitigate data contamination with the external benchmarks, we perform three checks: perceptual hashing to detect near-duplicate images, cross-referencing source dataset identifiers to exclude overlapping samples, and manual audit of shared source institutions. OmniMedVQA-

Table 1. OphReason-Vision dataset statistics.

Subset	N	Token Range	Imgs	Purpose
cold_start	3,418	1,609–2,113	1–2	Cold-start SFT
Lesion Localization	2,700	1,038–1,069	1	Stage 1
Multi-image Selection	2,700	1,115–2,240	1–2	Stage 2
Report Generation	3,600	1,012–5,985	1–5	Stage 3
Knowledge Q&A	1,000	1,012–5,985	1–5	Stage 4
eval_in_domain	2,000	1,060–1,138	1–2	Eval

Eye draws from 11 independent sources; we verify that none of these sources contribute training samples to OphReason-Vision.

3.2 OphVLM-R1 Model Architecture

OphVLM-R1 uses InternVL3.5-2B [20] as its base model. The 2B parameter count enables deployment on consumer-grade GPUs while retaining sufficient capacity for complex reasoning. Training proceeds in two stages as illustrated in Figure 2: the first injects ophthalmic domain knowledge through parameter-efficient supervised fine-tuning, and the second unlocks deep clinical reasoning through curriculum reinforcement learning with sequence-level policy optimization. The cold-start subset of 3,418 samples provides broad ophthalmic foundations, while the four curriculum subsets are ordered by increasing diagnostic complexity for progressive skill acquisition.

Stage 1 employs Low-Rank Adaptation for parameter-efficient fine-tuning, constraining the weight update to low-rank decomposition $W_0 + \Delta W = W_0 + BA$ with $h = W_0 x + \frac{\alpha}{r} BAx$, where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, $r \ll \min(d, k)$. The SFT loss is $\mathcal{L}_{\text{SFT}} = -\mathbb{E}[\sum_{t=1}^{|y|} \log \pi_\phi(y_t|x, y_{<t})]$. Key hyperparameters: LoRA rank $r = 64$, scaling factor $\alpha = 128$, applied to the W_q, W_k, W_v, and W_o attention projections. Learning rate 1×10^{-4} with cosine annealing, batch size 32, 3 epochs. Trainable parameters constitute approximately 0.5% of the total.

Stage 2 addresses GRPO instability in long reasoning chains, which stems from token-level ratio variance, optimization-reward misalignment, and variance accumulation across autoregressive steps. We adopt Group Sequence-level Policy Optimization, which computes importance ratios at the sequence level:

$$s_i(\phi) = \left(\frac{\pi_\phi(y_i|x)}{\pi_{\phi_{\text{old}}}(y_i|x)}\right)^{\frac{1}{|y_i|}} = \exp\left(\frac{1}{|y_i|}\sum_{t=1}^{|y_i|} \log \frac{\pi_\phi(y_{i,t}|x, y_{i,<t})}{\pi_{\phi_{\text{old}}}(y_{i,t}|x, y_{i,<t})}\right) \tag{2}$$

The objective uses PPO-style clipping with group-normalized advantages: $\mathcal{J}_{\text{GSPO}}(\phi) = \mathbb{E}[\frac{1}{G}\sum_{i=1}^{G} \min(s_i \hat{A}_i, \text{clip}(s_i, 1-\varepsilon, 1+\varepsilon)\hat{A}_i)]$. A key property is that $\text{Var}[\log s_i] = \sigma^2/|y_i|$ decreases with sequence length, ensuring stability for long chains.

The curriculum spans four stages of increasing complexity, each optimized with a mixed reward $r(x, y) = \lambda_1 \cdot r_{\text{rule}}(x, y) + \lambda_2 \cdot r_{\text{judge}}(x, y)$, where r_{rule} is

