FunBench: Benchmarking Fundus Reading Skills of MLLMs

Qijie Wei[1,2], Kaiheng Qian[1], and Xirong Li[1,2*]

[1] Renmin University of China, Beijing, China
[2] Beijing Key Laboratory of Fundus Diseases Intelligent Diagnosis & Drug/Device Development and Translation, Beijing, China
https://github.com/ruc-aimc-lab/FunBench

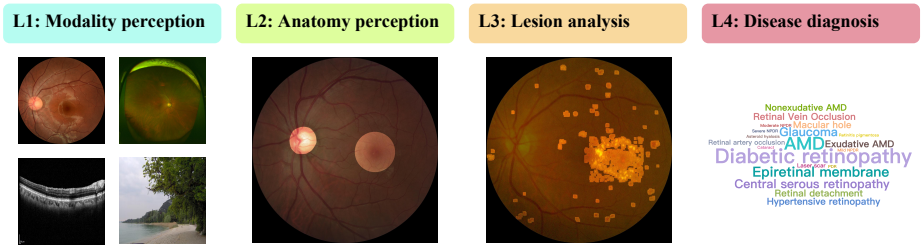
Abstract. Multimodal Large Language Models (MLLMs) have shown significant potential in medical image analysis. However, their capabilities in interpreting fundus images, a critical skill for ophthalmology, remain under-evaluated. Existing benchmarks lack fine-grained task divisions and fail to provide modular analysis of its two key modules, i.e., large language model (LLM) and vision encoder (VE). This paper introduces FunBench, a novel visual question answering (VQA) benchmark designed to comprehensively evaluate MLLMs' fundus reading skills. FunBench features a hierarchical task organization across four levels (modality perception, anatomy perception, lesion analysis, and disease diagnosis). It also offers three targeted evaluation modes: linear-probe based VE evaluation, knowledge-prompted LLM evaluation, and holistic evaluation. Experiments on nine open-source MLLMs plus GPT-4o reveal significant deficiencies in fundus reading skills, particularly in basic tasks such as laterality recognition. The results highlight the limitations of current MLLMs and emphasize the need for domain-specific training and improved LLMs and VEs.

Keywords: MLLM · ophthalmology · VQA benchmark.

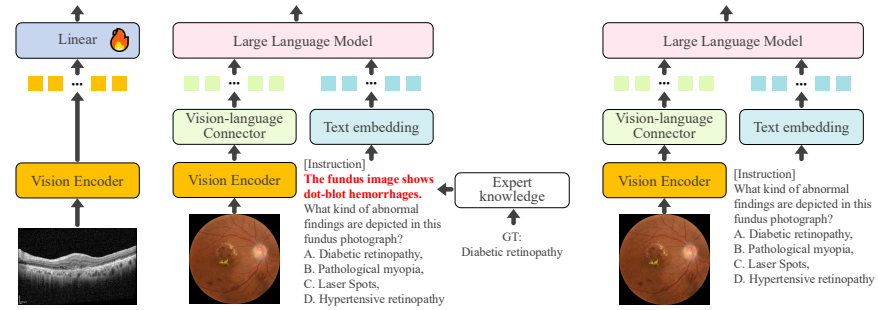
1 Introduction

Multimodal Large Language Models (MLLMs), with their strong capabilities in generic visual content understanding, are rocking the field of medical image analysis [8, 14] and consequently reshaping the research landscape of medical image-based disease diagnosis [6, 37]. Consider AI-enabled Ophthalmology for instance. The combination of remote MLLMs *and* locally deployed non-invasive fundus imaging devices such as color fundus photography (CFP) makes high-quality primary diabetes care possible at community clinics [15]. While research on medical MLLMs, including ophthalmology-focused studies [6, 15], is growing rapidly [32, 36], we observe that the development of ophthalmology-targeted benchmarks is lagging behind. This paper develops **FunBench**, a new benchmark for evaluating the efficacy of *open-source* MLLMs for **fundus reading** tasks of varied difficulties, see Fig. 1.

*Corresponding author (xirong@ruc.edu.cn)



(a) Tasks at four levels, from *modality perception* (L1) to *disease diagnosis* (L4).



