# OphthaVL-Ground: Interpretable Ophthalmic Diagnosis via Attention-Driven Weakly-Supervised Grounding

1st Yunfeng Liu
*College of Information Science and Technology*
*Beijing University of Chemical Technology*
Beijing, China
lyf1877583677@gmail.com

2rd Ruirui Li*
*College of Information Science and Technology*
*Beijing University of Chemical Technology*
City, Country
ilydouble@gmail.com

3nd Yuxin Xiao
*College of Information Science and Technology*
*Beijing University of Chemical Technology*
Beijing, China
xyx.xxxx1128@gmail.com

***Abstract*—Ophthalmic Multimodal Large Language Models (MLLMs) face a critical dilemma: generic models lack the fine-grained perception required for minute lesions, while specialized models are constrained by the scarcity of high-quality medical reports and bounding box annotations. To resolve this, we present OphthaVL-Ground, a novel framework designed for precise, interpretable, and scalable ophthalmic diagnosis. Our contributions are threefold: (1) We introduce the OphthaVL-W Data Engine, an automated multi-agent pipeline that synthesizes large-scale, pathology-aware diagnostic reports across Color Fundus Photography (CFP), Optical Coherence Tomography (OCT), and Lens Photography (LP), effectively bypassing the bottleneck of manual annotation. (2) We propose a Semantic-Structural Dual-Encoder that synergizes SigLIP's global semantic alignment with SAM2's fine-grained spatial features, ensuring the capture of subtle pathological structures often missed by standard encoders. (3) Crucially, we pioneer an Attention-Driven Self-Grounding strategy. By mining internal cross-attention maps to generate pseudo-bounding boxes, we enable the model to learn explicit lesion localization from weak supervision. This facilitates a clinical "localize-then-diagnose" inference paradigm that significantly mitigates hallucinations. Extensive experiments demonstrate that OphthaVL-Ground achieves state-of-the-art performance in both diagnostic accuracy and visual grounding, setting a new benchmark for interpretable ophthalmic AI.**

***Index Terms*—Multimodal Large Language Models, Ophthalmology, Visual Grounding, Weakly-Supervised Learning, Medical Image Analysis**

## I. Introduction

The advent of Multimodal Large Language Models (MLLMs) has revolutionized medical artificial intelligence [1], [2], offering promising capabilities in clinical question answering and report generation [3]. However, the transition from general-purpose multimodal understanding to specialized ophthalmic diagnosis remains fraught with challenges. Unlike natural images, ophthalmic imaging diagnosis relies heavily on identifying minute, morphologically variable lesions—such as microaneurysms in diabetic retinopathy or subtle fluid accumulation in OCT scans. In clinical practice, when reviewing ophthalmic imaging, ophthalmologists follow a strict "*localize-then-diagnose*" workflow: they first pinpoint suspicious regions and then apply medical knowledge to interpret pathological signs [4].

Regrettably, current ophthalmic MLLMs often fail to mimic this rigorous process, primarily due to two fundamental bottlenecks: data scarcity and visual disconnectedness.

**The Data Bottleneck: Lack of Fine-Grained Medical Supervision.** Constructing a model capable of generating physician-level reports requires high-quality, dense instruction data. Existing public datasets (e.g., OCTDL [5], IDRID [6]) are predominantly limited to single modalities with coarse-grained classification tags or sparse segmentation masks. They lack the comprehensive diagnostic reports describing lesion morphology, location, and severity that are essential for training reasoning capabilities. Manual annotation of such reports is prohibitively expensive and unscalable. Consequently, models trained on limited data struggle with long-tail diseases and lack the depth required for clinical utility.

**The Vision Bottleneck: The "Semantic-Spatial" Gap.** Most existing MLLMs rely on generic vision encoders like CLIP [7] or SigLIP [8]. While these encoders excel at aligning global semantics with text, they often sacrifice the high-frequency spatial details necessary for detecting subtle ophthalmic lesions. Without explicit spatial guidance, models suffer from severe hallucinations—generating plausible-sounding diagnoses without actually attending to the correct lesion area [9]. Although some methods introduce bounding box supervision [10], they rely on costly manual bounding box annotations, creating a chicken-and-egg problem for scalabil-

ity.

To bridge these gaps, we present **OphthaVL-Ground**, a novel framework that integrates a dual-encoder architecture and a weakly-supervised grounding mechanism to achieve precise, interpretable ophthalmic diagnosis.

Our contributions are summarized as follows:

- **Automated Multi-Agent Data Engine (OphthaVL-W Dataset).** Addressing the scarcity of medical reports, we introduce a scalable "*Generate-then-Refine*" pipeline. We leverage multiple foundation models to generate candidate captions, which are then rigorously cross-validated and synthesized by a **Meta-Reviewer** model. This approach constructs a large-scale, weakly-supervised dataset covering three modalities (CFP, OCT, LP) with rich diagnostic descriptions, effectively bypassing the need for manual curation.
- **Dual-Encoder Architecture for Semantic-Structural Synergy.** We propose a specialized visual backbone that fuses high-level semantics from SigLIP with fine-grained spatial details from the Segment Anything Model 2 (SAM2) [11]. This design addresses the limitations of traditional encoders, enabling the model to capture both the global context of the eye and the subtle structural edges of minute lesions.
- **Attention-Driven Self-Grounding Mechanism.** To enable interpretability without expensive bounding box annotations, we introduce a novel two-stage training strategy. We mine the model's internal cross-attention maps to automatically generate pseudo-bounding boxes, which serve as weak supervision signals. During inference, the model explicitly performs a "*localize-then-answer*" procedure, significantly reducing hallucinations and aligning the diagnostic process with clinical intuition.