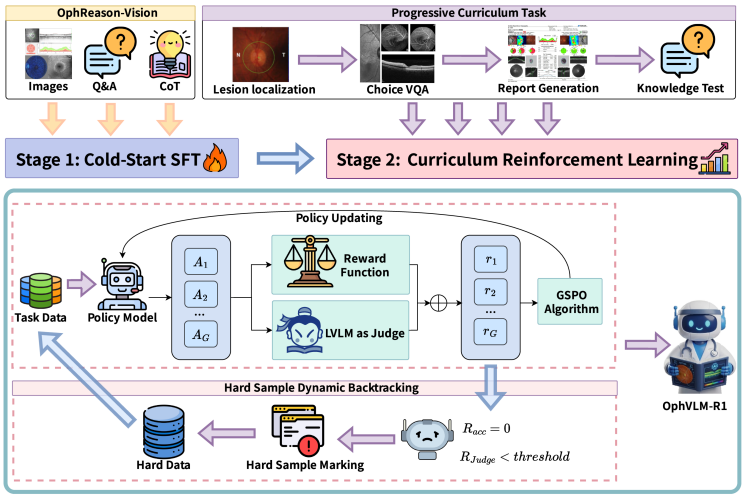

Fig. 2. Two-stage training pipeline. Stage 1 injects domain knowledge via LoRA fine-tuning. Stage 2 progressively increases task difficulty across four stages with sequence-level policy optimization.

rule-based verifiable reward and r_{judge} is an Intern-S1-mini judge reward. The stage ordering follows the clinical diagnostic pathway: lesion localization requires single-image visual perception, multi-image selection demands cross-image comparison, report generation calls for structured synthesis, and knowledge question answering necessitates integration of visual findings with clinical knowledge. This progression from perception to reasoning mirrors how ophthalmologists develop diagnostic competence. Stages are detailed in Table 1. Hyperparameters: $\varepsilon = 0.2$, $G = 8$, LR 5×10^{-6}, $\beta_{\text{KL}} = 0.04$, $\lambda_1 = 0.6$, $\lambda_2 = 0.4$. Each stage trains for 2 epochs.

To address the long tail of difficult clinical cases, an on-policy resampling mechanism increases the sampling probability of prompts whose rewards fall below a threshold in the most recent $k = 5$ rounds. The resampling probability follows $P_{\text{sample}}(x, y) \propto f(x, y)^{\beta}$, where $f(x, y)$ counts consecutive failures and $\beta = 2.0$ controls aggressiveness. The mechanism stores only prompt indices and failure counts, ensuring fresh on-policy generation for each resampled prompt. The resampling ratio is bounded to 30% of each training batch to prevent catastrophic forgetting of well-learned samples.

3.3 Ethics Statement

OphReason-Vision draws from two data sources with appropriate ethical authorization. The 30+ public datasets consist of published research with original IRB approval, used in accordance with their respective licenses. The 100K+ clinical cases were collected under IRB/Ethics Committee approval with research data

Table 2. Evaluation dataset statistics.

Dataset	Samples	Tasks	Type
In-Domain	2,000	4	Multiple-choice
Fundus-MMBench	620	31	Fine-grained fundus
OmniMedVQA-Eye	10,044	11 sources	Out-of-domain VQA

use authorization. Comprehensive de-identification was performed, including removal of protected health information (PHI), metadata scrubbing, and facial cropping to ensure patient privacy. A waiver of informed consent was approved by the IRB for this retrospective research. All expert reviewers involved in data quality assessment are board-certified ophthalmologists who participated under institutional review protocols.

4 Experiments

4.1 Experimental Setup

We evaluate across three complementary dimensions: the In-Domain benchmark comprising 2,000 held-out clinical cases from OphReason-Vision, Fundus-MMBench [9] with 620 samples spanning 31 fine-grained fundus analysis tasks, and OmniMedVQA-Eye [21] with 10,044 ophthalmic samples from 11 independent open-source datasets serving as the out-of-domain evaluation. Table 2 summarizes these datasets. In the following, we abbreviate Fundus-MMBench as Fundus and OmniMedVQA-Eye as Omni-Eye.

We compare against eight models spanning four categories. Generic MLLMs include InternVL3.5-2B and InternVL3.5-4B [20]. Medical RLVR models include MedVLM-R1-2B [11]. Large medical MLLMs include Lingshu-7B [10] and HuatuoGPT-Vision-7B [6]. Ophthalmic specialists include FundusExpert-8B [9], OphthaReason-Intern-2B, and OphthaReason-Qwen-3B [13]. The 7B/8B baselines are evaluated off-the-shelf without domain-specific fine-tuning; this asymmetric comparison confounds data exposure with parameter scale. Same-architecture/same-scale comparisons (e.g., vs. InternVL3.5-4B and domain-adapted OphthaReason-Qwen-3B) are less confounded. Training hardware consists of 8 NVIDIA GeForce RTX 4090 GPUs with 24 GB VRAM each. The supervised fine-tuning stage uses 4 GPUs; the reinforcement learning stage uses 6 GPUs for training and 2 for vLLM-based rollout generation. We use ms-swift with DeepSpeed ZeRO-3 and AdamW optimizer, and EvalScope for evaluation.