(b) E-mode I (c) E-mode II (d) E-mode III

Fig. 1: **Proposed FunBench for assessing an MLLM's funding reading skills** by (a) varied-level tasks and three distinct evaluation modes, *i.e.* (b) *E-mode I*: linear-probe based vision encoder (`VE`) evaluation, (c) *E-mode II*: knowledge-prompted large language model (`LLM`) evaluation and (d) *E-mode III*: holistic evaluation.

While medical benchmarks such as OmniMedVQA [9] and GMAI-MMBench [35] have retinal images included, they treat retinal image-based visual question answering (VQA) as a *single* task. A detailed and structured evaluation of how well a specific model can interpret retinal images is naturally absent from these *general-purpose* benchmarks. Towards filling the gap, LMOD [25] has been developed, evaluating the performance of MLLMs on recognizing major anatomical structures of the fundus, *e.g.* optic cup, optic disc and fovea, and on recognizing two diseases, *i.e.* glaucoma and macular hole. Therefore, LMOD enables a more detailed assessment as opposed to OmniMedVQA and GMAI-MMBench. Our FunBench technically differs from LMOD for its *hierarchical* task organization and *targeted* evaluation modes.

The construction of FunBench is driven by our quest to answer two fundamental questions related to the assessment of a MLLM's fundus reading skills. That is, *what to ask* and *how to ask*. On answering the first question, we consider four levels of tasks, ranging from *low-level* modality and anatomy perception to *high-level* lesion analysis and disease diagnosis. Such a task organization enables a comprehensive assessment of the level and extent to which an MLLM has mastered its fundus reading skills, an evaluation that prior benchmarks have not

adequately supported. For answering the second question, our evaluation is not only targeted at the MLLM as a whole, but also considers its two key modules, *i.e.* vision encoder (`VE`) and large language model (`LLM`). Such a design enables a joint evaluation that is both holistic and modular – an analytical approach not used in the previous work [9, 25].

To sum up, our contributions are three-fold as follows:

• **Dataset**. We introduce FunBench, a novel benchmark for evaluating fundus reading skills of MLLMs.

• **Evaluation**. We evaluate nine open-source MLLMs, released between 2023.10 to 2025.01 on HuggingFace. These models cover five `VE`s and seven `LLM`s. We include GPT-4o as a proprietary baseline and DIVOv2 [22] as a `VE` baseline.

• **Findings**. The MLLMs evaluated rely heavily on the internal `LLM`s to perform the fundus reading tasks. The models possess quite limited fundus reading skills. In particular, they lack some basic skills such as laterality recognition.

2 FunBench Construction

2.1 Dataset Curation

Hierarchical Task Organization Depending on the extent to which a professional fundus reading skill is required, we consider four levels of tasks, which mirrors the progressive complexity of fundus reading capabilities, ranging from basic modality perception to advanced fundus image interpretation.

• **Level 1 (L1): Modality perception**. A model possessing L1 skills shall identify the imaging technique used to produce a given fundus image. In a naive setting, one might consider selecting "fundus" from multiple choices such as "natural", "painting", and "remote sensing". A more difficult setting is to select among varied fundus modalities such as `CFP`, `OCT` and `UWF` plus other medical imaging such as `X-ray` and `MRI`. We name the two settings *coarse-grained* modality perception (**L1a**) and *fine-grained* modality perception (**L1b**), respectively.

• **Level 2 (L2): Anatomy perception**. The optic disc (OD) and the fovea are two major anatomical structures in the retina. Their visual patterns are relatively clear: OD typically appears as an oval bright object in a color fundus image, whilst the fovea is centered in the darkest area on OD's temporal side [34]. Moreover, the laterality of the fundus image, *i.e.* whether from a left or right eye, can be determined by the relative position of the OD in the given image [13]. Hence, a model possessing L2 skills shall tell the relative OD-fovea position (**L2a**) and recognize the laterality (**L2b**).

• **Level 3 (L3): Lesion analysis**. Lesions are pathological alterations caused by varied diseases and can be observed (to some extent) by specific fundus imaging techniques. Recognizing what the lesions are, where they occur, how large and how many they are id essential for reliable and explainable disease diagnosis. Consider diabetic retinopathy (DR) grading for instance. A sufficient criterion for severe nonproliferative DR is the presence of over 20 haemorrhages in each of the nasal, temporal, superior, and inferior quadrants of the fundus [33]. Therefore, a model mastering L3 skills shall be able to perform lesion *recognition*

Table 1: **Statistics of FunBench**: 16,348 fundus images and 91,810 visual questions *w.r.t.* 10 tasks in total.

