LMOD: A Large Multimodal Ophthalmology Dataset and Benchmark for Large Vision-Language Models

Anonymous ACL submission

Abstract

The prevalence of vision-threatening eye diseases is a significant global burden, with many cases remaining undiagnosed or diagnosed too late for effective treatment. Large vision-language models (LVLMs) have the potential to assist in understanding anatomical information, diagnosing eye diseases, and drafting interpretations and follow-up plans, thereby reducing the burden on clinicians and improving access to eye care. However, limited benchmarks are available to assess LVLMs' performance in ophthalmology-specific applications. In this study, we introduce LMOD, a large-scale multimodal ophthalmology benchmark consisting of 21,993 instances across (1) five ophthalmic imaging modalities: optical coherence tomography, color fundus photographs, scanning laser ophthalmoscopy, lens photographs, and surgical scenes; (2) free-text, demographic, and disease biomarker information; and (3) primary ophthalmology-specific applications such as anatomical information understanding, disease diagnosis, and subgroup analysis. In addition, we benchmarked 13 state-of-the-art LVLM representatives from closed-source, open-source, and medical domains. The results demonstrate a significant performance drop for LVLMs in ophthalmology compared to other domains. Systematic error analysis further identified six major failure modes: misclassification, failure to abstain, inconsistent reasoning, hallucination, assertions without justification, and lack of domain-specific knowledge. In contrast, supervised neural networks specifically trained on these tasks as baselines demonstrated high accuracy. These findings underscore the pressing need for benchmarks in the development and validation of ophthalmology-specific LVLMs.

1 Introduction

The ever-increasing prevalence of primary eye diseases poses a significant global burden, with more than 2.2 billion individuals suffering from vision impairment worldwide, particularly in low- and middle-income regions (Tham et al., 2014; Neely et al., 2017; Cavan et al., 2017; Organization, 2023). Limited access to comprehensive eye examinations and a shortage of clinicians result in a substantial proportion of major eye diseases being undiagnosed or diagnosed too late for effective treatment (Neely et al., 2017; Varma et al., 2011). To address this challenge, artificial intelligence (AI) applications, particularly Large Language Models (LLMs) and their multimodal extensions, have been developed to streamline clinical workflows by assisting in patient triage, disease diagnosis, prognosis prediction, and reducing documentation burdens (Ting et al., 2019; Peng et al., 2019; Keenan et al., 2022; De Fauw et al., 2018; Zhou et al., 2023; Kaushal et al., 2023; Elkholy and Marzouk, 2024; Tian et al., 2024; Chen et al., 2023; Qin et al., 2022). These large vision-language models (LVLMs), such as GPT-4o (Achiam et al., 2023) and LLaVA (Liu et al., 2024), combine LLMs with vision encoders to generate responses based on input images, which are particularly beneficial in medical imaging, where interpretations and follow-up plans are documented in free-text clinical notes

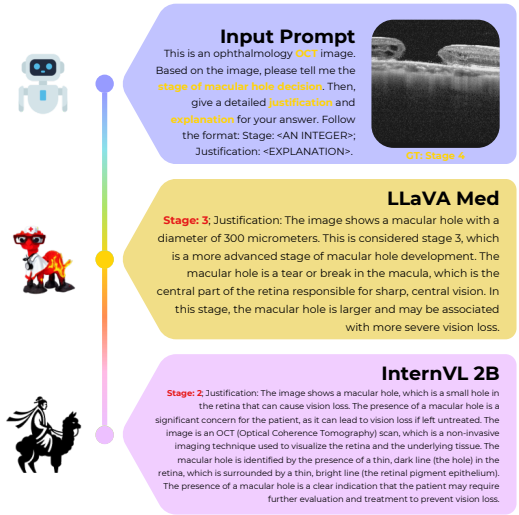


Figure 1: LVLM response examples for macular hole staging.

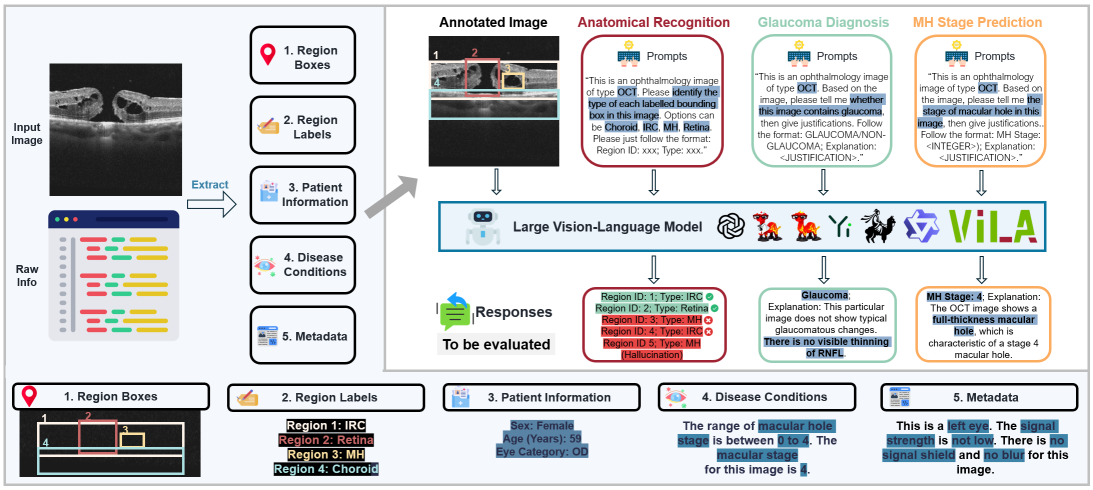


Figure 2: Overview of our data processing and evaluation pipeline for assessing the performance of LVLMs. The raw information is preprocessed to extract structured data such as bounding boxes and disease conditions. This aggregated information is then used to generate prompts for the LVLMs to identify the type of each labeled region, or conduct diagnosis analysis. The LVLMs processes the input image and prompt to generate responses categorizing each region or disease or describing diseases. Finally, the model's output is compared against the ground truth results using our proposed evaluation metrics.

(Tian et al., 2024; Xiao et al., 2024).

Specifically, in ophthalmology, LVLMs enable (1) automated disease diagnosis and classification, such as assessing the severity level of glaucoma; (2) region classification and segmentation, such as segmenting retinal nerve fiber layers in optical coherence tomography (OCT); and (3) documentation, such as generating free-text imaging interpretations (Betzler et al., 2023). Importantly, what distinguishes LVLMs is their ability to handle these tasks within a single model using zero- or few-shot learning. In contrast, previous applications required sophisticated fine-tuning on substantial manually curated instances, making it challenging to adapt fine-tuned models to different data types (Xiao et al., 2024; Betzler et al., 2023). This may improve the efficiency and accuracy of workflows in ophthalmology clinics (De Fauw et al., 2018).

Nevertheless, limited benchmarks are available to assess LVLMs' performance in ophthalmology-specific applications to date, and consequently, the potential and limitations of LVLMs are not clear. Existing studies in ophthalmology focus on evaluating the performance of LLMs on text-based tasks, including multiple-choice questions (Wu et al., 2024), general ophthalmology knowledge testing (Antaki et al., 2023), and free-text question answering on specific ophthalmology topics (Lim et al., 2023). While these studies are useful for demonstrating potential, ophthalmic images are arguably the most important data modality. Ophthalmologists require different imaging modalities for diagnosis and prognosis that are not derived from the text itself (Khan et al., 2021). Therefore, the capability of image analysis is crucial in the domain of ophthalmology. However, this capability poses challenges for existing LVLMs, which struggle to understand and analyze ophthalmic images, as illustrated in Figure 1. In addition, existing ophthalmic imaging datasets were designed for the development and evaluation of fine-tuning AI models; they may only contain a single imaging modality (e.g., OCT only), a specific task (e.g., region segmentation), and specific output types (e.g., a disease severity class rather than free text).

In response, this study proposes a systematic and reproducible data and evaluation pipeline that repurposes existing datasets to curate a dataset we refer to as LMOD (Large Multimodal Ophthalmology Dataset) for the development and evaluation of LVLMs in ophthalmology. LMOD consists of five imaging modalities: surgical scenes (SS), optical coherence tomography (OCT), scanning laser ophthalmoscopy (SLO), lens photographs (LP), and color fundus photographs (CFP), collectively comprising over 20K instances. It also provides multi-granular annotations, including region annotation and disease information. The pipeline can also be directly applied to new datasets.

Our contributions are as follows: (1) We introduce LMOD, a large-scale ophthalmology dataset that includes over 21K images across diverse imaging modalities. LMOD is richly annotated with disease labels and bounding boxes, supporting comprehensive evaluations from macro-level. (2) We systematically benchmark 13 state-of-the-art (SoTA) LVLMs, including models with diverse visual backbones and LLMs. The evaluation is con-

Benchmarks	Modalities		Image Types					Evaluation Perspectives	
	Images	Texts	Surgical Scenes	SLO	OCT	Eye Photos	Fundus Images	Anatomical Understanding	Diagnosis Analysis
General-Domain Benchmarks									
MMMU (Yue et al., 2024)	✓	✓	✗	✗	✗	✓	✓	✗	✗
MME-RealWorld (Zhang et al., 2024b)	✓	✓	✗	✗	✗	✗	✗	✗	✗
UNK-VQA (Guo et al., 2024)	✓	✓	✗	✗	✗	✗	✗	✗	✗
MMCBench (Zhang et al., 2024a)	✓	✓	✗	✗	✗	✗	✗	✗	✗
MathVista (Lu et al., 2023)	✓	✓	✗	✗	✗	✗	✗	✗	✗
SEED-Bench (Li et al., 2024a)	✓	✓	✗	✗	✗	✗	✗	✗	✗
Ophthalmology-Specific Benchmarks									
Bench-Nephrology (Wu et al., 2024)	✗	✓	✗	✗	✗	✗	✗	✗	✗
Eval-GPT-Ophth (Antaki et al., 2023)	✗	✓	✗	✗	✗	✗	✗	✗	✗
Bench-Myopia (Lim et al., 2023)	✗	✓	✗	✗	✗	✗	✗	✗	✗
OphNet (Hu et al., 2024)	✓	✓	✓	✗	✗	✗	✗	✗	✗
LMOD (ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison of existing general-domain and ophthalmology-specific benchmarks for evaluating large vision-language models, highlighting their supported modalities, coverage of image types, and evaluation perspectives.

ducted using a wide range of metrics, assessing strengths and weaknesses of LVLMs from various perspectives. (3) Through fine-tuning and supervised classification, we demonstrate that while the challenges posed by ophthalmic image analysis are intricate for LVLMs, they are insurmountable. Our comprehensive evaluations and error analysis provide both a high-level overview and detailed insights, presented through various result formats, including weighted averages, bar charts, radar charts, and visual illustrations, to highlight the key strengths and weaknesses.

