CAN MULTIMODAL LARGE LANGUAGE MODELS UNDERSTAND OCT?

Anonymous authors
Paper under double-blind review

ABSTRACT

Optical coherence tomography (OCT) is essential for the diagnosis and treatment of retinal diseases. Although multimodal large language models (MLLMs) have shown strong potential in medical image analysis, existing benchmarks mainly focus on coarse-grained disease classification or visual question answering, failing to comprehensively evaluate the cognitive process from visual perception to clinical reasoning. To address this gap, we present **OCT-Bench**, a comprehensive benchmark for OCT image understanding. OCT-Bench contains 10,076 high-quality multiple-choice questions derived from 4,137 OCT images across seven public datasets. Following the clinical interpretation workflow, we establish a hierarchical taxonomy of 20 fine-grained tasks spanning three capability dimensions: Perception, Cognition, and Reasoning. These tasks cover imaging attributes, retinal anatomy, lesion characteristics, spatial relationships, disease assessment, treatment decision-making, and prognosis. We evaluate 20 representative MLLMs, including close-source, open-source, and medical-domain models. Results show that current models remain far from reliable OCT understanding, while neither medical-domain adaptation nor larger model scale consistently improves performance across capability levels. OCT-Bench provides a comprehensive benchmark for evaluating OCT understanding and facilitates the development of clinically grounded MLLMs.

1 INTRODUCTION

Multimodal large language models (MLLMs) have recently achieved remarkable progress on general-purpose vision tasks Wu et al. (2024); Caffagni et al. (2024), driven by increasingly powerful visual understanding and cross-modal reasoning capabilities. This progress has also revealed considerable potential for assisting medical image analysis. Unlike natural-image understanding, however, medical image interpretation requires more than recognizing visual patterns Bai et al. (2025b); Lin et al. (2025a): models must integrate specialized medical knowledge to analyze disease, form clinical judgments, and support decision-making. This raises a fundamental question: *can current MLLMs genuinely understand complex medical images and complete the full process from visual perception to clinical reasoning?*

Optical coherence tomography (OCT), one of the most important imaging techniques in ophthalmic practice Huang et al. (1991), provides high-resolution cross-sectional views of retinal structures and is widely used for disease diagnosis, treatment planning, and longitudinal follow-up Gurumoorthy et al. (2025); Fang et al. (2025). OCT interpretation nevertheless presents a substantial technical and clinical challenge because scans contain complex layered anatomy, subtle variations in tissue reflectivity, and frequently coexisting abnormalities. In routine practice, clinicians first assess image quality and identify retinal anatomy, then examine lesion morphology, spatial relationships, and pathological characteristics, and finally integrate these findings with clinical knowledge to establish a diagnosis, determine treatment, and evaluate prognosis Wang et al. (2024); Chen et al. (2024a). This progression from low-level visual perception to high-level clinical reasoning makes OCT a particularly demanding test bed for evaluating medical image understanding in MLLMs.

Existing benchmarks, however, remain insufficient for comprehensively evaluating OCT understanding in two important respects. First, general-purpose multimodal benchmarks predominantly focus on natural images and generic visual capabilities Jiang et al. (2026); Fu et al. (2026),

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

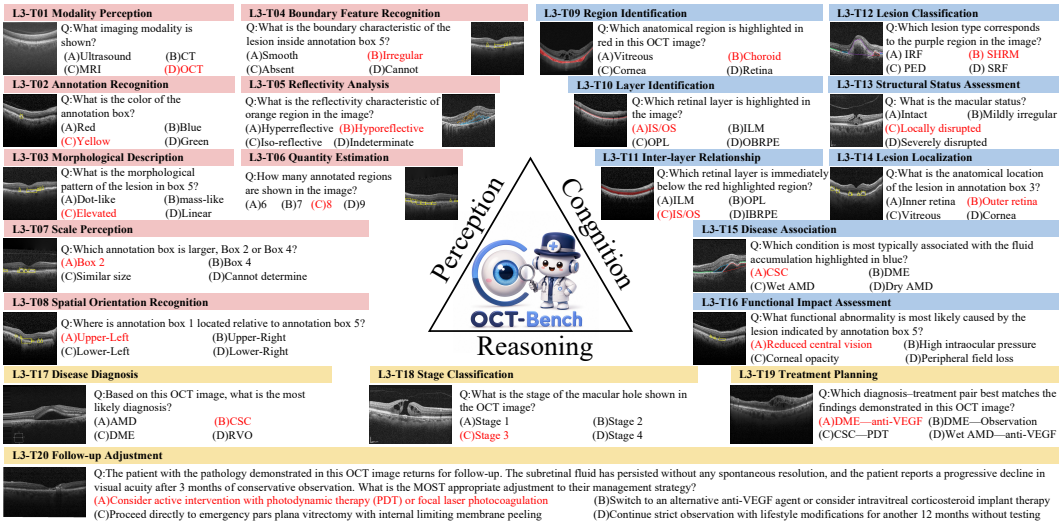


Figure 1: Overview of OCT-Bench.

offering little systematic evaluation of medical images and OCT in particular. Existing ophthalmic benchmarks Zou et al. (2025); Srinivasan et al. (2026) primarily address textual medical knowledge, fundus imagery, retinal image enhancement, or ophthalmic surgical videos. Although LMOD includes OCT among several ophthalmic modalities Qin et al. (2025), its evaluation remains limited to a small number of coarse-grained tasks and cannot capture the diverse capabilities required for clinical OCT interpretation.



Figure 2: Task taxonomy of OCT-Bench.