Level	#Visual questions		Sample question	Data sources
	Single-ans.	*Multi-ans.*		
L1 #Tasks: 2	32,696	0	What method was used to capture this image? A. Magnetic resonance imaging, B. Ultra-wide field fundus photography, C. Color fundus photography, D. Optical coherence tomography.	All datasets
L2 #Tasks: 2	10,980	0	Which eye is shown in this image, the left or the right? A. Right eye, B. Left eye.	`[CFP]` DDR, DeepDRiD, IDRiD, OIA-ODIR, Retinal-Lesions `[UWF]` TOP `[CFP+UWF]` DeepDRiD
L3 #Tasks: 4 #Subtasks: 39	15,606	7,237	What are the positions of the Haemorrhages in the fundus image? A. Nasal side of the optic disc center, B. Temporal side of the optic disc center, C. Superior side of the optic disc center, D. Inferior side of the optic disc center, E. Not observed in the image.	`[CFP]` DDR, IDRiD, Retinal-Lesion `[OCT]` RETOUCH
L4 #Tasks: 4	20,177	5,114	Which abnormalities can be seen in this fundus image? A. Glaucoma, B. Diabetic retinopathy, C. No abnormality, D. Age-related macular degeneration, E. Hypertensive retinopathy.	`[CFP]` DDR, IDRiD, OIA-ODIR, JSIEC, RFMiD, Retinal-Lesions `[OCT]` NEH, OCTDL, OCTID, UCSD `[UWF]` TOP `[CFP+OCT]` MMC-AMD `[CFP+UWF]` DeepDRiD

(**L3a**), *localization* (**L3b**), *size estimation* (**L3c**) and *counting* (**L3d**).

• **Level 4 (L4): Disease diagnosis**. Disease diagnosis typically requires making judgments based on a comprehensive consideration of the lesions presented, changes in anatomical structures and overall appearance about the fundus. Such a requirement naturally places L4 skills at the highest level, which typically takes medical students years to master. Different from tasks from the previous levels, the L4 tasks specifically evaluate an MLLM's ability of integrating imaging findings with clinical knowledge for final diagnosis. For example, one of the diagnostic criteria for age-related macular degeneration (AMD) is the presence of *drusen* in the macula. This necessitates the ability to simultaneously identify both the macula (L2) and *drusen* (L3), as well as accurately determine their spatial relationship.

Data Sources In order to instantiate the above four-level tasks, we adapt the following 14 public datasets: 1) six `CFP` datasets: IDRiD [24], DDR [17], JSIEC [2], RFMiD [23], OIA-ODIR [16] and Retinal-Lesions [33], 2) five `OCT` datasets: OCTDL [12], NEH [27], OCTID [7], UCSD [11], and RETOUCH [1], 3) one `UWF` dataset: TOP[3], and 4) two multimodal datasets: MMC-AMD (`CFP+OCT`) [31] and DeepDRiD [21] (`CFP+UWF`).

Subject to their original annotations, the use of the datasets in specific tasks is listed in Table 1. Note that we take from each dataset its test split[4] to form FunBench, with the remaining part preserved as a development set for future usage, *e.g.* supervised fine-tuning. Provided with the four lesion-annotated datasets, *i.e.* DDR, IDRiD, Retinal-Lesions, and RETOUCH, we subdivide each of the

[3] `https://github.com/DateCazuki/Fundus_Diagnosis`

[4] In case no official data split is provided, we randomly select 20% of the dataset.

four L3 tasks by distinct lesions whenever applicable, resulting in 39 subtasks in total. Using the multiple disease-annotated datasets, we now instantiate the L4 skills with 4 concrete tasks, namely *binary-condition (normal or abnormal) diagnosis* (**L4a**), *multi-condition diagnosis* (**L4b**), *DR grading* (**L4c**) and *fine-grained AMD categorization* (**L4d**).

From Annotations to VQA Quadruples Similar to OmniMedVQA [9], we generate multi-choice VQA quadruples of (`image`, `question`, `options`, `answer`) from given images and their associated labels by auto-completing a number of predefined *task-specific* question templates. See samples in Table 1.

