

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

Gujie Shao*, Zixun Xie*, Xuechun Xing*, Ruixiang Wang*, Ziyun Lan,
Yanlin Qi, Gangyi Zhang, Yuxin Yang, Dawei Li✉, Haiming Tang✉

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) [1,48,21,24,41] 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 [50,47,43]. Most available datasets were built for isolated tasks—such as diabetic retinopathy grading [27], glaucoma screening [31], or pathological myopia diagnosis [38], 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 [45]. 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 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

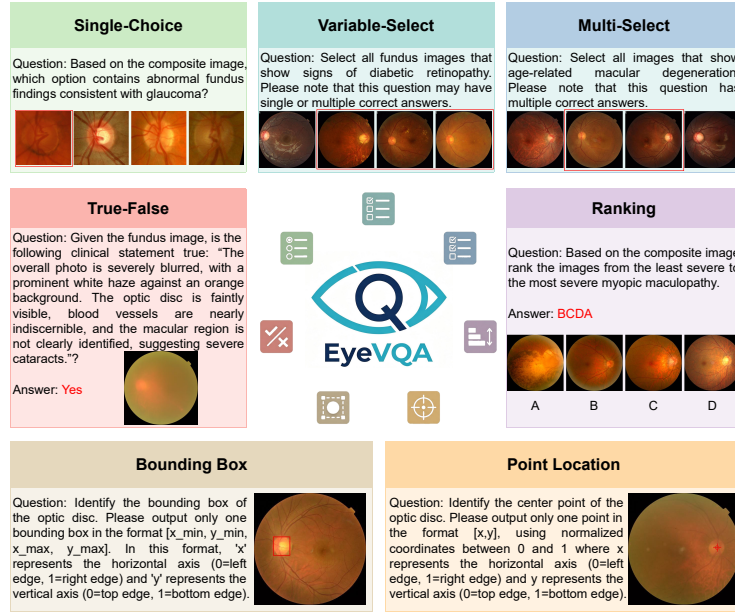


Fig. 1. Overview of the EyeVQA benchmark.

(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 [34,30,20], 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.

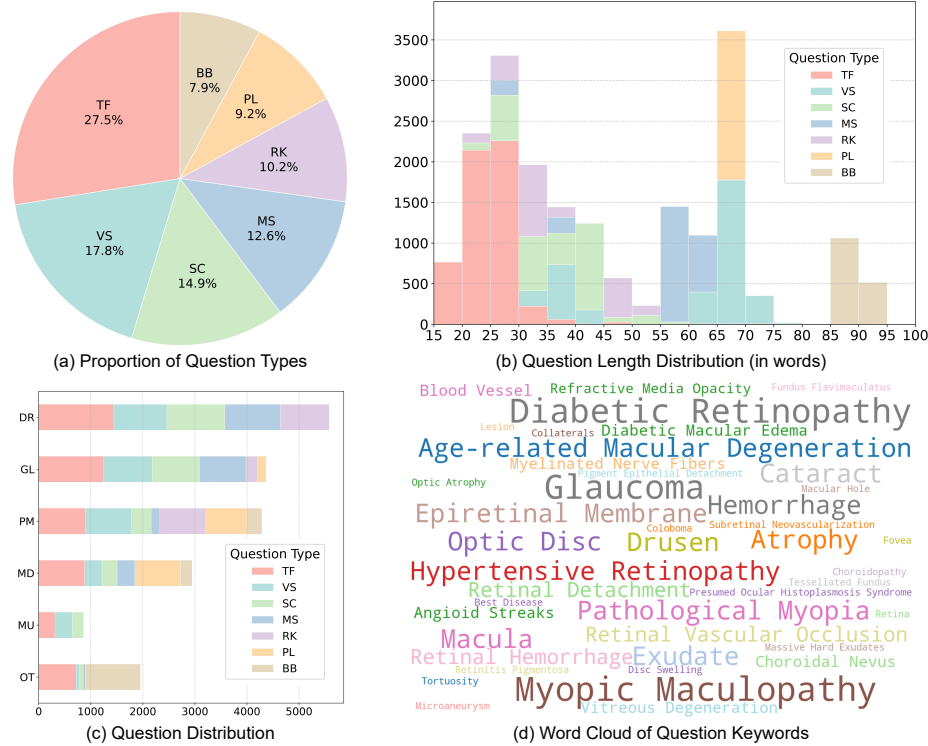


Fig. 2. Dataset overview of EyeVQA: (a) distribution of question types; (b) question length distribution by type; (c) question-type composition within each disease group; (d) clinical vocabulary frequency over the 20,000 questions.

