EyeVQA: Benchmarking Ophthalmic Vision-Language Models from Recognition to Spatial Grounding

Gujie Shao[1⋆][0009−0006−5015−9716], Zixun Xie[2⋆][0009−0000−5866−385X], Xuechun Xing[1⋆], Ruixiang Wang[1⋆], Ziyun Lan[1], Yanlin Qi[3][0000−0002−5482−3593], Gangyi Zhang[1], Yuxin Yang[1], Dawei Li[2✉], Haiming Tang[1,4✉][0009−0008−6666−2556]

[1] National University of Singapore, Singapore
[2] Peking University, Beijing, China
[3] Université Paris Cité, Paris, France
[4] Hong Kong Institute of Science and Innovation, Chinese Academy of Sciences, Hong Kong, China
haiming@comp.nus.edu.sg, lidawei@pku.edu.cn
⋆Equal contribution ✉Corresponding authors

Abstract. Vision-language models (VLMs) have shown increasing potential for medical image understanding, yet their capabilities in ophthalmic imaging remain insufficiently characterized. Existing ophthalmic datasets are typically designed for individual diseases or specialized tasks, making it difficult to systematically evaluate whether VLMs can move beyond disease recognition toward comparative reasoning and fine-grained spatial grounding. We introduce EyeVQA, a unified visual question answering benchmark for comprehensive evaluation of ophthalmic VLMs. EyeVQA is constructed from 21 available ophthalmic datasets and contains 20,000 question–answer pairs spanning six disease groups and seven question types: Single-Choice, Multi-Select, Variable-Select, True-False, Ranking, Point Location, and Bounding Box. Gold answers are deterministically derived from source-provided diagnoses, severity grades, clinical findings, segmentation masks, bounding boxes, and anatomical landmarks, enabling reproducible evaluation without relying on model-generated annotations. Notably, 44.5% of the questions require reasoning across multiple images, extending evaluation beyond conventional single-image medical VQA. We benchmark fourteen representative general-purpose, scientific, and medically specialized VLMs under a unified zero-shot protocol. The best-performing model only achieves an overall score of 62.8, while substantial gaps remain in spatial grounding and cross-task generalization. These results highlight the limitations of current VLMs in comprehensive ophthalmic visual understanding and establish EyeVQA as a diagnostic benchmark for developing more reliable and spatially grounded ophthalmic multimodal models. The project page is available at https://github.com/PKUTHM/EyeVQA.

Keywords: Ophthalmic VQA · VLM · Spatial Grounding.

1 Introduction

Recent breakthroughs in vision-language models (VLMs) [1,23,9,10,18] have marked a paradigm shift in multimodal artificial intelligence by integrating visual perception with complex natural-language reasoning. In medical imaging, this interactive paradigm is particularly transformative: clinical diagnosis inherently demands more than passive pattern recognition; it requires linking subtle pathological signs with medical concepts, performing cross-image comparative reasoning, and verifying fine-grained spatial evidence to support diagnostic conclusions.

Ophthalmic imaging, particularly fundus photography, serves as a paramount domain for evaluating multi-level visual understanding in multimodal AI. However, existing ophthalmic image analysis resources remain fragmented [24,22,20]. Most available datasets were built for isolated tasks—such as diabetic retinopathy grading [12], glaucoma screening [15], or pathological myopia diagnosis [17], offering insufficient disease diversity or task flexibility to holistically assess a VLM's clinical reasoning capabilities. Moreover, evaluation restricted to single-image classification fails to reflect real-world clinical workflows, where ophthalmologists routinely compare longitudinal images or bilateral fundus scans.

Visual question answering (VQA) provides a versatile framework for multimodal evaluation [21]. Nevertheless, current medical VQA paradigms suffer from three major limitations when applied to ophthalmology:

1. **Overemphasis on Single-Image Semantics:** Existing benchmarks primarily focus on single-image semantic identification or global disease classification, overlooking complex clinical reasoning patterns.
2. **Lack of Cross-Image Relational Reasoning:** Standard single-image evaluation cannot assess a model's ability to perform comparative analysis, such as ordinal severity assessment or anatomical progression across multiple visual inputs.
3. **Absence of Explicit Spatial Grounding Verification:** Semantic correctness alone does not guarantee that a VLM grounds its conclusions on correct anatomical evidence, making fine-grained spatial localization crucial for trustworthy medical AI.