To direct the MLLM to select directly from the provided options, we prepend specific instructions to each question. For single-answer questions, the instruction is "*Please choose the most suitable option based on the image and the question. Answer with the option's letter directly*". For multi-answer questions, the instruction is "*Please choose all the suitable options based on the image and the question. Answer with the option's letter directly. Please separate the answers with commas if needed.*".

2.2 Targeted Evaluation Modes

In order to assess a given MLLM and its two key modules, *i.e.* `LLM` and `VE`, we present three targeted evaluation modes (E-mode) in a bottom-up manner.

E-Mode I: Linear-probe based VE Evaluation. To assess the effectiveness of the `VE` in extracting visual features from a given fundus image, we employ the widely used linear probe (LP) technique [29]. As illustrated in Fig. 1(b), LP trains a `Linear`-layer based classification head per (sub-)task using the task-specific development dataset. As such, we omit tasks that cannot be directly tackled as a classification problem, *e.g.* **L3b**, **L3c** and **L3d**, and tasks trivial for LP, *e.g.* **L1a** and **L1b**. Comparing `VE`s used by different MLLMs in this manner helps reveal which `VE` is more suited for fundus feature extraction.

E-Mode II: Knowledge-prompted LLM evaluation. As the `LLM` module has been re-trained to handle multimodal tokens, evaluating the module by directly submitting a textual question is problematic. To reduce the influence of the `VE`, we propose a simple knowledge-prompted evaluation strategy as follows. Given a test image and its associated task-specific label, we convert the label to an indirect description by querying an expert-knowledge database (EyeWiki). Such a description is further formatted in a task-specific manner. Consider **L2b** *laterality recognition* for instance. A left-eye image will be described as "*fovea located to the right side of the optic disc*". As for *hard exudate recognition*, one of the subtasks of **L3a**, the corresponding description will be "*white or yellowish deposits with sharp margins*". As shown in Fig. 1(c), by placing the description before the question, we perform knowledge-prompted LLM evaluation. Note that for this evaluation mode, some of the tasks, *e.g.* **L1**, **L2a** and **L4a**, will be omitted as the provided prompts would make the tasks trivial to accomplish.

E-Mode III: Holistic Evaluation. This mode offers an end-to-end evaluation of the MLLM, see Fig. 1(d). By submitting a multimodal multi-option

Table 2: **Open-source MLLMs evaluated**. Medical models are marked with ∗.

MLLM	HuggingFace release	VE	LLM
LLaVA-v1.5-7B [18]	2023.10	CLIP-ViT	Vicuna-7B
∗Qilin-Med-VL-Chat [20]	2023.12	CLIP-ViT	Chinese-LLaMA2
LLaVA-v1.6-7B [19]	2024.01	CLIP-ViT	Vicuna-7B
∗LLaVA-Med-v1.5-7B [14]	2024.05	CLIP-ViT	Mistral-7B
∗HuatuoGPT-Vision-7B [3]	2024.06	CLIP-ViT	Qwen2-7B
Qwen2-VL-7B [30]	2024.09	Qwen2-ViT	Qwen2-7B
InternVL2.5-8B [5]	2024.12	InternViT	InternLM2.5-7B
Janus-Pro-7B [4]	2025.01	ViT-SigLIP	DeepSeek-LLM-7B
Qwen2.5-VL-7B [28]	2025.01	Qwen2.5-ViT	Qwen2.5-7B

question to the model followed by a string comparison between the model's answer and the ground truth, a binary output is obtained. We found in preliminary experiments that some MLLMs, *e.g.* Qilin-Med-VL-Chat [20], LLaVA-Med-v1.5-7B [14] and Janus-Pro-7B [4], cannot follow our instruction that requires them to produce a single-character response. Instead, they tend to respond with more extensive, open-ended text. In order to select the option that best matches with such text, we perform text-to-text semantic matching by a pre-trained Sentence-BERT [26], which encodes a given sentence into a 384-d embedding vector.

Performance Metrics An AI-assisted disease diagnosis system naturally aims for fewer missed detections and false alarms, which can be measured by *Sensitivity* and *Specificity*, respectively. We report their harmonic mean, *a.k.a. F1* score, as a combined metric. For a multi-class task such as DR grading (**L4c**), a task-level F1 is computed as the mean value of F1 scores across all its classes. Moreover, the overall performance is averaged over the four levels, whilst the per-level performance is obtained by averaging over the (sub-)tasks and in a hierarchical manner if subtasks exist as L3. Such a performance calculation effectively removes any bias caused by imbalanced numbers of subtasks across different levels.