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 [15], GLM-4.1V-Thinking-Flash [21,13], GLM-4.6V-Flash [14], 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 [16] and Gemini-3.5-Flash-Lite [17]. The science-oriented group contains Intern-S1-Pro, a trillion-parameter mixture-of-experts model [51], together with Intern-S2-Preview-35B [24] and Intern-S2-Preview-397B [25]. The medical group includes Lingshu-7B [48], MedGemma-4B-IT [42], and its successor MedGemma-1.5-4B-IT [41]. 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 [29], 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	<u>51.8</u>
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	<u>57.1</u>	55.6	60.8	<u>60.4</u>	<u>68.1</u>	<u>60.1</u>	63.3	50.5	70.5	<u>59.1</u>	<u>45.2</u>	49.4
Intern-S2-397B	62.8	63.2	57.9	57.9	73.4	62.5	67.7	60.2	<u>63.0</u>	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	<u>62.5</u>	52.2	47.2	48.7	57.0	30.1	47.6	53.6	41.9	<u>71.1</u>	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	<u>45.9</u>	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	<u>69.5</u>	53.3	56.5	53.1	58.9	45.1	70.2	49.8	<u>65.3</u>	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. Bazoge, A.: Mediqal: A french medical question answering dataset for knowledge and reasoning evaluation. *Scientific Data* **13**(1), 356 (2026)
4. Benítez, V.E.C., Matto, I.C., Román, J.C.M., Noguera, J.L.V., García-Torres, M., Ayala, J., Pinto-Roa, D.P., Gardel-Sotomayor, P.E., Facon, J., Grillo, S.A.: Dataset from fundus images for the study of diabetic retinopathy. *Data in brief* **36**, 107068 (2021)
5. Cen, L.P., Ji, J., Lin, J.W., Ju, S.T., Lin, H.J., Li, T.P., Wang, Y., Yang, J.F., Liu, Y.F., Tan, S., et al.: Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nature communications* **12**(1), 4828 (2021)
6. Choi, S.: retina_dataset: Fundus image dataset for eye disease classification. https://github.com/yiweichen04/retina_dataset (2021), accessed: Aug. 7, 2026
7. Cuadros, J., Bresnick, G.: Eyepacs: an adaptable telemedicine system for diabetic retinopathy screening. *Journal of diabetes science and technology* **3**(3), 509–516 (2009)
8. De Vente, C., Vermeer, K.A., Jaccard, N., Wang, H., Sun, H., Khader, F., Truhn, D., Aimyshev, T., Zhanibekuly, Y., Le, T.D., et al.: Airops: Artificial intelligence for robust glaucoma screening challenge. *IEEE transactions on medical imaging* **43**(1), 542–557 (2023)
9. Diaz-Pinto, A., Morales, S., Naranjo, V., Köhler, T., Mossi, J.M., Navea, A.: Cnns for automatic glaucoma assessment using fundus images: an extensive validation. *Biomedical engineering online* **18**(1), 29 (2019)
10. Fang, H., Li, F., Fu, H., Sun, X., Cao, X., Lin, F., Son, J., Kim, S., Quelled, G., Matta, S., et al.: Adam challenge: Detecting age-related macular degeneration from fundus images. *IEEE transactions on medical imaging* **41**(10), 2828–2847 (2022)
11. Fang, H., Li, F., Wu, J., Fu, H., Sun, X., Orlando, J.I., Bogunović, H., Zhang, X., Xu, Y.: Open fundus photograph dataset with pathologic myopia recognition and anatomical structure annotation. *Scientific data* **11**(1), 99 (2024)
12. Fang, H., Li, F., Wu, J., Fu, H., Sun, X., Son, J., Yu, S., Zhang, M., Yuan, C., Bian, C., et al.: Refuge2 challenge: A treasure trove for multi-dimension analysis and evaluation in glaucoma screening. arXiv preprint arXiv:2202.08994 (2022)
13. GLM Team: GLM-4.1V-Thinking-Flash. <https://docs.bigmodel.cn/cn/guide/models/vlm/glm-4.1v-thinking>, last accessed 2026/08/11
14. GLM Team: GLM-4.6V-Flash. <https://docs.bigmodel.cn/cn/guide/models/free/glm-4.6v-flash>, last accessed 2026/08/11
15. GLM Team: GLM-4V-Flash. <https://docs.bigmodel.cn/cn/guide/models/free/glm-4v-flash>, last accessed 2026/08/11
16. 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
17. 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

18. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6904–6913 (2017)
19. Hassan, T., Raja, H., Hassan, B., Akram, M.U., Dias, J., Werghi, N.: A composite retinal fundus and oct dataset to grade macular and glaucomatous disorders. In: 2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2). pp. 1–6. IEEE (2022)
20. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
21. 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)
22. Huang, J.H., Yang, C.H.H., Liu, F., Tian, M., Liu, Y.C., Wu, T.W., Lin, I., Wang, K., Morikawa, H., Chang, H., et al.: Deepoph: medical report generation for retinal images via deep models and visual explanation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2442–2452 (2021)
23. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6693–6702. IEEE (2019)
24. InternLM Team: Intern-S2-Preview-35B. <https://huggingface.co/internlm/Intern-S2-Preview>, last accessed 2026/08/11
25. InternLM Team: Intern-S2-Preview-397B. <https://huggingface.co/internlm/Intern-S2-Preview-397B>, last accessed 2026/08/11
26. Jin, K., Huang, X., Zhou, J., Li, Y., Yan, Y., Sun, Y., Zhang, Q., Wang, Y., Ye, J.: Fives: A fundus image dataset for artificial intelligence based vessel segmentation. Scientific data **9**(1), 475 (2022)
27. Kaggle: Diabetic Retinopathy Detection Challenge. <https://www.kaggle.com/c/diabetic-retinopathy-detection>, last accessed 2026/8/11
28. Kaggle: Diabetic retinopathy detection: Competition rules and data usage terms. <https://www.kaggle.com/competitions/diabetic-retinopathy-detection/rules>, accessed: Aug. 7, 2026
29. 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)
30. 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)
31. 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)
32. Li, N., Li, T., Hu, C., Wang, K., Kang, H.: A benchmark of ocular disease intelligent recognition: One shot for multi-disease detection. In: International symposium on benchmarking, measuring and optimization. pp. 177–193. Springer (2020)
33. Liu, R., Wang, X., Wu, Q., Dai, L., Fang, X., Yan, T., Son, J., Tang, S., Li, J., Gao, Z., et al.: Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. Patterns **3**(6) (2022)

34. 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)
35. Nakayama, L.F., Restrepo, D., Matos, J., Ribeiro, L.Z., Malerbi, F.K., Celi, L.A., Regatieri, C.S.: Brset: a brazilian multilabel ophthalmological dataset of retina fundus photos. *PLOS Digital Health* **3**(7), e0000454 (2024)
36. Niemeijer, M., Van Ginneken, B., Cree, M.J., Mizutani, A., Quellec, G., Sánchez, C.I., Zhang, B., Hornero, R., Lamard, M., Muramatsu, C., et al.: Retinopathy online challenge: automatic detection of microaneurysms in digital color fundus photographs. *IEEE transactions on medical imaging* **29**(1), 185–195 (2009)
37. Orlando, J.I., Fu, H., Breda, J.B., Van Keer, K., Bathula, D.R., Diaz-Pinto, A., Fang, R., Heng, P.A., Kim, J., Lee, J., et al.: Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis* **59**, 101570 (2020)
38. Pachade, S., Porwal, P., Thulkar, D., Kokare, M., Deshmukh, G., Sahasrabudhe, 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)
39. Qian, B., Sheng, B., Chen, H., Wang, X., Li, T., Jin, Y., Guan, Z., Jiang, Z., Wu, Y., Wang, J., et al.: A competition for the diagnosis of myopic maculopathy by artificial intelligence algorithms. *JAMA ophthalmology* **142**(11), 1006–1015 (2024)
40. Qin, L., Ou, S., Zhang, M., Wei, J., Zhang, Y., Song, X., Liu, Y., Wang, M., Xu, W.: Face-human-bench: A comprehensive benchmark of face and human understanding for multi-modal assistants. *Advances in Neural Information Processing Systems* **38** (2026)
41. 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)
42. 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)
43. 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)
44. Takahashi, H., Tampo, H., Arai, Y., Inoue, Y., Kawashima, H.: Applying artificial intelligence to disease staging: Deep learning for improved staging of diabetic retinopathy. *PloS one* **12**(6), e0179790 (2017)
45. 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)
46. Wu, J., Fang, H., Li, F., Fu, H., Lin, F., Li, J., Huang, Y., Yu, Q., Song, S., Xu, X., et al.: Gamma challenge: glaucoma grading from multi-modality images. *Medical Image Analysis* **90**, 102938 (2023)
47. 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)
48. 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)