Second, and more importantly, existing benchmarks commonly reduce OCT understanding to disease classification or isolated visual question answering, overlooking the hierarchical cognitive process underlying real-world interpretation Chen et al. (2026). OCT analysis does not proceed directly from an image to a diagnosis; rather, it follows a progressive pathway from visual perception through medical cognition to clinical reasoning. Existing evaluations often conflate these levels within a single task. When a model makes an incorrect prediction, it is therefore difficult to determine whether the failure arises from inadequate visual perception, deficient medical understanding, or unreliable clinical reasoningZhou et al. (2026). This entanglement not only hinders fair and informative model comparison but also provides limited guidance for targeted model improvement.

To address these limitations, we introduce OCT-Bench (Figure 1), a comprehensive benchmark for evaluating MLLMs on OCT image understanding. Rather than treating OCT interpretation as a conventional disease-classification problem, OCT-Bench follows the clinical interpretation workflow and establishes a hierarchical taxonomy comprising three primary dimensions, Perception, Cognition, and Reasoning, together with nine capability groups and 20 fine-grained tasks, as illustrated in Figure 2. Built from 4,137 carefully curated and quality-controlled OCT images collected from seven public datasets, OCT-Bench contains 10,076 expert-verified high-quality multiple-choice questions covering imaging quality, retinal anatomy, lesion characteristics, spatial relationships, disease assessment, therapeutic decision-making, and prognostic management. Compared with existing benchmarks (Table 1), OCT-Bench supports both comprehensive assessment of overall performance

Table 1: Comparison of OCT-Bench with existing general-domain and ophthalmology-specific benchmarks.

Benchmarks	OCT	Multimodal	Disease Classification	Anatomy Recognition	Lesion Recognition	Visual Perception	Medical Cognition	Clinical Reasoning	Tasks
General-Domain Benchmarks									
MMBench	✗	✓	✗	✗	✗	✓	✗	✗	20
MME-RealWorld	✗	✓	✗	✗	✗	✓	✗	✗	43
UNK-VQA	✗	✓	✗	✗	✗	✓	✗	✗	5
MMCBench	✗	✓	✗	✗	✗	✓	✗	✗	4
MathVista	✗	✓	✗	✗	✗	✓	✗	✗	6
SEED-Bench	✗	✓	✗	✗	✗	✓	✗	✗	12
Ophthalmology-Specific Benchmarks									
Eval-GPT-Ophth	✗	✗	✗	✗	✗	✗	✓	✗	13
Bench-Myopia	✗	✗	✗	✗	✗	✗	✓	✓	6
EyeBench	✗	✗	✗	✗	✓(2)	✓	✗	✗	7
OphNet	✗	✓	✗	✗	✗	✗	✓	✗	4
LMOD	✓	✓	✓(2)	✓(2)	✓(2)	✗	✗	✗	3
OCT-Bench (Ours)	✓	✓	✓**(10)**	✓**(7)**	✓**(10)**	✓	✓	✓	**20**

and precise diagnosis of capability bottlenecks across different stages, revealing how errors may propagate from low-level visual perception to high-level clinical reasoning.

Based on OCT-Bench, we systematically evaluate 20 representative MLLMs, including proprietary models, open-source general-purpose models, and medical-domain models. Our results demonstrate that current MLLMs remain far from reliable OCT understanding. The best-performing model achieves only 62.0% overall accuracy, and performance declines markedly as the required capability advances: the highest perception score reaches 75.8%, whereas the highest reasoning score is only 42.9%. Moreover, neither medical-domain adaptation nor increased model scale yields consistent improvements across capability levels. These results indicate that current MLLMs still struggle to perform reliable clinical reasoning grounded in OCT images.

The main contributions are summarized as follows:

- We introduce OCT-Bench, a large-scale benchmark dedicated to OCT image understanding, containing 10,076 questions over 4,137 images from seven public datasets and enabling comprehensive assessment beyond coarse disease classification.
- We develop a clinically grounded hierarchical taxonomy that decomposes OCT understanding into three primary dimensions, nine capability groups, and 20 fine-grained tasks, allowing capability deficiencies at different stages to be precisely identified and analyzed.
- We systematically evaluate 20 representative MLLMs and reveal substantial performance degradation from visual perception to clinical reasoning, while providing a fine-grained analysis of the strengths and limitations of different model families.

2 Related Work

2.1 Multimodal Large Language Models

Multimodal large language models (MLLMs) integrate visual encoders with large language models to enable image-grounded understanding and reasoning. Recent proprietary models, such as GPT-4o, Gemini, and Grok, have demonstrated strong multimodal capabilities, while open-source models, including LLaVA, Phi-3-Vision, InternVL, Qwen2.5-VL, and Gemma 3, have substantially advanced multimodal research through improved reproducibility and transparency Liu et al. (2024a); Li et al. (2024); Bilenko (2024); Chen et al. (2024b); Bai et al. (2025a); Team et al. (2024). Medical MLLMs, such as HealthGPT, MedGemma, Lingshu, Hulu-Med, MediX-R1, and Fleming-VL, further leverage biomedical knowledge and medical image-text data to improve medical image understanding and support clinical reasoning. However, most existing MLLMs are trained on natural images or broad medical corpora, whereas OCT interpretation requires precise perception of fine-grained retinal structures, subtle pathological features, and their clinical significance. Whether current proprietary, open-source, and medical MLLMs can effectively understand and reason over OCT images therefore remains an open question.

2.2 Ophthalmic Evaluation Benchmarks

In recent years, general-purpose multimodal benchmarks, such as MMBench, MME-RealWorld, MathVista, and SEED-Bench, have substantially advanced the evaluation of multimodal large language models (MLLMs). However, they are primarily designed for natural scenes and cannot effectively assess the understanding of ophthalmic anatomy, retinal lesions, and their clinical significance Liu et al. (2024b); Zhang et al. (2025); Lu et al. (2024); Li et al. (2023). Existing ophthalmology benchmarks mainly focus on medical knowledge, retinal image enhancement, surgical video understanding, or multimodal ophthalmic tasks Antaki et al. (2023); Lim et al. (2023); Zhu et al. (2025); Hu et al. (2024a). Nevertheless, a systematic benchmark dedicated to OCT image understanding remains unavailable. As a key imaging modality for retinal disease diagnosis, OCT requires models to recognize fine-grained retinal structures and lesions while integrating medical knowledge for clinical reasoning. To fill this gap, OCT-Bench systematically evaluates MLLMs on OCT image understanding across three progressive levels: visual perception, medical cognition, and clinical reasoning.

