OphthFlowBench: A Unified Ophthalmology Multimodal Benchmark with Workflow-Aligned Evaluation

Qiaojian Zheng[1], Xiaoling Luo[1(✉)], Meidan Ding[1], Ruli Zheng[1], Xiaoyan Dou[2], Yingying Wen[2], Chengliang Liu[3], and Linlin Shen[4,5]

[1] College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[2] Ophthalmology Department, Shenzhen Second People's Hospital, Shenzhen, China
[3] Department of Computer and Information Science, University of Macau, Macau
[4] School of Artificial Intelligence, Shenzhen University, Shenzhen, China
[5] Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen, China
✉ Corresponding author: xiaolingluoo@outlook.com

Abstract. Multimodal Large Language Models (MLLMs) demonstrate immense potential in medicine, yet existing ophthalmology benchmarks face significant limitations. They often neglect text-only clinical scenarios, restrict question formats, and lack structured evaluation systems that mirror real-world diagnostic workflows. To bridge these gaps, we introduce OphthFlowBench, the first comprehensive ophthalmology benchmark that integrates both multimodal VQA and text-only clinical QA pairs across open-ended and closed-form formats. Our benchmark features a pioneering workflow-aligned evaluation system that decomposes the clinical process into four sequential stages: visual perception, clinical interpretation, disease diagnosis, and clinical management, enabling granular, stage-by-stage assessment. Furthermore, to address the scarcity of domain-specific models, we developed Oph-RL, a specialized ophthalmic MLLM based on the Qwen3-VL-8B-Instruct architecture. Oph-RL is trained using a novel two-phase strategy comprising Supervised Fine-Tuning (SFT) and difficulty-aware Group Relative Policy Optimization (GRPO). Extensive evaluations on OphthFlowBench and the OphthalWeChat benchmark reveal that Oph-RL achieves exceptional accuracy and robust clinical utility. It markedly outperforms existing proprietary, general and medical MLLMs across multiple modalities, subspecialties, and languages, thereby paving the way for reliable ophthalmic decision-support systems. The benchmark and model are publicly available at `https://github.com/HapisQJ/OphthFlowBench`.

Keywords: MLLM · Ophthalmology benchmark · Workflow-aligned

1 Introduction

Multimodal Large Language Models (MLLMs) demonstrate immense potential across medical specialties [26,36,30]. However, their systematic evaluation

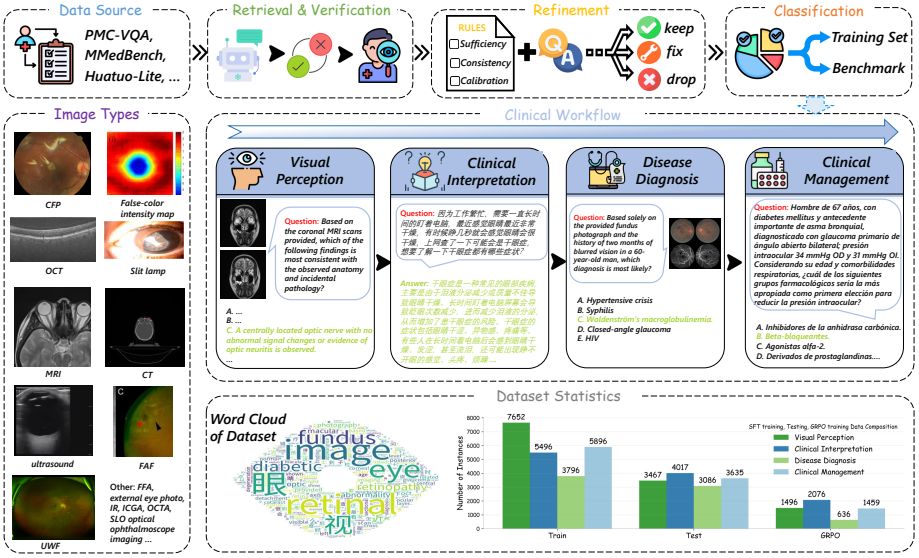


Fig. 1: **Overview of OphthFlowBench.** The upper section displays diverse data sources and the extraction process for ophthalmic data. The lower section presents the image types and examples of clinical workflows. The bottom-right corner shows a word cloud of the dataset and the composition of clinical workflow tasks across training and testing data.