2 Related Work

This section provides an overview of the advancements in LVLMs and highlights the lack of comprehensive benchmarks in ophthalmology.

2.1 Advances in LVLMs

The release of ChatGPT (Achiam et al., 2023) has sparked considerable interest in the potential of large language models (LLMs) across various domains (Liu et al., 2023; Tian et al., 2024; De Angelis et al., 2023). Building on the success of ChatGPT and other LLMs (Radford et al., 2018, 2019; Brown et al., 2020), researchers have developed LVLMs that integrate the strengths of vision encoders with LLMs. These models employ vision encoders, typically pretrained on vast image datasets in an unsupervised manner, to extract visual features from images and incorporate them into LLMs, enabling a combined understanding of both vision and language. Several notable LVLMs have been introduced in recent years, each with its unique architecture or training approach, such as GPT-4 (Achiam et al., 2023), LLaVA (Liu et al., 2024), InternVL (Chen et al., 2024), Qwen (Bai et al., 2023), and VILA (Lin et al., 2023). In the medical domain, representative LVLMs include LLaVA-Med (Li et al., 2024b) and its variants (Li et al., 2024b; Jiang et al., 2024; Xie et al., 2024), which demonstrate potential for disease diagnosis and medical question answering. The advent of LVLMs has opened up new possibilities for multimodal reasoning and comprehension, with applications spanning various fields, including the medical domain (Clusmann et al., 2023; Kung et al., 2023; Currie et al., 2023).

2.2 Lack of Benchmarks

In ophthalmology, domain-specific foundation models, such as vision encoders pretrained on ophthalmic images, have shown consistent improvements in diagnosing and predicting the prognosis of eye diseases (Zhou et al., 2023; Kaushal et al., 2023; Elkholy and Marzouk, 2024). However, these encoders lack the reasoning and conversational capabilities inherent to large language models (LLMs) and require task-specific fine-tuning with static inputs and outputs. Furthermore, existing research on LLMs in ophthalmology primarily focuses on text-based applications (Wu et al., 2024; Antaki et al., 2023; Lim et al., 2023; Gilson et al., 2024), neglecting ophthalmic images as a key data modality. To the best of our knowledge, few benchmarks exist for the development and evaluation of LVLMs in ophthalmology, posing a significant barrier to systematically evaluating the feasibility of applying existing LVLMs in this domain and hindering the development of ophthalmology-specific LVLMs. In contrast, several benchmarks have been established in both general and medical domains (Table 1). However, existing benchmarks in ophthalmology are primarily designed for

Models	Anatomical Recognition				Diagnosis Analysis	
	Precision ↑	Recall ↑	F1 ↑	HC ↑	Glaucoma Acc (%) ↑	MH Stage Acc (%) ↑
Random	-	-	-	-	50.00	25.00
Finetuned	Invalid	Invalid	Invalid	Invalid	Invalid	Invalid
GPT-4o	0.5609	**0.5896**	**0.5748**	0.9513	**54.09**	19.71
LLaVA-Med	0.0789	0.1163	0.0940	0.7435	50.00	25.00
LLaVA-1.5-7B	0.0567	0.0410	0.0475	0.2674	50.00	7.30
LLaVA-M-7B	0.1346	0.1450	0.1396	0.7569	50.00	0.00
LLaVA-V-7B	0.3095	0.2540	0.2790	0.7516	50.00	0.00
LLaVA-13B	0.0599	0.0803	0.0686	0.5993	50.00	0.00
Yi-6B	0.1952	0.1499	0.1695	0.8893	50.00	5.26
InternVL-2B	0.6025	0.3999	0.4807	**0.9809**	50.00	**30.26**
InternVL-4B	**0.7241**	0.4481	0.5536	0.9629	50.00	25.00
Qwen	0.0275	0.0372	0.0316	0.8418	50.00	18.42
VILA-3B	0.1429	0.1119	0.0633	0.5300	50.00	24.24
VILA-3B-S2	0.3340	0.2636	0.1255	0.7695	50.00	21.42
VILA-8B	N/A	N/A	N/A	N/A	50.00	22.53
Average	0.2688	0.2197	0.2189	0.7537	50.31	15.31

Table 2: Performance comparison of state-of-the-art large vision-language models on the LMOD benchmark, evaluating their capabilities in anatomical recognition and diagnosis analysis. Acc indicates accuracy. The best-performing model in each metric is highlighted in bold. LLaVA-Med-Finetuned consistently produced invalid outputs across both tasks.

Data Types	Num Images	Num Avg Boxes
Surgical Scenes (SS)	2,256	3.3
Optical Coherence Tomography (OCT)	3,859	2.4
Scanning Laser Ophthalmoscopy (SLO)	10,000	1.0
Eye Photos (EP)	2,432	1.9
Color Fundus Images (CFI)	3,386	1.6

Table 3: Overview of LMOD, including the number of images (Num Images) and average number of bounding boxes per image (Num Avg Boxes).

the development and evaluation of AI models under the fine-tuning paradigm, often focusing on a single imaging modality, a specific task, and restricted output types. Few benchmarks encompass diverse ophthalmic imaging modalities or support a broad range of downstream evaluations, such as anatomical understanding (Wang et al., 2021) and diagnostic analysis.

3 LMOD Curation

In this section, we present our methodology for curating LMOD: Large Multimodal Ophthalmology Dataset. We first describe our data curation pipeline, which involves selecting suitable datasets, generating consistent annotations across various image types, and designing standardized prompts for model evaluation. Next, we introduce the evaluation tasks and metrics used to assess the performance of LVLMs on LMOD, focusing on anatomical recognition and diagnostic analysis.

3.1 Data Curation Pipeline

The data curation pipline consists of two steps.

Step 1: Dataset Selection. We screened and selected the datasets for repurposing based on the following criteria: (1) **Accessibility:** The datasets are publicly accessible and non-commercial, ensuring that LMOD can be widely used by the research community for model evaluation and development. (2) **Coverage:** The datasets collectively need to cover key ophthalmic imaging modalities and primary eye diseases. (3) **Diversity:** Population diversity needs to be considered, enabling the benchmarks to quantify model effectiveness across subgroups, especially within disparity populations.

Algorithm 1 Anatomical Recognition Pipeline

Require: Original dataset $D = \{(I_1, R_1), (I_2, R_2), \ldots, (I_n, R_n)\}$, where I_i is an image and R_i is the corresponding raw data
Require: Minimum bounding box area threshold $\tau \in \mathbb{R}^+$
Ensure: Curated dataset $D' = \{(I_1, B'_1, P_1), (I_2, B'_2, P_2), \ldots, (I_n, B'_n, P_n)\}$, where B'_i is the set of curated bounding boxes and P_i is the set of corresponding prompts for image I_i

for each image-raw data pair $(I_i, R_i) \in D$ **do**
  $B_i \leftarrow$ ExtractBoundingBoxes(R_i), where $B_i = \{b_{i,1}, b_{i,2}, \ldots, b_{i,|B_i|}\}$ and $b_{i,j}$ is the j-th bounding box of image I_i
end for
$B \leftarrow \bigcup_{i=1}^{n} B_i$
$B' \leftarrow \{b \in B \mid \text{area}(b) \geq \tau\}$
for each image-raw data pair $(I_i, R_i) \in D$ **do**
  $B'_i \leftarrow \{b \in B' \mid b \text{ belongs to image } I_i\}$
  $P_i \leftarrow \emptyset$
  for each bounding box $b_{i,j} \in B'_i$ **do**
    $id_{i,j} \leftarrow$ GenerateUniqueID()
    $color_{i,j} \leftarrow$ AssignDistinctColor()
    $prompt_{i,j} \leftarrow$ GeneratePrompt$(b_{i,j})$
    $P_i \leftarrow P_i \cup \{(id_{i,j}, color_{i,j}, prompt_{i,j})\}$
  end for
end for
return D'

As a result, nine datasets were selected for curation. These datasets can be categorized into five ophthalmic imaging types, and the overall statistics can be found in Table 3:

SS: SS stands for surgical scenes. This category includes the Cataract-1K dataset (Ghamsarian et al., 2024), which contains 2,256 images extracted from cataract surgery videos. These images capture various stages and aspects of the surgical procedure, with an average of 3.3 bounding boxes per image.

OCT: OCT represents Optical Coherence Tomography. OIMHS (Ye et al., 2023) represents this category, comprising 3,859 OCT images. OCT is a non-invasive imaging technique that provides high-resolution cross-sectional retinal images. There 2.4 bounding boxes per image.

SLO: SLO indicates Scanning Laser Ophthalmoscopy. The Harvard FairSeg dataset (Luo et al., 2024) is included in this category, featuring 10,000 SLO fundus images. Each image contains an average of a single bounding box.