3 OCT-Bench

3.1 Hierarchical Capability Taxonomy

OCT-Bench models OCT understanding as a progressive process from visual perception to medical cognition and clinical reasoning. As shown in Figure 2, we establish a hierarchical taxonomy with three dimensions, nine capability groups, and 20 fine-grained tasks, following the clinical workflow of OCT interpretation: perceiving visual evidence, linking it to anatomical and pathological concepts, and supporting clinical decision-making.

3.1.1 Perception

The Perception dimension evaluates whether a model can accurately extract visual information from OCT images, which serves as the foundation for subsequent medical reasoning. This dimension encompasses fundamental image understanding, morphological feature perception, and spatial relationship perception, including tasks such as imaging attribute recognition, retinal structure recognition, reflectivity feature recognition, quantitative information understanding, and spatial relationship analysis. These tasks assess whether MLLMs can accurately capture fine-grained visual evidence that supports medical understanding.

3.1.2 Cognition

The Cognition dimension evaluates a model's ability to transform visual information into medical knowledge, with a focus on anatomical understanding, pathological recognition, and clinical association. Models are required to understand retinal anatomical structures and their relationships, identify and localize pathological abnormalities, and establish links between imaging manifestations, disease status, and functional impact. This dimension connects visual perception with clinical understanding and constitutes a critical stage of medical reasoning.

3.1.3 Reasoning

The Reasoning dimension evaluates whether a model can integrate multiple imaging findings with medical knowledge to support clinical decision-making. It encompasses disease reasoning, therapeutic decision-making, and prognostic management, requiring models to synthesize anatomical and pathological evidence, assess disease status, and further infer appropriate treatment and follow-up strategies. By explicitly separating reasoning from perception and cognition, OCT-Bench enables the identification of failures arising from insufficient visual grounding versus those caused by higher-level clinical inference.

3.2 Benchmark Construction

To construct a high-quality benchmark for OCT image understanding, we design a systematic five-stage pipeline, as illustrated in Figure 3.

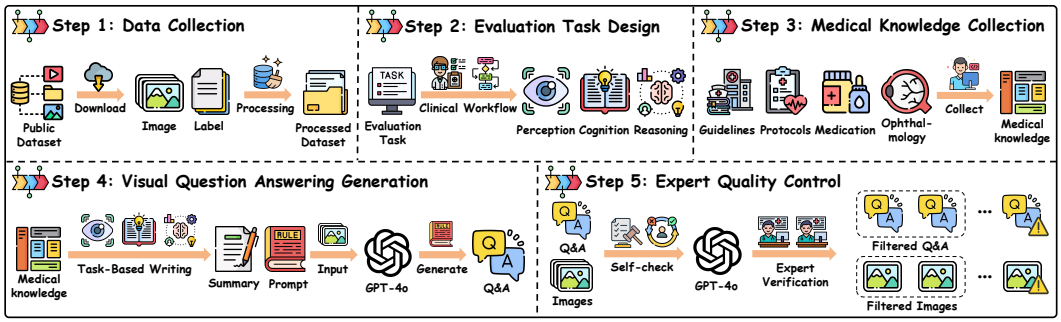


Figure 3: Construction pipeline of OCT-Bench. The pipeline comprises data collection, evaluation task design, medical knowledge collection, task-guided VQA generation, and expert quality control.

Step 1: Data Collection.

We collect OCT images from seven public datasets: OCT5k Arikan et al. (2025), OIMHS Ye et al. (2023), OCT-C8 Subramanian et al. (2022), AMD-SD Hu et al. (2024b), OCTDL Kulyabin et al. (2024), MMC-AMD Wang et al. (2022), and GOALS Fang et al. (2022). Since these datasets contain heterogeneous annotations, we establish a unified annotation schema and standardize different types of labels, including disease categories, anatomical regions, lesion types, bounding boxes, segmentation masks, and clinical attributes. Specifically, we map semantically equivalent labels into consistent categories, normalize annotation formats, and align spatial annotations with OCT images. The unified dataset covers diverse diseases, anatomical structures, and lesion information, supporting multi-level OCT understanding.

Step 2: Evaluation Task Design.

Following the progressive clinical OCT interpretation workflow from visual perception, medical cognition, to clinical reasoning, we establish a three-level capability evaluation framework consisting of Perception, Cognition, and Reasoning. This hierarchical design goes beyond traditional disease classification or visual question answering tasks by systematically evaluating the complete capability chain from image understanding to clinical decision-making and identifying stage-specific limitations. We further divide these three levels into nine capability groups and 20 fine-grained tasks, with each task defined according to its objective and knowledge boundary, enabling a comprehensive and precise evaluation of MLLMs for OCT image understanding.

Step 3: Medical Knowledge Collection.

For cognition and reasoning tasks, we collect medical knowledge from evidence-based guidelines, expert consensus statements, and authoritative ophthalmic references from organizations such as the American Academy of Ophthalmology (AAO) and Chinese Medical Association (CMA), along with OCT reference books Xun & Xiaoxin (2023); Vemulakonda et al. (2025); Lim et al. (2025); Kim et al. (2025); Kovach et al. (2025); Duker et al. (2021). These materials are organized into task-specific knowledge constraints, including disease characteristics, anatomical relationships, lesion interpretations, and clinical management principles. These constraints provide medical guidance for question generation, ensuring that generated questions are clinically relevant, visually grounded, and consistent with established ophthalmic knowledge.

