Constructing Ophthalmic MLLM for Positioning-diagnosis Collaboration Through Clinical Cognitive Chain Reasoning

Supplementary Material

I. FundusGen Details

I.1. Data Sources and Annotation

We collect approximately 200K fundus images and their corresponding annotations from both open-source datasets and in-house data.

MM-Retinal[12]. MM-Retinal is a multimodal dataset comprising high-quality image-text pairs collected from professional ophthalmology textbooks.

BRSET[9]. BRSET is the first Brazilian multi-label ophthalmic dataset. It consists of retinal fundus photographs centered on the macula, providing extensive global diagnostic disease labels.

IDRiD[10]. IDRiD is the first dataset representing the Indian population. It includes pixel-level annotations for typical diabetic retinopathy lesions and normal retinal structures. The dataset provides severity grading for diabetic retinopathy and diabetic macular edema for each image.

APTOS2019[1]. This dataset focuses on the severity grading of diabetic retinopathy.

MESSIDOR2[3]. The Messidor-2 dataset is a collection for diabetic retinopathy (DR) screening, where each examination consists of two macula-centered fundus images, one for each eye.

PAPILA[7]. This dataset contains medical records and binocular fundus images from the same patient. It also provides segmentation annotations for the optic cup and optic disc, along with patient-level labels based on clinical assessment.

Retina[5]. This dataset consists of normal and cataract fundus images for cataract detection.

Glaucoma_fundus[6]. This dataset includes glaucoma annotations, providing grading labels for different stages of glaucoma.

In-house Data. A collection of high-quality color fundus images annotated by professional ophthalmologists, including comprehensive annotations of overall disease diagnoses and characteristic lesions.

I.2. Curated Instruction Fine-tuning Data Scheme

This section provides an expanded description of FundusGen. FundusGen is developed to overcome the limitations of conventional ophthalmic datasets and to support the development of domain-specific multimodal large language models (MLLMs) with enhanced clinical reasoning capabilities.

In addition to the annotation process, we curate instruction fine-tuning data tailored to the diverse needs of ophthalmic clinical tasks. We design different types of instructional prompts based on clinical task formats and semantic emphasis:

1. **General Report:** Instructions to generate standardized diagnostic reports (e.g., "Generate a diagnostic report based on the fundus image"). This data serves as startup data during fine-tuning.
2. **Regional QA:** Rule-based instructions that focus on localization and identification tasks (e.g., "Label the location of hard exudates").
3. **Grounding Report:** Prompts that require the report content to directly correspond to image regions (e.g., "Describe the fundus image with reference to its location information").
4. **Multi-turn Diagnostic Reasoning:** Simulated multi-turn dialogues that mimic the clinical inquiry process, where the model integrates information from abnormal regions to generate diagnostic conclusions (e.g., "Based on the characteristics of the fundus image, provide a diagnostic conclusion").
5. **Multi-turn Confirmation Analysis:** Multi-turn dialogues that verify and deepen the evidence chain (e.g., "Describe and analyze the fundus features indicative of glaucoma in this image").

Tasks (4) and (5) explicitly construct the cognitive chain, ensuring that the dataset not only covers high-incidence conditions such as diabetic retinopathy and macular edema but also addresses complex diseases like hypertensive retinopathy and age-related macular degeneration.

II. Fundus-MMBench

To address the clinical requirements for fundus photography, we construct a multimodal evaluation framework specifically dedicated to fundus images, termed Fundus-MMBench. In Fundus-MMBench, each task category comprises 20 test samples. It consists of 31 fine-grained tasks covering three core clinical domains: region-based object recognition (e.g., optic disc identification), disease classification (e.g., glaucoma versus non-glaucoma diagnosis), and severity grading (e.g., assessment of diabetic retinopathy severity). In the disease classification tasks, we implement a case-control balancing strategy to ensure that the number of positive samples is equal to that of negative samples for each disease, thereby mitigating the impact of data distribution bias on evaluation results. Our training data are strictly isolated from Fundus-MMBench, and all evaluation categories in Fundus-MMBench are represented in the training data, allowing us