and application in ophthalmology face notable limitations. Current benchmarks are highly fragmented, predominantly concentrating on either isolated VQA tasks [27,14] or text-only queries [29]. This dichotomy fails to reflect real-world clinical practice, which demands the seamless integration of diverse information formats. Furthermore, existing resources treat medical queries in isolation rather than as a continuous process, lacking a structured evaluation system that mirrors the authentic clinical diagnostic workflow. Consequently, it remains exceedingly difficult to comprehensively evaluate a model's practical clinical utility and pinpoint its specific capability deficits in realistic scenarios.

To bridge these gaps, we introduce OphthFlowBench, the first unified benchmark dedicated to comprehensive ophthalmic multimodal and textual analysis. Unlike previous datasets, it systematically integrates both multimodal VQA and text-only clinical questions across open-ended and closed-form formats. Crucially, our benchmark features a pioneering workflow-aligned assessment system to move beyond isolated task evaluations. We decompose the evaluation into four sequential stages mirroring actual diagnostic workflows [24]: visual perception, clinical interpretation, disease diagnosis, and clinical management. This structured paradigm enables a highly granular and transparent assessment, ensuring models are rigorously tested across the clinical continuum.

Although preliminary studies [11] explore the application of MLLMs to ophthalmic imaging, their performance remains unsatisfactory. Recognizing the scarcity of domain-specific models capable of navigating complex clinical workflows, we developed Oph-RL, a specialized ophthalmic MLLM tailored for comprehensive imaging and clinical analysis. We assess Oph-RL on the OphthFlowBench and OphthalWeChat [35] benchmarks. Experimental results demonstrate that Oph-RL achieves exceptional accuracy and robust clinical utility. Specifically, Oph-RL exhibits highly balanced and superior language capabilities. Furthermore, it achieves the best performance across 8 distinct ophthalmic subspecialties and markedly outperforms existing models across over 13 common imaging modalities. By bridging isolated visual tasks and continuous clinical workflows, our work paves the way for reliable ophthalmic decision-support systems.

In summary, our main contributions are as follows:

- **Dataset.** We present OphthFlowBench, the first comprehensive benchmark integrating VQA and text-only scenarios for ophthalmology, comprising 34,261 open-ended and closed-form question-answer pairs and 22,216 images extracted from high-quality medical benchmarks.
- **Evaluation.** We evaluated 23 representative MLLMs, including 3 proprietary models, 13 open-source models, and 7 medical-specific models.
- **Model.** We introduce Oph-RL, a specialized ophthalmic MLLM that demonstrates superior accuracy and clinical viability across multiple benchmarks.

2 Method

2.1 Data Curation

Data Source. To construct OphthFlowBench, we systematically curate data from a diverse array of established, high-quality medical benchmarks. Our data collection spans both multimodal and text-only formats to thoroughly evaluate varied clinical capabilities. Specifically, the VQA data is systematically sourced from Asclepius [19], GMAI-MMBench [38], MedFrameQA [39], MedXpertQA [42], OmniMedVQA [7], and PMC-VQA [40]. For the text-only modalities, closed-form questions are extracted from MedMCQA [25], MedQA [8], MedXpertQA [42], and MMedBench [28], whereas open-ended questions are derived from Huatuo-Lite [33] and PubMedQA [9]. These benchmarks encompass not only diverse ophthalmic imaging modalities but also multiple languages and question forms.

Retrieval and Verification. Given that these source benchmarks encompass the entirety of the medical field, isolating a high-purity ophthalmic subset requires a precise and scalable curation strategy. We implement a rigorous, two-stage extraction and validation pipeline. In the first phase, we utilize GPT-4o-mini as an initial semantic filter to process the massive pool of general medical queries. By leveraging its strong instruction-following capabilities, we efficiently identify and extract questions that are highly relevant to ophthalmology.