Step 4: Visual Question Answering Generation. We design dedicated generation instructions for each task and provide GPT-4o Hurst et al. (2024) with the task description, OCT image, annotation information, and relevant medical knowledge to generate four-option multiple-choice questions. This task-driven strategy aligns each question with its target capability and produces candidate VQA samples covering visual attributes, anatomical structures, lesion characteristics, disease status, therapeutic decisions, and prognostic management.

Step 5: Expert Quality Control.

We adopt a rigorous two-stage quality-control strategy to ensure the reliability and quality of OCT-Bench. Initially, 12,895 candidate multiple-choice questions are generated through the task-guided VQA generation process. First, GPT-4o automatically filters potential errors based on question qual-

ity, answer uniqueness, and image–text consistency. Then, three experienced ophthalmologists independently review each sample from multiple perspectives, including medical correctness, visual answerability, task relevance, and ambiguity. Only questions with consistent judgments from all three experts are retained, while inconsistent or problematic samples are removed or revised. After this process, 10,076 high-quality, clinically consistent, and expert-verified multiple-choice questions are retained to form the final OCT-Bench dataset.

3.3 Data Analysis

We analyze the constructed benchmark from two complementary perspectives: coverage and reliability. In terms of coverage, OCT-Bench spans clinically relevant OCT understanding across 10 disease categories, 2 types of region recognition, 5 types of retinal layer recognition, and 10 types of lesion recognition. We also inspect distributions across tasks, diseases, anatomical regions, retinal layers, and lesion types to reduce label concentration and annotation artifacts. Detailed statistics and distribution analyses are provided in the appendix.

In terms of reliability, we further evaluate the quality of the finalized benchmark from multiple aspects. We examine whether each question is supported by sufficient visual evidence, whether the answer is uniquely determined, and whether the difficulty level aligns with the intended capability. We also analyze the proportion of ambiguous or invalid samples and verify the consistency between questions and their corresponding annotations. These analyses demonstrate the reliability and clinical validity of OCT-Bench.

4 Experiment

4.1 Evaluation Strategy

All OCT-Bench tasks are formulated as multiple-choice questions (MCQs) to enable standardized, objective, and reproducible evaluation. Each question contains four options (A–D) with a single correct answer. Compared with open-ended responses, MCQ format reduces evaluation ambiguity caused by diverse language expressions and avoids the influence of output variations across models. Given an OCT image and question, models are required to output only the selected option, which is compared with the ground truth. Invalid responses are counted as incorrect. Accuracy is adopted as the primary metric for fair comparison of MLLMs across different tasks and capability levels.

4.2 Experimental Setup

We evaluate 20 representative MLLMs on OCT-Bench, covering proprietary, general-purpose, and medical-domain models. Proprietary models include GPT-5.4-mini, Gemini-2.5-flash Comanici et al. (2025), and Grok-4-fast. General-purpose open-source models include Phi-3-Vision-128K Bilenko (2024); Abdin et al. (2024), InternVL2.5 series Chen et al. (2024b), Qwen2.5-VL series Bai et al. (2025a), Gemma-3-12B Team et al. (2024), and LLaVA series Liu et al. (2024a); Li et al. (2024). Medical-domain models include HealthGPT-M3 Lin et al. (2025b), MedGemma-4B Sellergren et al. (2025), Lingshu series Xu et al. (2025), Hulu-Med series Jiang et al. (2025), MediX-R1-8B Mullappilly et al. (2026), Fleming-VL-8B Shu et al. (2025), and HealthGPT-Pro-8B. This selection enables comprehensive comparison across general and specialized MLLMs. All models are evaluated under a unified zero-shot setting without fine-tuning or in-context demonstrations.

Table 2: Overall and dimension-average performance on OCT-Bench (%). The best and second-best results in each column are highlighted in bold and underlined, respectively.

Model	Overall	Perception	Cognition	Reasoning
Closed-Source Models				
GPT-5.4-mini	**62.01**	**75.82**	<u>59.98</u>	38.45
Gemini-2.5-flash	<u>60.46</u>	67.74	**64.22**	38.40
Grok-4-fast	52.42	63.48	51.44	32.23
Open-Source Models				
Phi-3-Vision-128k	42.05	60.95	32.18	23.99
InternVL2.5-4B	46.62	61.19	42.29	26.17
Qwen2.5-VL-7B	49.49	66.74	42.66	28.65
Gemma-3-12B	44.73	62.62	36.56	25.29
LLaVA-1.5-13B	39.60	54.21	28.76	32.07
InternVL2.5-26B	54.54	<u>72.46</u>	49.36	29.05
Qwen2.5-VL-32B	52.44	65.98	48.62	33.00
LLaVA-1.6-34B	50.53	65.31	46.22	29.60
Medical-Domain Models				
HealthGPT-M3	45.07	61.55	38.06	26.16
MedGemma-4B	42.18	55.28	35.18	29.98
Lingshu-7B	51.53	64.47	44.76	39.18
Hulu-Med-7B	54.13	65.97	50.25	38.22
MediX-R1-8B	51.68	65.11	48.79	30.59
Fleming-VL-8B	53.62	67.84	48.76	34.93
HealthGPT-Pro-8B	55.58	72.40	51.61	29.88
Lingshu-32B	56.58	62.53	58.75	<u>40.35</u>
Hulu-Med-32B	58.40	67.49	57.07	**42.89**

Table 3: Performance of different models on tasks T01–T20 (%).