4.2 Main Results

Table 3 reports accuracy across all three benchmarks. The following paragraphs analyze these results from complementary perspectives.

The three benchmarks exhibit markedly different absolute scores, reflecting their distinct characteristics. OmniMedVQA-Eye uses binary yes/no questions

Table 3. Main experimental results across three benchmarks (Avg. is reference-only). Best in **bold**, second underlined.

Model	In-Domain	Fundus	Omni-Eye	Avg.*
Generic MLLMs				
InternVL3.5-2B [20]	34.50	36.61	55.47	42.19
InternVL3.5-4B [20]	36.23	42.10	77.51	51.95
Medical MLLMs (RLVR)				
MedVLM-R1-2B [11]	27.80	20.81	68.06	38.89
Large Medical MLLMs (SFT, off-the-shelf)				
Lingshu-7B [10]	**44.20**	41.29	87.42	**57.64**
HuatuoGPT-Vision-7B [6]	38.30	28.06	71.78	46.05
Ophthalmic MLLMs (SFT, off-the-shelf)				
FundusExpert-8B [9]	31.20	**54.84**	64.71	50.25
Ophthalmic MLLMs (RLVR)				
OphthaReason-Intern-2B [13]	31.00	35.48	79.61	48.70
OphthaReason-Qwen-3B [13]	36.60	38.87	86.86	54.11
OphVLM-R1-2B (Ours)	38.40	42.58	**88.24**	56.41

*Average across benchmarks is for reference only due to differing task scales, evaluation formats, and random baselines; per-benchmark results are primary.

across 11 sources with a 50% random baseline, yielding high absolute scores. The in-domain benchmark employs four-option clinical reasoning questions with a 25% random baseline, testing deeper diagnostic chains. Fundus-MMBench spans 31 specialized tasks with varying difficulty. These differences in evaluation format and task complexity mean that absolute scores are not directly comparable across benchmarks; within-benchmark ranking is the meaningful comparison.

On the out-of-domain OmniMedVQA-Eye benchmark, our 2B model achieves 88.24%, surpassing domain-adapted OphthaReason-Qwen-3B at 86.86% and the off-the-shelf 7B Lingshu-7B at 87.42%. Same-architecture and same-scale comparisons confirm the advantage of curriculum training: our model outperforms InternVL3.5-4B (same architecture, 2× parameters) by +4.46% average, and domain-adapted OphthaReason-Qwen-3B by +2.30% average. On Fundus-MMBench, our model achieves 42.58%, slightly above the off-the-shelf Lingshu-7B at 41.29% while trailing the 8B FundusExpert at 54.84%, which employs a specialized positioning-diagnosis architecture. On in-domain evaluation, Lingshu-7B leads at 44.20% owing to its larger capacity, while our 2B model achieves 38.40%. The strong out-of-domain performance demonstrates that curriculum RL enhances generalizable reasoning patterns rather than overfitting to the training distribution.

Our 2B model outperforms the general-purpose InternVL3.5-4B on all three benchmarks with an average of 56.41% versus 51.95%, demonstrating that

Table 4. Ablation study results. ΔOmni shows the drop from the full model on Omni-Eye.

Configuration	In-Domain	Fundus	Omni-Eye	ΔOmni
OphVLM-R1 (Full)	**38.40%**	**42.58%**	**88.24%**	—
Training Strategy				
SFT Only	37.52%	34.47%	62.03%	−26.21
SFT + RL One-shot	37.96%	38.62%	78.14%	−10.10
SFT + RL Shuffled	37.73%	37.85%	76.48%	−11.76
Stage Ablation				
w/o Stage 1 (Lesion Loc.)	38.14%	40.43%	85.62%	−2.62
w/o Stage 2 (Multi-image)	38.02%	40.17%	85.13%	−3.11
w/o Stage 3 (Report Gen.)	37.88%	39.72%	84.38%	−3.86
w/o Stage 4 (Knowledge QA)	38.07%	40.31%	85.47%	−2.77
Component Ablation				
w/ GRPO (token-level)	37.84%	39.76%	84.52%	−3.72
w/o Hard Sample BT	38.11%	40.83%	86.12%	−2.12

domain-specific curriculum training can compensate for a 2× parameter deficit. Compared to MedVLM-R1-2B, which uses GRPO-based RL, our model achieves +30.18% on Omni-Eye, validating the advantage of curriculum design. The in-domain to out-of-domain gap is only +4.01%, the smallest among all models. General-purpose models exhibit negative transfer while medical SFT models show large gaps, indicating that curriculum RL effectively balances specialization and generalization.