LP: LP means Lens Photographs. This category includes two datasets: CAU001 (PupiUp, 2023) and Cataract Detection 2 (Ramapuram, 2023). CAU001 contains 1,417 RGB photographs of human eye regions, with bounding box annotations indicating

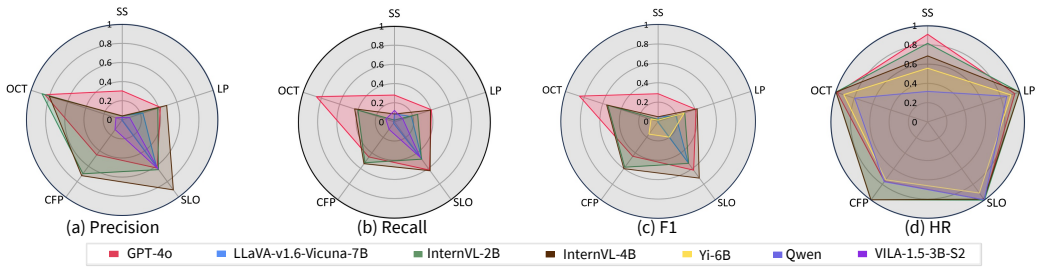


Figure 3: Performance comparison of top-performing LVLMs across different ophthalmic imaging modalities. The radar charts display the performance of the top-F1-performing models, for each evaluation metric (Precision, Recall, F1, and HR) across five different imaging modalities: surgical scenes (SS), optical coherence tomography (OCT), color fundus photographs (CFP), scanning laser ophthalmoscopy (SLO), and lens photographs (LP).

the locations of the left and right eyes, pupils, and irises. Cataract Detection 2 consists of 1,015 photographs of eyes with and without cataracts, each containing a single bounding box annotation. Each image contains 1.9 bounding boxes in average.

CFP: CFP implies Color Fundus Photographs. This category includes four datasets: REFUGE (Orlando et al., 2020), IDRiD (Prasanna et al., 2018), ORIGA (Zhang et al., 2010), and G1020 (Bajwa et al., 2020). REFUGE contains 1,200 retinal fundus images, including both glaucoma and normal eyes, with detailed annotations of optic disc and cup segmentations. IDRiD includes 516 images with pixel-level annotations of typical diabetic retinopathy lesions and normal retinal structures. ORIGA consists of 650 retinal images annotated by trained professionals, containing a comprehensive set of image features critical for glaucoma diagnosis. G1020 contains 1,020 high-resolution color fundus photographs, accompanied by detailed ground-truth annotations, including glaucoma diagnosis, optic disc and cup segmentations, and other clinically relevant measurements. On average, each image in this category contains 1.6 bounding boxes.

Step 2: Multi-granular Annotation. We further curated the datasets to support the development and evaluation of LVLMs in ophthalmology applications. These applications are categorized into two main parts: (1) anatomical understanding, which involves the accurate observation and identification of ocular structures (Wang et al., 2021; De Fauw et al., 2018), and (2) diagnostic analysis, which requires the interpretation of visual features and patterns to assess the presence and severity of ocular diseases. The curation procedures for each application is detailed below.

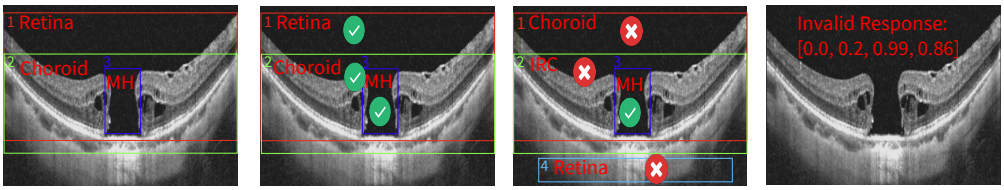
Anatomical Recognition: Anatomical recognition refers to the ability of models to accurately identify various anatomical structures in ophthalmic images. The algorithmic pipeline is outlined in Algorithm 1. In brief, the steps are as follows: First, we generate bounding boxes and their associated labels. The coordinates of the bounding boxes are present in the open-source datasets. To ensure consistency and standardization across the dataset, we map the region types provided in the open-source datasets to a predefined set of ophthalmology-specific region types, such as optic disk, macula, lesion, and tumor. This mapping process allows us to handle variations in terminology and granularity used in the original datasets.

To balance between the comprehensiveness of annotations and the clarity of the images, we establish a threshold and remove bounding boxes whose areas fall below this threshold (10% for ours). This step is crucial as an excessive number of bounding boxes overlaid on the images can lead to significant occlusion, hindering the visibility and interpretability of the underlying image content.

In the final step, using the extracted bounding box coordinates, we generate visual markers in the form of bounding boxes on the images to highlight the annotated regions. These visual markers are assigned unique labels (e.g., letters or numbers) to clearly identify each region. To increase the differentiation between regions, we assign different colors to the bounding boxes.

Diagnostic Analysis: We evaluate models' ability to determine the presence and severity of eye diseases, using glaucoma and macular holes as case studies. We formulate glaucoma detection as a binary classification problem. Macular holes are classified into four distinct stages, ranging from 1 to 4 to indicate severity increase.

To ensure reliability and validity, we extract glaucoma and macular hole labels from the original datasets when available. In cases of imbalanced label distribution, we employ a balanced sampling strategy to mitigate potential biases and ensure fair evaluation. To be more specific, we determine the minimum number of samples across all classes and randomly select an equal number of samples from



(a) GT (b) GPT-4o (c) LLaVA-M-7B (d) VILA-3B

Figure 4: Visual examples of LVLM predictions for anatomical recognition in OCT images. The figure presents a comparison of ground truth (GT) annotations and predictions from three representative LVLMs: GPT-4o, LLaVA-M-7B, and VILA-3B. Green ticks indicate correct predictions, while red crosses mark incorrect ones. VILA-3B generates an invalid response consisting of a sequence of numbers unrelated to the task.

	Anatomical Recognition (F1)						Diagnosis Analysis (Accuracy)	
Model	**Macro Avg**	**SS**	**OCT**	**SLO**	**LP**	**CFP**	**Glaucoma**	**MH Stage**
NNC	94.36	93.76	98.42	94.92	88.85	95.86	82.69	98.17

Table 4: Supervised-trained neural network classifier (NNC) performance on anatomical recognition and diagnosis analysis tasks.

each class to create a balanced dataset. This approach guarantees that models are evaluated on a representative and unbiased sample, preventing them from exploiting class imbalance.

4 Benchmarking Results

We here present the results of benchmarking 13 state-of-the-art LVLMs on the LMOD benchmark. We conclude with an error analysis to highlight common failure modes of LVLMs.

4.1 Benchmarked LVLMs

We benchmarked 13 LVLMs on the LMOD benchmark, including several variations of LLaVA, such as LLaVA-7B, LLaVA-M-7B (Mistral), LLaVA-V-7B (Vicuna), LLaVA-13B, and the domain-specific LLaVA-Med. Additionally, we evaluated InternVL models, including InternVL-2B and InternVL-4B, as well as VILA models, including VILA-3B, VILA-3B-S2, and VILA-8B. Other models in the benchmark included GPT-4o, Yi-6B, and Qwen. These models represent a diverse range of architectures, parameter scales, and training methodologies, highlighting the breadth of current LVLM development. See appendix for details.

4.2 Evaluation Metrics

To comprehensively assess the performance of the LVLMs on the LMOD benchmark, we employed several evaluation metrics that captured different aspects of their capabilities. These metrics provided a holistic view of the models' strengths and weaknesses in analyzing ophthalmic images.

For anatomical recognition, we employed a comprehensive set of metrics to evaluate the performance of LVLMs:

Precision: Measure the proportion of correctly predicted region types among all predicted regions. A high precision indicates that the model is more likely to be correct when predicting region types.

Recall: Quantify the proportion of correctly predicted region types among all ground truth regions. A high recall indicates that the model is able to identify a larger fraction of the relevant regions.

F1 Score: The harmonic mean of precision and recall, providing a balanced measure.

Hallucination Resistance (HR): The Hallucination Resistance (HR) metric is a new metric that quantifies a model's ability to avoid hallucinations:

$$\text{HR} = 1 - \frac{|\{r \in \mathcal{P}_i \mid r \notin \mathcal{T}_i\}|}{|\{r \in \mathcal{P}_i\}|},$$

where $\mathcal{P}_i$ represents set of all predicted region IDs for image i, and $\mathcal{T}_i$ indicates set of all ground truth region IDs for image i. Higher HR values indicating fewer hallucinations.

These metrics collectively assessed the models' ability to accurately identify and localize anatomical structures in ophthalmic images.

For diagnostic analysis, we focused on the models' performance in glaucoma detection and macular hole staging using the metric of accuracy. It quantified the proportion of correctly classified glaucoma cases and macular hole stages, measuring the models' ability to determine the presence of glaucoma and the severity of macular holes based on visual characteristics.

4.3 Anatomical Recognition

Overall Results. The results for anatomical recognition in Table 2 revealed that the performance of SoTA LVLMs was far from optimal. The average F1 score across all models was a mere 0.2189, highlighting a significant discrepancy between the models' predictions and the ground truth annotations. The best-performing model, GPT-4o, achieved an F1 score of 0.5748, which, while notably higher than the average, still indicated substantial room for improvement on anatomical recognition.

Nuanced View. Figure 3 provided a more nuanced view of the top five models' F1 performance across

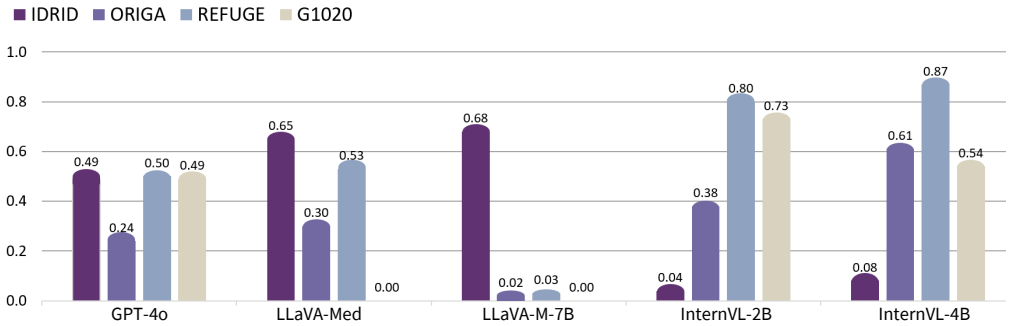


Figure 5: Robustness analysis of LVLMs across different color fundus photograph datasets. The bar chart displays the F1 scores of the five models (GPT-4o, LLaVA-Med, LLaVA-M-7B, InternVL-2B, and InternVL-4B) on four different color fundus photograph datasets: IDRID, ORIGA, REFUGE, and G1020.

different image types. Note that the performance of all models, including GPT-4o, was markedly lower for complex images with a higher number of bounding boxes, such as those in the SS category. This trend suggested that the models struggled to accurately recognize and localize anatomical structures in more intricate and information-dense images. Figure 4 presented visual examples of various responses, including correct and wrong answers, hallucinations, as well as invalid responses.

