This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Xinyao Liu[2,1], Diping Song[†1]
[1]Shanghai Artificial Intelligence Laboratory [2]University of Science and Technology of China
liuxinyao@mail.ustc.edu.cn, songdiping@pjlab.org.cn

## Abstract

*Multimodal large language models (MLLMs) demonstrate significant potential in the field of medical diagnosis. However, they face critical challenges in specialized domains such as ophthalmology, particularly the fragmentation of annotation granularity and inconsistencies in clinical reasoning logic, which hinder precise cross-modal understanding. This paper introduces* ***FundusExpert****, an ophthalmology-specific MLLM with integrated positioning-diagnosis reasoning capabilities, along with* ***FundusGen****, a dataset constructed through the intelligent* ***Fundus-Engine*** *system. Fundus-Engine automates localization and leverages MLLM-based semantic expansion to integrate global disease classification, local object detection, and fine-grained feature analysis within a single fundus image. Additionally, by constructing a clinically aligned cognitive chain, it guides the model to generate interpretable reasoning paths. FundusExpert, fine-tuned with instruction data from FundusGen, achieves the best performance in ophthalmic question-answering tasks, surpassing the average accuracy of the 40B MedRegA by 26.6%. It also excels in zero-shot report generation tasks, achieving a clinical consistency of 77.0%, significantly outperforming GPT-4o's 47.6%. Furthermore, we reveal a scaling law between data quality and model capability ($L \propto N^{0.068}$), demonstrating that the cognitive alignment annotations in FundusGen enhance data utilization efficiency. By integrating region-level localization with diagnostic reasoning chains, our work develops a scalable, clinically-aligned MLLM and explores a pathway toward bridging the visual-language gap in specific MLLMs. Our project can be found at* https://github.com/MeteorElf/FundusExpert.

## 1. Introduction

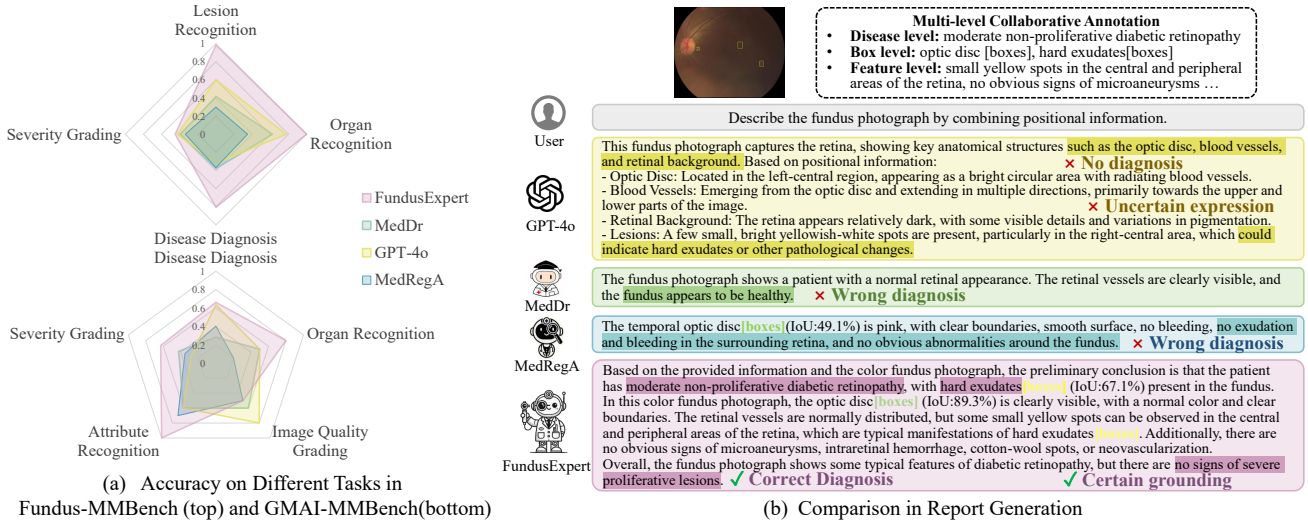
In recent years, multimodal large language models (MLLMs) [3, 8, 21, 23, 37]demonstrate remarkable performance and generalization potential in cross-modal understanding and reasoning, making them promising tools for computer-aided medical diagnosis.

As a uniquely valuable type of medical imaging, fundus images contain rich lesion features that play a crucial role in ophthalmic disease diagnosis[17]. Existing general-purpose models [5, 13, 28] explore tasks such as medical image-based report generation and disease diagnosis, demonstrating the advantages of MLLMs in cross-modal data processing. However, as illustrated in Fig. 1, these general models often exhibit inferior precision and specificity compared to domain-specific models when applied to specialized medical fields due to the need to handle multiple data modalities simultaneously.

Although current mainstream multimodal ophthalmology models achieve marginal performance gains through large-scale training with millions of images [25, 26], their reliance on discretized annotation-based supervision creates misalignment with clinical reasoning processes. These models are not guided to establish a cognitive chain of *region localization* $\rightarrow$ *feature analysis* $\rightarrow$ *diagnostic reasoning* during training. As shown in Fig. 1, such representational deficiencies lead to spatial perception inaccuracies (e.g., optic disc localization errors in MedRegA [29]) and disrupted diagnostic correlations (e.g., wrong diagnosis in MedDr [13] ).