To bridge these gaps, we present **EyeVQA**, a unified benchmark specifically designed to evaluate the multi-faceted reasoning and spatial grounding capabilities of VLMs in ophthalmology, as shown in Fig. 1. EyeVQA aggregates 21 ophthalmic datasets into 20,000 clinically grounded question–answer (QA) pairs across six primary disease categories: Diabetic Retinopathy (DR), Glaucoma (GL), Pathological Myopia (PM), Macula Disease (MD), Multiple (MU), and Other (OT).

To systematically evaluate complementary reasoning dimensions, EyeVQA formulates seven standardized question types: Single-Choice (SC), Multi-Select (MS), Variable-Select (VS), True-False (TF), Ranking (RK), Point Location (PL), and Bounding Box (BB). Together, these question formats span core clinical problems, including disease presence identification, severity grading, image–description correspondence, comparative disease progression, and precise

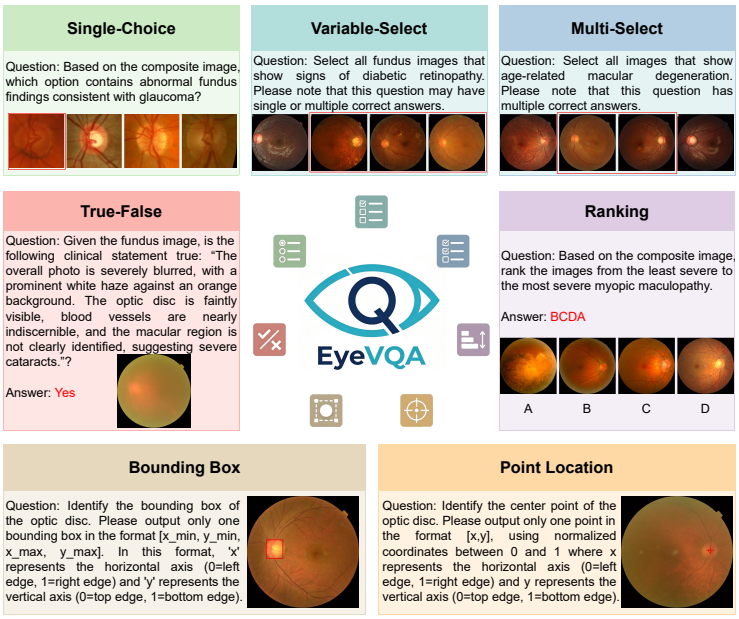

Fig. 1. Overview of the EyeVQA benchmark.

anatomical or lesion grounding. Crucially, all gold answers and spatial targets are deterministically derived from expert annotations rather than generated by secondary LLMs, eliminating hallucinated references. Furthermore, EyeVQA incorporates 10,344 four-panel composites, making **44.5%** of the benchmark (8,902 questions) require comparative reasoning across multiple images—a key setting absent from prior medical VQA benchmarks.

We conduct a standardized evaluation of fourteen representative VLMs spanning open-source general-purpose, close-source general-purpose, scientific, and medically specialized model families under a zero-shot, inference-only protocol. Our experimental findings reveal crucial insights into current multimodal architectures:

- Current state-of-the-art VLMs still fall short of robust clinical utility: the leading model, Intern-S2-Preview-397B, achieves an overall performance of only 62.8%.
- High semantic accuracy does not guarantee spatial grounding precision, as models often excel at verbal disease recognition but fail significantly at coordinate-based localization tasks (PL and BB).
- Specialized medical VLMs do not consistently outperform top-tier generalist or scientific models, underscoring the need for balanced training on both domain knowledge and precise spatial alignment.

Our main contributions are summarized as follows:

1. **A Large-Scale, Unified Benchmark:** We introduce EyeVQA, unifying 21 datasets into 20,000 QA pairs across six disease groups and seven standardized question types, establishing a multi-dimensional testbed from visual perception to clinical reasoning.
2. **Deterministic Grounding & Multi-Image Design:** All QA pairs are deterministically constructed from verified expert annotations, with 44.5% of the benchmark featuring four-panel composite images to evaluate cross-image comparative reasoning.
3. **Extensive Benchmarking & Diagnostic Insights:** We systematically evaluate fourteen leading general, scientific, and medical VLMs under unified zero-shot constraints, uncovering critical bottlenecks in fine-grained visual grounding and cross-modal alignment.