Experiments demonstrate that OphthaVL-Ground achieves state-of-the-art performance in both diagnostic accuracy and visual grounding, setting a new benchmark for interpretable ophthalmic AI.

## II. Related Work

### A. General Multimodal Large Language Models

Recent advancements in MLLMs have bridged the gap between vision and language by aligning visual encoders (e.g., CLIP, SigLIP) with LLMs [12]. Models like LLaVA-OneVision [13] and InternVL [14] have set benchmarks in general visual understanding by scaling up resolution and data diversity. However, a critical domain gap persists when applying these generalists to ophthalmology. Standard visual encoders, pre-trained on natural images, often struggle to preserve the high-frequency spatial details required to detect minute lesions like microaneurysms or retinal fluid [15], [16]. While recent works have attempted to increase input resolution, they lack the specialized structural priors necessary for medical imaging. This limitation motivates our Dual-Encoder design, which synergizes semantic features with fine-grained spatial cues from SAM2 to capture subtle pathological evidence.

### B. Ophthalmic Foundation Models

To address domain-specific needs, ophthalmic foundation models have emerged. Early discriminative models like VisionFM [17] and RetiZero [18] excel at disease classification by pre-training on millions of fundus images. To further capture complex 3D structures, MIRAGE [19] employs masked autoencoding to learn robust representations from OCT volumes, while MultiEYE [20] leverages cross-modal knowledge distillation to enhance diagnostic precision. However, they lack the generative capability to produce descriptive diagnostic reports. Conversely, Generative models (e.g., EyeGPT [21]) attempt to generate reports but are severely constrained by the scarcity of high-quality image-text pairs. Most public datasets (e.g., OCTDL [5]) provide only coarse labels, forcing models to rely on limited or synthetic data that lacks clinical depth. Unlike these approaches, we introduce the OphthaVL-W Data Engine, an automated pipeline that synthesizes large-scale, pathology-aware reports, fundamentally resolving the data bottleneck.

### C. Visual Grounding in Medical VQA

Visual grounding—the ability to link text to image regions—is pivotal for trustworthy AI. In general domains, grounding is typically supervised by massive datasets with bounding box annotations. In the medical field, however, such annotations are prohibitively expensive. Existing methods like R-LLaVA [10] attempt to bypass this by injecting external bounding boxes (detected by separate models or humans) as prompts to guide the VQA process. While effective, this "passive grounding" relies heavily on the quality of external detectors and fails if the detector misses the lesion. In contrast, OphthaVL-Ground proposes an active, self-supervised grounding paradigm. Instead of relying on external boxes, we mine the model's internal cross-attention maps to autonomously discover and localize lesions. This "localize-then-diagnose" approach enables interpretability without the need for dense manual supervision.

## III. Dataset Construction: The OphthaVL-W Data Engine

In this section, we introduce **OphthaVL-W**, a large-scale multimodal ophthalmic dataset constructed via an automated, consistency-checked annotation engine. Unlike previous datasets that rely on simplistic tags, OphthaVL-W provides dense, pathology-aware medical reports.

### A. Multi-Modal Data Aggregation

To ensure clinical breadth, we aggregate data from diverse open-access sources covering three key modalities: Color Fundus Photography (CFP), Optical Coherence Tomography (OCT), and Lens Photography (LP). Data sources include REFUGE2 [22] and ORIGA [23] for CFP; OCTID [24] and OCT2017 [25] for OCT; and Eye-Sehoy [26] for LP. Post-deduplication, the dataset comprises **171,125 samples** (64,677 CFP, 64,734 OCT, 41,714 LP) spanning 29 disease categories.

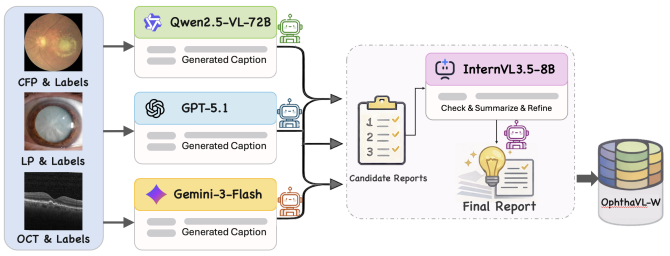


Fig. 1. The Automated Multi-Agent Annotation Pipeline. Three VLM agents generate initial hypotheses, which are then consolidated by a Meta-Reviewer to produce high-fidelity medical reports.