Robustness Analysis. The performance distribution of the five best-F1 LVLMs across different CFP datasets, as depicted in Figure 5, revealed variability and inconsistency in model performance. Even for the same image modality, the models exhibited notable differences in performance depending on the specific dataset. For instance, InternVL-4B achieved a high F1 score of 0.87 on REFUGE but experienced a substantial drop in performance on IDRID, with an F1 score of only 0.04. These observations suggested that the models' robustness and generalization ability might be limited when confronted with variations in image quality, acquisition protocols, and patient populations.

4.4 Diagnosis Analysis

Overall Results. The evaluation results for diagnosis analysis in Table 2 demonstrated that the performance was far from perfect. For glaucoma detection, all models achieved accuracies close to random guessing, indicating that they struggled to distinguish between glaucoma and non-glaucoma cases based on the provided ophthalmic images. In the case of macular hole staging, the average accuracy across all models was only 15.31%, suggesting a significant gap between the models' predictions and the ground truth labels. InternVL-2B, the best-performing model, achieved an accuracy of 30.26% for macular hole staging. While this number was higher than the average, still fell short of the desired performance for clinical use.

Error Analysis. To better understand the failure modes of LVLMs, we conducted error analysis. Six primary categories emerged as follows. These categories illustrated the models' need for better robustness, domain-specific integration, and uncertainty handling. See supplement for more details.

E1: Misclassification. Models misidentified conditions or stages, leading to incorrect predictions. For example, LLaVA-Med incorrectly classified non-glaucoma images as glaucoma.

E2: Failure to Abstain. Some models failed to abstain from making predictions when presented with irrelevant or insufficient data. For example, LLaVA-Med diagnosed glaucoma from an image of a cat, which was clearly not relevant to the medical task. This error type occurred because the model should have recognized that the image was not suitable and refrained from making a prediction.

E3: Inconsistent Reasoning. Models provided contradictory explanations within their own predictions. For example, InternVL initially predicted a macular hole stage 1 but subsequently stated there was no visible macular hole or other abnormalities in the macular region, contradicting its earlier assessment. This inconsistency reflected conflicting reasoning, as the model simultaneously described both a healthy and diseased state.

E4: Hallucination. Models generated details that did not exist in the input image. As illustrated in Figure 4, LLaVA-M-7B hallucinated an additional region, labeled as region 4, which did not correspond to any ground truth regions. This demonstrates the model's tendency to hallucinate features not present in the actual data.

E5: Assertion. Models presented assertive predictions without detailed explanations. For example, Yi-6B simply output the glaucoma condition without further explanation, despite being explicitly instructed to provide one. This lack of justification

weakens the credibility of the prediction.

E6: Lack of Domain-Specific Knowledge. The models exhibited a lack of medical knowledge or produced medically inaccurate explanations. For instance, VILA-3B erroneously referred to a macular hole stage 5, claiming that it represents the final stage where the hole has healed and vision is fully restored, a stage that does not exist in the established staging system. This error demonstrated a lack of domain-specific knowledge because the model made incorrect medical assumptions.

Demographic Subgroup Analysis. We conducted demographic analyses based on age and gender to assess whether the predictive performance of LVLMs is influenced by these factors. Results showed statistically significant differences in accuracy across age and gender subgroups for several models, indicating that certain LVLMs may perform differently across diverse demographic groups. For example, InternVL 2B demonstrated significant variance across both age and gender. See appendix for details.

4.5 Benchmark Justifications

In this section, we justified the design of our benchmark from two perspectives: (1) feasibility and (2) intricacy. These results demonstrated that existing LVLMs faced significant challenges in handling ophthalmic images, even when exposed to medical literature or fine-tuned on relevant datasets.

Feasibility. We demonstrated that our proposed benchmark was feasible, as evidenced by the results in Table 4. We showed that supervised neural networks could effectively address both anatomical recognition and diagnosis analysis tasks. For anatomical recognition, we cropped each region and assigned a corresponding label, formulating the task as a multi-class classification problem. For diagnosis analysis, the input consisted of the entire image, with binary labels for glaucoma detection or 4-class labels for MH staging. As shown in Table 4, the classifiers achieved accuracies far above random baselines, justifying the feasibility of our dataset and task formulation. See supplement for neural network details.

Intricacy. The poor performance of existing LVLMs on LMOD was not due to the absence of ophthalmic images during pretraining. As shown in Table 2, LLaVA-Med performed poorly on both anatomical recognition and diagnostic tasks. However, LLaVA-Med had been pre-trained on PubMed, which included extensive medical literature, likely covering ophthalmology topics. More directly, when we presented LLaVA-Med with an optical coherence tomography (OCT) image and a color fundus photograph (CFP), the model correctly identified the image types as ophthalmic images and described the medical conditions, although the diagnoses were incorrect.

Furthermore, fine-tuning did not significantly improve performance. We fine-tuned LLaVA-Med on a combined dataset of OCT and CFP images. To avoid modality dominance, we balanced the number of OCT and CFP images. Following the official fine-tuning protocol (Liu et al., 2024; Li et al., 2024b), we froze the visual encoder and fine-tuned the MLP adapters and the language model. Despite this, the fine-tuned LLaVA-Med failed to produce meaningful responses as Table 2 showed. For example, it output a series of "opt" for anatomical recognition and empty strings for diagnostic analysis. These findings highlighted that the complexity of ophthalmic images posed inherent challenges that went beyond simple fine-tuning.

5 Conclusion

In this paper, we introduced LMOD, a comprehensive benchmark designed to evaluate the performance of LVLMs on ophthalmic images, free-text, demographic, and disease biomarker information. Spanning a wide range of image modalities and enriched with annotations for anatomical structures and diagnostic labels, LMOD offers a framework for assessing LVLM capabilities in ophthalmology. Our evaluation of 13 state-of-the-art LVLMs revealed significant shortcomings in understanding ophthalmic images, with models struggling in anatomical recognition, showing inconsistent performance across datasets, and performing close to random in diagnostic tasks such as glaucoma detection and macular hole staging. Fine-tuning on ophthalmic data failed to improve results, highlighting the complexity of these images for LVLMs. In contrast, supervised neural networks trained on the same tasks achieved high accuracy, demonstrating that the challenges are not insurmountable. Our error analysis uncovered six key failure modes of the LVLMs, emphasizing the need for more robust models, better integration of domain-specific knowledge, and improved uncertainty handling. We also showed inconsistencies between subgroups of ages and genders. These discoveries suggest Pressing need for benchmarks in the development and validation of ophthalmology-specific LVLMs.

6 Limitations

Modality: While LMOD covers five ophthalmic imaging modalities, free-text, and demographic information—making it by far the most comprehensive—it is inevitable that the selected nine datasets may not include all data modalities in ophthalmology. Additionally, longitudinal data is essential for a more thorough assessment of VLLMs in tracking disease progression. These challenges, which are critical, remain open issues in ophthalmology(Khan et al., 2021), and new datasets under development are addressing them(Rajesh et al., 2023). We plan to further enhance our data and evaluation pipelines and leverage new datasets to enrich the dataset of LMOD.

Continuing Update: While we evaluated 13 representative SoTA LVLMs, covering a range of architectures and training approaches, the rapid pace of development in this field means that new and improved models are continually emerging. As a result, our benchmark may not encompass all the latest advancements in LVLM technology. To address this limitation and promote ongoing evaluation, we will make LMOD and our evaluation code fully available to the research community. This will enable researchers to benchmark new models as they become available, contributing to a continuously evolving understanding of LVLM performance in ophthalmology.

Environmental Impact The evaluation of LVLMs in this study was conducted using NVIDIA H100 GPUs. The energy consumption and carbon footprint associated with training and deploying large-scale AI models have become a growing concern in the research community (Schwartz et al., 2020). It is essential to consider the environmental implications of such practices. The electricity consumed by GPUs during model evaluation contributes to greenhouse gas emissions, depending on the energy mix of the power grid (Schwartz et al., 2020). It is crucial to develop and adopt sustainable practices, as well as promoting the sharing and reuse of pretrained models (Schwartz et al., 2020).

7 Ethical and Risk Considerations

We place a high priority on patient privacy by exclusively using open-access datasets in our study, ensuring that there are no concerns related to privacy leakage. All datasets employed in our research are openly accessible and licensed under various Creative Commons (CC) licenses, allowing for ethical and transparent use. For clarity and transparency, we provide a comprehensive list of the licenses associated with each dataset used in our study in Table 5. For future work involving private data, we will implement strict protocols to prevent data retention and adhere to relevant regulations and ethical guidelines. Any collaborations with healthcare institutions will involve robust data sharing agreements and security measures.

Dataset	Copyright License
Cataract-1k	CC BY 4.0
CAU001	CC BY 4.0
G1020	CC BY-NC 4.0
Harvard-FairSeg	CC BY-NC-ND 4.0
IDRID	CC BY 4.0
OIMHS	CC BY 4.0
ORIGA	CC BY 4.0
REFUGE	CC BY 4.0
Cataract Detection 2	CC BY 4.0

Table 5: Summary of copyright licenses for various publicly available ophthalmology datasets used in LMOD.

Additionally, to mitigate potential biases in datasets and model performance across diverse populations, we have collected datasets that represent a wide range of ethnicities, including Hispanic, Indian, Austrian, and others. By incorporating data from various populations, we aim to ensure that our models are representative of diverse individuals. However, we acknowledge that biases may still persist, and when deploying our method in real-world settings, it is crucial to collect a comprehensive test dataset that includes diverse ethnicities and other factors to thoroughly validate the model's performance and identify any potential biases. We remain committed to ongoing efforts to ensure dataset diversity and fairness in our research.