2 Construction of EyeVQA

2.1 Dataset Collection

To comprehensively evaluate ophthalmic vision-language models (VLMs), we collected **21 available ophthalmic datasets** covering major retinal conditions across diverse platforms, such as *Scientific Data*, Kaggle, Hugging Face, Zenodo, and GitHub. These datasets exhibit rich visual and textual heterogeneity, ranging from disease classification and lesion segmentation to clinical descriptions and paired multimodal imaging. A comprehensive description of the data sources, annotation formats, and standardization procedures is detailed in Section B in the Supplementary Material (with a complete overview summarized in Table S1).

2.2 Question–Answer Design

Following established benchmark designs in general and medical VQA [16,14,8], EyeVQA formulates seven standardized question types spanning diverse reasoning granularities, as illustrated in Fig. 1.

All QA pairs are deterministically constructed from verifiable ground truths, including diagnostic labels, severity grades, clinical findings, segmentations, and landmark annotations. Specifically, **SC**, **MS**, and **VS** evaluate diagnostic recognition and candidate selection with varying answer cardinalities. **TF** verifies clinical statement correctness, while **RK** assesses ordinal disease progression or anatomical metrics across composite images. For spatial grounding, **PL** and **BB** require precise coordinate predictions for clinical landmarks and lesion areas.

Detailed prompt templates, distractor sampling strategies, and coordinate transformation specifications are deferred to Section C in the Supplementary Material.

2.3 Statistics and Analysis

Scale and Composition. EyeVQA contains **20,000** standardized question–answer pairs referencing **27,935** fundus images from 21 source datasets. Beyond

single-image VQA, it incorporates 10,344 four-panel composite images, making **8,902** questions (44.5%) require multi-image relational reasoning. The benchmark spans seven standardized question types (SC, MS, VS, TF, RK, PL, BB) and six disease groups (DR, GL, PM, MD, OT, MU), ensuring comprehensive coverage from visual perception to multi-disease reasoning (Fig. 2).

Text Characteristics and Balanced Options. Question lengths average **46.0** words (median 40, range 16–170), exhibiting a multi-modal distribution aligned with task output complexity (Fig. 2b): TF prompts are the most concise (mean 24.3 words), whereas spatial grounding tasks (e.g., BB, mean 89.7 words) require explicit coordinate specifications. To prevent shortcut learning and language priors, options are carefully balanced across tasks: SC options are nearly uniform (A: 24.2%, B: 24.7%, C: 26.6%, D: 24.5%), TF maintains a well-balanced binary distribution, and MS/VS cover diverse answer cardinalities.

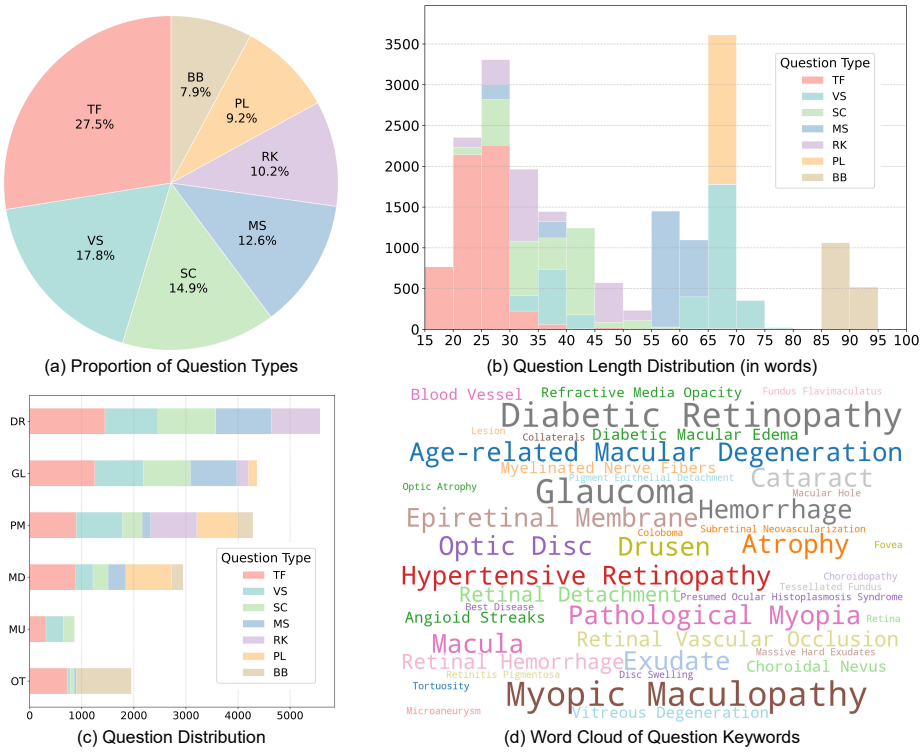

Fig. 2. Dataset overview of EyeVQA: (a) distribution of question types; (b) question length distribution by type; (c) question-type composition within each disease group; (d) clinical vocabulary frequency over the 20,000 questions.