To guarantee domain accuracy of the filtered subset, we introduce a rigorous human-in-the-loop verification step. A specialized panel comprising four professional ophthalmologists systematically reviews every extracted instance. They carefully cross-verify clinical accuracy and data consistency, strictly filtering out any residual non-ophthalmic items or clinically invalid cases.

Refinement. To bridge the semantic gap between visual perception and precise textual answering, and to mitigate the hallucinations prevalent in medical VQA, we implement a data refinement phase. Existing medical VQA datasets frequently suffer from underspecified queries and granularity inconsistencies, where a vague question might validly prompt multiple semantically distinct answers. Such ambiguity not only confounds evaluation but also severely destabilizes reinforcement-based policy learning, which requires precise reward signals [20].

To address this, we employ GPT-5 to meticulously refine the extracted QA pairs. Drawing inspiration from recent VQA consistency auditing protocols [20], our refinement strategy strictly adheres to three core principles: (1) **Discriminative sufficiency**: ensuring that the question alone provides enough semantic constraints to allow exactly one answer to be selected through valid reasoning. (2) **Task-Intent consistency**: maintaining alignment between the question's semantic intent and the decision dimension represented by the answer and options. (3) **Granularity calibration**: aligning the level of semantic specificity in the question with the discriminative granularity of the answer and options. Ambiguous cases were corrected under these rules, unresolved cases were discarded.

Classification. To reflect real-world ophthalmic practice, we organize OphthFlowBench along a continuous clinical workflow comprising four sequential stages [24]. Stages are distinguished by the dominant cognitive operation required: visual perception identifies visible findings; clinical interpretation explains their clinical meaning; disease diagnosis determines the disease category; and clinical management selects diagnostic, referral, follow-up, or treatment actions. Each item is assigned to the single stage that best matches its primary reasoning objective, while text-only questions bypass visual perception. We employ GPT-5 to automatically categorize the refined dataset into these specific stages, after which four expert ophthalmologists review and correct all labels to ensure clinical validity and reproducibility.

2.2 Model and Training Strategy

To tailor the foundational multimodal capabilities to the stringent requirements of ophthalmic diagnosis, we developed Oph-RL based on the Qwen3-VL-8B-Instruct [3] architecture. A two-phase training strategy was employed to progressively enhance its multimodal understanding. The training data included ophthalmic data extracted from 11 medical benchmarks, as well as multi-view ophthalmic images and corresponding reports from a private hospital. The in-house data was processed into training-ready data by leveraging GPT-5 models and a specialized team of ophthalmologists.

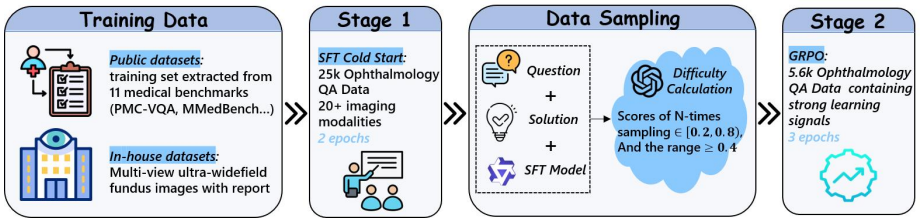


Fig. 2: Training strategy for Oph-RL

As shown in Fig. 2, in the first phase, we performed SFT on the Qwen3-VL-8B-Instruct model. The fine-tuning process used 22,840 high-quality ophthalmic question-answer pairs, encompassing the full four-stage clinical workflow. The training corpus spanned over 20 different imaging modalities and text-only question-answer data, ensuring comprehensive coverage across a wide range of ophthalmic applications.