### B. Automated Multi-Agent Refinement Pipeline

Manual annotation of medical reports is prohibitively expensive. To address this, we design a **Multi-Agent Consensus Pipeline** (Fig. 1) that leverages the collective intelligence of foundational models to synthesize high-fidelity reports. **1. Ensemble Caption Generation:** We employ three distinct state-of-the-art MLLMs—Qwen2.5-VL-72B, GPT-5.1, and Gemini-3-Flash—as the initial annotator agents. Each agent $A_i$ receives the image $I$ and a pathology-conditioned prompt(e.g., "Severe NPDR" or "DME") $P$ to generate a candidate report $C_i$.

**2. Meta-Reviewer Consolidation:** To mitigate individual hallucinations, we introduce a Meta-Reviewer mechanism driven by InternVL3.5-8B. The Meta-Reviewer takes the tuple $\{I, C_1, C_2, C_3\}$ as input and performs a logical cross-check. It filters out inconsistent findings (noise) and retains pathological descriptions confirmed by the majority of agents (signal). The final output is a consolidated report $R_{final}$ that is clinically grounded and stylistically consistent.

This pipeline achieved a **96.5%** clinical acceptance rate in a blind review by board-certified ophthalmologists, confirming its reliability for downstream training.

## IV. Methodology

In this section, we introduce **OphthaVL-Ground**, an ophthalmic vision-language framework(Fig. 2) designed to bridge the gap between global semantic understanding and fine-grained lesion localization. Built upon the LLaVA-OneVision architecture, our method features a dual-encoder visual representation, a two-stage training strategy with attention-driven self-grounding, and a grounding-enhanced two-step inference paradigm.

### A. Dual-Encoder: Semantic-Structural Synergy

Ophthalmic diagnosis requires perceiving both the forest" (global context) and thetrees" (minute pathological cues). Standard single-branch encoders often fail to balance these vastly different scales.To achieve this synergy, we propose a dual-encoder architecture that processes the input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ through two complementary pathways:

- **Semantic Encoder ($E_{sem}$):** We employ **SigLIP-SO400M**to extract global, high-level semantic features $\mathbf{F}_{sem} \in \mathbb{R}^{H_1\times W_1\times C_1}$, where $H_1 \times W_1$ is the grid size of visual patches and $C_1 = 1152$ is the channel dimension.
- **Spatial Encoder ($E_{spat}$):** We incorporate the image encoder from **SAM2-Hiera-L**. Unlike classification-based encoders, SAM2 yields fine-grained spatial features $\mathbf{F}_{spat} \in \mathbb{R}^{H_2\times W_2\times C_2}$, where $C_2 = 256$, retaining high-frequency structural details and edge information.

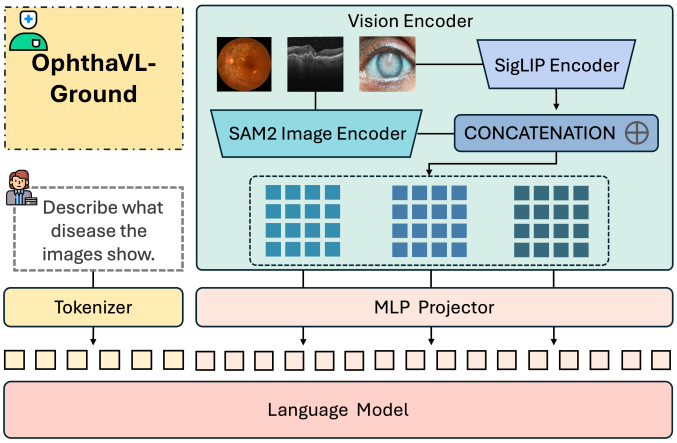


Fig. 2. Overview of the Dual-Encoder Vision-Language Architecture.

Since $H_1 \neq H_2$, we align the spatial dimensions via bilinear interpolation $\mathcal{T}(\cdot)$. Specifically, we upsample $\mathbf{F}_{spat}$ to match the resolution of $\mathbf{F}_{sem}$. The features are then concatenated along the channel dimension:

$$\mathbf{F}_{fused} = \text{Concat}\left(\mathbf{F}_{sem}, \mathcal{T}(\mathbf{F}_{spat})\right) \in \mathbb{R}^{H_1\times W_1\times (C_1+C_2)} \tag{1}$$

This fused representation is flattened into a sequence of $N_v = H_1 \times W_1$ visual tokens. A learnable Multi-Layer Perceptron (MLP) projector $\phi$ then maps them into the LLM's embedding space:

$$\mathbf{H}_v = \phi(\mathbf{F}_{fused}) \in \mathbb{R}^{N_v\times D_{llm}} \tag{2}$$

where $D_{llm}$ denotes the hidden dimension of the language model. This design ensures the LLM receives a visual sequence $\mathbf{H}_v$ rich in both semantic context and structural edges.

### B. Two-Stage Training Strategy

To progressively endow the model with biomedical alignment and spatial grounding capabilities without relying on manual bounding-box annotations, we employ a two-stage training strategy.

*1) Stage I: Language-Image Alignment.:* The first stage aligns the foundational visual representations with biomedical text concepts. We freeze both visual encoders ($E_{sem}$,$E_{spat}$) and the LLM backbone, training only the projector $\phi$. Given the visual tokens $\mathbf{H}_v$ and the ground-truth text tokens $\mathbf{Y} = \{y_t\}_{t=1}^{T}$ , the objective is standard autoregressive log-likelihood::

$$\mathcal{L}_{align} = -\sum_{t=1}^{T} \log P(y_t \mid \mathbf{H}_v, y_{<t}; \theta_\phi) \tag{3}$$

where $y_{<t}$ represents the preceding tokens and $\theta_\phi$ denotes the trainable parameters of the projector.