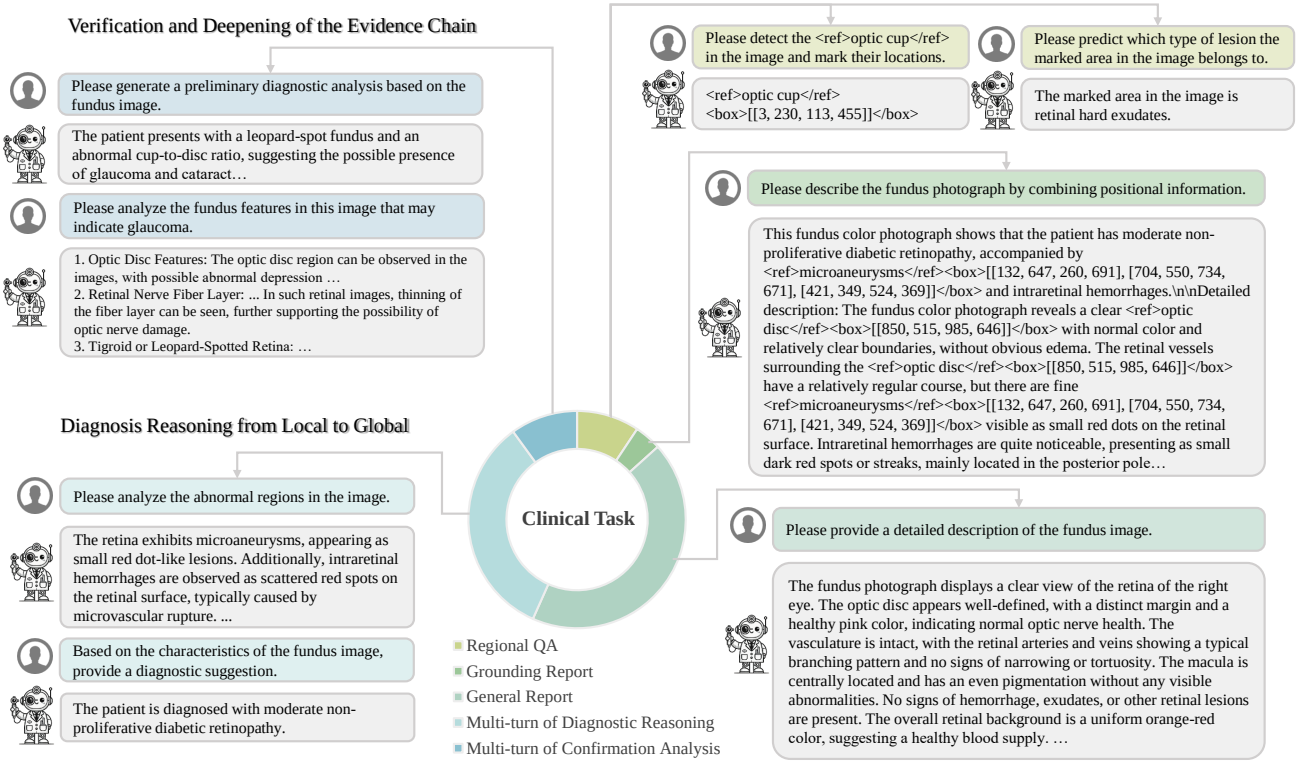


Figure S1. Data types in FundusGen.

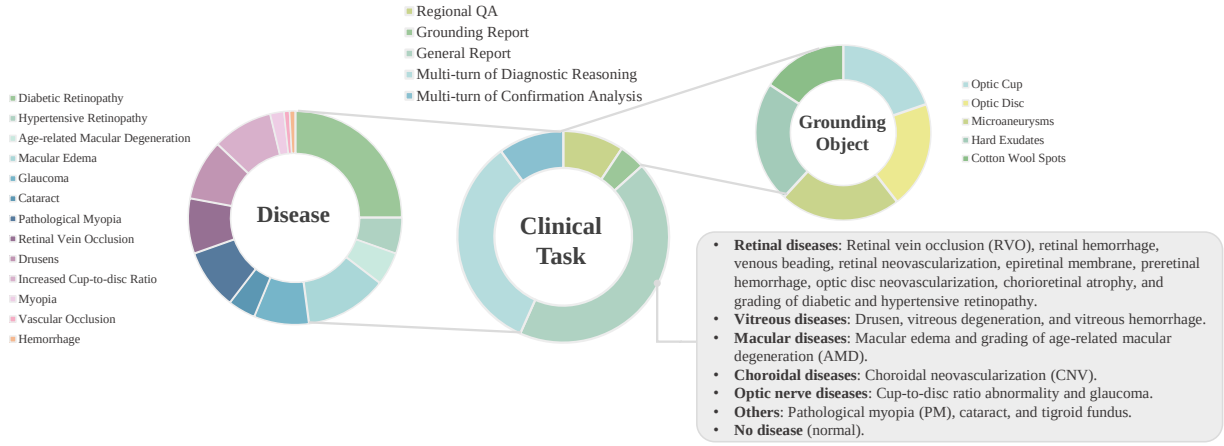


Figure S2. The Composition of FundusGen.

to quantify the performance boundaries of FundusExpert on in-distribution tasks.