3 Evaluating MLLMs on FunBench

3.1 Choices of MLLMs

For reproducible research, we focus on open-source MLLMs. Subject to our GPU computation capability, we select MLLMs at about **7B**/**8B** scales, compiling a list of six general-purpose and three medical models, see Table 2. In addition, we include GPT-4o [10] as a proprietary baseline[5]. We adopt DINOv2-large [22], a strong vision foundation model, as a `VE` baseline.

3.2 Results

VE Comparison. The performance of the different `VE`s is shown in the `E-mode I` part of Table 3. DINOv2 is the best, though not used by the MLLMs. By

[5] API version: gpt-4o-2024-08-06.

Table 3: **Results on FunBench**. Per evaluation mode, best numbers per column are highlighted in bold.

Model	Overall performance					L1: Modality		L2: Anatomy		L3: Lesion				L4: Disease			
	MEAN	*L1*	*L2*	*L3*	*L4*	*L1a*	*L1b*	*L2a*	*L2b*	*L3a*	*L3b*	*L3c*	*L3d*	*L4a*	*L4b*	*L4c*	*L4d*
Random guess	0.393	0.313	0.500	0.361	0.397	0.250	0.375	0.500	0.500	0.468	0.458	0.250	0.269	0.500	0.398	0.319	0.372
`E-mode I`:																	
CLIP-ViT	0.578	-	0.820	0.373	0.541	-	-	-	0.820	0.373	-	-	-	0.814	0.175	0.556	0.621
Qwen2-ViT	0.597	-	0.892	0.338	0.560	-	-	-	0.892	0.338	-	-	-	0.814	0.225	0.506	**0.694**
InternViT	0.614	-	0.932	0.392	0.517	-	-	-	0.932	0.392	-	-	-	0.815	0.230	0.530	0.491
ViT-SigLIP	0.619	-	0.880	0.398	0.581	-	-	-	0.880	0.398	-	-	-	0.833	0.293	0.528	0.669
Qwen2.5-ViT	0.651	-	0.932	**0.449**	0.573	-	-	-	0.932	**0.449**	-	-	-	0.812	0.319	0.514	0.645
`DINOv2-large`	**0.655**	-	**0.939**	0.408	**0.616**	-	-	-	**0.939**	0.408	-	-	-	**0.858**	**0.371**	**0.568**	0.669
`E-mode II`:																	
Janus-Pro	0.409	-	**0.560**	0.213	0.454	-	-	-	**0.560**	0.213	-	-	-	-	0.517	0.446	0.401
Qilin-Med-VL	0.473	-	0.473	0.519	0.426	-	-	-	0.473	0.519	-	-	-	-	0.552	0.391	0.334
LLaVA-v1.5	0.487	-	0.386	0.560	0.515	-	-	-	0.386	0.560	-	-	-	-	0.521	0.557	0.467
LLaVA-Med-v1.5	0.529	-	0.483	0.507	0.597	-	-	-	0.483	0.507	-	-	-	-	0.638	0.640	0.513
LLaVA-v1.6	0.604	-	0.502	0.717	0.594	-	-	-	0.502	0.717	-	-	-	-	0.606	0.647	0.529
Qwen2.5-VL	0.654	-	0.497	0.735	0.729	-	-	-	0.497	0.735	-	-	-	-	0.757	0.726	0.704
Qwen2-VL	0.673	-	0.488	0.747	0.783	-	-	-	0.488	0.747	-	-	-	-	0.750	0.855	0.743
HuatuoGPT-V	0.692	-	0.541	0.706	**0.829**	-	-	-	0.541	0.706	-	-	-	-	**0.793**	0.927	**0.768**
InternVL2.5	**0.703**	-	0.488	**0.831**	0.789	-	-	-	0.488	**0.831**	-	-	-	-	0.792	**0.929**	0.647
`E-mode III`:																	
Janus-Pro	0.355	0.777	0.167	0.108	0.369	0.943	0.611	0.189	0.145	0.115	0.203	0.060	0.055	0.523	0.319	0.278	0.358
LLaVA-v1.5	0.418	0.722	0.489	0.165	0.296	0.974	0.470	0.492	0.485	0.276	0.266	0.013	0.104	0.364	0.292	0.187	0.341
Qilin-Med-VL	0.426	0.651	0.496	0.211	0.347	0.910	0.392	0.504	0.489	0.463	0.361	0.000	0.019	0.491	0.354	0.214	0.328
LLaVA-v1.6	0.435	0.777	0.449	0.194	0.319	0.970	0.585	0.400	0.497	0.461	0.243	0.058	0.013	0.357	0.314	0.289	0.315
LLaVA-Med-v1.5	0.440	0.813	0.507	0.215	0.223	0.961	0.665	0.502	0.512	0.166	0.349	0.261	0.086	0.000	0.255	0.290	0.345
Qwen2.5-VL	0.480	0.942	0.364	0.231	0.384	0.978	0.906	0.310	0.417	0.386	0.242	0.218	0.077	0.532	0.394	0.303	0.308
InternVL2.5	0.506	0.946	**0.564**	0.244	0.269	**0.997**	0.894	**0.640**	0.488	0.391	0.144	0.227	0.213	0.031	0.400	0.294	0.351
Qwen2-VL	0.520	0.925	0.497	0.316	0.343	0.995	0.855	0.506	0.488	0.503	**0.432**	0.073	**0.255**	0.508	0.352	0.202	0.311
HuatuoGPT-V	0.523	0.933	0.309	**0.374**	**0.477**	0.961	0.905	0.140	0.477	**0.548**	0.379	0.338	0.231	**0.622**	**0.504**	0.414	0.369
`GPT-4o`	**0.542**	**0.961**	0.535	0.341	0.331	0.965	**0.957**	0.557	**0.514**	0.362	0.412	**0.361**	0.228	0.018	0.452	**0.476**	**0.378**