49. Zhang, Z., Yin, F.S., Liu, J., Wong, W.K., Tan, N.M., Lee, B.H., Cheng, J., Wong, T.Y.: Origa-light: An online retinal fundus image database for glaucoma analysis and research. In: 2010 Annual international conference of the IEEE engineering in medicine and biology. pp. 3065–3068. IEEE (2010)
50. 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)
51. 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)

Supplementary Material for EyeVQA

A Overview

This supplementary material complements the main text of EyeVQA by providing additional technical details and formal specifications. Specifically, Section B details the collection, institutional origins, disease coverage, and preprocessing procedures of the 21 ophthalmic datasets summarized in Table S1. Section C elaborates on the design protocols, distractor sampling strategies, and coordinate system conventions across all seven standardized question types. Section D presents the explicit mathematical formulations of the evaluation metrics, score normalization rules, and refusal handling mechanisms.

B Supplementary Details for Dataset Collection

B.1 Public Ophthalmic Dataset Collection

The rapid development of ophthalmic image analysis has led to the release of numerous retinal image datasets over the past decade. Most of these datasets were designed for individual diseases or specific clinical applications, such as diabetic retinopathy grading, glaucoma screening, pathological myopia diagnosis, and retinal lesion segmentation. Although these datasets have significantly promoted the development of computer-aided diagnosis systems, few single datasets provide sufficient disease diversity or annotation richness for comprehensively evaluating ophthalmic vision-language models (VLMs). Therefore, instead of relying on a single data source, we collected multiple available ophthalmic datasets to support the construction of a comprehensive ophthalmic VLM benchmark.

The final collection consists of **21 available ophthalmic datasets** covering major retinal diseases, including diabetic retinopathy, glaucoma, pathological myopia, macula disease, cataract, retinal vascular abnormalities, and other common retinal disorders. We collected these datasets from diverse sources, including *Scientific Data* journal publications, Kaggle challenges, Hugging Face datasets, Figshare, Zenodo, PhysioNet, Mendeley Data, CodaLab platforms, Grand Challenge platforms, and GitHub repositories. These datasets have been widely adopted in ophthalmic artificial intelligence research. An overview of all collected datasets is presented in Table S1.

The collected datasets exhibit substantial diversity in both clinical content and annotation protocols. Disease-specific datasets mainly focus on a single ophthalmic disorder with carefully designed grading criteria or lesion annotations, whereas multi-disease datasets include multiple retinal abnormalities within a unified labeling framework. Besides image-level disease classification, several

datasets provide anatomical structure segmentation, retinal vessel segmentation, lesion localization, paired fundus–OCT images, demographic information, and clinical descriptions. Such heterogeneous annotations enable the construction of evaluation tasks ranging from visual perception to clinically oriented reasoning, providing a much more comprehensive assessment of ophthalmic vision-language models than conventional image classification benchmarks.

Another important characteristic of the collected datasets is their diversity in imaging conditions. Since the datasets were collected by different medical institutions using various fundus cameras under different clinical environments, considerable differences exist in image resolution, field of view, illumination, image quality, annotation standards, and metadata organization. These discrepancies increase the difficulty of cross-dataset evaluation while better reflecting real-world clinical scenarios, thereby improving the robustness and generalizability of the benchmark.

To ensure consistency during benchmark construction, all datasets were standardized before downstream processing. Image formats, annotation schemas, disease terminology, metadata fields, and directory structures were unified through a preprocessing pipeline, thereby reducing inconsistencies introduced by heterogeneous data sources. The standardized datasets were subsequently used for automatic question–answer generation.

B.2 Data Distribution and Licensing Policy

To respect the intellectual property rights and raw data terms of service of the original dataset creators, EyeVQA strictly adheres to a dual-track data distribution policy based on the underlying dataset licenses (detailed in Table S1):

1. **Direct Image Redistribution (Permissive Licenses):** For datasets released under permissive open-access licenses that explicitly allow redistribution and adaptation (e.g., CC BY 4.0, CC BY-SA 4.0, such as CSDI [47], ACRIMA [9], and FIVES [26]), we directly distribute the standardized fundus images alongside our generated QA pairs in the public benchmark repository for maximum ease of evaluation.
2. **Metadata-Only Distribution (Restrictive / Non-Redistributable Licenses):** For datasets governed by restrictive licenses, NDA requirements, or non-redistributable terms (e.g., CC BY-NC-ND 4.0, PhysioNet Credentialed Licenses, Kaggle Competition Rules, or proprietary terms, such as ADAM [10], BRSET [35], EyePACS [7], and PALM [11]), **we do not host or redistribute the raw fundus images**. Instead, we release only the deterministic QA pairs mapped to the original image filenames/identifiers. Users can obtain the raw images directly from the official repositories via the provided references/URLs and align them with our VQA benchmark using our provided automated pairing scripts.