Given the pervasive issue of class imbalance in medical dataespecially where abnormal samples far outnumber normal ones, leading to suboptimal model performance on normal samples and an increased risk of misdiagnosiswe implement a sample balancing strategy in the disease diagnosis evaluation on Fundus-MMBench. For each disease, the number of samples exhibiting the condition is maintained at parity with the number of samples not exhibiting it. By balancing positive and negative samples, we aim to preserve robust disease detection performance while enhancing the model's ability to recognize normal cases, thereby reducing the false positive rate and the risk of misdiagnosis in clinical applications.

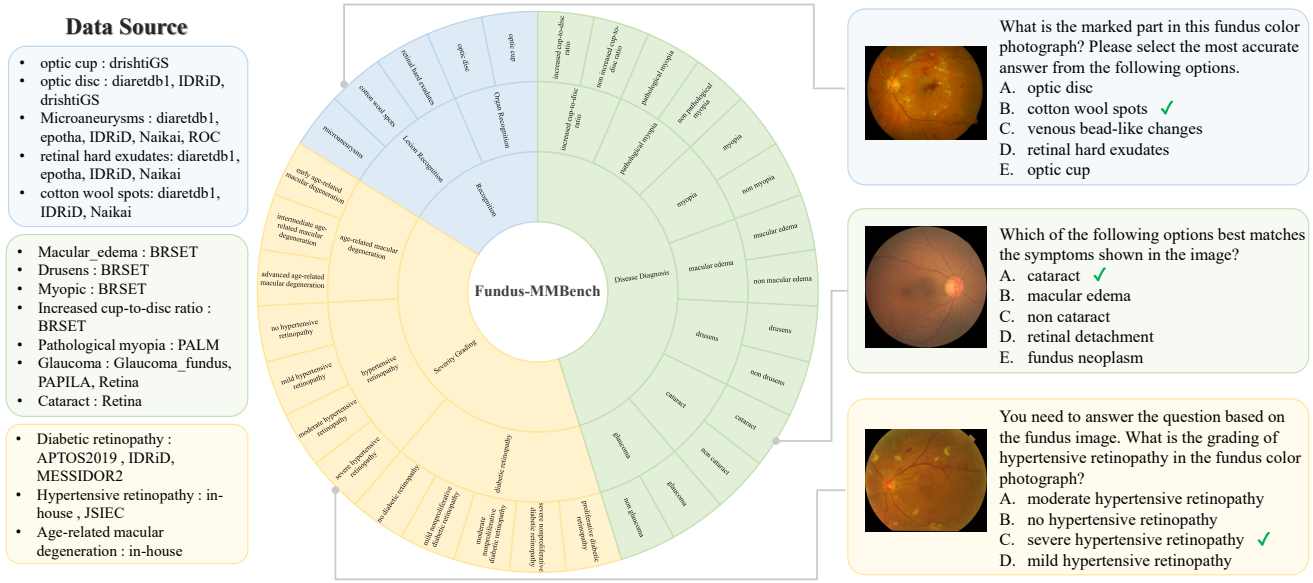


Figure S3. The Composition and Presentation of Fundus-MMBench.

Dataset	Microaneurysms	Hard Exudates	Cotton-wool Spots	Optic Cup	Optic Disc
Open-source Dataset (True Labels)	882	642	291	901	1070
Annotated Dataset (Pseudo-labels)	5357	10089	1876	16551	16720

Table S1. Comparison of label quantities between open-source and annotated datasets.

III. Training Details

Implementation Details For FundusExpert. We employ InternVL2.5[2] as the base model for full-scale instruction tuning. Its vision encoder consists of a 300M InternViT, while its language encoder is a 7B InternLM. During instruction tuning, we unfreeze the vision encoder, MLP, and LLM, optimizing the entire model using 300,000 samples from FundusGen. Training is conducted on four NVIDIA A100 GPUs, with fine-tuning hyperparameters following the official InternVL settings. The per-device batch size is set to 4, with a gradient accumulation step of 8. A cosine learning rate schedule is used, starting at 4e-5, for training over one epoch. We utilize DeepSpeed ZeRO Stage 2 optimization for efficient training.