*2) Stage II: Grounding-Aware Instruction Refinement.:* In Stage II, we fully fine-tune the LLM and the projector $\phi$ while keeping the visual encoders frozen. This stage involves a novel **self-grounding refinement cycle** comprising three consecutive phases:

*a) Phase A: Intermediate Tuning.:* We first fine-tune the model on the original ophthalmic instruction-following datasets (OphthaVL-W and a classification subset) to create an intermediate model capable of generating general ophthalmic reports. This step is crucial, as it establishes the fundamental text-image alignment necessary for accurate attention mining.

*b) Phase B: Attention-Driven Grounding Mining.:* To obtain spatial supervision, we extract signals from the causal self-attention mechanism of the intermediate LLM decoder. During text generation, visual tokens act as prefixes. For a specific disease-related text token at index $i$ (e.g., "*edema*"), its attention distribution over visual tokens $j \in \{1, \ldots, N_v\}$ at layer $l$ and head $h$ is:

$$\alpha_{i,j}^{(l,h)} = \text{Softmax}_j \left( \frac{\mathbf{q}_i^{(l,h)} \cdot (\mathbf{k}_j^{(l,h)})^\top}{\sqrt{d_k}} \right) \tag{4}$$

where $\mathbf{q}_i$ is the query of the text token, $\mathbf{k}_j$ is the key of the visual prefix token, and $d_k$ is the head dimension. We aggregate these weights across the middle-to-late layers $l \in [L_{start}, L_{end}]$ and all attention heads $|H_{all}|$, as deeper layers capture high-level semantic alignment rather than irrelevant low-level saliency:

$$\mathbf{M}_{raw} = \frac{1}{|L| \cdot |H_{all}|} \sum_{l=L_{start}}^{L_{end}} \sum_{h=1}^{|H_{all}|} \alpha_i^{(l,h)} \tag{5}$$

The aggregated vector $\mathbf{M}_{raw} \in \mathbb{R}^{N_v}$ is reshaped to $H_1 \times W_1$ and upsampled to the original image dimensions $H \times W$ to form a spatial heatmap $\mathbf{M}$.

To convert $\mathbf{M}$ into a discrete Region of Interest (RoI), we apply adaptive statistical thresholding. A binary mask $\mathbf{B}_{mask}$ is generated via $\mathbf{B}_{mask} = \mathbb{I}(\mathbf{M} > \mu + \lambda \cdot \sigma)$, where $\mu$ and $\sigma$ are the mean and standard deviation of $\mathbf{M}$, and $\lambda$ controls sensitivity. After applying morphological filtering (Connected Component Analysis) to suppress sporadic noise, we extract the coordinates $B_{pseudo} = [x_1, y_1, x_2, y_2]$ of the largest connected component as the pseudo-bounding box. This acts as an automated self-denoising mechanism, mining spatial supervision for a subset (10%) of the training data.

*c) Phase C: RoI-Augmented Refinement.:* The mined pseudo-bounding boxes $B_{pseudo}$ are visually overlaid onto the original images (e.g., drawing explicit rectangles) to create visually prompted inputs. The model is then retrained on a dual objective: explicit coordinate prediction and lesion-focused diagnosis. This forces the model to associate its diagnostic logic explicitly with the highlighted pathological regions.

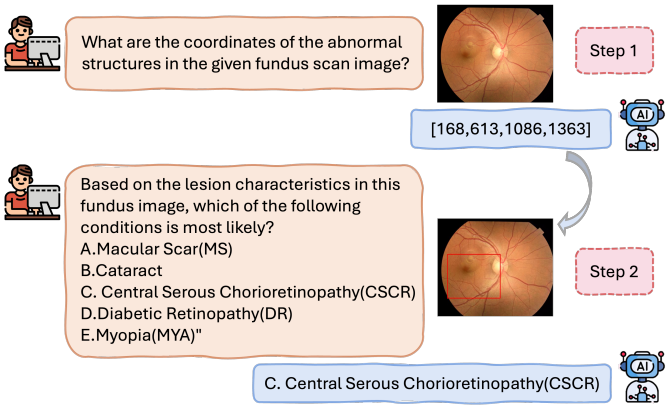


Fig. 3. Two-Step Inference Workflow. The model first predicts lesion coordinates, then uses the localized region to generate the final diagnosis.

### C. Grounding-Enhanced Two-Step Inference

During inference, we adopt a two-step reasoning strategy (Fig. 3) that mirrors the clinical workflow of ophthalmologists: "localize first, diagnose second." Let $\Omega = \{\theta_{LLM}, \theta_\phi\}$ denote the parameters of the finalized model.