Furthermore, we emphasize that our proposed method is designed to assist and support medical professionals, rather than replace them. We acknowledge the indispensable role of human expertise in healthcare and believe that AI tools should complement, not replace, the knowledge and judgment of trained clinicians. Our approach aims to provide valuable insights and guidance, empowering doctors to make more informed decisions with the help of advanced data analysis and pattern recognition capabilities. However, the final diagnosis and treatment plan should always be determined by the medical professional, whose experience and understanding of the patient's unique context are essential. In real-world clinical settings, AI should act as a trusted assistant, offering reliable insights that enhance healthcare accuracy and efficiency, while preserving the clinician's central role.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *OpenAI*.

Fares Antaki, Samir Touma, Daniel Milad, Jonathan El-Khoury, and Renaud Duval. 2023. Evaluating the performance of chatgpt in ophthalmology: an analysis of its successes and shortcomings. *Ophthalmology science*, 3(4):100324.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Muhammad Naseer Bajwa, Gur Amrit Pal Singh, Wolfgang Neumeier, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. 2020. G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma detection. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Bjorn Kaijun Betzler, Haichao Chen, Ching-Yu Cheng, Cecilia S Lee, Guochen Ning, Su Jeong Song, Aaron Y Lee, Ryo Kawasaki, Peter van Wijngaarden, Andrzej Grzybowski, et al. 2023. Large language models and their impact in ophthalmology. *The Lancet Digital Health*, 5(12):e917–e924.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

D Cavan, L Makaroff, J da Rocha Fernandes, M Sylvanowicz, P Ackland, J Conlon, D Chaney, A Malhi, and J Barratt. 2017. The diabetic retinopathy barometer study: global perspectives on access to and experiences of diabetic retinopathy screening and treatment. *Diabetes research and clinical practice*, 129:16–24.

Qingyu Chen, Jingcheng Du, Yan Hu, Vipina Kuttichi Keloth, Xueqing Peng, Kalpana Raja, Rui Zhang, Zhiyong Lu, and Hua Xu. 2023. Large language models in biomedical natural language processing: benchmarks, baselines, and recommendations. *arXiv preprint arXiv:2305.16326*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *IEEE/CVF Conference on Computer Visision Pattern Recognition (CVPR)*.

Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, et al. 2023. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141.

G Currie, C Singh, T Nelson, C Nabasenja, Y Al-Hayek, and K Spuur. 2023. Chatgpt in medical imaging higher education. *Radiography*, 29(4):792–799.

Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. 2023. Chatgpt and the rise of large language models: the new ai-driven infodemic threat in public health. *Frontiers in public health*, 11:1166120.

Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature Medicine*, 24(9):1342–1350.

Mohamed Elkholy and Marwa A Marzouk. 2024. Deep learning-based classification of eye diseases using convolutional neural network for oct images. *Frontiers in Computer Science*, 5:1252295.

Negin Ghamsarian, Yosuf El-Shabrawi, Sahar Nasirihaghighi, Doris Putzgruber-Adamitsch, Martin Zinkernagel, Sebastian Wolf, Klaus Schoeffmann, and Raphael Sznitman. 2024. Cataract-1k dataset for deep-learning-assisted analysis of cataract surgery videos. *Scientific Data*, 11(1):373.

Aidan Gilson, Xuguang Ai, Qianqian Xie, Sahana Srinivasan, Krithi Pushpanathan, Maxwell B Singer, Jimin Huang, Hyunjae Kim, Erping Long, Peixing Wan, et al. 2024. Language enhanced model for eye (leme): An open-source ophthalmology-specific large language model. *arXiv preprint arXiv:2410.03740*.

Yangyang Guo, Fangkai Jiao, Zhiqi Shen, Liqiang Nie, and Mohan Kankanhalli. 2024. Unk-vqa: A dataset and a probe into the abstention ability of multi-modal large models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ming Hu, Peng Xia, Lin Wang, Siyuan Yan, Feilong Tang, Zhongxing Xu, Yimin Luo, Kaimin Song, Jurgen Leitner, Xuelian Cheng, et al. 2024. Ophnet: A large-scale video benchmark for ophthalmic surgical workflow understanding. *arXiv preprint arXiv:2406.07471*.

Songtao Jiang, Tuo Zheng, Yan Zhang, Yeying Jin, and Zuozhu Liu. 2024. Moe-tinymed: Mixture of experts for tiny medical large vision-language models. *arXiv preprint arXiv:2404.10237*.

Vishal Kaushal, Nishant Singh Hada, and Sangeeta Sharma. 2023. Eye disease detection through image classification using federated learning. *SN Computer Science*, 4(6):836.

Tiarnan DL Keenan, Qingyu Chen, Elvira Agrón, Yih-Chung Tham, Jocelyn Hui Lin Goh, Xiaofeng Lei, Yi Pin Ng, Yong Liu, Xinxing Xu, Ching-Yu Cheng, et al. 2022. Deeplensnet: deep learning automated diagnosis and quantitative classification of cataract type and severity. *Ophthalmology*, 129(5):571–584.

Saad M Khan, Xiaoxuan Liu, Siddharth Nath, Edward Korot, Livia Faes, Siegfried K Wagner, Pearse A Keane, Neil J Sebire, Matthew J Burton, and Alastair K Denniston. 2021. A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *The Lancet Digital Health*, 3(1):e51–e66.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024a. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024b. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. 36.

Zhi Wei Lim, Krithi Pushpanathan, Samantha Min Er Yew, Yien Lai, Chen-Hsin Sun, Janice Sing Harn Lam, David Ziyou Chen, Jocelyn Hui Lin Goh, Marcus Chun Jin Tan, Bin Sheng, et al. 2023. Benchmarking large language models' performances for myopia care: a comparative analysis of chatgpt-3.5, chatgpt-4.0, and google bard. *EBioMedicine*, 95.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. 2023. Vila: On pretraining for visual language models. *arXiv preprint arXiv:2312.07533*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. 2023. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, page 100017.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *International Conference on Learning Representation (ICLR)*.

Yan Luo, Yu Tian, Min Shi, Louis R Pasquale, Lucy Q Shen, Nazlee Zebardast, Tobias Elze, and Mengyu Wang. 2024. Harvard glaucoma fairness: a retinal nerve disease dataset for fairness learning and fair identity normalization. *IEEE Transactions on Medical Imaging*.

David C Neely, Kevin J Bray, Carrie E Huisingh, Mark E Clark, Gerald McGwin, and Cynthia Owsley. 2017. Prevalence of undiagnosed age-related macular degeneration in primary eye care. *JAMA ophthalmology*, 135(6):570–575.

World Health Organization. 2023. Blindness and vision impairment. Accessed: [Insert access date here].

José Ignacio Orlando, Huazhu Fu, João Barbosa Breda, Karel Van Keer, Deepti R Bathula, Andrés Diaz-Pinto, Ruogu Fang, Pheng-Ann Heng, Jeyoung Kim, JoonHo Lee, et al. 2020. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis*, 59:101570.

Yifan Peng, Shazia Dharssi, Qingyu Chen, Tiarnan D Keenan, Elvira Agrón, Wai T Wong, Emily Y Chew, and Zhiyong Lu. 2019. Deepseenet: a deep learning model for automated classification of patient-based age-related macular degeneration severity from color fundus photographs. *Ophthalmology*, 126(4):565–575.

Porwal Prasanna, Pachade Samiksha, Kamble Ravi, Kokare Manesh, D Girish, S Vivek, and Meriaudeau Fabrice. 2018. Indian diabetic retinopathy image dataset (idrid). *IEEE Dataport*, 2.

PupiUp. 2023. cau001 dataset. *Roboflow Universe*. Visited on 2024-06-03.

Ziyuan Qin, Huahui Yi, Qicheng Lao, and Kang Li. 2022. Medical image understanding with pretrained vision language models: A comprehensive study. *arXiv preprint arXiv:2209.15517*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Anand E Rajesh, Oliver Q Davidson, Cecilia S Lee, and Aaron Y Lee. 2023. Artificial intelligence and diabetic retinopathy: Ai framework, prospective studies, head-to-head validation, and cost-effectiveness. *Diabetes care*, 46(10):1728–1739.

SRM University Ramapuram. 2023. Cataract detection 2 dataset. *Roboflow Universe*. Visited on 2024-06-03.

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. 2020. Green ai. *Communications of the ACM*, 63(12):54–63.

Yih-Chung Tham, Xiang Li, Tien Y Wong, Harry A Quigley, Tin Aung, and Ching-Yu Cheng. 2014. Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology*, 121(11):2081–2090.

Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, et al. 2024. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493.

Daniel Shu Wei Ting, Louis R Pasquale, Lily Peng, John Peter Campbell, Aaron Y Lee, Rajiv Raman, Gavin Siew Wei Tan, Leopold Schmetterer, Pearse A Keane, and Tien Yin Wong. 2019. Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology*, 103(2):167–175.

Rohit Varma, Paul P Lee, Ivan Goldberg, and Sameer Kotak. 2011. An assessment of the health and economic burdens of glaucoma. *American journal of ophthalmology*, 152(4):515–522.

Vijaya Kumar Velpula and Lakhan Dev Sharma. 2023. Multi-stage glaucoma classification using pre-trained convolutional neural networks and voting-based classifier fusion. *Frontiers in Physiology*, 14:1175881.

Shanshan Wang, Cheng Li, Rongpin Wang, Zaiyi Liu, Meiyun Wang, Hongna Tan, Yaping Wu, Xinfeng Liu, Hui Sun, Rui Yang, et al. 2021. Annotation-efficient deep learning for automatic medical image segmentation. *Nature Communications*, 12(1):5915.

Sean Wu, Michael Koo, Lesley Blum, Andy Black, Liyo Kao, Zhe Fei, Fabien Scalzo, and Ira Kurtz. 2024. Benchmarking open-source large language models, gpt-4 and claude 2 on multiple-choice questions in nephrology. *NEJM AI*, 1(2):AIdbp2300092.

Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. 2024. A comprehensive survey of large language models and multimodal large language models in medicine. *arXiv preprint arXiv:2405.08603*.

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

Xin Ye, Shucheng He, Xiaxing Zhong, Jiafeng Yu, Shangchao Yang, Yingjiao Shen, Yiqi Chen, Yaqi Wang, Xingru Huang, and Lijun Shen. 2023. Oimhs: An optical coherence tomography image dataset based on macular hole manual segmentation. *Scientific Data*, 10(1):769.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *IEEE/CVF Conference on Computer Visision Pattern Recognition (CVPR)*.