contrast, the most popular CLIP-ViT, with mean score of 0.578, has turned out to be the least effective. Checking its performance per task, we see the largest performance gap at **L4b** multi-condition diagnosis, 0.175 *versus* 0.319 (from Qwen2.5-ViT). Recall that the CLIP series were pre-trained on large-scale web data for image-text semantic matching. Hence, the CLIP features might lack fine-grained details required for discriminating dozens of fundus diseases which typically bear large inter-class similarities. Indeed we notice that for many diseases at the long tail, the LP-based classifiers built on top of the varied `VE`s fail to recognize them, yielding *Sensitivity* of 0 and consequently zero *F1* score. As such, the performance of the `VE`s is even worse than chance on **L4b**. The result suggests the limitation of pure-vision solutions for fundus image analysis.

LLM Comparison. The `LLM` result is shown in the `E-mode II` part of Table 3. The superior performance of InternVL2.5, HuatuoGPT-V and Qwen against the `VE` counterpart suggests that their `LLM`s possess certain ophthalmic knowledge pertinent to fundus reading. Also notice how their performance varies over tasks, see for instance **L4c** *DR grading* and **4d** *AMD categorization*. The `LLM`s perform clearly better on **L4c**. Our hypothesis is compared to AMD, DR-related materials are abundant online, making the `LLM`s more "familiar" with DR. Another empirical evidence supporting this hypothesis is `LLM`s' near-to-chance performance on

Table 4: **Correlation analysis in terms of MEAN-performance ranks**.

Module	Spearman-correlation to MLLM
LLM	0.917
VE	0.055

L2b *laterality recognition*. Such a skill is too basic to be discussed, making the related training data rare, and consequently making it a "novel" task for the big models. The results suggest a fundamental limitation of the current data-driven paradigm: it produces "powerful" models that lack basic fundus reading skills.

Our setup also supports a decomposition study that evaluates different `LLM`s with the same VE, see the following five models in Tab. 2, *i.e.* LLaVA-v1.5, Qilin-Med-VL-Chat, LLaVA-v1.6, LLaVA-Med-v1.5 and HuatuoGPT-Vision, all using the same-weights CLIP-ViT as their `VE`. HuatuoGPT-Vision is the best, probably due to its targeted fine-tuning on medical data, particularly excelling in lesion analysis and disease diagnosis.