The root cause of these deficiencies in spatial perception and diagnostic correlation lies in the significant granularity fragmentation within the multimodal learning framework for ophthalmology. Current mainstream fundus datasets contain global diagnostic labels (e.g., diabetic retinopathy grading) and scattered fine-grained annotations (e.g., local structure segmentation masks), but these annotations often exist in isolation across different dataset subsets. This "single-image-single-dimension" flat mapping paradigm imposes two critical limitations on the models ability to analyze complex fundus manifestations. First, the model struggles to establish cross-scale semantic associations, such as mapping microaneurysm distribution at the microscopic level to diabetic retinopathy staging at the macroscopic level. Second, fragmented training objectives fail to simulate the progressive

†Corresponding author.




(a) Accuracy on Different Tasks in Fundus-MMBench (top) and GMAI-MMBench(bottom)

(b) Comparison in Report Generation

Figure 1. Comparison of FundusExpert with other models.

diagnostic pathway adopted by clinicians, which involves analyzing the spatial distribution of fundus structures and local features to infer disease progression.

To mitigate the fragmentation across different annotation granularities and better align with clinical workflows, we construct the FundusGen dataset using the intelligent data system Fundus-Engine. This dataset incorporates a cognitive chain that progresses from local-to-global diagnostic reasoning to in-depth validation of the evidential chain. Fundus-Engine enhances annotation efficiency through automated annotation mechanism, achieving a three-stage collaborative annotation process on a single image. At the model training level, we adopt a task-oriented approach that prioritizes data quality over sheer data scale. We investigate a selective strategy for instruction-tuning data, systematically analyzing how data selection patterns and scale influence model performance to better adapt to ophthalmic tasks.

Based on the above methodology, we successfully develop a specialized ophthalmic MLLM, FundusExpert. We conduct a comprehensive evaluation of FundusExpert across multiple dimensions of medical-related tasks, including region recognition and detection, clinical question answering, and medical report generation. In clinical question-answering evaluations, FundusExpert outperforms other domain-specific models and proprietary commercial models in both in-distribution and out-of-distribution benchmarks. Specifically, in the GMAI-MMBench[6], FundusExpert(8B) surpasses the accuracy of the 40B MedRegA by 26.6%.

This paper presents three key contributions:

1. **Proposal of the Fundus-Engine and the construction of a collaborative annotation framework with a clinical cognitive reasoning chain, resulting in the FundusGen dataset.** We achieve collaborative annotation of disease classification, regional localization, and lesion characteristics within a single fundus image, enhancing the association of ophthalmic data across different granularity levels. By aligning with clinical reasoning, we establish a semantic-associative cognitive chain in FundusGen, which expands the models disease analysis capabilities and improves interpretability.
2. **Development of an ophthalmic MLLM with integrated positioning-diagnosis reasoning: FundusExpert.** This MLLM demonstrates the ability of region-semantic self-reference without the need for external tools. It can associate the spatial location of regions in medical images with corresponding statements in the conversation. Additionally, it implements multi-task progressive inference on fundus color images, offering promising potential as a foundational model for the ophthalmology field.
3. **Revealing the scaling law properties of medical multimodal data.** Through subset sampling experiments on FundusGen, we observe a significant scaling law between model performance and data volume. This finding not only quantitatively supports the data quality of FundusGen but also reveals that a granularity-fusion annotation aligned with clinical cognitive logic enhances data utilization efficiency. This provides a reference for constructing specific MLLMs.

## 2. Related Works

**Ophthalmic Supervised Learning Paradigms.** VisionFM [25] constructs a multimodal foundation model based on 3.4 million ophthalmic images, achieving marginal improvements in performance. RetiZero [30] employs contrastive pretraining on 340,000 eye fundus image-text pairs, covering over 400 diseases, but its global alignment strategy

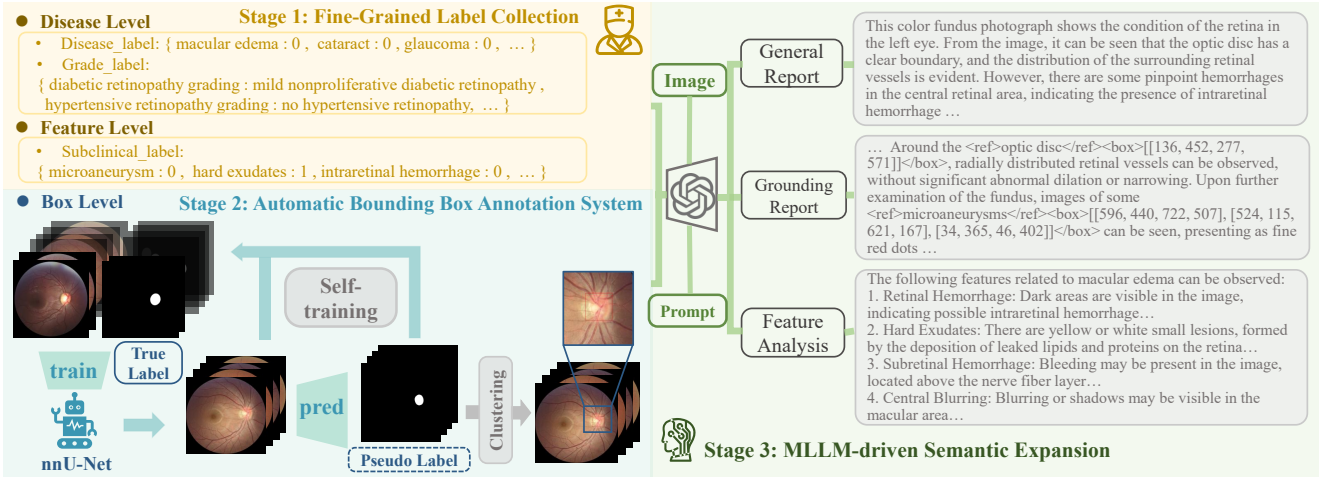



Figure 2. Collaborative Annotation of Fundus-Engine.

is susceptible to interference from false negative samples (e.g., images and texts that are semantically identical but misclassified as negative pairs). To address this issue, ViLReF [34] proposes the Weighted Similarity Coupling Loss and dynamic memory queue, which guide label extraction and compensate for the absence of false negative samples using expert knowledge. However, such methods still rely on a single image-single label coarse-grained alignment, lacking hierarchical semantic connections in the annotation system and depending on large-scale supervised learning paradigms.