In the second phase, we applied reinforcement learning within the GRPO framework to further enhance the decision-making ability of Oph-RL. Considering that overly simple or extremely difficult samples can generate ineffective gradient signals and hinder learning efficiency [5], we introduced a difficulty-aware data selection mechanism [6]. Specifically, for each instance, we generated N samples using the SFT-initialized model and evaluated them using a large language model acting as a judge. For a given set of scores $S = \{S_1, S_2, ..., S_N\}$, for closed-form questions, we retained those with average scores (i.e., accuracy) in the range of $[0.2, 0.8)$. For open-ended questions, we retained those with an average score in the range $[0.2, 0.8)$ and a score range of 0.4 or greater. In practice, we set $N = 5$, ultimately selecting 5,667 medium-difficulty questions from the original set of 25,624 QA pairs. In line with standard reinforcement learning reward calculation methods, both answer reward R_{answer} and format reward R_{format} were integrated into the final reward signal. GPT-5-Nano served as the reward evaluator, assigning scores between 0 and 1 for the answer. The overall reward is then determined using the following formula, which combines both answer and format contributions:

$$R_{\text{total}} = \alpha \cdot R_{\text{answer}} + \beta \cdot R_{\text{format}}, \tag{1}$$

where $\alpha + \beta = 1$. Note that for closed questions $R_{\text{answer}} = 1$ if the candidate answer matches the gold solution exactly, and $R_{\text{answer}} = 0$ if it does not. For open-ended questions, R_{answer} is further decomposed as:

$$R_{\text{answer}} = \omega_1 \cdot \text{core} + \omega_2 \cdot \text{coverage} + \omega_3 \cdot \text{consistency}, \tag{2}$$

where $\omega_1 + \omega_2 + \omega_3 = 1$, core measures whether the main conclusion or claim in the candidate's response matches the gold solution, coverage evaluates how many key points from the gold solution are present in the candidate's response, and consistency ensures the response is free from contradictions and hallucinated facts.

3 Experiments and Results

3.1 Experimental Setup

Benchmarked MLLMs. We systematically evaluated 23 representative MLLMs and Oph-RL on the OphthFlowBench benchmark: 3 proprietary models accessed via API [23,34,31], 13 open-source large models [3,21,32,17,18,1,12,22,31,37], and 7 medical MLLMs [13,36,4,30,11,26]. We also assessed the performance of Oph-RL across different dimensions on the OphthalWeChat benchmark [35].

Performance Metrics. For closed-form questions, we use accuracy. For open-ended questions on OphthFlowBench, we report ROUGE-L as a reproducible lexical-overlap metric. As a complementary semantic evaluation, the open-ended subsets of OphthalWeChat are assessed using DeepSeek-V3 [16] as an LLM judge for factual correctness.

Implementation Details. All models were evaluated using the vLLM framework and on Hugging Face transformers library [10]. Our model was initialized using Qwen3-VL-8B-Instruct, followed by SFT and GRPO. Both training stages were conducted via LoRA using the ms-swift framework [41]. All evaluation and training were conducted on a single server equipped with six NVIDIA A100 80GB GPUs. For the GRPO policy, eight candidate rationales were generated per sample with a sampling temperature of 0.8.

3.2 Evaluation on OphthFlowBench

The performance of various MLLMs on OphthFlowBench is summarized in Table 1. Overall, our model achieves the highest macro-averaged score of 0.532, which is the average of 10 stage-specific metrics. This score outperforms GPT-5.2 (0.472) and the strongest open-source baseline, Qwen3-VL-32B-Instruct (0.483), highlighting the effectiveness of in-domain post-training for workflow-aligned ophthalmic reasoning. The advantage is particularly pronounced in multimodal perception, interpretation, and diagnosis, where our model achieves accuracies of 0.740, 0.763, and 0.779, respectively. It also outperforms other models in closed-form text QA, such as diagnosis, and open-ended QA, where the ROUGE-L score reaches up to 0.267. However, it is important to note that our model underperforms compared to proprietary MLLMs in the multimodal management stage. We attribute this discrepancy to the differing capability requirements, as management questions involve guideline-aware planning, including referral urgency, follow-up, and treatment recommendations, which depend on broader clinical knowledge and safety considerations. In contrast, our training is primarily focused on image-based perception and diagnostic reasoning. Additionally, we observe a cascading dependency across the multimodal clinical workflow. Models such as LLaVA-1.5-7B and Chiron-o1-8B show progressively lower performance