This protocol ensures full compliance with medical data governance and original licensing restrictions while providing a standardized, reproducible evaluation framework for the community.

Table S1. Overview of the collected ophthalmic datasets.

Dataset	Images	Source (institution/country)	Disease	License
ACRIMA [9] (2019)	705	FISABIO Oftalmología Médica, Valencia, Spain	Glaucoma	CC BY 4.0
ADAM [10] (2018)	1,200	Zhongshan Ophthalmic Center, Sun Yat-sen University, China	Macula Disease	CC BY-NC-ND 4.0
AIROGS [8] (2021)	101,442	500 screening centers across the United States of America	Glaucoma	CC BY-NC-ND 4.0
BRSET [35] (2024)	16,266	Department of Ophthalmology, São Paulo Federal University, Brazil	Diabetic Retinopathy, Macula Disease, etc.	PhysioNet Credentialed Health Data License 1.5.0
CSDI [47] (2026)	187	Peking University Third Hospital, China	Cataract	CC BY 4.0
Retina Dataset [6] (2016)	601	8 open access databases (specific sources undisclosed)	Glaucoma, Cataract, etc.	Undisclosed
DEN [22] (2021)	15,708	Undisclosed (NDA required)	Diabetic Retinopathy, Glaucoma, Macula Disease, etc.	Undisclosed
DeepDRiD [33] (2022)	2,256	Department of Ophthalmology, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai, China	Diabetic Retinopathy	CC BY-SA 4.0
EyePACS Dataset [7,28] (2015)	88,702	EyePACS platform (USA) and 3 eye hospitals in India (Aravind Eye Hospital, Sankara Nethralaya, Narayana Nethralaya)	Diabetic Retinopathy	Kaggle Competition Rules
FIVES [26] (2022)	800	Ophthalmology Centre at the Second Affiliated Hospital of Zhejiang University, China	Diabetic Retinopathy, Glaucoma, Macula Disease, etc.	CC BY 4.0
FUND-OCT [19] (2021)	200	Undisclosed	Glaucoma, Macula Disease, etc.	CC BY 4.0
GAMMA [46] (2021)	300	Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China	Glaucoma	CC BY-NC-ND
JICHI Dataset [44] (2017)	9,940	Jichi Medical University Hospital, Tochigi, Japan	Diabetic Retinopathy	CC BY 4.0
JSIEC1000 [5] (2019)	1,000	Joint Shantou International Eye Centre (JSIEC), Shantou, Guangdong, China	Diabetic Retinopathy, Glaucoma, Macula Disease, etc.	DbCL v1.0
MMAC [39] (2023)	1,391	Shanghai Health and Medical Center (SHMC), China; Shanghai Sixth People's Hospital (SSPH), China	Pathological Myopia	CC BY 4.0
ODIR [32] (2019)	8,000	Multiple hospitals and medical centers in China	Diabetic Retinopathy, Glaucoma, Cataract, Macula Disease, etc.	Undisclosed
ORIGA-light [49] (2010)	650	Singapore Eye Research Institute, Singapore	Glaucoma	Undisclosed
PALM [11] (2019)	1,200	Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China	Pathological Myopia	Non-redistributable
PARAGUAY [4] (2021)	1,437	Department of Ophthalmology, Hospital de Clínicas, Facultad de Ciencias Médicas, Universidad Nacional de Asunción, San Lorenzo, Paraguay	Diabetic Retinopathy	CC BY 4.0
REFUGE2 [37,12] (2020)	1,200	Multiple medical centers (specific sources not fully disclosed)	Glaucoma	Non-redistributable
ROC [36] (2009)	100	Undisclosed (from a large diabetic retinopathy screening program)	Diabetic Retinopathy	Non-redistributable

C Detailed Question–Answer Generation Guidelines

This section provides comprehensive specifications for the seven standardized question types in EyeVQA, including distractor sampling, ordering criteria, coordinate system conventions, and metric definitions.