3 Experiments

3.1 Experimental Setting

We evaluate fourteen representative vision-language models (VLMs), spanning open-source general-purpose, closed-source general-purpose, science-oriented, and medically specialized model families. The open-source general-purpose group comprises GLM-4V-Flash [5], GLM-4.1V-Thinking-Flash [9,3], GLM-4.6V-Flash [4], Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct [2], and Qwen3-VL-8B-Instruct [1], while the closed-source general-purpose group includes Gemini-3.1-Flash-Lite [6] and Gemini-3.5-Flash-Lite [7]. The science-oriented group contains Intern-S1-Pro, a trillion-parameter mixture-of-experts model [25], together with Intern-S2-Preview-35B [10] and Intern-S2-Preview-397B [11]. The medical group includes Lingshu-7B [23], MedGemma-4B-IT [19], and its successor MedGemma-1.5-4B-IT [18]. This selection enables comparisons across model scales and between general-purpose, science-oriented, and medically specialized VLMs.

All models are evaluated in a zero-shot, inference-only setting without training or fine-tuning. Open-source VLMs are deployed with vLLM [13], while API-based models are accessed through their official services. Each model receives the same question and standardized image input. Task-specific format-enforcing prompts require concise structured responses, which are parsed with regular expressions to extract the final answers. This action minimizes the bias introduced by the generation style of VLMs.

3.2 Evaluation Metrics

We evaluate model performance across seven distinct question types: Single-Choice (SC), True-False (TF), Multi-Select (MS), Variable-Select (VS), Ranking (RK), Bounding Box (BB), and Point Location (PL). Detailed mathematical formulations for each metric—including exact-match accuracy, Hamming metric, Kendall's Tau, Intersection over Union (IoU), and Euclidean distance tolerance—are provided in Section D in the Supplementary Material. For cross-task comparison, all raw metric scores are linearly normalized to $[0, 100]$, and refused responses are assigned a score of zero.

3.3 Results

Disease-category results Table 1 compares performance across the six disease categories. Intern-S2-397B achieves the highest overall score of 62.8 and leads in five categories, including DR, GL, PM, MD, and MU, supported by its particularly strong performance on MD (73.4). Intern-S2-35B follows with the second-highest overall score of 59.3 and remains competitive across disease categories, while Gemini-3.1-Flash-Lite secures the top rank in the OT category with 73.4 (surpassing Intern-S2-35B which ranks second with 68.1). Among the medically specialized models, MedGemma-4B performs strongly on DR, ranking second with 62.5, whereas Lingshu-7B and MedGemma-1.5-4B obtain more

moderate overall scores. These results indicate that the Intern-S2 models provide the most consistent cross-disease performance, while lightweight closed-source models deliver exceptional category-specific results.

Question-type results The question-type breakdown reveals complementary model strengths. Intern-S2-397B ranks first on SC, VS, TF, and PL, and is only 0.3 points behind the best result on MS. Intern-S2-35B leads on MS, while nearly matching Intern-S2-397B on SC and VS. In contrast, Intern-S1-Pro achieves the highest RK score (60.0) despite its relatively low BB score (17.0), suggesting strong ordering ability but limited fine-grained spatial grounding. Qwen3-VL-8B is another notable exception: although its overall score is 48.9, it ranks third on BB with 46.2. MedGemma-4B attains the second-best TF score (71.1) but performs poorly on BB (6.2), further showing that recognition and spatial localization capabilities do not necessarily improve together. Furthermore, the lightweight closed-source models exhibit remarkable spatial and grounding capabilities, with Gemini-3.1-Flash-Lite achieving the highest Bounding Box (BB) score of 58.1 and Gemini-3.5-Flash-Lite securing the second-best Point Location (PL) score of 65.3, highlighting their strong potential in fine-grained visual localization tasks.

Table 1. Performance comparison of the evaluated VLMs. **Bold** indicates the best result, and Underline indicates the second-best result.