*a) Step 1: Lesion Localization.:* Given an unannotated image $\mathbf{I}$, we query the model with a spatial instruction $Q_{loc}$ (e.g., *"What are the coordinates of the abnormal structures?"*). The model predicts the bounding box:

$$\hat{B} = \underset{B}{\text{argmax}}\, P(B \mid \mathbf{I}, Q_{loc}; \Omega) \tag{6}$$

This allows the model to perform zero-shot detection without relying on any external bounding-box networks.

*b) Step 2: RoI-Conditioned Diagnosis.:* The predicted box $\hat{B}$ is overlaid onto the original image, yielding a grounded image $\mathbf{I}^{\hat{B}}$ where the suspected region is visually emphasized. We feed $\mathbf{I}^{\hat{B}}$ back into the model along with a diagnostic query $Q_{diag}$ (e.g., *"What disease is present?"*):

$$\hat{Y} = \underset{Y}{\text{argmax}}\, P(Y \mid \mathbf{I}^{\hat{B}}, Q_{diag}; \Omega) \tag{7}$$

By explicitly conditioning the final prediction on the RoI-highlighted image, the model is forced to focus on disease-relevant structures. This drastically reduces hallucinations caused by spurious background correlations, ensures diagnostic stability, and provides a highly interpretable visual explanation of the model's decision-making process.

## V. Experiments

### A. Experimental Setup

**Datasets & Protocols.** We evaluated the model using a held-out test set independent of the OphthaVL-W training data. This set comprises 3,000 samples balanced across CFP, OCT, and LP modalities, with explicit disease labels. We adopt Disease Classification Accuracy as the primary quantitative metric. For qualitative evaluation, we assess the model's ability to generate accurate lesion descriptions and precise grounding boxes.

**Baselines.** We benchmark OphthaVL-Ground against a comprehensive suite of SOTA models, categorized into three groups: 1) Closed-Source MLLMs: GPT-5.2, Gemini-3-Flash, Claude-Sonnet-4.5, etc. 2) Open-Source MLLMs: InternVL3.5-8B, Qwen2.5-VL, LLaVA-OneVision. 3) Medical MLLMs: LLaVA-Med-v1.5-Mistral-7B. All baselines are evaluated under a consistent zero-shot or few-shot protocol with identical prompt templates to ensure fair comparison.

**Implementation Details.** The dual-encoder backbone combines SigLIP-SO400M and SAM2-Hiera-L. The LLM backbone is Qwen2.5-3B-Instruct. Stage I (Alignment) is trained for 1 epoch with a learning rate of $1 \times 10^{-3}$. Stage II (Refinement) is fine-tuned for 1 epoch with a learning rate of $1 \times 10^{-5}$. All experiments are conducted on a single NVIDIA A800 (80GB) GPU with BF16 precision.

### *B. Comparative Analysis*

*a) Quantitative Superiority:* As presented in Table I, OphthaVL-Ground achieves state-of-the-art performance across all three modalities, recording accuracies of **83.45% (CFP)**, **85.75% (OCT)**, and **84.91% (LP)**. It surpasses the strongest closed-source competitor, Gemini-3-Flash, by a significant margin of **+19.94%** on average (84.70% vs. 64.76%).

The breakdown reveals critical insights:

- **General MLLMs struggle with structural imaging.** While models like GPT-5.2 perform acceptably on color photography (CFP/LP), their performance drops sharply on OCT. This confirms that general visual encoders lack the frequency-domain sensitivity required for cross-sectional tomographic analysis.
- **Domain gap in Medical MLLMs.** Surprisingly, LLaVA-Med performs poorly (Avg: 21.17%), akin to random guessing. This attributes to its training data being dominated by radiology (X-ray/MRI) rather than ophthalmic imagery, highlighting the necessity of our domain-specific OphthaVL-W dataset.

*b) Qualitative Analysis:* We present a qualitative comparison in Fig. 4 to analyze diagnostic reasoning behaviors.

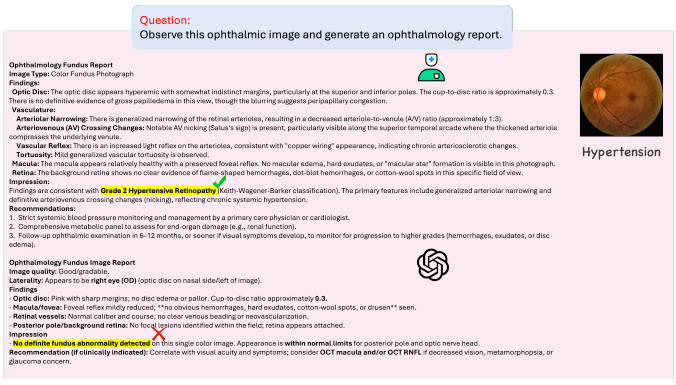


Fig. 4. Reports on CFP Generated by GPT-5.2 and OphthaVL-Ground. Our model (bottom) correctly localizes vascular abnormalities and diagnoses Hypertensive Retinopathy, whereas GPT-5.2 (top) hallucinates a normal impression.