Model	Perception								Cognition								Reasoning			
	A1		A2			A3			B1			B2			B3		C1		C2	C3
	T01	T02	T03	T04	T05	T06	T07	T08	T09	T10	T11	T12	T13	T14	T15	T16	T17	T18	T19	T20
Closed-Source Models																				
GPT-5.4-mini	**100.00**	98.38	43.31	69.94	**61.10**	**90.70**	**67.05**	**76.10**	61.17	**56.79**	46.89	35.01	**81.18**	**92.06**	34.66	72.09	38.42	30.68	40.87	43.83
Gemini-2.5-flash	**100.00**	99.54	32.56	60.13	57.96	87.79	49.42	54.52	92.42	37.26	44.69	37.87	78.71	85.98	52.15	**84.65**	19.31	**34.00**	49.36	**50.94**
Grok-4-fast	**100.00**	95.36	33.14	73.31	54.22	62.50	42.46	46.87	51.89	30.53	35.62	37.73	66.92	81.54	**58.01**	49.30	23.75	**34.00**	29.56	41.62
Open-Source Models																				
Phi-3-Vision-128k	95.41	97.91	9.01	74.28	36.15	54.65	60.79	59.40	40.91	23.03	36.01	28.89	62.36	12.15	45.97	8.14	17.18	32.45	19.79	26.54
InternVL2.5-4B	**100.00**	97.45	24.13	55.95	38.51	77.62	54.06	41.76	70.45	30.27	32.64	36.89	50.19	26.17	47.75	43.95	24.52	33.11	21.85	25.20
Qwen2.5-VL-7B	98.06	**99.77**	43.31	43.09	46.56	90.12	44.08	68.91	44.89	27.68	34.97	31.84	58.75	63.55	42.62	36.98	23.36	32.23	30.33	28.69
Gemma-3-12B	99.47	**99.77**	25.58	81.35	48.33	58.72	40.14	47.56	51.70	25.49	29.15	34.22	40.87	42.99	31.10	36.98	24.13	25.83	25.71	25.47
LLaVA-1.5-13B	68.25	93.27	11.34	81.67	36.35	62.79	56.61	23.43	10.80	20.44	23.19	30.15	35.17	28.50	40.63	41.16	25.68	33.11	36.25	33.24
InternVL2.5-26B	**100.00**	99.30	**45.35**	76.85	39.49	86.92	61.02	70.77	63.07	28.98	24.87	32.96	69.39	74.53	47.12	53.95	24.52	31.35	37.28	23.06
Qwen2.5-VL-32B	99.12	99.30	30.81	41.48	52.85	90.12	45.48	68.68	65.15	21.99	33.16	36.89	60.46	68.69	52.15	50.47	21.62	33.11	36.25	41.02
LLaVA-1.6-34B	**100.00**	97.22	35.17	81.67	36.54	47.97	**67.05**	56.84	50.38	23.80	34.97	25.11	62.93	87.62	47.96	36.98	32.05	28.26	27.25	30.83
Medical-Domain Models																				
HealthGPT-M3	**100.00**	93.04	0.87	**82.32**	**61.10**	75.00	64.73	15.31	72.73	32.21	38.99	23.28	11.79	33.64	43.46	48.37	32.63	33.11	18.51	20.38
MedGemma-4B	99.65	83.99	7.27	74.60	33.01	80.81	38.05	24.83	53.98	24.45	31.09	39.55	5.13	41.36	44.92	40.93	39.77	33.11	16.20	30.83
Lingshu-7B	99.29	99.30	2.62	70.42	50.49	90.12	51.97	51.51	61.17	29.11	31.09	**39.97**	51.90	67.76	28.48	48.60	30.31	28.48	48.33	49.60
Hulu-Med-7B	**100.00**	99.54	5.81	80.39	59.92	84.88	46.64	50.58	89.58	28.33	30.31	30.72	64.45	70.09	39.69	48.84	44.59	32.89	42.42	32.98
MediX-R1-8B	**100.00**	99.30	33.14	81.35	46.17	63.95	55.92	41.07	50.95	29.11	39.12	27.63	79.66	62.85	48.69	52.33	26.83	33.33	19.02	43.16
Fleming-VL-8B	**100.00**	99.30	44.19	30.87	55.99	86.34	64.04	61.95	69.13	31.95	33.03	39.41	64.45	70.09	38.74	43.26	41.12	32.67	38.56	27.35
HealthGPT-Pro-8B	**100.00**	**99.77**	21.80	78.46	59.14	87.50	66.36	66.13	91.10	36.87	43.39	31.42	51.14	63.79	45.86	49.30	30.89	32.01	29.82	26.81
Lingshu-32B	98.94	**99.77**	8.43	74.28	31.24	90.12	30.63	66.82	89.96	43.60	**59.20**	38.99	61.98	73.36	45.45	57.44	34.36	32.67	**56.04**	38.34
Hulu-Med-32B	99.65	**99.77**	9.30	80.71	60.12	85.47	60.79	44.08	**97.35**	36.48	47.28	34.22	60.84	76.40	50.26	53.72	**47.68**	32.89	53.47	37.53

4.3 Main Results

Table 2 summarizes the overall performance and average results across the three capability dimensions. Four key observations emerge from these results.

Current MLLMs remain far from reliable OCT understanding. GPT-5.4-mini achieves the best overall accuracy (62.0%), followed by Gemini-2.5-flash (60.5%) and Hulu-Med-32B (58.4%). Although all models outperform the 25% random-guess baseline, even the best model answers nearly 38% of questions incorrectly, indicating that reliable clinical OCT interpretation remains out of reach. Moreover, no model consistently leads across all capability dimensions: GPT-5.4-mini performs best on Perception (75.8%), Gemini-2.5-flash on Cognition (64.2%), and Hulu-Med-32B on Reasoning (42.9%). This divergence shows that overall accuracy alone can mask substantial differences in capability profiles.

Performance degrades as capability requirements progress from perception to reasoning. Using the best result in each dimension, accuracy declines from 75.8% on Perception to 64.2% on Cognition and 42.9% on Reasoning, a drop of 32.9 points from the first to the final stage. This trend is also evident within individual models: strong visual perception does not necessarily translate into clinical reasoning. For example, GPT-5.4-mini achieves the highest Perception score but reaches only 38.5% on Reasoning. Even the best Reasoning result remains below 50% and only 17.9 points above chance, suggesting that visual evidence alone is insufficient without robust medical knowledge and image-grounded reasoning.