Model	Total	DR	GL	PM	MD	MU	OT	SC	MS	VS	TF	RK	PL	BB
GLM-4V-Flash	39.3	37.5	45.0	41.6	41.5	35.7	25.4	37.1	48.9	31.8	42.1	53.2	51.3	4.0
GLM-4.1V-Flash	49.5	49.1	46.3	46.9	55.5	46.6	55.7	43.1	51.2	38.5	60.3	57.3	50.6	34.5
GLM-4.6V-Flash	45.6	44.8	44.0	44.6	45.2	38.5	57.3	37.6	54.0	36.4	51.6	57.4	27.9	51.8
Intern-S1-Pro	54.1	58.7	51.9	51.9	58.6	54.0	44.4	49.6	59.0	44.5	70.2	**60.0**	50.7	17.0
Intern-S2-35B	59.3	60.0	57.1	55.6	60.8	60.4	68.1	60.1	**63.3**	50.5	70.5	59.1	45.2	49.4
Intern-S2-397B	**62.8**	**63.2**	**57.9**	**57.9**	**73.4**	**62.5**	67.7	**60.2**	63.0	**51.6**	**76.2**	55.8	**70.7**	45.5
Lingshu-7B	50.9	58.6	46.3	48.6	52.2	47.7	43.8	53.0	59.5	45.0	61.8	56.8	35.1	19.2
MedGemma-1.5-4B	49.9	59.8	53.3	40.3	51.0	56.0	31.1	48.7	57.8	40.8	64.6	50.0	51.3	7.5
MedGemma-4B	51.5	62.5	52.2	47.2	48.7	57.0	30.1	47.6	53.6	41.9	71.1	58.1	46.5	6.2
Qwen2.5-VL-3B	41.0	44.4	34.4	43.5	38.2	54.3	39.0	39.2	51.2	39.7	45.9	52.1	23.2	20.1
Qwen2.5-VL-7B	47.1	45.5	49.4	48.4	51.1	54.9	34.6	45.2	58.2	44.2	54.4	54.8	44.6	7.3
Qwen3-VL-8B	48.9	46.9	50.8	48.5	40.7	47.5	64.7	44.7	58.8	39.4	60.1	49.5	29.0	46.2
Gemini-3.1-Flash	57.0	55.7	49.5	53.0	66.8	54.1	**73.4**	53.4	58.6	44.3	67.3	54.6	56.5	**58.1**
Gemini-3.5-Flash	55.9	55.2	55.4	48.4	69.5	53.3	56.5	53.1	58.9	45.1	70.2	49.8	65.3	27.8

3.4 Key Findings

We summarize four key findings from the benchmark, as supported by Table 1.

Larger models perform better in most tasks. Performance generally increases with model scale across the benchmark. The scaling trend is clearest within matched model families and is also reflected in the overall ranking, where

larger models achieve stronger aggregate results and broader performance across disease categories and question types.

Medical specialization offers parameter-efficient benefits. Compact medically specialized VLMs achieve competitive performance relative to substantially larger science-oriented models across clinically relevant recognition, reasoning, and localization tasks. This recurring trend indicates that domain-specific medical knowledge improves the effective use of limited model capacity.

Reasoning and spatial grounding are partially decoupled. Across the evaluated models, performance rankings on recognition and ordering tasks show limited correspondence with rankings on bounding-box and point-localization tasks. The systematic difference between these task groups indicates that semantic reasoning and fine-grained spatial grounding capture distinct dimensions of ophthalmic visual competence. Their weak co-variation supports evaluating both capabilities independently.

Performance varies substantially across disease categories. Model performance exhibits substantial and recurring variation across disease categories, with cross-model rankings shifting according to the evaluated pathology. This trend reflects differences in disease-specific visual features, imaging characteristics, and diagnostic demands. Disease-stratified evaluation is therefore essential for characterizing category-level strengths and weaknesses beyond aggregate performance.

4 Conclusion

In this work, we introduce **EyeVQA**, a unified benchmark for comprehensively evaluating vision-language models in ophthalmic image understanding. EyeVQA integrates 21 available ophthalmic datasets and contains 20,000 question-answer pairs spanning six disease groups and seven question types, covering recognition, verification, comparative reasoning, and fine-grained spatial grounding. By deterministically deriving gold answers from source-provided clinical annotations and incorporating a substantial proportion of multi-image questions, EyeVQA enables reproducible and diagnostic evaluation beyond conventional single-image disease recognition.