Jiawei Zhang, Tianyu Pang, Chao Du, Yi Ren, Bo Li, and Min Lin. 2024a. Benchmarking large multimodal models against common corruptions. *arXiv preprint arXiv:2401.11943*.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. 2024b. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? *arXiv preprint arXiv:2408.13257*.

Zhuo Zhang, Feng Shou Yin, Jiang Liu, Wing Kee Wong, Ngan Meng Tan, Beng Hai Lee, Jun Cheng, and Tien Yin Wong. 2010. 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*, pages 3065–3068. IEEE.

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. 2023. A foundation model for generalizable disease detection from retinal images. *Nature*, 622(7981):156–163.

A Prompts for Benchmarking

Below are the specific prompts used for the evaluation of large vision-language models (LVLMs) in the tasks of anatomical recognition and diagnosis analysis:

A.1 Anatomical Recognition Prompts

The following prompt was used to evaluate the models' ability to recognize and classify anatomical regions in ophthalmic images based on labeled bounding boxes:

- **Prompt 1:** "This is an image [IMAGE DESCRIPTION] of type [IMAGE TYPE]. Please identify the type of each labeled bounding box in this image. Options can be: [REGION TYPE 1], [REGION TYPE 2], ... Please just follow the format: Region ID: xxx; Type: xxx."

A.2 Diagnosis Analysis Prompts

The following prompts were used to assess the models' diagnostic reasoning for conditions such as glaucoma and macular hole staging:

- **Prompt 1:** "This is <IMAGE TYPE>. Based on the image, please tell me whether this image contains glaucoma, then give justifications. Follow the format: GLAUCOMA / NON-GLAUCOMA; Explanation: <JUSTIFICATION>."
- **Prompt 2:** "This is <IMAGE TYPE>. Based on the image, please tell me the stage of the macular hole, then give justifications. Follow the format: Stage: <AN INTEGER>; Explanation: <JUSTIFICATIONS>."

B Intrinsic Difficulties of Ophthalmology

While the low performance of LVLMs on the LMOD benchmark might suggest a lack of exposure to ophthalmology images during pretraining, our investigation reveals that these models can indeed recognize such images. As demonstrated in Figure 6, when presented with a retinal fundus photograph and asked to identify the image, both GPT-4o and LLaVA-7B correctly recognize it as an ophthalmology-related image. GPT-4o specifically identifies the image as a "retinal fundus photograph, commonly used in ophthalmology to examine the interior surface of the eye," highlighting its potential applications in detecting and monitoring various eye conditions. Similarly, LLaVA-7B recognizes the image as a close-up view of a human eye, albeit focusing more on superficial features such as the iris and surrounding tissue. These findings suggest that LVLMs have been exposed to ophthalmology data during training and can identify such images when encountered. However, the performance of LVLMs on more complex tasks, such as glaucoma classification, remains a challenge, as evidenced by the results on the LMOD benchmark. This discrepancy between image recognition and task-specific performance underscores the inherent challenges posed by ophthalmology data for LVLMs.

Regarding the first point, we find that LVLMs like GPT-4o and LLaVA-7B can correctly identify color fundus photographs (CFPs) from LMOD, indicating that these models have encountered ophthalmology data during training. This suggests that the low performance is not due to a complete lack of exposure to the domain.

As for the second point, we fine-tuned LLaVA-7B on the glaucoma detection task using the training set of LMOD. Notably, even on the same training data, the fine-tuned model still achieves only 50% accuracy, always predicting "non-glaucoma" regardless of the input. This indicates that simple fine-tuning is insufficient to overcome the random accuracy in glaucoma detection, underscoring the inherent challenges of ophthalmology data for LVLMs.

Additionally, some may question the suitability of the ophthalmology images in LMOD for glaucoma classification. However, this concern is unfounded. As demonstrated in (Velpula and Sharma, 2023), a deep learning model achieves an accuracy of 93.86% on the ORIGA dataset, which is the same dataset used for evaluating glaucoma detection in our benchmark. This high performance indicates that the images in ORIGA contain sufficient discriminative information for distinguishing between glaucoma and non-glaucoma cases. Consequently, the poor performance of LVLMs on this task cannot be attributed to the quality or suitability of the dataset itself, but rather highlights the inherent challenges that ophthalmology data poses for these models.

C Experimental Setup

We developed a general framework based on PyTorch, providing a unified interface for performing inference across various vision-language models (VLMs). This framework ensures consistent evalu-

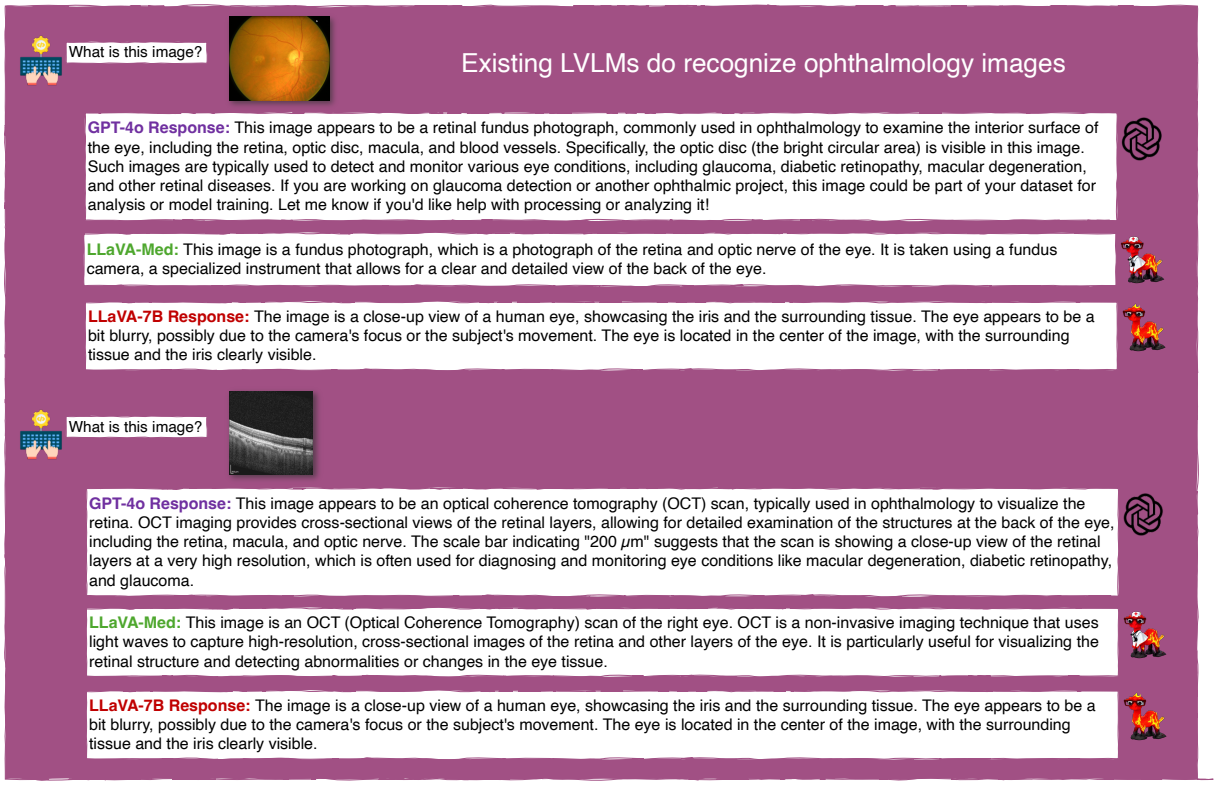


Figure 6: Responses from GPT-4o and LLaVA-7B when presented with a retinal fundus photograph and asked, "What is this image?"

ation and smooth integration with different models.

For each VLM, we used the same computing infrastructure—specifically, two RTX 6000 GPUs—to perform the inference. We evaluated the models using nine different ophthalmology datasets, with identical prompts and inputs provided to each VLM. Moreover, we applied the default hyperparameters for each model during the evaluation. This approach allowed us to fairly compare the performance of the different models.

D Computational Resource

The computing infrastructure includes 11 GPU nodes, each equipped with 2x AMD EPYC 7742 processors (128 cores), 1TB of RAM, and 8 Quadro RTX 6000 GPUs per node. Additionally, there are 7 GPU nodes with 2x Intel Icelake Xeon Platinum 8358 processors.

For vision-language model inference tasks on various ophthalmology datasets, the runtime typically ranges from two to four hours, depending on the specific dataset.

E Use Of AI Assistants

We used AI tools to assist with coding tasks, such as debugging and optimizing code during the development phase. Additionally, we leveraged AI to help polish the manuscript, addressing grammar issues and ensuring clarity and coherence in our presentation. However, all critical decisions such as the research design, methodology, and conclusions were made independently by the authors.

F Hyperparameters

This section outlines the essential hyperparameters that were chosen for the large vision-language models (LVLMs) in our experiments, listed in Table 6.

1. **Image Resolution**: The image resolution defines the size of the visual input processed by each LVLM. Higher resolutions capture finer details. Table 6 summarizes the exact values of these hyperparameters.

2. **Top-p Sampling**: Top-p, also known as nucleus sampling, is a hyperparameter that influences the randomness of a language model's output. It defines a probability threshold and selects the smallest set of tokens whose cumulative probability exceeds this threshold. The model then samples randomly from this subset to generate the output. This approach allows for more diverse and creative results compared to methods that randomly sample from the entire vocabulary.

3. **Temperature**: The temperature hyperparameter influences the randomness of the model's output by scaling logits before applying softmax. Higher temperatures (e.g., >1) encourage more diverse outputs by flattening the probability distribution, making it suitable for creative tasks. Lower temperatures (e.g., <1) concentrate the distribution, resulting in more focused outputs, which is critical in medical domains to ensure reliable, deterministic responses. Temperature is disabled when setting to be 0.