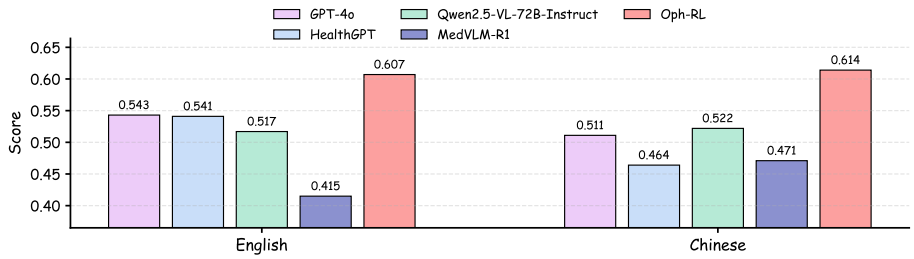
Table 1: **Results on OphthFlowBench.** The best values for each assessment task are highlighted in bold. Specifically, *T1* denotes visual perception, *T2* denotes clinical interpretation, *T3* denotes disease diagnosis, and *T4* denotes clinical management.

Model	Multimodal				Closed-form Text			Open-ended Text		
	T1	*T2*	*T3*	*T4*	*T2*	*T3*	*T4*	*T2*	*T3*	*T4*
Proprietary MLLMs										
GPT-5.2 [23]	0.6755	0.7124	0.6183	**0.6780**	0.5395	0.6190	0.5339	0.1378	0.1063	0.0957
Grok-4 [34]	0.6040	0.6471	0.5703	0.5593	0.5330	0.6095	0.5400	0.1582	0.1225	0.1116
GLM-4.6v [31]	0.6432	0.6669	0.5905	0.5254	0.5027	0.5238	0.4805	0.2288	0.2014	0.1935
Open-Source MLLMs										
Qwen3-VL-8B-Instruct [3]	0.6222	0.6566	0.6659	0.4407	0.4249	0.4762	0.4535	0.1621	0.1694	0.1850
Qwen3-VL-32B-Instruct [3]	0.6674	0.7175	0.7057	0.5932	0.5114	0.5619	0.4969	0.2176	0.1783	0.1781
Ovis2.5-9B [21]	0.6660	0.6583	0.6899	0.5763	0.3427	0.4310	0.3943	0.2274	0.1919	0.1900
InternVL3.5-4B [32]	0.6654	0.7139	0.7321	0.5593	0.3265	0.3833	0.3121	0.1297	0.1469	0.1621
InternVL3.5-8B [32]	0.6746	0.6823	0.6817	0.4915	0.3362	0.3381	0.3450	0.1268	0.1335	0.1588
LLaVA-1.5-7B [17]	0.5001	0.4923	0.3634	0.3559	0.2649	0.2786	0.3203	0.0720	0.1338	0.1346
LLaVA-V1.6-Mistral-7B [18]	0.4788	0.5591	0.4537	0.4407	0.2865	0.3143	0.3224	0.2095	0.2037	0.1914
Phi-3.5-vision-instruct [1]	0.5307	0.5517	0.4854	0.3729	0.3449	0.3952	0.3696	0.2101	0.1919	0.1874
Phi-4-multimodal-instruct [2]	0.5777	0.6119	0.4882	0.4746	0.2865	0.3690	0.3737	0.2246	0.2131	0.2111
Idefics3-8B-Llama3 [12]	0.6022	0.6273	0.5934	0.4746	0.4562	0.5452	0.5154	0.1872	0.1057	0.1407
Mistral-Small-3.1-24B [22]	0.5942	0.6178	0.5295	0.5593	0.4832	0.5214	0.4990	0.1938	0.1801	0.1956
GLM-4.1V-9B-Thinking [31]	0.3533	0.3213	0.3380	0.2373	0.3308	0.3548	0.3388	0.0313	0.0594	0.0617
R-4B [37]	0.5907	0.6038	0.5588	0.4746	0.3395	0.3643	0.3758	0.2191	0.1465	0.1509
Medical-Specific MLLMs										
LLaVA-Med [13]	0.4732	0.4806	0.5378	0.3793	0.3092	0.3833	0.3368	0.1875	0.1105	0.0978
Lingshu-7B [36]	0.6870	0.6867	0.6553	0.5593	0.4259	0.4429	0.4209	0.0557	0.1248	0.1306
HuatuoGPT-Vision-7B [4]	0.6262	0.6515	0.6500	0.4746	0.3135	0.3476	0.3593	0.0476	0.0524	0.0608
Chiron-o1-8B [30]	0.6709	0.6537	0.6519	0.5424	0.3968	0.4333	0.4004	0.0322	0.0622	0.0878
Med-R1-Fundus [11]	0.6053	0.5855	0.5619	0.5345	0.3059	0.2690	0.2505	0.2030	0.1234	0.1244
Med-R1-OCT [11]	0.6036	0.6471	0.5619	0.5345	0.2843	0.2714	0.2669	0.2082	0.1287	0.1306
MedVLM-R1 [26]	0.5669	0.5723	0.4636	0.5172	0.2822	0.2905	0.2485	0.2049	0.1288	0.1280
Qwen3-VL-8B-Instruct (SFT)	0.7367	0.7417	0.7465	0.4576	0.4941	0.5690	0.5175	0.2605	0.2617	0.2475
Oph-RL	**0.7401**	**0.7630**	**0.7792**	0.4746	**0.5492**	**0.6476**	**0.5873**	**0.2610**	**0.2674**	**0.2501**