MLLM Comparison. The performance of the MLLMs is summarized in the last part of Table 3. Among the open-source models, HuatuoGPT-Vision is the best, followed by Qwen2-VL and InternVL2.5. Note that HuatuoGPT-Vision and Qwen2-VL adopt `LLM` of the same structure (Qwen2-7B), yet the former's `VE` (CLIP-ViT) is shown to be less effective than that of the latter (Qwen2-ViT). Such a difference shows the importance of domain-specific fine-tuning. Note that the relatively inferior performance of HuatuoGPT-Vision on localization-related tasks, see **L2a** and **L3b**. Based on our evaluation, we believe that its performance is likely to be improved when a stronger `VE` is used.

As shown in Table 4, the high rank correlation between MLLMs and their `LLM`s clearly suggests that the former heavily rely on the latter for performing the fundus reading tasks. While both `VE` and `LLM` are important, the correlation analysis underscores the urgent need of developing a strong ophthalmic `LLM`.

In general, the MLLMs evaluated lack fundus reading skills. While the models perform reasonably well on L1, their effectiveness on L2 is close to random guess. Given that the L2 tasks are quite basic, this deficiency clearly indicates the lack of related ophthalmic anatomy knowledge within the `LLM`s. Model performance on L3 and L4 is also rather limited. Moreover, the relatively lower performance under E-mode III than E-mode II suggests that the `LLM`s possess ophthalmic knowledge to some extent, but fail to correctly interpret visual features presented in the input image. For more evaluations, please refer to the FunBench website.

4 Conclusions

Our evaluation of varied MLLMs on the new FunBench benchmark supports conclusions as follows. First, the MLLMs evaluated remain weak for performing fundus reading tasks related to anatomy perception, lesion analysis and disease diagnosis. Second, they rely much more on their LLMs rather than their VEs.

Third, the VEs are less effective than DINOv2. Lastly, the overall best performance of HuatuoGPT-Vision shows the importance of domain-specific training. The future design of the training procedure needs to consider the big picture, or we risk developing an MLLM that lacks basic fundus reading skills.

Acknowledgments. This research was supported by Beijing Natural Science Foundation (L254039) and NSFC (62172420).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bogunović, H., Venhuizen, F., et al.: RETOUCH: The retinal OCT fluid detection and segmentation benchmark and challenge. TMI **38**(8), 1858–1874 (2019)
2. Cen, L.P., Ji, J., Lin, J.W., Ju, S.T., Lin, H.J., Li, T.P., Wang, Y., Yang, J.F., Liu, Y.F., Tan, S., et al.: Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. NComms. **12**(1), 4828 (2021)
3. Chen, J., Gui, C., et al.: HuatuoGPT-Vision, towards injecting medical visual knowledge into multimodal LLMs at scale. arXiv (2024)
4. Chen, X., Wu, Z., et al.: Janus-Pro: Unified multimodal understanding and generation with data and model scaling. arXiv (2025)
5. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv (2024)
6. Deng, Z., Gao, W., Chen, C., Niu, Z., Gong, Z., Zhang, R., Cao, Z., Li, F., Ma, Z., Wei, W., et al.: OphGLM: An ophthalmology large language-and-vision assistant. Artificial Intelligence in Medicine **157**, 103001 (2024)
7. Gholami, P., Roy, P., Parthasarathy, M.K., Lakshminarayanan, V.: OCTID: Optical coherence tomography image database. Computers & Electrical Engineering **81**, 106532 (2020)
8. He, R., Xu, M., Das, A., Khan, D.Z., Bano, S., Marcus, H.J., Stoyanov, D., Clarkson, M.J., Islam, M.: PitVQA: Image-grounded text embedding llm for visual question answering in pituitary surgery. In: MICCAI (2024)
9. Hu, Y., Li, T., et al.: OmniMedVQA: A new large-scale comprehensive evaluation benchmark for medical LVLM. In: CVPR (2024)
10. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: GPT-4o system card. arXiv (2024)
11. Kermany, D.S., Goldbaum, M., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. cell **172**(5), 1122–1131 (2018)
12. Kulyabin, M., Zhdanov, A., et al.: OCTDL: Optical coherence tomography dataset for image-based deep learning methods. Scientific data **11**(1), 365 (2024)
13. Lai, X., Li, X., Qian, R., Ding, D., Wu, J., Xu, J.: Four models for automatic recognition of left and right eye in fundus images. In: MMM (2019)
14. Li, C., Wong, C., et al.: LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. NeurIPS (2024)
15. Li, J., Guan, Z., et al.: Integrated image-based deep learning and language models for primary diabetes care. Nature medicine **30**(10), 2886–2896 (2024)