**Ophthalmic-Specific Multimodal Models.** RETFound [36] extracts general representations from 1.6 million unlabeled eye fundus images through self-supervised learning, but its pretraining process does not integrate the text modality, making it difficult for the model to establish fine-grained associations between visual features and clinical descriptions. DeepDR-LLM [16] combines 372,000 primary care chronic disease diagnosis and management data to optimize LLM training. However, the model's training tasks primarily focus on diabetic retinopathy grading, making it challenging to extend to other fundus diseases. VisionUnite [19] fine-tunes with 296,379 high-quality eye fundus image-text pairs. However, its text data is largely generated by GPT-4V, leading to limitations in data quality and the accuracy of medical knowledge.

**Region-Aware Medical MLLMs.** Enhancing the model's region-aware capabilities for medical images has become a critical research direction[33]. General medical MLLMs such as MedDr [13] improve disease diagnosis accuracy through expert collaboration mechanisms, but their region-awareness relies on external tools, making semantic self-reference challenging. In the ophthalmic domain, VisionFM [25] adopts a two-stage "detection followed by description" paradigm, which results in a disconnection between region analysis and diagnostic decisions. This decoupled design deviates from the integrated cognitive logic of clinical doctors, limiting the model's reliability and interpretability in complex scenarios.

## 3. Fundus-Engine

We construct the Fundus-Engine achieving the collaborative annotation of disease classification, regional localization, and lesion characteristics within a single fundus image. By simulating the progressive cognitive process of human clinicians during diagnosis, we establish an explicit cognitive chain, resulting in FundusGen.

The Fundus-Engine system employs a dual-path automated annotation framework to enhance the efficiency: (1) The automatic bounding box annotation system utilizes a semi-supervised self-training nnU-Net to perform regional segmentation, followed by a spatial clustering algorithm to aggregate pixel-level predictions into bounding boxes. (2) An MLLM-driven semantic expansion algorithm transforms discrete medical annotations into clinically standardized descriptive texts.

### 3.1. Three-Stage Collaborative Annotation

**Stage 1: Fine-Grained Label Collection.** We collect 20w fundus images with fine-grained annotations from open-source datasets (e.g.[1, 9, 15, 22, 24, 32]) and in-house datasets. These images are expert-annotated by ophthalmologists, providing granular labels for both global disease classification and detailed lesion annotations.

**Stage 2: Automated Bounding Box Annotation System.** To address the lack of bounding box annotations in existing datasets, we propose an automatic bounding box annotation system. Since open-source datasets generally lack direct bounding box annotations, we generate them by clustering segmentation labels. Based on both clinical guidelines and data availability, we select two representative fundus

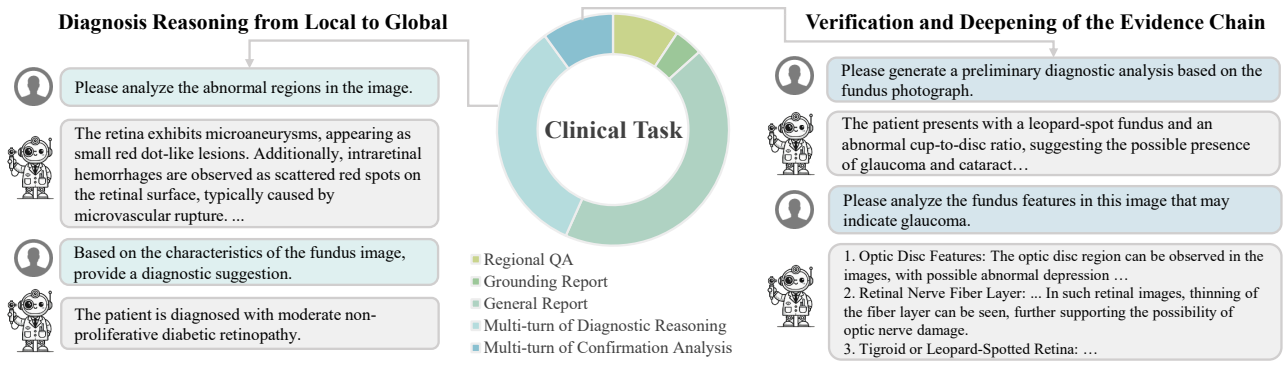



Figure 3. Two Typical Interaction Patterns of Clinically Aligned Cognitive Chains and Curated Instruction Fine-tuning Data Scheme.

structures (optic cup and optic disc) and three lesion types (hard exudates, microaneurysms, and cotton-wool spots) for bounding box annotation. The automated bounding box annotation system consists of the following steps:

- **Data Preprocessing:** We integrate segmentation annotations from publicly available datasets and apply the filtering method from [35] to remove low-quality images, constructing a foundational training set (<1,000 samples per category).
- **Model Training:** For different categories, we build separate nnU-Net [14] segmentation networks, leveraging a feature pyramid structure to extract multi-scale features.
- **Semi-Supervised Expansion:** We design an iterative self-training process that generates pseudo-labels for unannotated lesion regions in images labeled during Stage One. These pseudo-labels are incorporated into the training set for iterative optimization. We evaluate the performance of the automated segmentation annotation on cross-domain datasets and demonstrate the feasibility of pseudo-labels through experimental results presented in the appendix.
- **Bounding Box Generation:** The DBSCAN clustering algorithm [11] is applied to convert pixel-level segmentation results into standardized bounding box annotations.