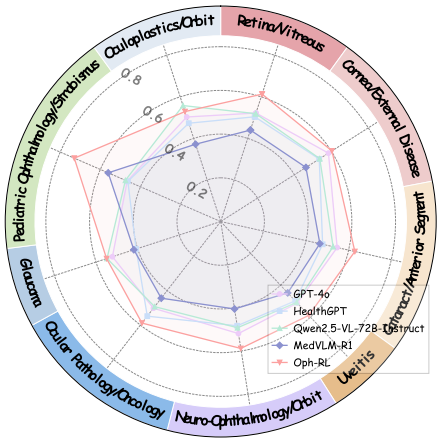
from T1 to T4, suggesting that weaknesses in upstream perception and interpretation may propagate to diagnosis and management. This stage-wise degradation highlights the importance of evaluating MLLMs as an interconnected clinical process rather than as isolated tasks.

3.3 Comparisons on OphthalWeChat

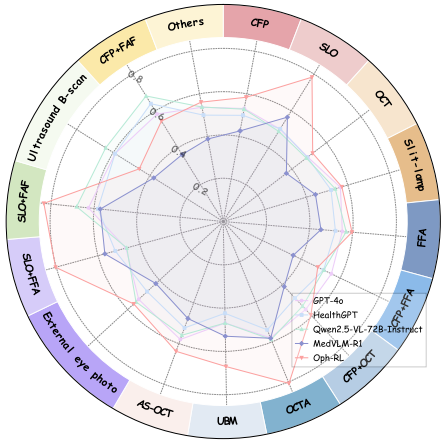
We evaluated our model on the OphthalWeChat benchmark, focusing specifically on its performance across language, ophthalmic subspecialties, and different modalities. As shown in Fig. 3, experimental results on the OphthalWeChat benchmark demonstrate that our model significantly outperforms strong models such as HealthGPT [15] across three key dimensions. Linguistically, our model exhibits the best balanced language capabilities, achieving top scores of 0.607 and 0.614 in English and Chinese, effectively mitigating prevalent language biases in MLLMs. Across clinical subspecialties, the model maintains near-universal leadership. It achieves the best performance in 8 domains, with particularly outstanding performance in pediatric *ophthalmology/strabismus* and *cataract/anterior* segment domains. Furthermore, the model demonstrates exceptional adaptability, outperforming existing models across more than 13 imaging modalities. It excels in single-modality diagnosis and complex multimodal combinations, reaching unprecedented accuracy in *SLO+FAF* and *SLO+FFA*. This fully highlights its profound capabilities in multimodal clinical analysis.