Model specialization does not ensure uniform superiority. Closed-source models are competitive overall, with GPT-5.4-mini and Gemini-2.5-flash ranking first and second. However, the best performance remains distributed across model types: Gemini-2.5-flash leads Cognition, while a medical-domain model leads Reasoning. Medical models also vary substantially. Although Hulu-Med-32B and Lingshu-32B rank third and fourth overall, HealthGPT-M3 and MedGemma-4B perform worse than several general-purpose models. These results suggest that domain specialization benefits specific capabilities but does not guarantee comprehensive OCT understanding.

Scaling produces uneven gains across capability levels. Scaling consistently improves overall accuracy within the InternVL, Qwen, Lingshu, and Hulu-Med families, but the gains are uneven across capability levels. For example, scaling Lingshu from 7B to 32B increases Cognition by 14 points, while Perception decreases slightly and Reasoning improves by only 1.2 points. Similar trends are observed for Qwen2.5-VL. These results indicate that scaling strengthens specific capabilities but does not resolve the bottleneck in clinically grounded reasoning, highlighting the need for better visual grounding, image–knowledge alignment, and multi-step clinical inference.

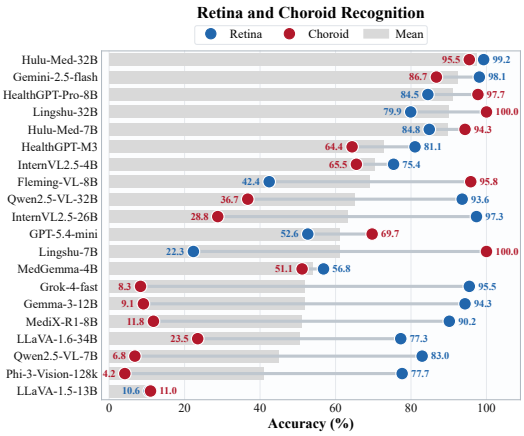


Figure 4: Retina and choroid recognition accuracy.

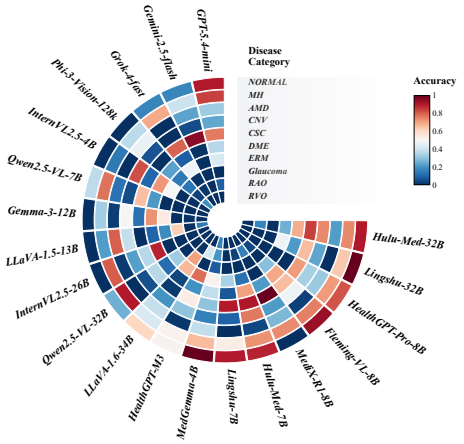


Figure 5: Diagnosis accuracy by disease category.

4.4 Fine-grained Analysis

Table 3 further decomposes model performance into 20 fine-grained tasks, revealing that the difficulty of OCT understanding is highly task-dependent even within the same capability level.

Low-level recognition is nearly saturated, but fine-grained visual description remains fragile. Most models perform well on modality perception (T01) and annotation recognition (T02), achieving average accuracies of 97.9% and 97.6%, demonstrating reliable recognition of imaging attributes and explicit annotations. However, performance drops sharply for subtle morphological description, averaging only 23.4% on T03 with a best score of 45.4%. This suggests that MLLMs can recognize OCT images but lack precise representations of retinal shape, deformation, and lesion morphology. Similar limitations appear in scale perception (T07, 53.4%) and spatial orientation recognition (T08, 51.9%), indicating insufficient geometric grounding.

Anatomical cognition exposes strong structure-specific bias. Although region identification (T09) achieves a high best score of 97.4%, Figure 4 reveals substantial variation across anatomical regions. Retina recognition averages 74.8%, while choroid recognition averages only 53.0%. Several models show strong asymmetry, achieving over 80% on retina recognition but below 21% on choroid recognition, whereas Lingshu-7B and Fleming-VL-8B exhibit the opposite trend. These results suggest that high anatomical scores may rely on region-specific shortcuts rather than balanced understanding. The limitation is more evident in layer identification (T10), where the average accuracy is only 30.9% and the best result reaches 56.8%, indicating persistent difficulty in distinguishing fine retinal layers and boundaries.

Disease reasoning is dominated by category-dependent shortcuts. Disease diagnosis (T17) is one of the most clinically important tasks, yet the average accuracy is only 30.1% and the best result is 47.7%. Figure 5 reveals substantial variation across disease categories. Models are relatively more successful on MH and CSC, whose average accuracies are 58.4% and 53.0%, but perform poorly on AMD (16.7%), Glaucoma (7.8%), RAO (5.5%), and RVO (2.5%). Such category imbalance suggests that models tend to rely on salient or frequently observed patterns, while failing on diseases that require subtle vascular, glaucomatous, or differential diagnostic reasoning.

Clinical reasoning remains weak even when visual cues are available. The reasoning tasks show consistently low ceilings. Stage classification (T18) is tightly clustered between 25.8% and 34.0%, indicating that nearly all models struggle to map OCT findings to disease severity. Treatment planning (T19) and follow-up adjustment (T20) achieve best scores of 56.0% and 50.9%, respectively, but their average accuracies remain only 33.8% and 33.9%. These results imply that models may occasionally recover common clinical priors, yet they cannot reliably integrate anatomical evidence, lesion status, and disease context into coherent decisions.