TABLE I
DIAGNOSIS ACCURACY OF MLLMS ON OPHTHALMOLOGICAL IMAGING MODALITIES

| Model | Size | Diagnosis Analysis Acc(%) | | | |
|---|---|---|---|---|---|
| | | CFP | OCT | LP | AVG |
| Random | - | 20.00 | 20.00 | 20.00 | 20.00 |
| **Close-Source MLLMs** | | | | | |
| GPT-5.2 | - | 54.31 | 39.25 | 44.44 | 46.00 |
| Doubao-seed-1.8 | - | 53.54 | 27.75 | 56.00 | 45.76 |
| Gemini-3-Flash | - | <u>68.15</u> | <u>61.25</u> | <u>64.89</u> | <u>64.76</u> |
| Grok-4.1 | - | 35.23 | 18.75 | 54.00 | 35.99 |
| Claude-Sonnet-4.5 | - | 41.69 | 33.50 | 55.11 | 43.43 |
| **Open-Source MLLMs** | | | | | |
| InternVL3.5 | 8B | 38.15 | 53.00 | 41.56 | 44.24 |
| Qwen2.5-VL | 72B | 35.08 | 24.00 | 51.11 | 36.73 |
| GLM-4.6V | 106B | 33.85 | 27.50 | 55.33 | 38.89 |
| LLaVA-OneVision | 7B | 34.72 | 23.18 | 54.92 | 37.61 |
| LLaVA-Med-v1.5-Mistral | 7B | 21.08 | 17.75 | 24.67 | 21.17 |
| **OphthaVL-Ground** | 3B | **83.45** | **85.75** | **84.91** | **84.70** |

Note: **Bold** indicates the best performance in each column; <u>underline</u> indicates the second best; "-" indicates unavailable data.

The baseline (GPT-5.2) produces a well-structured and fluent report, correctly identifying anatomical landmarks (optic disc, macula). However, it fails to identify the key pathological cues of hypertensive retinopathy, concluding with a "normal" impression. This case highlights a common limitation of general VLMs: the hallucination of normalcy, where models prioritize descriptive completeness over lesion detection. In contrast, OphthaVL-Ground explicitly localizes the vascular abnormalities via its grounding mechanism. This spatial anchor forces the language decoder to attend to signs like "arteriolar narrowing" and "AV nicking", resulting in a clinically accurate diagnosis. This confirms that our Attention-Driven Grounding effectively bridges the gap between seeing pixels and understanding pathology.

### *C. Ablation Studies*

We conduct controlled ablation studies to validate the contribution of our core architectural innovations.

*a) Impact of Dual-Encoder Architecture:* We investigate whether the integration of SAM2 features provides tangible benefits over a standard SigLIP encoder. As shown in Table II, the Dual-Encoder (SigLIP+SAM2) strategy improves average accuracy from 77.40% to **81.74%**. The gain is most pronounced in OCT (+4.96%), a modality rich in fine-grained

TABLE II
DIAGNOSIS ACCURACY OF DIFFERENT VISUAL ENCODERS ON OPHTHALMOLOGICAL IMAGING MODALITIES

| Visual Encoder | Diagnosis Analysis Acc(%) | | | |
|---|---|---|---|---|
| | CFP | OCT | LP | AVG |
| Single-Encoder(SigLIP) | 76.65 | 78.31 | 77.24 | 77.40 |
| Dual-Encoder(SigLIP+SAM2) | 80.32 | 83.27 | 81.64 | 81.74 |

TABLE III
DIAGNOSIS ACCURACY OF DIFFERENT INFERENCE METHODS ON OPHTHALMOLOGICAL IMAGING MODALITIES

| Inference Method | Diagnosis Analysis Acc(%) | | | |
|---|---|---|---|---|
| | CFP | OCT | LP | AVG |
| One-Step | 80.77 | 80.93 | 81.68 | 81.13 |
| Two-Step | 83.45 | 85.75 | 84.91 | 84.70 |

edges and layer separations. This empirically proves our hypothesis: while SigLIP captures semantic concepts, SAM2 provides the necessary structural spatial priors to detect subtle morphological changes that are otherwise lost in global pooling.

*b) Efficacy of Grounding-Enhanced Inference:* We validate the "Localize-then-Diagnose" inference paradigm compared to direct prediction. As reported in Table III, the Two-Step Inference boosts performance from 81.13% to **84.70%**. This improvement reflects a shift in reasoning logic. By conditioning the diagnosis on the self-predicted bounding box, the model effectively filters out background noise and visual artifacts. This confirms that explicitly mining weak supervision signals from attention maps transforms the model from a "black-box classifier" into an "interpretable reasoner."

### D. Quantitative Evaluation of Visual Grounding

To rigorously validate the reliability of our framework, accurate lesion localization is paramount. A quantitative assessment is essential to verify that the model explicitly attends to correct pathological features rather than relying on dataset biases or statistical correlations.

**Grounding Test Set Setup.** Since public datasets lack pixel-level lesion annotations for report generation tasks, we constructed a dedicated evaluation subset, *OphthaVL-Ground-500*. We randomly sampled 500 cases from the test set (200 CFP, 200 OCT, 100 LP) covering diverse pathologies. Three senior ophthalmologists manually annotated the bounding boxes for the primary lesions described in the reports. These human annotations serve as the Ground Truth (GT).

**Metrics.** We employ two standard metrics:

- **Mean IoU (mIoU):** The average Intersection over Union between the predicted box and the GT box.
- **Pointing Accuracy (Acc@0.5):** The percentage of samples where the IoU exceeds 0.5.