(a) Comparison on language



(b) Comparison on subspecialties



(c) Comparison on modality

Fig. 3: Performance comparison on OphthalWeChat.

4 Conclusion

In this paper, we introduce OphthFlowBench, the first unified ophthalmic benchmark that comprehensively covers multimodal VQA and text-only scenarios. To address the limitations of existing evaluation frameworks, we propose a novel assessment protocol aligned with clinical workflows, decomposing the evaluation process into four sequential stages: visual perception, clinical interpretation, disease diagnosis, and clinical management. This provides a comprehensive evaluation suite for digital ophthalmology. Furthermore, we employed a two-stage training strategy incorporating SFT and difficulty-aware GRPO to develop Oph-RL, a powerful ophthalmology-specific multimodal large language model. Extensive experimental results demonstrate that Oph-RL significantly outperforms existing general and medical multimodal large models on both OphthFlowBench and OphthalWeChat benchmarks. Overall, our work provides a workflow-oriented evaluation framework for ophthalmic MLLMs and offers empirical insights into their capabilities and remaining challenges in clinically relevant scenarios.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant No. 62502320, the Natural Science Foundation of Guangdong Province under Grant No. 2025A1515010184, the project of Shenzhen Science and Technology Innovation Committee under Grant No. JCYJ20240813141424032, and the Scientific Foundation for Youth Scholars of Shenzhen University under Grant No. 827-0001083.

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

References

1. Abdin, M., Aneja, J., et al.: Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219 (2024)
2. Abdin, M., Aneja, J., et al.: Phi-4 technical report. arXiv preprint arXiv:2412.08905 (2024)
3. Bai, S., Cai, Y., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Chen, J., Gui, C., et al.: Towards injecting medical visual knowledge into multimodal LLMs at scale. In: EMNLP. pp. 7346–7370 (2024)
5. Chen, Z., Zhang, J., et al.: Scale down to speed up: Dynamic data selection for reinforcement learning. In: Findings of EMNLP. pp. 7806–7817 (2025)
6. Hao, J., Liang, Y., et al.: OralGPT-Omni: A versatile dental multimodal large language model. In: CVPR. pp. 38509–38519 (2026)
7. Hu, Y., Li, T., et al.: OmniMedVQA: A new large-scale comprehensive evaluation benchmark for medical LVLM. In: CVPR. pp. 22170–22183 (2024)
8. Jin, D., Pan, E., et al.: What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Applied Sciences **11**(14), 6421 (2021)
9. Jin, Q., Dhingra, B., et al.: PubMedQA: A dataset for biomedical research question answering. In: EMNLP-IJCNLP. pp. 2567–2577 (2019)
10. Kwon, W., Li, Z., et al.: Efficient memory management for large language model serving with PagedAttention. In: SOSP (2023)
11. Lai, Y., Zhong, J., et al.: Med-R1: Reinforcement learning for generalizable medical reasoning in vision-language models. IEEE Transactions on Medical Imaging (2026)
12. Laurençon, H., Marafioti, A., et al.: Building and better understanding vision-language models: insights and future directions. arXiv preprint arXiv:2408.12637 (2024)
13. Li, C., Wong, C., et al.: LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. NeurIPS **36**, 28541–28564 (2023)
14. Li, S., Lin, T., et al.: EyecareGPT: Boosting comprehensive ophthalmology understanding with tailored dataset, benchmark and model. In: ACM MM. pp. 3893–3902 (2025)
15. Lin, T., Zhang, W., et al.: HealthGPT: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. In: ICML. pp. 37975–37995. PMLR (2025)
16. Liu, A., Feng, B., et al.: DeepSeek-V3 technical report. arXiv preprint arXiv:2412.19437 (2024)
17. Liu, H., Li, C., et al.: Visual instruction tuning. NeurIPS **36**, 34892–34916 (2023)