4. **Beams Number**: Beam search is a decoding strategy that retains multiple candidate sequences at each generation step. A higher number of beams (e.g., 5 or 10) explores more possibilities, potentially yielding better results at the cost of increased computation. A lower beams number (e.g., 1) favors efficiency and speed but risks missing better sequences, which may be a concern in domains requiring high-quality outputs.

5. **Number of Parameters**: The number of parameters refers to the total count of learnable weights in a model, directly influencing its capacity and performance. Larger models tend to perform better due to increased capacity, but at the cost of higher memory usage and slower inference times.

6. **Max New Tokens**: This hyperparameter limits the number of tokens generated by the model during inference.

G Response Examples

Input:

This is an ophthalmology OCT image. Based on the image, please tell me the stage of macular hole decision. Then, give a detailed justification and explanation for your answer. Follow the format: {Stage: <AN INTEGER>; Justification: <EXPLANATION>}.

Ground Truth (GT):

Stage: 4

GPT-4o Response:

Stage: 4; Justification: The OCT image displays a full-thickness macular hole, where there is a complete defect involving all the retinal layers from the inner surface to the retinal pigment epithelium. The presence of this full-thickness defect, along with the evident central gap extending through the macula and alterations in the surrounding retina, clearly indicates a Stage 4 macular hole. Additionally, the involvement of the posterior vitreous detachment supports this stage classification.

LLaVA Med Response:

Stage: 3; Justification: The image shows a macular hole with a diameter of 300 micrometers. This is considered stage 3, which is a more advanced stage of macular hole development. The macular hole is larger and may be associated with more severe vision loss.

LLaVA-1.5-7B Response:

Stage: 2; Justification: The image shows a macular hole with a well-defined, clean, and clear appearance. The hole is surrounded by a well-organized retinal structure, including the fovea and the outer retinal layers. The presence of a clear and well-defined macular hole indicates that it is in an early stage of development, which is typically characterized by a clean and organized appearance.

LLaVA-M-7B Response:

The image you've provided appears to be an Ophthalmology OCT (Optical Coherence Tomography) scan, which is a non-invasive imaging technique

Model	Image Resolution	Top-p Sampling	Temperature	Beams Number	Number of Parameters	Max New Tokens
GPT-4o	512 x 512	-	-	-	-	512
LLaVA-Med	336 x 336	1.0	0	1	7.56 B	512
LLaVA-1.5-7B	336 x 336	1.0	0	1	7.06 B	512
LLaVA-M-7B	336 x 336	1.0	0	1	7.56 B	512
LLaVA-V-7B	336 x 336	1.0	0	1	7.06 B	512
LLaVA-13B	336 x 336	1.0	0	1	13.35 B	512
Yi 6B	336 x 336	1.0	0	1	6.71 B	512
InternVL 2B	448 x 448	-	0	1	2.20 B	512
InternVL 4B	448 x 448	-	0	1	4.14 B	512
QWen	448 x 448	-	0	-	9.65 B	512
VILA 3B	384 x 384	1.0	0	1	3.14 B	512
VILA 3B-S2	768 x 768	1.0	0	1	3.16 B	512
VILA 8B	384 x 384	1.0	0	1	8.49 B	512

Table 6: Hyperparameters for the large vision-language models evaluated in this work. The "Image Resolution" column shows the spatial dimensions of the input images processed by each model. The "Top-p Sampling" column indicates the value of the top-p sampling hyperparameter used during text generation. The "Temperature" column specifies the value of the temperature hyperparameter used in the softmax function during text generation. It is disabled given value 0. The `Beams Number" column indicates the number of beams used in the beam search decoding algorithm. The "Number of Parameters" column reports the model size in billions (B) of parameters. The "Max New Tokens" column indicates the maximum number of new tokens the model can generate during inference. Symbol "-" denoting cases where it is not applicable or not reported.

```
used to visualize the layers of the
retina...
```

LLaVA-V-7B Response:

```
Stage: 200 µm; Justification: The
macular hole is not clearly visible
in this image. The hole is typically
smaller than 200 µm, which is the
scale indicated by the measurement
on the left side of the image.
```

LLaVA-13B Response:

```
Stage: 2; Justification: The image
shows a macular hole in the retina,
which is a stage 2 macular hole. In
stage 2, the hole is small and has
not yet begun to contract.
```

Yi-6B Response:

```
Stage: 4
```

InternVL-2B Response:

```
Stage: 2; Justification: The image
shows a macular hole, which is a
small hole in the retina that can
cause vision loss.
```

InternVL-4B Response:

```
Stage: 1; Justification: The
image shows a clear view of the
retinal layers with a distinct,
well-defined macular hole.
```

QWen Response:

```
Stage: 3; Justification: The image
shows a large, full-thickness
macular hole with posterior
vitreous detachment and subretinal
fluid.
```

VILA-8B Response:

```
[0.00, 0.00, 1.00, 1.00]
```

H Supervised Training Settings

To evaluate the feasibility of our proposed benchmark, we implemented neural network classifiers for anatomical recognition and diagnosis analysis. For both tasks, we used 80%, 15%, and 5% data for training, validation, and test. For anatomical recognition, we employed a CNN visual encoder whose architecture is like below:
The CNN was trained with the following settings:

- Image resolution: 128×128
- Batch size: 512
- Learning rate: 0.001

```python
class RegionClassifier(nn.Module):
    def __init__(self, num_classes):
        self.conv1 = nn.Conv2d(3, 32, kernel_size=3, stride=1, padding=1)
        self.pool = nn.MaxPool2d(2, 2)
        self.conv2 = nn.Conv2d(32, 64, kernel_size=3, stride=1, padding=1)
        self.fc1 = nn.Linear(64 * 32 * 32, 128)
        self.fc2 = nn.Linear(128, num_classes)
        self.relu = nn.ReLU()

    def forward(self, x):
        x = self.pool(self.relu(self.conv1(x)))
        x = self.pool(self.relu(self.conv2(x)))
        x = x.view(-1, 64 * 32 * 32)
        x = self.relu(self.fc1(x))
        x = self.fc2(x)
        return x
```

Figure 7: The RegionClassifier model implementation in PyTorch.

- Epochs: 20

For diagnostic analysis, we fine-tuned RETFound as the visual encoder. RETFound is a foundation model for retinal images, built on a large Vision Transformer (ViT) architecture with 24 Transformer blocks and an embedding vector size of 1,024 (Zhou et al., 2023). The RETFound model offers two variations designed for different image types: CFP and OCT. For macular hole (MH) stage classification, we employed the OCT variation, while the CFP model was used for glaucoma classification (according to the dataset's image type). For both tasks, we fine-tuned RETFound using the default parameter settings:

- Image resolution: 224 × 224
- Batch size: 16
- Base learning rate:: 5e-3
- Epochs: 50
- Layer decay: 0.65
- Weight decay: 0.05

The model's performance on anatomical recognition and diagnosis analysis tasks served as a baseline for the complexity of our dataset, and is compared with the performance of LVLMs in subsequent sections.

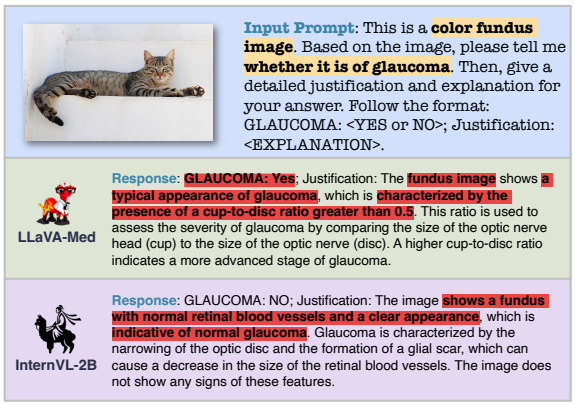


Figure 8: **Failure to abstain from making predictions on out-of-domain images.** The figure illustrates the responses generated by LLaVA-Med and InternVL-2B when presented with a non-ophthalmology image of a cat and queried about the presence of glaucoma. Despite the image being outside the domain of fundus photography, LLaVA-Med incorrectly classifies the cat image as showing signs of glaucoma, citing a typical cup-to-disc ratio greater than 0.5. Similarly, InternVL-2B misinterprets the cat image as a fundus image but concludes that there are no signs of glaucoma based on the visible features.

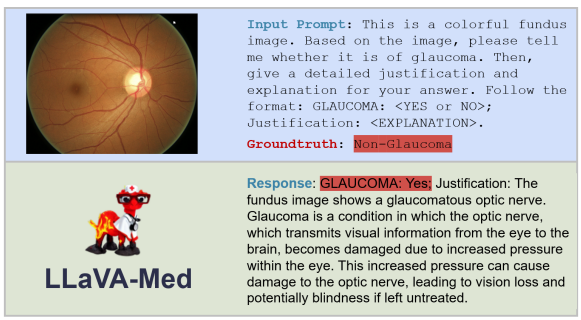


Figure 9: LLaVA-Med incorrectly predicted the image as glaucoma, providing a detailed but inaccurate explanation.

I Detailed Error Analysis

I.1 Misclassification

As shown in Figure 9, LLaVA-Med incorrectly predicted the image as glaucoma, providing a detailed but inaccurate explanation. Experimental results in

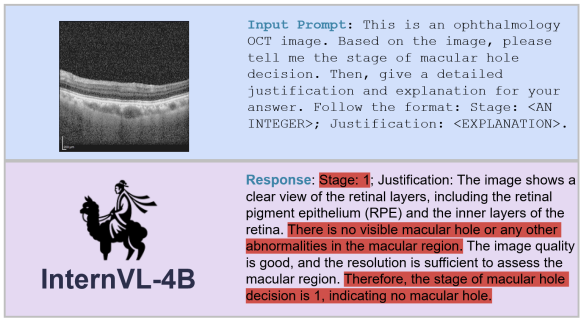


Figure 10: InternVL-4B presented contradictory explanations within its prediction.