To characterize remaining failure modes, two ophthalmology residents reviewed 200 randomly sampled errors with inter-rater agreement $\kappa = 0.81$. Perceptual errors account for 35%, reasoning errors for 45%, and knowledge gaps for 20%. Since curriculum RL targets reasoning errors, this distribution is consistent with the observed out-of-domain improvement.

4.3 Ablation Studies

We conduct ablation studies to validate each component's contribution. Table 4 presents the full results, and the following paragraphs interpret the key findings.

The ablation reveals that curriculum reinforcement learning is essential. The SFT-only baseline trains on all 13,418 samples for the same total number of epochs as the full pipeline, but without curriculum staging or RL, achieving 62.03% on Omni-Eye versus 88.24% for the full pipeline. The large RL gain on OmniMedVQA-Eye partly reflects its binary evaluation format, which provides dense reward signals that RL exploits more effectively than SFT; the gains on the four-option in-domain benchmark and the multi-task Fundus-MMBench are smaller but consistent. One-shot RL training on the full corpus yields 78.14%,

and a shuffled curriculum yields 76.48%, both far below the ordered progressive curriculum, confirming that difficulty ordering is critical.

All four curriculum stages are indispensable. Stage 3 contributes the most with a −3.86% drop on Omni-Eye when removed, reflecting the importance of long-form clinical reasoning. Stage 2 follows at −3.11%, indicating multi-image reasoning is critical for comprehensive diagnosis. Sequence-level optimization and hard sample backtracking each contribute independently: replacing GSPO with standard token-level GRPO reduces Omni-Eye by 3.72%, and removing hard sample backtracking causes a −2.12% drop, validating both mechanisms.

5 Discussion

5.1 Limitations

We identify several limitations. The 7B and 8B baselines are evaluated off-the-shelf without domain-specific fine-tuning, while our model is trained on OphReason-Vision, confounding comparisons: performance differences may reflect data exposure rather than architectural advantage. Same-architecture/same-scale comparisons are less confounded. All results are from single runs without confidence intervals or significance tests; while key effect sizes are large, we cannot claim statistical significance without multi-seed replication. The 0.82 percentage point margin over Lingshu-7B on OmniMedVQA-Eye should be interpreted with caution.

The GSPO variance reduction guarantee $\mathrm{Var}[\log s_i] = \sigma^2/|y_i|$ holds strictly under i.i.d. per-token log-ratio assumptions. With correlated tokens in natural language reasoning chains, this guarantee is approximate; our +3.72% empirical improvement demonstrates practical benefit, but the underlying mechanism warrants further study. Final-answer accuracy alone cannot verify reasoning quality—we lack step-level correctness evaluation and hallucination rate measurement. A staged SFT-without-RL baseline would more cleanly disentangle curriculum and RL effects.

5.2 Future Directions

Future work includes multi-center clinical validation, multi-seed statistical replication, staged-SFT-without-RL baselines, step-level reasoning evaluation, and hallucination rate measurement using EH-Benchmark [12].

5.3 Data Availability

The OphVLM-R1 model weights and OphReason-Vision dataset are publicly available at https://qizishi.github.io/OphVLM-R1.

6 Conclusion

We have presented OphVLM-R1, a 2B-parameter ophthalmic reasoning model trained through a complete pipeline of curated reasoning data (OphReason-Vision, 15,418 trajectories), supervised fine-tuning, and curriculum reinforcement learning grounded in clinical diagnostic pathways. Our primary contribution is methodological: demonstrating that a progressive curriculum from perceptual lesion localization through multi-image comparison to long-form clinical reasoning, combined with sequence-level policy optimization (GSPO) and hard-sample dynamic backtracking, can effectively cultivate ophthalmic reasoning ability in lightweight models. Within the same architecture class, our model outperforms InternVL3.5-4B (2× parameters) by +4.46% average and domain-adapted OphthaReason-Qwen-3B by +2.30%. On out-of-domain OmniMedVQA-Eye, it achieves 88.24%, comparable to the off-the-shelf 7B Lingshu-7B at 87.42%. On Fundus-MMBench, it achieves 42.58%. Ablation studies confirm that each component—curriculum structure (+10.10% over one-shot RL), sequence-level optimization (+3.72% over token-level GRPO), and hard sample backtracking (+2.12%)—contributes measurably. We view this work as evidence that principled curriculum RL can substantially improve efficient medical multimodal reasoning. Future directions include strengthening statistical rigor through multi-seed replication, isolating RL contributions via staged-SFT-without-RL baselines, and validating reasoning quality through ophthalmologist-annotated step-level evaluation and hallucination rate measurement, ultimately enabling multi-center clinical deployment.