Implementation Details of Other Architectures. For the fine-tuning of LLaVA-v1.5[8] and Qwen2VL[11], we adhere to the official InternVL hyperparameter settings, conducting training on four NVIDIA A100 GPUs.

IV. Automated Methods in Fundus-Engine

IV.1. Bounding Box Generation

We apply the DBSCAN clustering algorithm[4] to convert pixel-level segmentation labels into bounding box annotations (Table S1). The epsilon value for DBSCAN clustering is set to 160, and the minimum samples parameter is set to 10. If a bounding box has a pixel area greater than the threshold ($>$100), it is added to the candidate list. The bounding boxes are then sorted by area, and the top three largest bounding boxes are retained.

V. Experiment

V.1. Clinical Deployment Efficiency

As shown in Table S2, FundusExpert-mini(1B), optimized for consumer GPUs, excels in accuracy and efficiency over models like InternerVL2.5-38B. On an RTX 4090, it achieves 0.20 img/s (2.0GB VRAM, bs=1), scaling to 2.34 img/s (max BS 128) (Table S2). In contrast, larger models like InternerVL2.5-38B are less accurate, require high-end A100 GPUs, support very limited batch sizes (Max BS 2), and have slow inference. FundusExpert-mini provides an optimal balance for widespread clinical adoption.

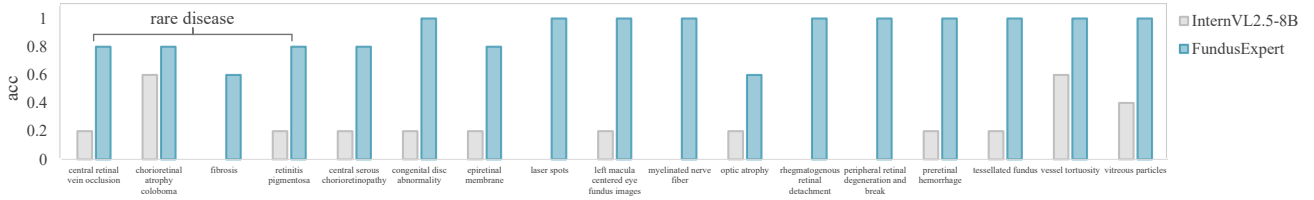
V.2. Performance Evaluation

V.2.1. Extrapolation Ability of FundusExpert

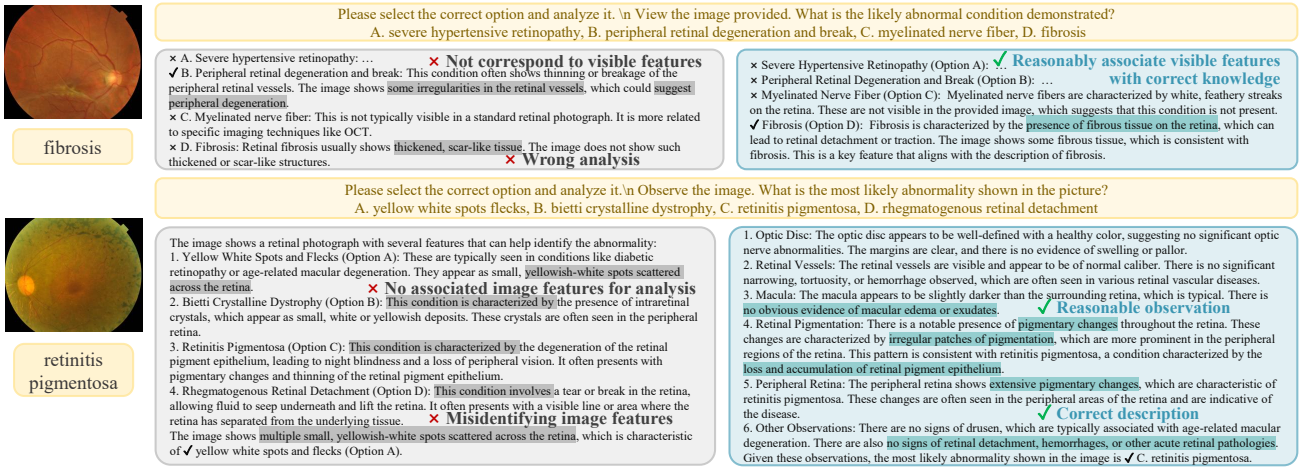
FundusExpert demonstrates extrapolation reasoning ability in out-of-domain tasks on GMAI-MMBench. As shown

Table S2. Model deployment efficiency comparison. Metrics include Accuracy (Acc.) on Fundus-MMBench and GMAI-MMBench, Throughput (Thrpt.) in images per second (img/s), VRAM Memory (Mem.) in GB at batch size 1 (bs=1), Maximum deployable Batch Size (Max BS), and Throughput at Max BS. Results highlighted in gray were obtained on an RTX 4090; all other results were obtained on an A100 GPU.