16. Li, N., Li, T., Hu, C., Wang, K., Kang, H.: A benchmark of ocular disease intelligent recognition: One shot for multi-disease detection. In: BMO (2021)
17. Li, T., Gao, Y., Wang, K., Guo, S., Liu, H., Kang, H.: Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. Information Sciences **501**, 511–522 (2019)
18. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: CVPR (2024)
19. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: LLaVA-NeXT: Improved reasoning, OCR, and world knowledge (2024), `https://llava-vl.github.io/blog/2024-01-30-llava-next/`
20. Liu, J., Wang, Z., Ye, Q., Chong, D., Zhou, P., Hua, Y.: Qilin-Med-VL: Towards Chinese large vision-language model for general healthcare. arXiv (2023)
21. Liu, R., Wang, X., Wu, Q., Dai, L., Fang, X., Yan, T., Son, J., Tang, S., Li, J., Gao, Z., et al.: DeepDRiD: Diabetic retinopathy—grading and image quality estimation challenge. Patterns **3**(6) (2022)
22. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research Journal pp. 1–31 (2024)
23. Pachade, S., Porwal, P., Thulkar, D., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Giancardo, L., Quellec, G., Mériaudeau, F.: Retinal fundus multi-disease image dataset (RFMiD): A dataset for multi-disease detection research. Data **6**(2), 14 (2021)
24. Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., Meriaudeau, F.: Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. Data **3**(3), 25 (2018)
25. Qin, Z., Yin, Y., Campbell, D., Wu, X., Zou, K., Tham, Y.C., Liu, N., Zhang, X., Chen, Q.: LMOD: A large multimodal ophthalmology dataset and benchmark for large vision-language models. In: NAACL (2025)
26. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: EMNLP (2019)
27. Sotoudeh-Paima, S., Jodeiri, A., Hajizadeh, F., Soltanian-Zadeh, H.: Multi-scale convolutional neural network for automated AMD classification using retinal OCT images. Computers in biology and medicine **144**, 105368 (2022)
28. Team, Q.: Qwen2.5-vl (2025), `https://qwenlm.github.io/blog/qwen2.5-vl/`
29. Tu, C.H., Mai, Z., Chao, W.L.: Visual query tuning: Towards effective usage of intermediate representations for parameter and memory efficient transfer learning. In: CVPR (2023)
30. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. arXiv (2024)
31. Wang, W., Li, X., Xu, Z., Yu, W., Zhao, J., Ding, D., Chen, Y.: Learning two-stream CNN for multi-modal age-related macular degeneration categorization. IEEE Journal of Biomedical and Health Informatics **26**(8), 4111–4122 (2022)
32. Wang, Z., Jiang, X., Gao, C., Dong, F., Dai, W., Wang, B., Yan, B., Chen, Q., Huang, W., Zhang, T., et al.: Eyegraphgpt: Knowledge graph enhanced multimodal large language model for ophthalmic report generation. In: BIBM (2024)
33. Wei, Q., Li, X., Yu, W., Zhang, X., Zhang, Y., Hu, B., Mo, B., Gong, D., Chen, N., Ding, D., et al.: Learn to segment retinal lesions and beyond. In: ICPR (2021)
34. Yang, Z., Li, X., He, X., Ding, D., Wang, Y., Dai, F., Jin, X.: Joint localization of optic disc and fovea in ultra-widefield fundus images. In: MLMI@MICCAI (2019)

35. Ye, J., Wang, G., Li, Y., Deng, Z., Li, W., Li, T., Duan, H., Huang, Z., Su, Y., Wang, B., et al.: GMAI-MMBench: A comprehensive multimodal evaluation benchmark towards general medical AI. In: NeurIPS (2024)
36. Yeh, C.H., Wang, J., Graham, A.D., Liu, A.J., Tan, B., Chen, Y., Ma, Y., Lin, M.C.: Insight: A multi-modal diagnostic pipeline using llms for ocular surface disease diagnosis. In: MICCAI (2024)
37. Zhang, T., Lin, M., Guo, H., Zhang, X., Chiu, K.F.P., Feragen, A., Dou, Q.: Incorporating clinical guidelines through adapting multi-modal large language model for prostate cancer PI-RADS scoring. In: MICCAI (2024)