**Stage 3: MLLM-Driven Semantic Expansion.** To bridge the semantic gap between discrete annotations and the training requirements of multimodal models, we employ an MLLM-based cross-modal alignment framework.

- **Structured Parsing:** The annotation results from Stages One and Two are transformed into structured labels.
- **Prompt Engineering:** A constraint-driven prompt framework is designed to ensure that the generated text meets the following criteria: (1) *Observational Objectivity*: All descriptions are traceable to image pixel features. (2) *Clinical Relevance*: Implicit diagnostic clues are incorporated while avoiding conclusive statements.
- **Text Generation:** The structured annotations are mapped to natural language descriptions via the MLLM(GPT-4o), producing compound texts that integrate localization information (e.g., "optic disc located at [box]") and diagnostic reasoning (e.g., "fundus disease grading inferred based on lesion distribution").
- **Quality Control:** Clinical experts review the generated texts in a double-blind manner to ensure compliance with clinical standards. Texts that do not meet the required quality are discarded or regenerated.

## 3.2. Construction of Clinically Aligned Cognitive Chains

To simulate the progressive cognitive process of human clinicians from lesion observation to comprehensive diagnosis, in the training, we construct cognitive chains within MLLMs using multi-turn dialogues. This approach guides the model in generating interpretable reasoning paths, as shown in Figure 3. The core logic follows a *region localization* $\rightarrow$ *feature analysis* $\rightarrow$ *diagnostic reasoning* cognitive pathway, with typical interaction patterns as follows:

### 3.2.1. Diagnosis Reasoning from Local to Global

The first dialogue round focuses on abnormal regions, with a prompt such as "Please analyze the abnormal regions in the image." The model's response includes an analysis of the abnormal areas, incorporating positional information if available. In the second round, the model is prompted to integrate clinical knowledge for analysis, such as "Based on the characteristics of the fundus image, provide a diagnostic suggestion." At this stage, the model may integrate previously mentioned abnormal region characteristics and spatial distributions, obtaining the diagnostic conclusion that "The patient is diagnosed with ..."

### 3.2.2. Verification and Deepening of the Evidence Chain

The initial instruction prompts the model to generate a coarse-grained diagnostic overview based on image features: "Please generate a preliminary diagnostic analysis based on the fundus image." The model synthesizes the image features to provide an initial assessment: "...suggesting a potential presence of glaucoma and cataracts...". Subsequent

dialogues refine the evidence chain by prompting verification of fine-grained features for specific diseases. For example, a follow-up query may ask: "Please analyze the fundus features in this image that may indicate glaucoma." The model then expands on its reasoning based on image features: "1. Optic Disc Features: ..."

# 4. FundusExpert

FundusExpert is a positioning-diagnosis collaborative multimodal model designed to address visual-language tasks related to ophthalmology. In this section, we first introduce the curated scheme for instruction fine-tuning data. Then, we present the training process of the model. Finally, we introduce Fundus-MMBench, a standardized multimodal evaluation framework focused on fundus images.

## 4.1. Curated Instruction Fine-tuning Data Scheme

To meet the core needs of clinical ophthalmology, we design instructions for various tasks to enhance the model's profound understanding of fundus images, including: (1) **General Report:** Generate standardized diagnostic reports (e.g., "Generate a diagnostic report based on the fundus image"). This part of the data serves as the initialization data in fine-tuning experiments, and the ablation results are presented in the Experiment section; (2) **Regional QA:** These instruction data can be generated through rules, focusing on localization and identification (e.g., "Label the location of hard exudates"); (3) **Grounding Report:** The report content must be directly associated with image regions (e.g., "Describe the fundus image with positional information"); (4) **Multi-turn Diagnostic Reasoning:** Simulate the doctor's questioning process, diagnose diseases based on characteristics of abnormal regions and their location distribution (e.g., "Provide a diagnostic suggestion based on the characteristics of the fundus image"); (5) **Multi-turn Confirmation Analysis:** Verify and deepen the evidence chain through multi-turn dialogues (e.g., "Describe and analyze the fundus features related to glaucoma in this image"). Instructions (4) and (5) explicitly demonstrate the construction of cognitive chains.

## 4.2. FundusExpert Training

Given the powerful multimodal understanding capabilities of InternVL2.5 [7] and the extensive medical knowledge pre-training, we use it as the base model for instruction fine-tuning. The performance of InternVL2.5, as shown in Table 1, surpasses even larger medical domain-specific models like MedDr in terms of its performance on the ophthalmic fundus image modality. We fully fine-tune the entire network on the 8B version to obtain FundusExpert, which has a visual encoder of 300M parameters (InternViT) and a language encoder of 7B parameters (InternLM). Since our training data consists of curated instruction fine-tuning data, the entire training process involves only the instruction fine-tuning.