Model	Params Num	Fundus Acc(%)	GMAI Acc(%)	Thrpt. (img/s)	Mem.(bs1) (GB)	Max BS	Thrpt.(Max) (img/s)
InternerVL2.5-8B	8B	30.6	37.8	0.14	16.3	32	1.43
InternerVL2.5-38B	38B	44.0	42.3	0.03	74.0	2	0.07
FundusExpert-mini	**1B**	63.5	58.3	**0.24**	**2.0**	**512**	**3.54**
FundusExpert	8B	**69.7**	**66.7**	0.14	16.3	32	1.43
FundusExpert-mini	**1B**	63.5	58.3	**0.20**	**2.0**	**128**	**2.34**
FundusExpert	8B	**69.7**	**66.7**	0.10	16.3	4	0.31



a) Zero-shot Performance Improvement for Out-of-domain Testing. (Each category has five samples)



b) Case Study Comparing Responses From InternVL2.5-8B(left) and FundusExpert(right).

Figure S4. Comparison of the foundation model and FundusExpert on out-of-distribution representative categories.

in Table 1, it achieves a 66.7% accuracy rate in zero-shot tasks on GMAI-MMBench, surpassing the base model InternVL2.5 by 30.2%. This is primarily attributed to FundusGen's explicit modeling of clinical feature inference logic. Case comparisons in Figure S4 further validate this ability.

For the "retinitis pigmentosa" diagnostic task in Figure S4(b), FundusExpert locks onto the correct diagnosis based on the peripheral retinal pigment deposition pattern through extensive feature analysis and exclusion of other options, while the pretrain model incorrectly identifies the features.

V.2.2. Evaluation of Zero-shot Ability in Open-domain Tasks

Localization Ability Evaluation.

Under evaluation, the IoU calculation formula is:

$$IoU = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}, \tag{1}$$

where TP is the number of intersection pixels between the predicted box and the ground truth region, FP is the number of redundant pixels in the predicted box that exceed the ground truth region, and FN is the number of missed pixels in

the ground truth region that are not covered by the predicted box.

Clinical Consistency Evaluation in Medical Report Generation.

We propose a multi-granularity semantic matching framework to compute the accuracy of medical report generation tasks. It uses VLM(GPT-4o) to decouple the structured evaluation of clinical logical consistency in generated reports.

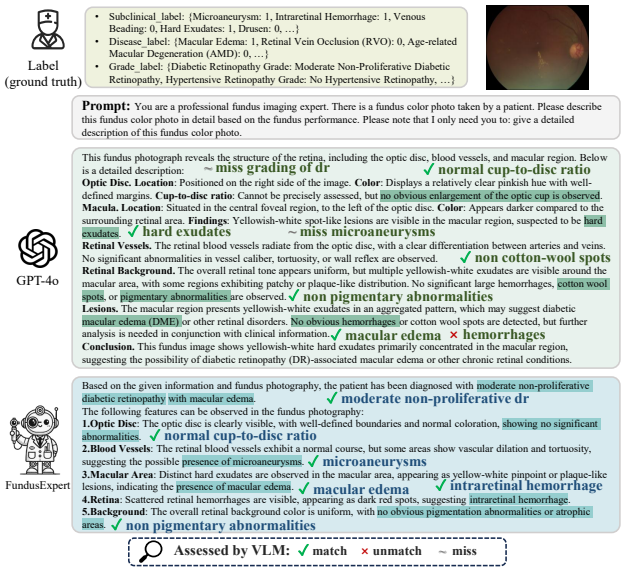


Figure S5. Example of Clinical Consistency Evaluation

Existing likelihood-based benchmarks for medical text generation, such as BLEU and ROUGE, inadequately assess semantic plausibility. To overcome this, we introduce a multi-granularity semantic matching framework that evaluates the accuracy of generated medical reports. This framework leverages a Vision Language Model (VLM), specifically GPT-4o, to perform a structured evaluation of clinical logical consistency.