**Baselines.** We compare OphthaVL-Ground against: 1) **Grad-CAM Baselines:** We apply Grad-CAM visual explanations to *LLaVA-Med* and *InternVL3.5* to generate heatmaps, thresholding them to obtain bounding boxes. 2)**Open-Set Detection (GroundingDINO):** Since specific ophthalmic detectors are unavailable, we utilize the state-of-the-art open-set detector GroundingDINO. It is prompted with specific disease names (e.g., "cataract", "hemorrhage") to generate zero-shot region proposals.

**Results Analysis.** As shown in Table IV, our framework significantly outperforms existing approaches.

- **Limitations of General Attention:** Standard MLLMs like LLaVA-Med obtain low IoU scores. Their attention mechanisms are often dispersed over the entire image or focused on salient but non-pathological features (e.g., the optic disc in healthy eyes), failing to pinpoint minute lesions.
- **Performance of Open-Set Detectors:** GroundingDINO shows variable performance. It achieves relatively decent results on LP (mIoU 46.55%) because external eye photos resemble natural images, allowing it to leverage its pre-training. However, it fails significantly on OCT (mIoU 28.90%), lacking the domain knowledge to interpret grayscale cross-sectional scans.
- **Effectiveness of Self-Grounding:** OphthaVL-Ground achieves a substantial improvement. This confirms that our *Attention-Driven Self-Grounding* strategy successfully translates weak supervision from diagnostic reports into precise spatial localization. The high performance on OCT (65.88%) specifically highlights the contribution of the Dual-Encoder in capturing fine-grained structural edges.

TABLE IV
VISUAL GROUNDING PERFORMANCE (IOU & ACCURACY) ON OPHTHAVL-GROUND-500 TEST SET.

| Model | CFP | | OCT | | LP | | Average | |
|---|---|---|---|---|---|---|---|---|
| | mIoU | Acc | mIoU | Acc | mIoU | Acc | mIoU | Acc |
| LLaVA-Med (Grad-CAM) | 24.15 | 12.50 | 19.82 | 8.40 | 22.45 | 11.20 | 22.14 | 10.70 |
| InternVL3.5 (Attn-Map) | 35.60 | 28.10 | 31.45 | 22.35 | 37.80 | 29.50 | 34.95 | 26.65 |
| GroundingDINO (Zero-shot) | 41.20 | 34.60 | 28.90 | 18.50 | 46.55 | 39.80 | 38.88 | 30.97 |
| **OphthaVL-Ground (Ours)** | **62.14** | **58.30** | **65.88** | **61.20** | **63.45** | **59.10** | **63.82** | **59.53** |

## VI. CONCLUSION

In this work, we presented OphthaVL-Ground to bridge the persistent gap between data scalability and diagnostic interpretability in ophthalmic AI. By constructing the multi-modal OphthaVL-W dataset via a meta-reviewed data engine and leveraging a SigLIP-SAM2 Dual-Encoder, our framework effectively captures both global semantics and fine-grained spatial features. Most notably, our Attention-Driven Self-Grounding mechanism demonstrates that explicit lesion localization can be achieved without expensive human annotations. This transforms the model from a "black-box" classifier into an interpretable reasoner that follows a verifiable "localize-then-diagnose" workflow, significantly enhancing clinical trust. Experimental results confirm that OphthaVL-Ground outperforms existing baselines in diagnostic reasoning and visual explanation. Future work will focus on extending this framework to video-based modalities and validating its efficacy in real-world clinical settings.

## ACKNOWLEDGMENT

The preferred spelling of the word "acknowledgment" in America is without an "e" after the "g". Avoid the stilted expression "one of us (R. B. G.) thanks ...". Instead, try "R. B. G. thanks...". Put sponsor acknowledgments in the unnumbered footnote on the first page.

## REFERENCES


[1] R. AlSaad, A. Abd-Alrazaq, S. Boughorbel, A. Ahmed, M.-A. Renault, R. Damseh, and J. Sheikh, "Multimodal large language models in health care: applications, challenges, and future outlook," *Journal of medical Internet research*, vol. 26, p. e59505, 2024.

[2] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature medicine*, vol. 29, no. 8, pp. 1930–1940, 2023.

[3] Z. Wang, Y. Sun, Z. Li, X. Yang, F. Chen, and H. Liao, "Llm-rg4: Flexible and factual radiology report generation across diverse input contexts," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 8, 2025, pp. 8250–8258.

[4] A. Bora, S. Balasubramanian, B. Babenko, S. Virmani, S. Venugopalan, A. Mitani, G. de Oliveira Marinho, J. Cuadros, P. Ruamviboonsuk, G. S. Corrado *et al.*, "Predicting the risk of developing diabetic retinopathy using deep learning," *The Lancet Digital Health*, vol. 3, no. 1, pp. e10–e19, 2021.

[5] M. Kulyabin, A. Zhdanov, A. Nikiforova, A. Stepichev, A. Kuznetsova, M. Ronkin, V. Borisov, A. Bogachev, S. Korotkich, P. A. Constable *et al.*, "Octdl: Optical coherence tomography dataset for image-based deep learning methods," *Scientific data*, vol. 11, no. 1, p. 365, 2024.