## 4.3. Fundus-MMBench

In the objective evaluation of our experiments, we analyze model performance using two standardized evaluation frameworks: GMAI-MMBench [6] and Fundus-MMBench. These two frameworks are forming a complementary evaluation system.

GMAI-MMBench [6], as a general medical multimodal benchmark, covers various medical imaging modalities, including X-ray, CT, and fundus image. This study focuses on its fundus image subset, which comprehensively includes 60 task categories, primarily involving ophthalmic disease diagnosis and grading, including over ten rare ophthalmic diseases. Notably, approximately 45% of the disease categories are not explicitly annotated in our training data, making this benchmark effective in assessing FundusExperts generalization ability in out-of-distribution scenarios. However, as a general medical evaluation system, GMAI-MMBench exhibits a sample distribution imbalance in the fundus photography modality, where over 90% of the categories contain only around 5 test samples, potentially leading to significant variability in evaluation results due to small sample effects.

To address the clinical needs of fundus image, we construct Fundus-MMBench, a multimodal evaluation framework dedicated to fundus imaging. Fundus-MMBench increases the number of test samples per task category to 20. 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 vs. non-glaucoma diagnosis), and severity grading (e.g., diabetic retinopathy severity assessment). Our training dataset FundusGen is strictly isolated from Fundus-MMBench. All evaluation categories in Fundus-MMBench are represented in the training data, enabling the quantification of FundusExperts performance boundaries in in-distribution tasks.

# 5. Experiment

## 5.1. Performance Evaluation

### 5.1.1. Clinical Question Answering Evaluation

In the clinical QA task, we evaluate the model performance using two objective benchmarks: Fundus-MMBench and GMAI-MMBench [6](fundus image subset), following the evaluation setup in VLMEvalKit[10]. We employ a deterministic sampling strategy with a temperature parameter of 0 to ensure response stability. The semantic matching mechanism allows for fault-tolerant answers (e.g., "A. optic cup" being equivalent to option A), making the evaluation more aligned with real-world scenarios. For the evaluation of commercial closed-source models such as GPT-4o, we repeatedly prompt the model until it generates a response, if the model initially refuses to respond.

The evaluation results are shown in Table 1. The experimental results indicate that FundusExpert achieves optimal

| Type | Model | Params | Fundus-MMBench | GMAI-MMBench[6]† |
|---|---|---|---|---|
| Specialist | MedDr[13] | 40B | 34.8% | 33.7% |
| | MedRegA[29] | 40B | 40.3% | 40.1% |
| | GMAI-VL[18] | 8B | 44.5% | 56.1% |
| | RetiZero[30]‡ | - | 42.0% | 59.2% |
| Generalist | LLaVA-v1.5[20] | 7B | 21.1% | 32.7% |
| | Qwen2-VL[31] | 7B | 33.5% | 35.9% |
| | Qwen2.5-VL[4] | 7B | 30.6% | 37.8% |
| | InternVL2.5[7] | 8B | 40.6% | 36.5% |
| | GPT-4o[2] | - | 41.6% | 57.4% |
| | Gemini-2.0-pro[27] | - | 46.1% | 59.0% |
| Fine-tuned by FundusGen | LLaVA-v1.5* | 7B | 38.2%(+17.1%) | 41.7%(+9.0%) |
| | Qwen2-VL* | 7B | 57.4%(+23.9%) | 56.4%(+20.5%) |
| | FundusExpert-mini | 1B | 63.5%(+30.0%) | 58.3%(+28.5%) |
| | **FundusExpert** | 8B | **69.7%**(+29.1%) | **66.7%**(+30.2%) |

† Only fundus images are selected for evaluation. GMAI-MMBench mentioned in the following text refers to the same subset.
‡ CLIP-base vision-language foundation model.
* Fine-tuned by FundusGen.

Table 1. Performance comparison on clinical QA tasks.

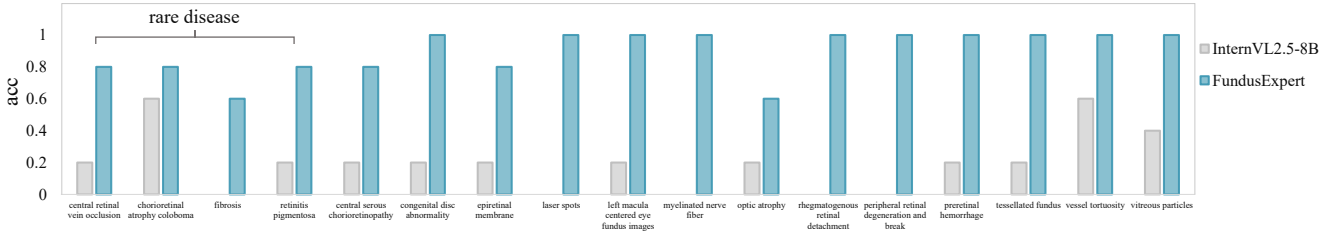



Figure 4. Zero-shot Performance Improvement for Out-of-domain Testing. (Each category has five samples)

accuracy in both evaluation frameworks, surpassing the ophthalmic vision-language foundation model RetiZero [30] by an average of 17.6%. RetiZero [30] employs a purely contrastive learning framework (CLIP-based), with a large collection of fundus images and pure text data covering over 400 diseases. While it outperforms other models in the GMAI-MMBench evaluation, its pretraining task focuses on zero-shot classification and cross-domain recognition. It can only perform global text matching and cannot directly generate texts.