Let the set of ground-truth labels be $\mathcal{L} = \{l_1, l_2, ..., l_N\}$, which includes both positive and negative findings. Let the set of semantic features extracted from the generated report be $\mathcal{S} = \{s_1, s_2, ..., s_M\}$. The clinical consistency score is defined as:

$$\text{Clinical Consistency} = \frac{\sum_{i=1}^{N} \mathbb{I}(\text{match}(l_i, \mathcal{S}))}{|\mathcal{L} \cup \mathcal{S}|}$$

Where:

- The function $\text{match}(l_i, \mathcal{S})$ checks for a bidirectional semantic correspondence between a label l_i and the set of generated features $\mathcal{S}$, as determined by the VLM.
  For a positive label l_i, a match occurs if the report's semantics $\mathcal{S}$ correctly describe the finding. And for a negative label l_i, a match occurs if the report's semantics $\mathcal{S}$ correctly state the absence of the finding.
- $\mathbb{I}(\cdot)$ is the indicator function, which is 1 if the condition is true and 0 otherwise.
- The denominator $|\mathcal{L} \cup \mathcal{S}|$ is the size of the union of the ground-truth labels and the generated features(determined by the VLM), which normalizes the score.

FundusExpert achieves 77.0% in clinical consistency evaluation, significantly outperforming GPT-4o, which scores 47.6% (+29.4%). This advantage stems from the model's ability to model multi-level pathological associations. For example, in diabetic retinopathy report generation, the model not only accurately identifies microaneurysms and macular edema but also verifies the stage of the lesion through contextual semantics, such as the distribution of retinal hemorrhages.

V.3. Supplement to the ablation experiment results

Cognitive Chain Construction Data Ablation. The results of the clinical QA task evaluation are shown in Table 3. The average diagnostic accuracy for diseases in the GMAI-MMBench decreases by 3.5% for the (2) Cognitive Chain Degradation group compared to (1).

Further analysis reveals that the average diagnostic accuracy for 21 complex diseases, such as retinitis pigmentosa, in the GMAI-MMBench decreases by 4.8% (75.2% $\rightarrow$ 70.4%) for the (2) Cognitive Chain Degradation group compared to (1), indicating that reasoning by constructing a progressive chain, enhances the model's logical deduction ability for complex pathologies. These 21 diseases include 17 of the rarer disease categories in Figure S4 as well as bietti crystalline dystrophy, fundus neoplasm, vkh disease, and pathological myopia. The prevalence of these diseases is relatively low, or they represent more severe or specific pathological conditions than common diseases (such as common myopia, cataracts).

Startup Data Ablation. Startup data enhances the model's basic understanding of different diseases by providing diverse disease descriptions. In addition to the performance degradation in Table 3 (training for 1 epoch), further experiments show that (6) requires 0.5 additional epochs (training for 1.5 epochs) to achieve the same accuracy as (5) on Fundus-MMBench, indicating that there is a delay in convergence without startup data. At the same time, its performance on out-of-distribution GMAI-MMBench worsens, with the gap increasing from $\downarrow$4.5% to $\downarrow$5.9%.

References

[1] Aptos 2019 blindness detection. https://www.kaggle.com/c/aptos2019-blindness-detection. Accessed: 2025-03-08. 1

[2] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 3

[3] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, John-Richard Ordóñez-Varela, Pascale Massin, Ali Erginay, et al. Feedback on a publicly distributed image database: the messidor database. *Image Analysis & Stereology*, pages 231–234, 2014. 1

[4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, pages 226–231, 1996. 3

[5] jr2ngb. Cataract dataset. https://www.kaggle.com/datasets/jr2ngb/cataractdataset. Accessed: 2025-03-08. 1

[6] Ungsoo Kim. Machine learn for glaucoma, 2018. 1

[7] Oleksandr Kovalyk, Juan Morales-Snchez, Rafael Verd-Monedero, Inmaculada Sells-Navarro, Ana Palazn-Cabanes, and Jos-Luis Sancho-Gmez. PAPILA. 2022. 1

[8] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 3

[9] Luis Filipe Nakayama, David Restrepo, João Matos, Lucas Zago Ribeiro, Fernando Korn Malerbi, Leo Anthony Celi, and Caio Saito Regatieri. Brset: a brazilian multilabel ophthalmological dataset of retina fundus photos. *PLOS Digital Health*, 3(7):e0000454, 2024. 1

[10] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018. 1

[11] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

[12] Ruiqi Wu, Chenran Zhang, Jianle Zhang, Yi Zhou, Tao Zhou, and Huazhu Fu. Mm-retinal: Knowledge-enhanced foundational pretraining with fundus image-text expertise. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 722–732. Springer, 2024. 1