This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Supplementary Material for EyeVQA

A Introduction

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 [27], ACRIMA [8], and FIVES [16]), 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 [9], BRSET [19], EyePACS [6], and PALM [10]), **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 [8] (2019)	705	FISABIO Oftalmología Médica, Valencia, Spain	Glaucoma	CC BY 4.0
ADAM [9] (2018)	1,200	Zhongshan Ophthalmic Center, Sun Yat-sen University, China	Macula Disease	CC BY-NC-ND 4.0
AIROGS [7] (2021)	101,442	500 screening centers across the United States of America	Glaucoma	CC BY-NC-ND 4.0
BRSET [19] (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 [27] (2026)	187	Peking University Third Hospital, China	Cataract	CC BY 4.0
Retina Dataset [5] (2016)	601	8 open access databases (specific sources undisclosed)	Glaucoma, Cataract, etc.	Undisclosed
DEN [14] (2021)	15,708	Undisclosed (NDA required)	Diabetic Retinopathy, Glaucoma, Macula Disease, etc.	Undisclosed
DeepDRiD [18] (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 [6, 1] (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 [16] (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 [13] (2021)	200	Undisclosed	Glaucoma, Macula Disease, etc.	CC BY 4.0
GAMMA [26] (2021)	300	Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China	Glaucoma	CC BY-NC-ND
JICHI Dataset [24] (2017)	9,940	Jichi Medical University Hospital, Tochigi, Japan	Diabetic Retinopathy	CC BY 4.0
JSIEC1000 [4] (2019)	1,000	Joint Shantou International Eye Centre (JSIEC), Shantou, Guangdong, China	Diabetic Retinopathy, Glaucoma, Macula Disease, etc.	DbCL v1.0
MMAC [22] (2023)	1,391	Shanghai Health and Medical Center (SHMC), China; Shanghai Sixth People's Hospital (SSPH), China	Pathological Myopia	CC BY 4.0
ODIR [17] (2019)	8,000	Multiple hospitals and medical centers in China	Diabetic Retinopathy, Glaucoma, Cataract, Macula Disease, etc.	Undisclosed
ORIGA-light [28] (2010)	650	Singapore Eye Research Institute, Singapore	Glaucoma	Undisclosed
PALM [10] (2019)	1,200	Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China	Pathological Myopia	Non-redistributable
PARAGUAY [3] (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 [21, 11] (2020)	1,200	Multiple medical centers (specific sources not fully disclosed)	Glaucoma	Non-redistributable
ROC [20] (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 [12, 15], 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 [25].

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:

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

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) [2]:

$$\mathrm{HM}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} \frac{|Y_i \cap \hat{Y}_i|}{|Y_i \cup \hat{Y}_i|}. \tag{2}$$

where $t \in \{\mathrm{MS}, \mathrm{VS}\}$, 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 [23]:

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

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):

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

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:

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

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.

References

1. Diabetic retinopathy detection: Competition rules and data usage terms. https://www.kaggle.com/competitions/diabetic-retinopathy-detection/rules, accessed: Aug. 7, 2026
2. Bazoge, A.: Mediqal: A french medical question answering dataset for knowledge and reasoning evaluation. Scientific Data **13**(1), 356 (2026)
3. 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)
4. 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)
5. Choi, S.: retina_dataset: Fundus image dataset for eye disease classification. https://github.com/yiweichen04/retina_dataset (2021), accessed: Aug. 7, 2026
6. Cuadros, J., Bresnick, G.: Eyepacs: an adaptable telemedicine system for diabetic retinopathy screening. Journal of diabetes science and technology **3**(3), 509–516 (2009)
7. 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.: Airogs: Artificial intelligence for robust glaucoma screening challenge. IEEE transactions on medical imaging **43**(1), 542–557 (2023)
8. 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)
9. Fang, H., Li, F., Fu, H., Sun, X., Cao, X., Lin, F., Son, J., Kim, S., Quellec, 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)
10. 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)
11. 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)
12. 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)
13. 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)
14. 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.: Deepopht: 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)
15. 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)

16. 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)
17. 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)
18. 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)
19. 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)
20. 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)
21. 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)
22. 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)
23. 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)
24. 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)
25. 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)
26. 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)
27. 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)
28. 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)