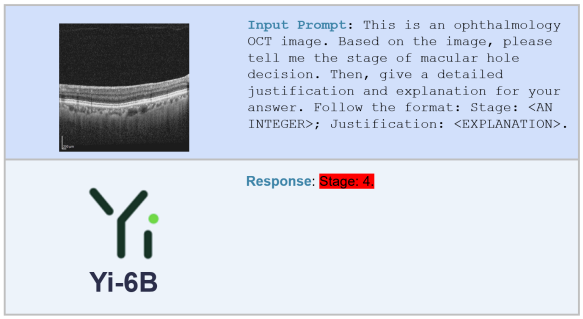


Figure 11: Yi-6B provided assertive predictions without offering detailed explanations.

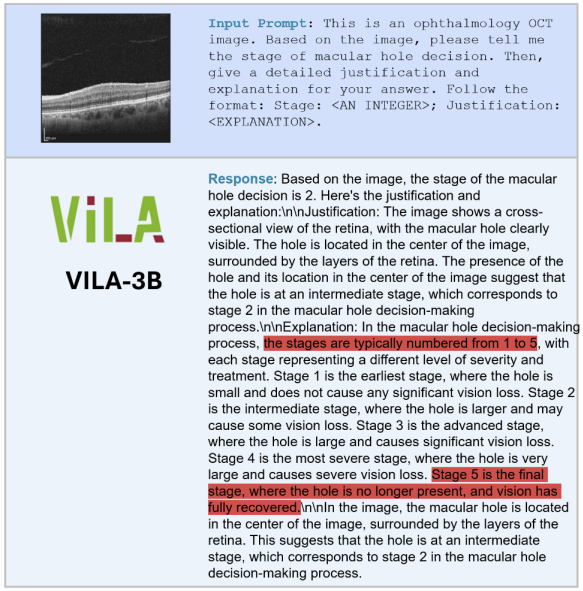


Figure 12: LVLMs display a lack of medical knowledge and produce medically inaccurate explanations. The figure illustrates the responses generated by VILA-3B

Table 2 indicate that all LVLMs achieve suboptimal accuracy in both glaucoma and macular hole stage classification tasks. Additionally, models in the LLaVA series, such as LLaVA-Med, LLaVA-13B, and LLaVA-M-7B, consistently generate the same response, predicting glaucoma disease across different input images. These findings suggest that current LVLMs still lack the capability to accurately interpret images and make reliable predictions.

I.2 Failure to Abstain

Some models failed to recognize when the input image was irrelevant to the medical task at hand and continued to provide diagnostic predictions regardless of the mismatch. As Figure 8 showed, LLaVA-Med and InternVL-2B misdiagnosed the image of a cat as showing signs of glaucoma. In the provided scenario, the prompt explicitly asked whether the image depicted a case of glaucoma, which typically involves human ophthalmic images, such as retinal scans or fundus photographs. However, the input image was clearly a photograph of a cat, an object entirely outside the scope of the medical context. The inability of models like LLaVA-Med to abstain from making predictions on out-of-domain inputs points to their limitations in robustly handling uncertainty or recognizing when data does not meet the conditions required for valid predictions. It underscores the need for integrating mechanisms into LVLMs that can detect when the input data is irrelevant, ensuring that models avoid generating misleading or incorrect medical diagnoses.

I.3 Inconsistent Reasoning

As observed in 10, InternVL-4B initially predicted the macular hole stage of the input image as stage 1 but then stated in its explanation that there was no visible macular hole or other abnormalities in the macular region, ultimately concluding with no macular hole. Consistent reasoning is proven to be crucial for providing accurate answers, yet this explanation contradicts the initial prediction, highlighting the intrinsic issue of inconsistent model reasoning.

I.4 Assertion

Some LVLMs, such as Yi-6B, tend to provide direct answers without offering detailed explanations. As shown in 11, Yi-6B simply predicted the presence of glaucoma without providing any justification, despite the prompt explicitly requesting one. In

practical clinical scenarios, reasoning steps are critically important. AI models should not only assist clinicians by streamlining their workflow but also ensure decision-making transparency. This allows clinicians to validate both the final prediction and the intermediate reasoning steps, helping to identify potential issues and fostering trust in AI-assisted clinical systems.

I.5 Lack of Domain-Specific Knowledge

Some Large Vision-Language Models (LVLMs) demonstrate inadequate medical knowledge and produce medically inaccurate explanations. As illustrated in Figure 12, VILA-3B inaccurately performs an initial analysis by misclassifying the stages of macular holes, erroneously numbering them from 1 to 5. Specifically, it incorrectly identifies Stage 5 as the final stage, where the macular hole is presumed to be resolved and vision fully restored. In reality, the established staging system for macular holes ranges from Stage 1 to Stage 4, without Stage 5. These inaccuracies highlight the persistent issue of LVLMs lacking domain-specific expertise.

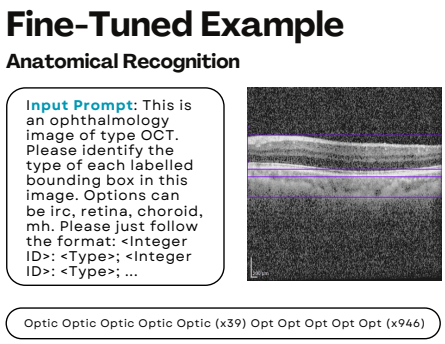


Figure 13: Fine-tuned example response for anatomical recognition. The model was tasked with identifying the type of each labeled region in an OCT image. The fine-tuned LVLM failed to produce meaningful responses, outputting repetitive text such as "Optic Optic Optic Optic Optic (x39)" followed by "Opt Opt Opt Opt Opt (x946)."

J Details of Fine-tuning LVLM

We provide the fine-tuning configuration used for training the large vision-language model (LVLM) with the script in the following script. During fine-tuning for anatomical recognition, the model failed to produce coherent outputs, generating the repetitive response: "Optic Optic Optic Optic Optic (x39) Opt Opt Opt Opt Opt (x946)". For diagnostic tasks, such as glaucoma and macular hole staging, the model returned no responses.

Model	P-value
GPT4o	0.3395
LLaVA-Med	5.1625e-15
LLaVA-1.5-7B	9.3736e-06
LLaVA-M-7B	0.1766
LLaVA-V-7B	N/A
LLaVA-13B	N/A
Yi 6B	0.1022
InternVL 2B	0.0472
InternVL 4B	0.0754
Qwen	9.0541e-16
VILA 3B	0.5818
VILA 3B-S2	0.7814
VILA 8B	0.9269

Table 7: P-values from the age group analysis of various LVLMs.

K Demographic analysis

Tables 7 and 8 provide detailed p-value statistics from the age and gender subgroup analyses of various large vision-language models (LVLMs). Specifically, we evaluated the models across different age groups (18-40, 40-60, 60+) and gender categories (male, female). This two-tiered evaluation approach allowed us to examine how demographic factors such as age and gender influence the predictive accuracy of the models. Statistically significant differences were observed for several models, including the following examples:

- **InternVL 2B**: $p = 0.0472$ for age, and $p = 6.94 \times 10^{-24}$ for gender.
- **LLaVA-1.5-7B**: $p = 9.3736 \times 10^{-6}$ for age, and $p = 8.5040 \times 10^{-7}$ for gender.
- **LLaVA-Med** and **Qwen**: significant p-values in the age subgroup analysis ($p = 5.16 \times 10^{-15}$ and $p = 9.05 \times 10^{-16}$, respectively).
- **InternVL 4B**: significant p-value in the gender subgroup analysis ($p = 0.0082$).

L Detailed benchmarked LVLMs

We benchmarked 13 LVLMs on the LMOD benchmark. These models included:
GPT-4o: A proprietary model developed by OpenAI (Achiam et al., 2023).
LLaVA Variations: LLaVA (Liu et al., 2024) leveraged a pre-trained vision encoder and a large

```
deepspeed llava/train/train_mem.py \
    --model_name_or_path llava-med-7b-delta \
    --data_path <DATA_PATH> \
    --vision_tower openai/clip-vit-large-patch14 \
    --mm_vision_select_layer -2 \
    --mm_use_im_start_end True \
    --bf16 True \
    --output_dir <OUTPUT_PATH> \
    --num_train_epochs 3 \
    --per_device_train_batch_size 1 \
    --per_device_eval_batch_size 4 \
    --gradient_accumulation_steps 8 \
    --evaluation_strategy "no" \
    --save_strategy "steps" \
    --save_steps 5000 \
    --save_total_limit 3 \
    --learning_rate 2e-5 \
    --weight_decay 0. \
    --warmup_ratio 0.03 \
    --lr_scheduler_type "cosine" \
    --logging_steps 1 \
    --tf32 True \
    --fsdp "full_shard auto_wrap" \
    --fsdp_transformer_layer_cls_to_wrap 'LlamaDecoderLayer' \
    --model_max_length 2048 \
    --gradient_checkpointing True \
    --lazy_preprocess True \
    --report_to wandb
```

Model	P-value
GPT4o	0.3734
LLaVA-Med	0.4411
LLaVA-1.5-7B	8.5040e-07
LLaVA-M-7B	0.1937
LLaVA-V-7B	N/A
LLaVA-13B	N/A
Yi 6B	0.1681
InternVL 2B	6.9497e-24
InternVL 4B	0.0082
Qwen	0.8096
VILA 3B	0.9345
VILA 3B-S2	0.1090
VILA 8B	0.2258

Table 8: P-values from the gender group analysis of various LVLMs.

language model to achieve state-of-the-art performance on various vision-language tasks. We evaluated several variations based on different language models and sizes, including LLaVA-7B, LLaVA-M-7B (M for Mistral), LLaVA-V-7B (V for Vicuna), LLaVA-13B, and LLaVA-Med (Li et al., 2024b), which was fine-tuned on a large-scale medical image-text dataset.

Yi-6B: A vision-language model supporting both Chinese and English (Young et al., 2024).

InternVL Variations: InternVL (Chen et al., 2024) aligned a scaled-up vision foundation model with a LLM using web-scale image-text data. We evaluated two variants with 2B and 4B parameters.

QWen: A vision-language model that used a query-based approach to align visual and textual representations (Bai et al., 2023).

VILA Variations: VILA explored different pre-training strategies for LVLMs. We evaluated three variants: VILA-3B, VILA-3B-S2, and VILA-8B, with parameter counts ranging from 3 billion to 8 billion (Lin et al., 2023).