**Validation of the FundusGen Dataset Effectiveness.** The experiments (Table 1) show that when using a unified instruction fine-tuning scheme, multiple Vision-Language models achieve significant performance improvements. After fine-tuning with FundusGen, LLaVA-v1.5 [20] shows a 13.1% average accuracy improvement (26.9% $\rightarrow$ 40.0%), and Qwen2-VL [31] shows a 22.2% average accuracy improvement (34.7% $\rightarrow$ 56.9%), validating the dataset's value in adapting general multimodal models to the domain.

**Extrapolation Ability of FundusExpert.** FundusExpert demonstrates significant extrapolation reasoning ability in out-of-domain tasks on GMAI-MMBench. As shown 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 4 further validate this ability.

#### 5.1.2. Zero-shot Ability in Open-domain Tasks

**Localization Ability Evaluation.** For the localization boxes output by the model, we perform a quantitative analysis using Intersection over Union (IoU). The spatial alignment at the pixel level is used to assess the localization accuracy. The zero-shot results are shown in Table 2.

| Model | $IoU_{OC}$ | $IoU_{OD}$ | $IoU_{EX}$ | $IoU_{CWS}$ | $IoU_{MA}$ |
|---|---|---|---|---|---|
| LLaVA-1.5-7B[20] | 0.021 | 0.064 | 0.005 | 0.002 | 0.002 |
| InternVL2.5-8B[7] | 0.036 | 0.077 | 0.035 | 0.004 | 0.007 |
| MedRegA[29] | 0.302 | 0.543 | 0.038 | 0.006 | 0.011 |
| **FundusExpert** | **0.632** | **0.738** | **0.194** | **0.141** | **0.116** |

*Note:* OC: Optic Cup, OD: Optic Disc, EX: Hard Exudates, CWS: Cotton Wool Spots, and MA: Microaneurysms.

Table 2. Performance on the object detection task.

**Clinical Consistency Evaluation in Medical Report Gen-**

**eration.** 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 VLM(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 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. $\mathbb{I}(\cdot)$ 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.

In the evaluation of 200 reports generation, FundusExpert achieves 77.0% in clinical consistency, significantly outperforming GPT-4o, which scores 47.6% (+29.4%).

## 5.2. Verification of Dataset Scaling Law

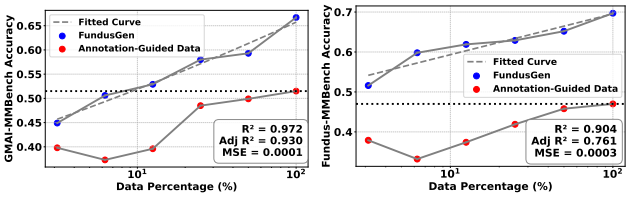



Figure 5. The change of model accuracy on GMAI-MMBench(left) and Fundus-MMBench(right) under different data percentage. The blue dots represent the model trained with FundusGen, and the red dots represent the model trained with Classification Annotation-Guided Data. The gray dotted line is the fitting curve, showing scaling law.

In the subset sampling experiment of the FundusGen dataset (fine-tuning based on InternVL2.5-8B), the model trained with FundusGen demonstrates a scaling law with respect to the percentage of data, as shown in Figure 5. This indicates that the FundusGen dataset, built on clinical cognitive chains, possesses high information density, strong scalability, and low semantic noise outperforming comparison datasets such as Classification Annotation-Guided Data. The latter is generated by directly combining fine-grained annotations(the same disease labels and feature labels as FundusGen) with images and using GPT-4o to produce comparative data. However, Classification Annotation-Guided Data does not have localization information and lacks the explicit instructions necessary to guide the model in generating interpretable reasoning paths.

For performance in GMAI-MMBench, The fitted curve shows that models trained with FundusGen (blue points) exhibit a significant power-law scaling behavior, which can be expressed as $L \propto N^{\alpha}, \alpha = 0.068$, with a correlation coefficient of $R^2 = 0.972$, adjusted $R^2 = 0.930$, and MSE = 0.0001. As the data volume increases, the model trained with FundusGen diagnostic accuracy shows a stable and predictable upward trend. In contrast, models trained with Classification Annotation-Guided Data (red points) show no significant performance improvement as the data size increases, and in some data intervals, performance even declines. For performance in GMAI-MMBench, a comparative experiment reveals that fine-tuning on a 10% subset of FundusGen produces results comparable to those obtained using 100% of Classification Annotation-Guided Data.

## 5.3. Ablation on Dataset

The ablation experiments (Table 3) aim to validate the effectiveness of the data selection scheme in Section 4.1, with the experimental setup being the same as in Section 4.2.

| Data | Modification | Fundus-MMBench | GMAI-MMBench |
|---|---|---|---|
| (1) Complete FundusGen | Cognitive | 69.7% | 66.7% |
| (2) Cognitive Degradation | Chain | 67.1% (↓2.6%) | 63.2% (↓3.5%) |
| (3) Sampled FundusGen 1 | Region | 68.9% | 64.7% |
| (4) Region Data Removal | Perception | 65.3% (↓3.6%) | 59.3% (↓5.4%) |
| (5) Sampled FundusGen 2 | Startup | 66.2% | 60.9% |
| (6) Startup Data Removal | Data | 62.9% (↓3.3%) | 56.4% (↓ 4.5%) |

Table 3. Ablation Experiment. Comparison of Clinical Question Answering Results under Different Data Conditions and the Same Training Setup.