18. Liu, H., Li, C., et al.: LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/ (2024), [Accessed 02-26-2026]
19. Liu, J., Wang, W., et al.: Asclepius: A spectrum evaluation benchmark for medical multi-modal large language models. In: ACL. pp. 24181–24201 (2025)
20. Liu, Y., Yang, D., et al.: Adaptive reinforcement for open-ended medical reasoning via semantic-guided reward collapse mitigation. In: CVPR Findings. pp. 8651–8660 (2026)
21. Lu, S., Li, Y., et al.: Ovis2.5 technical report. arXiv preprint arXiv:2508.11737 (2025)
22. Mistralai: Mistral-Small-3.1-24b-Instruct-2503. https://huggingface.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503 (2025), [Accessed 02-26-2026]
23. OpenAI: Introducing gpt-5.2. https://openai.com/index/introducing-gpt-5-2/ (2025), [Accessed 02-26-2026]
24. Ozkaynak, M., Unertl, K., et al.: Clinical workflow analysis, process redesign, and quality improvement. In: Clinical informatics study guide: Text and review, pp. 103–118. Springer (2022)
25. Pal, A., Umapathi, L.K., et al.: MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In: Conference on health, inference, and learning. pp. 248–260. PMLR (2022)
26. Pan, J., Liu, C., et al.: MedVLM-R1: Incentivizing medical reasoning capability of vision-language models (VLMs) via reinforcement learning. In: MICCAI. pp. 337–347. Springer (2025)
27. Qin, Z., Yin, Y., et al.: LMOD: A large multimodal ophthalmology dataset and benchmark for large vision-language models. In: NAACL. pp. 2501–2522 (2025)
28. Qiu, P., Wu, C., et al.: Towards building multilingual language model for medicine. Nature Communications **15**(1), 8384 (2024)
29. Srinivasan, S., Ai, X., et al.: Benchmarking LLMs for ophthalmology (BELO): An expert-curated dataset and evaluation framework for knowledge and reasoning. Ophthalmology Science p. 101050 (2025)
30. Sun, H., Jiang, Y., et al.: Chiron-o1: Igniting multimodal large language models towards generalizable medical reasoning via mentor-intern collaborative search. NeurIPS **38**, 104180–104217 (2026)
31. Team, V., Hong, W., et al.: GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv preprint arXiv:2507.01006 (2025)
32. Wang, W., Gao, Z., et al.: InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
33. Wang, X., Li, J., et al.: Huatuo-26M, a large-scale Chinese medical QA dataset. In: NAACL. pp. 3828–3848 (2025)
34. xAI: Grok-4. https://x.ai/grok (2025), [Accessed 02-26-2026]
35. Xu, P., Gong, X., et al.: Benchmarking large multimodal models for ophthalmic visual question answering with OphthalWeChat. Advances in Ophthalmology Practice and Research pp. 33–41 (2025)
36. Xu, W., Chan, H.P., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)
37. Yang, Q., Ni, B., et al.: R-4B: Incentivizing general-purpose auto-thinking in MLLMs via bi-mode annealing and reinforce learning. In: CVPR. pp. 7891–7900 (2026)

38. Ye, J., Wang, G., et al.: GMAI-MMBench: A comprehensive multimodal evaluation benchmark towards general medical AI. NeurIPS **37**, 94327–94427 (2024)
39. Yu, S., Wang, H., et al.: MedFrameQA: A multi-image medical VQA benchmark for clinical reasoning. arXiv preprint arXiv:2505.16964 (2025)
40. Zhang, X., Wu, C., et al.: PMC-VQA: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415 (2023)
41. Zhao, Y., Huang, J., et al.: Swift: a scalable lightweight infrastructure for fine-tuning. In: AAAI. vol. 39, pp. 29733–29735 (2025)
42. Zuo, Y., Qu, S., et al.: MedXpertQA: Benchmarking expert-level medical reasoning and understanding. In: ICML. pp. 80961–80990. PMLR (2025)