Single-Choice (SC), Multi-Select (MS), and Variable-Select (VS) Questions. SC questions present exactly four candidate options (A–D) with one deterministic gold answer, testing disease identification, severity grading, or image–text correspondence. MS questions explicitly specify the requirement of multiple correct options. VS questions leave the answer cardinality unstated, requiring models to dynamically determine whether one or several options are correct. To construct challenging distractors and mitigate language priors [18,23], incorrect options are preferentially drawn from clinically related disease classes, adjacent severity levels, or visually similar cases within the same dataset.

True-False (TF) and Ranking (RK) Questions. TF questions prompt the model to verify whether a given clinical statement, diagnostic claim, or qualitative descriptor matches the visual evidence. RK questions evaluate ordinal perception by requiring the model to sort a sequence of fundus images according to a designated clinical attribute (e.g., disease severity, lesion burden, or cup-to-disc ratio). To avoid ambiguity, RK instances are included only when ground-truth ordering is strictly monotonic and clinically indisputable.

Point Location (PL) and Bounding Box (BB) Questions. PL and BB assess spatial perception and visual grounding in fundus photography:

- **Point Location (PL):** Predicts a single normalized coordinate

$$[x, y] \in [0, 1]^2$$

representing key anatomical landmarks, such as the optic disc center or macula fovea.

- **Bounding Box (BB):** Predicts a normalized bounding box

$$[x_{\min}, y_{\min}, x_{\max}, y_{\max}] \in [0, 1]^4$$

circumscribing specific lesions or structural entities.

All spatial targets are extracted from expert-annotated segmentation masks, boxes, or landmarks. During dataset preprocessing, target coordinates undergo exact geometric transformations aligned with image cropping and resizing; samples with inconsistent or ambiguous spatial annotations are systematically excluded [45].

D Detailed Formulations of Evaluation Metrics

Let N_t denote the number of questions of type t . Metrics are reported separately for each question type, with invalid responses assigned a value of zero.

D.1 Single-Choice and True-False Questions

We evaluate SC and TF questions using exact-match accuracy:

$$z_i = \begin{cases} 1, & \hat{y}_i = y_i, \\ 0, & \text{otherwise,} \end{cases} \quad \text{Acc}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} z_i, \quad t \in \{\text{SC}, \text{TF}\}. \quad (1)$$

where y_i and $\hat{y}_i$ are the reference and parsed answers, respectively.

D.2 Multi-Select and Variable-Select Questions

We evaluate MS and VS questions using the Hamming metric (HM) [3]:

$$\text{HM}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} \frac{|Y_i \cap \hat{Y}_i|}{|Y_i \cup \hat{Y}_i|}. \quad (2)$$

where $t \in \{\text{MS}, \text{VS}\}$, and Y_i and $\hat{Y}_i$ are the reference and predicted option sets, respectively. The metric assigns partial credit according to set overlap.

D.3 Ranking Questions

We evaluate RK questions using Kendall's Tau [40]:

$$\text{Tau} = \frac{1}{N_t} \sum_{i=1}^{N_t} \frac{C_i - D_i}{K(K-1)/2}. \quad (3)$$

where $K = 4$ is the number of ranked options, and C_i and D_i are the number of concordant and discordant pairs among K items.

D.4 Bounding Box Questions

We evaluate BB questions using intersection over union (IoU):

$$\text{IoU} = \frac{1}{N_t} \sum_{i=1}^{N_t} \frac{\text{Area}(B_i \cap \hat{B}_i)}{\text{Area}(B_i \cup \hat{B}_i)}. \quad (4)$$

where B_i is normalized reference box and $\hat{B}_i$ is normalized predicted box.

D.5 Point Location Questions

We evaluate PL questions using accuracy with a normalized Euclidean-distance tolerance of 0.05:

$$z_i^L = \begin{cases} 1, & \sqrt{(\hat{x}_i - x_i)^2 + (\hat{y}_i - y_i)^2} \leq 0.05, \\ 0, & \text{otherwise,} \end{cases} \quad \text{Acc}_{L@0.05} = \frac{1}{N_t} \sum_{i=1}^{N_t} z_i^L. \quad (5)$$

where $p_i = (x_i, y_i)$ and $\hat{p}_i = (\hat{x}_i, \hat{y}_i)$ are the normalized reference and predicted points, respectively.

Score Normalization and Refusal Handling. For comparison, all metrics are linearly normalized so that the final scores lie in $[0, 100]$: values in $[0, 1]$ are multiplied by 100, while Kendall’s Tau is mapped from $[-1, 1]$ to $[0, 100]$. If a VLM refuses to answer a question, that question receives a score of zero and remains included in the average.