**Cognitive Chain Construction Data Ablation.** To validate the effectiveness of the explicit reasoning mechanism in multi-turn dialogues, we compare: (1) Full FundusGen group: directly using FundusGen; (2) Cognitive Chain Degradation group: splitting multi-turn diagnostic instructions into independent single-turn tasks, which disrupt the continuity of the cognitive chain, with random sampling of instructions of the same scale as (1). 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), indicating that reasoning by constructing a progressive chain, enhances the model's logical deduction ability for complex pathologies.

**Region-Aware Data Ablation.** To verify the effectiveness of the positioning-diagnosis coordination mechanism, we compare: (3) FundusGen Control group 1: randomly sampling from FundusGen with the same scale of instructions as (4);

Table 4. Out-of-domain data test results. The same number of image-true labels and image-pseudo labels of different categories are used as training data to train different segmentation models and test them on Messidor[9].

| Category | True Labels | | Pseudo Labels$_{1st_round_prediction}$ | | Iterative Pseudo Labels$_{2nd_round_prediction}$ | |
|---|---|---|---|---|---|---|
| | $Dice$ | $IoU_{pixel_level}$ | $Dice$ | $IoU_{pixel_level}$ | $Dice$ | $IoU_{pixel_level}$ |
| Hard Exudates | 29.1% | 19.6% | 29.9% | 20.3% | 29.6% | 20.2% |
| Microaneurysms | 21.1% | 12.4% | 21.3% | 12.7% | 25.0% | 15.2% |
| Cotton-wool Spots | 28.3% | 20.3% | 25.9% | 18.3% | 28.6% | 20.2% |
| Optic Cup | 57.3% | 45.9% | 61.4% | 50.3% | 60.1% | 49.4% |
| Optic Disc | 75.3% | 66.5% | 82.0% | 73.4% | 80.0% | 71.2% |

(4) Region Data Removal group: removing all instructions in FundusGen containing region annotations (e.g., bounding box annotations, region-text alignment tasks) while retaining other types of tasks. The experimental results show that the (4) Region Data Removal group experiences an overall decrease of 3.6% in Fundus-MMBench. Additionally, group (4) experiences an overall decline of 5.4% on the out-of-distribution GMAI-MMBench, indicating that region-level annotations contribute to the model's spatial semantic understanding and enhance its generalization ability.

**Startup Data Ablation.** To quantify the value of using General Report(enhances the model's basic understanding of different diseases) as startup data, we compare: (5) FundusGen Control group 2: randomly sampling from FundusGen with the same scale of instructions as (6); (6) No Startup Data group: removing all standardized diagnostic report data. 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.

### 5.4. Data Generator

| Model | Fundus-MMBench | GMAI-MMBench |
|---|---|---|
| Qwen2-VL-7B$^{\dagger}$ | 35.6% | 39.4% |
| Qwen2-VL-7B* | 40.1% | 41.7% |
| InternVL2.5-8B$^{\dagger}$ | 48.2% | 46.2% |
| InternVL2.5-8B* | 50.8% | 48.4% |

$^{\dagger}$ Fine-tuned by GPT-4o-generated data.
* Fine-tuned by FundusExpert-generated data.

Table 5. Comparison of synthetic data fine-tuning performance.

Experiment results(5.1.2) from clinical consistency evaluation show that FundusExpert outperforms GPT-4o in zero-shot ophthalmic medical report generation and domain knowledge comprehension. This finding suggests a paradigm shift in data generation from commercial closed-source models (e.g., GPT-4o) to lightweight domain-specific models, enabling more efficient iterative data acquisition.

We replace GPT-4o with FundusExpert as the MLLM for semantic expansion in Fundus-Engine. By leveraging the multi-level annotation labels in the dataset, we generate medical texts for 100K images(outside the training data) and construct single-turn General Report. Instruction fine-tuning experiments are subsequently conducted on Qwen2-VL[31] and InternVL2.5[7]. As shown in Table 5, the fine-tuning results using synthetic data from the lightweight domain-specific model surpass those obtained from GPT-4o-generated data.

### 5.5. Pseudo-Label Accuracy Evaluation

This experiment aims to assess the quality and usability of the bounding box pseudo-labels(Table 4). In this experiment, Messidor[9] serves as an out-of-domain(OOD) test set to evaluate the cross-domain predictive performance of models trained with real labels versus pseudo-labels. Our primary objective is to demonstrate the segmentation model's performance in OOD tasks, as the process of annotating in-house data with bounding box pseudo-labels using a model trained on open-source data inherently represents a OOD task.

## 6. Conclusion and Discussion

We present FundusExpert, an ophthalmic MLLM with integrated positioning-diagnosis reasoning, along with the FundusGen dataset featuring hierarchical semantic fusion. Our experiments show that FundusExpert achieves 69.7% diagnostic accuracy on the Fundus-MMBench exceeding GPT-4o by 28.1% and attains 98% accuracy in regional identification tasks. We also reveal scaling law in medical multimodal data ($L \propto N^{0.068}$) and investigate that cognitive-aligned annotations enhance data utilization efficiency.

The performance of FundusExpert provides a foundation for the next stage of development, which involves broadening its adaptability for dynamic reasoning and diverse scenarios. Promising methodologies, including test-time scaling and reinforcement learning-based post-training like Deepseek-r1[12], can guide this evolution. Future research will integrate reinforcement learning with FundusGen's semantic hierarchy to enhance model adaptability in low-annotation settings, and expand its performance envelope in specialized medical applications.