4.5 Error Analysis

To characterize model failures beyond aggregate accuracy, we select the best-performing closed-source model, GPT-5.4-mini, and the best-performing medical-domain open-source model, Hulu-Med-32B, and randomly sample 50 incorrect responses from each for error categorization. As shown in Figure 6, medical knowledge association errors and clinical reasoning errors are the two dominant failure types for both models, jointly accounting for 60% of GPT-5.4-mini errors and 56% of Hulu-Med-32B errors. This result suggests that complex integrative diagnosis and clinical decision-making remain major bottlenecks that limit the clinical applicability of current MLLMs. Meanwhile, incomplete medical knowledge and incorrect associations between OCT evidence and relevant medical concepts continue to hinder both models, highlighting the need for stronger domain knowledge and more reliable image–knowledge alignment.

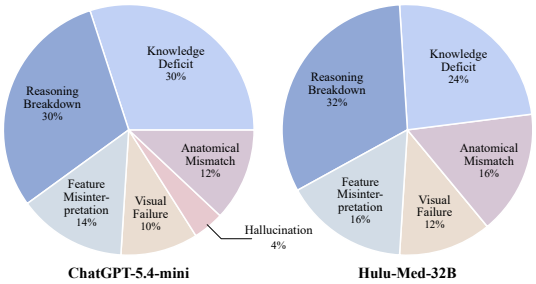


Figure 6: Distribution of error types.

5 Case Study

To further analyze model behavior, Figure 7 presents a representative case of anatomical region identification. Although all three models correctly identify the highlighted region as the choroid, their reasoning quality differs substantially. ChatGPT-5.4-mini first localizes the highlighted region with respect to surrounding retinal structures, then identifies it as the choroid based on its anatomical position beneath the retinal pigment epithelium (RPE) and above the sclera. MediX-R1-8B adopts a more concise but still visually grounded strategy, recognizing the vitreous cavity, retinal layers, and highlighted region before inferring its anatomical identity from its location beneath the retina. In contrast, Hulu-Med-32B reaches the correct answer but introduces unsupported medical knowledge, claiming that the choroid is "often highlighted in red in OCT images due to its rich vasculature." In fact, the red region is a manual annotation rather than an intrinsic imaging characteristic of OCT. This case illustrates that correct answers do not necessarily imply faithful reasoning: a model may rely on hallucinated medical knowledge while still arriving at the correct prediction.

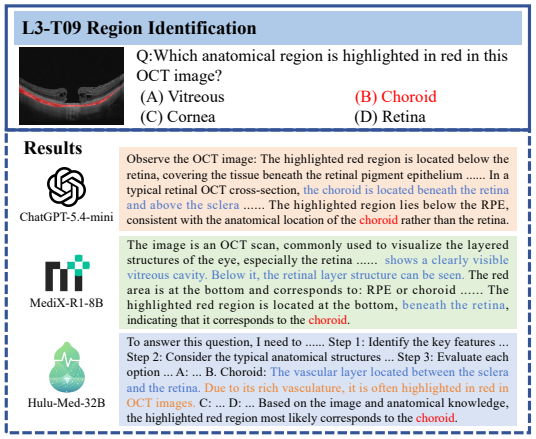


Figure 7: Case study of Region Identification.

6 Conclusion

In this paper, we introduced **OCT-Bench**, a benchmark for evaluating multimodal large language models on OCT image understanding. Following the clinical interpretation workflow, OCT-Bench organizes OCT understanding into three progressive dimensions: Perception, Cognition, and Reasoning. Built from 4,137 OCT images across seven public datasets, it contains 10,076 expert-verified multiple-choice questions spanning 20 fine-grained tasks, enabling comprehensive evaluation of both overall performance and capability bottlenecks. Evaluations of 20 representative MLLMs reveal that reliable OCT understanding remains challenging. The best model achieves only 62.0% overall accuracy, with performance dropping markedly from perception to reasoning. Models consistently struggle with fine-grained retinal structures, disease diagnosis, and image-grounded clinical reasoning, while larger model scale and medical-domain adaptation provide limited gains. We hope OCT-Bench will promote more reliable ophthalmic MLLM development.

AI USE STATEMENT

In this work, we used generative AI tools for OCT-Bench construction, quality control assistance, and manuscript polishing. GPT-4o was used to generate candidate multiple-choice questions based on OCT images, annotations, task descriptions, and medical knowledge, and to assist preliminary checks of question quality, answer uniqueness, and image–text consistency. All AI-assisted content was reviewed and validated by the authors and ophthalmology experts, and only expert-verified samples were retained in the final benchmark. We also used LLM assistance for grammar and wording refinement of the manuscript.

ETHICS STATEMENT

OCT-Bench is constructed using publicly available OCT datasets and does not involve additional human subject recruitment or personally identifiable information collection. All generated questions and clinical reasoning tasks are reviewed by experienced ophthalmology experts to ensure medical reliability. OCT-Bench is designed solely for evaluating OCT image understanding capabilities of multimodal large language models and should not replace professional medical diagnosis or treatment. We recommend appropriate human oversight for any future clinical applications.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the OCT-Bench construction pipeline, task taxonomy, data processing procedures, and evaluation protocols in this paper. All OCT images are collected from publicly available datasets, with their sources and preprocessing procedures documented. We will release the benchmark, evaluation code, and associated resources after the review process to enable the community to reproduce our experiments and verify our findings.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

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

Mustafa Arikan, James Willoughby, Sevim Ongun, Ferenc Sallo, Andrea Montesel, Hend Ahmed, Ahmed Hagag, Marius Book, Henrik Faatz, Maria Vittoria Cicinelli, et al. Oct5k: A dataset of multi-disease and multi-graded annotations for retinal layers. *Scientific data*, 12(1):267, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025a. URL https://arxiv.org/abs/2502.13923.

Yu Bai, Liang Bai, Xian Yang, and Jiye Liang. Label-semantic-based prompt tuning for vision transformer adaptation in medical image analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025b.