Our systematic evaluation of representative general-purpose, scientific, and medically specialized VLMs shows that current models remain far from robust ophthalmic visual understanding. Although larger models generally achieve stronger overall performance, substantial variation persists across diseases and task types, while strong semantic reasoning does not necessarily translate into accurate spatial grounding. These findings highlight the importance of evaluating ophthalmic VLMs along multiple complementary dimensions rather than relying on a single aggregate score. We hope EyeVQA can serve as a standardized testbed

for future research toward more reliable ophthalmic multimodal models with stronger cross-task generalization, multi-image reasoning, and fine-grained visual grounding capabilities.

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

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. GLM Team: GLM-4.1V-Thinking-Flash. https://docs.bigmodel.cn/cn/guide/models/vlm/glm-4.1v-thinking, last accessed 2026/08/11
4. GLM Team: GLM-4.6V-Flash. https://docs.bigmodel.cn/cn/guide/models/free/glm-4.6v-flash, last accessed 2026/08/11
5. GLM Team: GLM-4V-Flash. https://docs.bigmodel.cn/cn/guide/models/free/glm-4v-flash, last accessed 2026/08/11
6. Google DeepMind: Gemini 3.1 Flash-Lite. https://ai.google.dev/gemini-api/docs/models/gemini-3.1-flash-lite, last accessed 2026/08/11
7. Google DeepMind: Gemini 3.5 Flash-Lite. https://ai.google.dev/gemini-api/docs/models/gemini-3.5-flash-lite, last accessed 2026/08/11
8. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
9. Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., et al.: Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv preprint arXiv:2507.01006 (2025)
10. InternLM Team: Intern-S2-Preview-35B. https://huggingface.co/internlm/Intern-S2-Preview, last accessed 2026/08/11
11. InternLM Team: Intern-S2-Preview-397B. https://huggingface.co/internlm/Intern-S2-Preview-397B, last accessed 2026/08/11
12. Kaggle: Diabetic Retinopathy Detection Challenge. https://www.kaggle.com/c/diabetic-retinopathy-detection, last accessed 2026/8/11
13. Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C.H., Gonzalez, J., Zhang, H., Stoica, I.: Efficient memory management for large language model serving with pagedattention. In: Proceedings of the 29th symposium on operating systems principles. pp. 611–626 (2023)
14. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 180251 (2018)
15. Li, L., Xu, M., Wang, X., Jiang, L., Liu, H.: Attention based glaucoma detection: A large-scale database and cnn model. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10563–10572. IEEE (2019)
16. Liu, Y., Tang, H., Peng, J., Zhang, J., Ji, X., He, Q., Wu, W., Luo, D., Gan, Z., Zhu, J., et al.: Human-mme: A holistic evaluation benchmark for human-centric multimodal large language models. arXiv preprint arXiv:2509.26165 (2025)
17. Pachade, S., Porwal, P., Thulkar, D., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Giancardo, L., Quellec, G., Mériaudeau, F.: Retinal fundus multi-disease image dataset (rfmid): A dataset for multi-disease detection research. Data **6**(2), 14 (2021)
18. Sellergren, A., Gao, C., Mahvar, F., Kohlberger, T., Jamil, F., Traverse, M., Tono, A., Sadjad, B., Yang, L., Lau, C., et al.: Medgemma 1.5 technical report. arXiv preprint arXiv:2604.05081 (2026)

19. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
20. Silva-Rodriguez, J., Chakor, H., Kobbi, R., Dolz, J., Ayed, I.B.: A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. Medical Image Analysis **99**, 103357 (2025)
21. Wang, X., Liu, B., Wang, M., Zhang, Z., Zhu, C., Fu, H., Liu, J.: Towards clinically interpretable ophthalmic vqa via spatially-grounded lesion evidence. arXiv preprint arXiv:2605.22414 (2026)
22. Xie, Z., Ao, M., Tang, H., Li, X., Bai, X., Zhang, S., Li, D.: A fine-grained fundus image dataset for cataract severity assessment and diagnosis. Scientific Data **13**(1), 418 (2026)
23. 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)
24. Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., et al.: A foundation model for generalizable disease detection from retinal images. Nature **622**(7981), 156–163 (2023)
25. Zou, Y., Zhu, D., Zhu, L., Zhu, T., Zhou, Y., Zhou, P., Zhou, X., Zhou, D., Zhou, Z., Zhou, Y., et al.: Intern-s1-pro: Scientific multimodal foundation model at trillion scale. arXiv preprint arXiv:2603.25040 (2026)