## 7. Acknowledgment

This work was supported by Shanghai Artificial Intelligence Laboratory. We would like to thank Professor Yu Qiao for his support for this project. We also thank Professor Xiaodong Sun, Professor Huixun Jia, Yanlin Qu, and Guanran Zhang from Shanghai General Hospital, as well as Zefeng Yang from the Zhongshan Ophthalmic Center of Sun Yat-sen University, for their contributions in labeling and reviewing the experimental data. We are also grateful to Guang Liang for completing parts of the paper writing.

## References

[1] Aptos 2019 blindness detection. https://www.kaggle.com/c/aptos2019-blindness-detection. Accessed: 2025-03-08. 3

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 6

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023. 1

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6

[5] Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, et al. Biomedlm: A 2.7 b parameter language model trained on biomedical text. *arXiv preprint arXiv:2403.18421*, 2024. 1

[6] Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *arXiv preprint arXiv:2408.03361*, 2024. 2, 5, 6

[7] 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. 5, 6, 8

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1

[9] 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. 3, 8

[10] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201, 2024. 5

[11] 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. 4

[12] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 8

[13] Sunan He, Yuxiang Nie, Zhixuan Chen, Zhiyuan Cai, Hongmei Wang, Shu Yang, and Hao Chen. Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. *arXiv preprint arXiv:2404.15127*, 2024. 1, 3, 6

[14] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 4

[15] Oleksandr Kovalyk, Juan Morales-Snchez, Rafael Verd-Monedero, Inmaculada Sells-Navarro, Ana Palazn-Cabanes, and Jos-Luis Sancho-Gmez. PAPILA. 2022. 3

[16] Jiajia Li, Zhouyu Guan, Jing Wang, Carol Y Cheung, Yingfeng Zheng, Lee-Ling Lim, Cynthia Ciwei Lim, Paisan Ruamviboonsuk, Rajiv Raman, Leonor Corsino, et al. Integrated image-based deep learning and language models for primary diabetes care. *Nature medicine*, 30(10):2886–2896, 2024. 3

[17] Tao Li, Wang Bo, Chunyu Hu, Hong Kang, Hanruo Liu, Kai Wang, and Huazhu Fu. Applications of deep learning in fundus images: A review. *Medical Image Analysis*, 69: 101971, 2021. 1

[18] Tianbin Li, Yanzhou Su, Wei Li, Bin Fu, Zhe Chen, Ziyan Huang, Guoan Wang, Chenglong Ma, Ying Chen, Ming Hu, et al. Gmai-vl & gmai-vl-5.5 m: A large vision-language model and a comprehensive multimodal dataset towards general medical ai. *arXiv preprint arXiv:2411.14522*, 2024. 6

[19] Zihan Li, Diping Song, Zefeng Yang, Deming Wang, Fei Li, Xiulan Zhang, Paul E Kinahan, and Yu Qiao. Visionunite: A vision-language foundation model for ophthalmology enhanced with clinical knowledge. *arXiv preprint arXiv:2408.02865*, 2024. 3

[20] 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. 6

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 1

[22] 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. 3

[23] OpenAI. Chatgpt. https://openai.com/blog/chatgpt, 2023. Accessed: 2023-12-22. 1

[24] 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. 3

[25] Jianing Qiu, Jian Wu, Hao Wei, Peilun Shi, Minqing Zhang, Yunyun Sun, Lin Li, Hanruo Liu, Hongyi Liu, Simeng Hou, et al. Development and validation of a multimodal multitask vision foundation model for generalist ophthalmic artificial intelligence. *NEJM AI*, 1(12):AIoa2300221, 2024. 1, 2, 3

[26] Danli Shi, Weiyi Zhang, Jiancheng Yang, Siyu Huang, Xiaolan Chen, Mayinuer Yusufu, Kai Jin, Shan Lin, Shunming Liu, Qing Zhang, et al. Eyeclip: A visual-language foundation model for multi-modal ophthalmic image analysis. *arXiv preprint arXiv:2409.06644*, 2024. 1

[27] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 6

[28] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024. 1

[29] Lehan Wang, Haonan Wang, Honglong Yang, Jiaji Mao, Zehong Yang, Jun Shen, and Xiaomeng Li. Interpretable bilingual multimodal large language model for diverse biomedical tasks. *arXiv preprint arXiv:2410.18387*, 2024. 1, 6

[30] Meng Wang, Tian Lin, Aidi Lin, Kai Yu, Yuanyuan Peng, Lianyu Wang, Cheng Chen, Ke Zou, Huiyu Liang, Man Chen, et al. Common and rare fundus diseases identification using vision-language foundation model with knowledge of over 400 diseases. *arXiv preprint arXiv:2406.09317*, 2024. 2, 6

[31] 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. 6, 8

[32] 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. 3

[33] Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, et al. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*, 2024. 3

[34] Shengzhu Yang, Jiawei Du, Jia Guo, Weihang Zhang, Hanruo Liu, Huiqi Li, and Ningli Wang. Vilref: An expert knowledge enabled vision-language retinal foundation model. *arXiv preprint arXiv:2408.10894*, 2024. 3

[35] Yukun Zhou, Siegfried K Wagner, Mark A Chia, An Zhao, Moucheng Xu, Robbert Struyven, Daniel C Alexander, Pearse A Keane, et al. Automorph: automated retinal vascular morphology quantification via a deep learning pipeline. *Translational vision science & technology*, 11(7):12–12, 2022. 4

[36] Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163, 2023. 3

[37] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1