References

1. World Health Organization: World report on vision. https://www.who.int/publications/i/item/9789241516570, last accessed 2026/05/12
2. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. In: Advances in Neural Information Processing Systems 36 (NeurIPS 2023) (2023)
3. Alayrac, J.B., Donahue, J., Luc, P., et al.: Flamingo: a Visual Language Model for Few-Shot Learning. In: Advances in Neural Information Processing Systems 35 (NeurIPS 2022) (2022)
4. Li, C., Wong, C., Zhang, S., et al.: LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. In: Advances in Neural Information Processing Systems 36 (NeurIPS 2023) (2023)
5. Tu, T., Azizi, S., Driess, D., et al.: Towards Generalist Biomedical AI. NEJM AI **1**, AIoa2300138 (2024). https://doi.org/10.1056/AIoa2300138
6. Chen, J., Ouyang, R., Gao, A., et al.: HuatuoGPT-Vision, Towards Injecting Medical Visual Knowledge into Multimodal LLMs at Scale. arXiv preprint arXiv:2406.19280 (2024)
7. Li, W., et al.: OphGLM: Training an Ophthalmology Large Language-and-Vision Assistant based on Instructions and Dialogue. arXiv preprint arXiv:2306.12174 (2023)
8. Li, X., et al.: EyecareGPT: Boosting Comprehensive Ophthalmology Understanding with Tailored Dataset, Benchmark and Model. arXiv preprint arXiv:2504.13650 (2025)

9. Liu, X., Song, D.: Constructing Ophthalmic MLLM for Positioning-diagnosis Collaboration Through Clinical Cognitive Chain Reasoning. arXiv preprint arXiv:2507.17539 (2025)
10. Xu, W., Chan, H.P., Li, L., et al.: Lingshu: A Generalist Foundation Model for Unified Multimodal Medical Understanding and Reasoning. arXiv preprint arXiv:2506.07044 (2025)
11. Deng, Y., Wei, Q., Qian, K., et al.: Fundus-R1: Training a Fundus-Reading MLLM with Knowledge-Aware Reasoning on Public Data. arXiv preprint arXiv:2604.08322 (2026)
12. Pan, X., Bai, Y., Zou, K., et al.: EH-Benchmark: Ophthalmic Hallucination Benchmark and Agent-Driven Top-Down Traceable Reasoning Workflow. arXiv preprint arXiv:2507.22929 (2025)
13. Wu, Y., et al.: Bridging the Gap in Ophthalmic AI: MM-Retinal-Reason Dataset and OphthaReason Model toward Dynamic Multimodal Reasoning. arXiv preprint arXiv:2508.16129 (2025)
14. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th Annual International Conference on Machine Learning (ICML 2009), pp. 41–48 (2009)
15. Radford, A., Kim, J.W., Hallacy, C., et al.: Learning Transferable Visual Models From Natural Language Supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML 2021), pp. 8748–8763 (2021)
16. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In: Proceedings of the 40th International Conference on Machine Learning (ICML 2023), pp. 19730–19742 (2023)
17. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved Baselines with Visual Instruction Tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024), pp. 26296–26306 (2024)
18. Singhal, K., Azizi, S., Tu, T., et al.: Large Language Models Encode Clinical Knowledge. Nature **620**(7972), 172–180 (2023). https://doi.org/10.1038/s41586-023-06291-2
19. Matiisen, T., Oliver, A., Cohen, T., Schulman, J.: Teacher-Student Curriculum Learning. IEEE Transactions on Neural Networks and Learning Systems **31**(9), 3369–3380 (2020)
20. Chen, Z., et al.: InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024) (2024)
21. Hu, Y., Li, T., Lu, Q., et al.: OmniMedVQA: A New Large-Scale Comprehensive Evaluation Benchmark for Medical LVLM. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024), pp. 22170–22183 (2024)

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant 62573204 and 62173153.

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