[6] P. Porwal, S. Pachade, R. Kamble, M. Kokare, G. Deshmukh, V. Sahasrabuddhe, and F. Meriaudeau, "Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research," *Data*, vol. 3, no. 3, p. 25, 2018.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[8] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 11 975–11 986.

[9] Z. Zhao, Y. Liu, H. Wu, M. Wang, Y. Li, S. Wang, L. Teng, D. Liu, Z. Cui, Q. Wang *et al.*, "Clip in medical imaging: A survey," *Medical Image Analysis*, vol. 102, p. 103551, 2025.

[10] X. Chen, Z. Lai, K. Ruan, S. Chen, J. Liu, and Z. Liu, "R-llava: Improving med-vqa understanding through visual region of interest," in *2025 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2025, pp. 1–10.

[11] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[12] J. Zhou, X. He, L. Sun, J. Xu, X. Chen, Y. Chu, L. Zhou, X. Liao, B. Zhang, S. Afvari *et al.*, "Pre-trained multimodal large language model enhances dermatological diagnosis using skingpt-4," *Nature Communications*, vol. 15, no. 1, p. 5649, 2024.

[13] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu *et al.*, "Llava-onevision: Easy visual task transfer," *arXiv preprint arXiv:2408.03326*, 2024.

[14] W. Wang, Z. Gao, L. Gu, H. Pu, L. Cui, X. Wei, Z. Liu, L. Jing, S. Ye, J. Shao *et al.*, "Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency," *arXiv preprint arXiv:2508.18265*, 2025.

[15] Z. Qin, Y. Yin, D. Campbell, X. Wu, K. Zou, N. Liu, Y. C. Tham, X. J. Zhang, and Q. Chen, "Lmod: A large multimodal ophthalmology dataset and benchmark for large vision-language models," in *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025, pp. 2501–2522.

[16] Z. Qin, Y. Liu, Y. Yin, J. Ding, H. Zhang, A. Li, D. Campbell, X. Wu, K. Zou, T. D. Keenan *et al.*, "Lmod+: A comprehensive multimodal dataset and benchmark for developing and evaluating multimodal large language models in ophthalmology," *arXiv preprint arXiv:2509.25620*, 2025.

[17] J. Qiu, J. Wu, H. Wei, P. Shi, M. Zhang, Y. Sun, L. Li, H. Liu, H. Liu, S. Hou *et al.*, "Visionfm: a multi-modal multi-task vision foundation model for generalist ophthalmic artificial intelligence," *arXiv preprint arXiv:2310.04992*, 2023.

[18] M. Wang, T. Lin, K. Yu, A. Lin, Y. Peng, L. Wang, C. Chen, K. Zou, H. Liang, M. Chen *et al.*, "Common and rare fundus diseases identification using vision-language foundation model with knowledge of over 400 diseases," *arXiv e-prints*, pp. arXiv–2406, 2024.

[19] J. Morano, B. Fazekas, E. Sükei, R. Fecso, T. Emre, M. Gumpinger, G. Faustmann, M. Oghbaie, U. Schmidt-Erfurth, and H. Bogunović, "Mirage: Multimodal foundation model and benchmark for comprehensive retinal oct image analysis," *arXiv preprint arXiv:2506.08900*, 2025.

[20] L. Wang, C. Qi, C. Ou, L. An, M. Jin, X. Kong, and X. Li, "Multieye: Dataset and benchmark for oct-enhanced retinal disease recognition from fundus images," *IEEE Transactions on Medical Imaging*, 2024.

[21] X. Chen, Z. Zhao, W. Zhang, P. Xu, L. Gao, M. Xu, Y. Wu, Y. Li, D. Shi, and M. He, "Eyegpt: Ophthalmic assistant with large language models," *arXiv preprint arXiv:2403.00840*, 2024.

[22] H. Fang, F. Li, J. Wu, H. Fu, X. Sun, J. Son, S. Yu, M. Zhang, C. Yuan, C. Bian *et al.*, "Refuge2 challenge: A treasure trove for multi-dimension analysis and evaluation in glaucoma screening," *arXiv preprint arXiv:2202.08994*, 2022.

[23] Z. Zhang, F. S. Yin, J. Liu, W. K. Wong, N. M. Tan, B. H. Lee, J. Cheng, and T. Y. Wong, "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*. IEEE, 2010, pp. 3065–3068.

[24] P. Gholami, P. Roy, M. K. Parthasarathy, and V. Lakshminarayanan, "Octid: Optical coherence tomography image database," *Computers & Electrical Engineering*, vol. 81, p. 106532, 2020.

[25] D. Kermany, K. Zhang, and M. Goldbaum, "Labeled optical coherence tomography (oct) and chest x-ray images for classification (2018)," *Mendeley Data, v2 https://doi. org/10.17632/rscbjbr9sj https://nihcc. app. box. com/v/ChestXray-NIHCC*, 2018.

[26] Kyudi, "Eye Dataset," Roboflow Universe, May 2025, accessed: Jan. 27, 2026. [Online]. Available: https://universe.roboflow.com/kyudi/eye-sehoy