Misha Bilenko. New models added to the phi-3 family, available on microsoft azure. *Microsoft GenAI*, 2024.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: A survey. *Findings of the association for computational linguistics: ACL 2024*, pp. 13590–13618, 2024.

Xiaolan Chen, Ruoyu Chen, Pusheng Xu, Weiyi Zhang, Xianwen Shang, Mingguang He, and Danli Shi. Visual question answering in ophthalmology: A progressive and practical perspective. *arXiv preprint arXiv:2410.16662*, 2024a.

Xiaolan Chen, Ruoyu Chen, Pusheng Xu, Xiaojie Wan, Weiyi Zhang, Bingjie Yan, Xianwen Shang, Mingguang He, and Danli Shi. From visual question answering to intelligent ai agents in ophthalmology. *British Journal of Ophthalmology*, 110(1):1–7, 2026.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Jay S Duker, Nadia K Waheed, and Darin Goldman. *Handbook of Retinal OCT: Optical Coherence Tomography E-Book*. Elsevier Health Sciences, 2021.

Huihui Fang, Fei Li, Huazhu Fu, Junde Wu, Xiulan Zhang, and Yanwu Xu. Dataset and evaluation algorithm design for goals challenge. In *International Workshop on Ophthalmic Medical Image Analysis*, pp. 135–142. Springer, 2022.

Pinqi Fang, Yiting Wu, Yufeng He, Haoxuan Li, Zhouyu Guan, Xiangning Wang, Tingli Chen, and Jie Shen. Research progress on ai-assisted screening and prediction of systemic diseases based on retinal images. *The Visual Computer*, 41(12):9509–9537, 2025.

Baochen Fu, Yuntao Du, Cheng Chang, Baihao Jin, Wenzhi Deng, Muhao Xu, Hongmei Yan, Weiye Song, and Yi Wan. Mmku-bench: A multimodal update benchmark for diverse visual knowledge. *arXiv preprint arXiv:2603.15117*, 2026.

Sushmitha Gurumoorthy, NV Meena Lochana Nayaki, and K Ezhil Vendhan. The role of artificial intelligence in monitoring glaucoma progression using optical coherence tomography. *Indian Journal of Clinical and Experimental Ophthalmology*, 11(4):631–640, 2025.

Ming Hu, Peng Xia, Lin Wang, Siyuan Yan, Feilong Tang, Zhongxing Xu, Yimin Luo, Kaimin Song, Jurgen Leitner, Xuelian Cheng, et al. Ophnet: A large-scale video benchmark for ophthalmic surgical workflow understanding. In *European Conference on Computer Vision*, pp. 481–500. Springer, 2024a.

Yunwei Hu, Yundi Gao, Weihao Gao, Wenbin Luo, Zhongyi Yang, Fen Xiong, Zidan Chen, Yucai Lin, Xinjing Xia, Xiaolong Yin, et al. Amd-sd: an optical coherence tomography image dataset for wet amd lesions segmentation. *Scientific Data*, 11(1):1014, 2024b.

David Huang, Eric A Swanson, Charles P Lin, Joel S Schuman, William G Stinson, Warren Chang, Michael R Hee, Thomas Flotte, Kenton Gregory, Carmen A Puliafito, et al. Optical coherence tomography. *science*, 254(5035):1178–1181, 1991.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Kailin Jiang, Yuntao Du, Yukai Ding, Yuchen Ren, Ning Jiang, Zhi Gao, Zilong Zheng, Lei Liu, Bin Li, and Qing Li. When large multimodal models confront evolving knowledge: Challenges and explorations. In *The Fourteenth International Conference on Learning Representations*, 2026.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

Stephen J Kim, Jennifer I Lim, Steven T Bailey, Jaclyn L Kovach, G Atma Vemulakonda, Guishuang Ying, Christina J Flaxel, et al. Idiopathic macular hole preferred practice pattern®. *Ophthalmology*, 132(4):P234–P269, 2025.

Jaclyn L Kovach, Steven T Bailey, Stephen J Kim, Jennifer I Lim, G Atma Vemulakonda, GuiShuang Ying, Christina J Flaxel, et al. Retinal and ophthalmic artery occlusions preferred practice pattern®. *Ophthalmology*, 132(4):P270–P302, 2025.

Mikhail Kulyabin, Aleksei Zhdanov, Anastasia Nikiforova, Andrey Stepichev, Anna Kuznetsova, Mikhail Ronkin, Vasilii Borisov, Alexander Bogachev, Sergey Korotkich, Paul A Constable, et al. Octdl: Optical coherence tomography dataset for image-based deep learning methods. *Scientific data*, 11(1):365, 2024.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

Jennifer I Lim, Stephen J Kim, Steven T Bailey, Jaclyn L Kovach, G Atma Vemulakonda, Guishuang Ying, and Christina J Flaxel. Diabetic retinopathy preferred practice pattern®. *Ophthalmology*, 132(4):P75–P162, 2025.

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. Benchmarking large language models' performances for myopia care: a comparative analysis of chatgpt-3.5, chatgpt-4.0, and google bard. *EBioMedicine*, 95, 2023.

Haoneng Lin, Cheng Xu, and Jing Qin. Taming vision-language models for medical image analysis: A comprehensive review. *arXiv preprint arXiv:2506.18378*, 2025a.

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, et al. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *arXiv preprint arXiv:2502.09838*, 2025b.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024.

Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Omair Mohamed, Mohamed Zidan, Fahad Khan, Salman Khan, Rao Anwer, and Hisham Cholakkal. Medix-r1: Open ended medical reinforcement learning. *arXiv preprint arXiv:2602.23363*, 2026.

Zhenyue Qin, Yu Yin, Dylan Campbell, Xuansheng Wu, Ke Zou, Ninghao Liu, Yih Chung Tham, Xiuzhen Jenny Zhang, and Qingyu Chen. Lmod: A large multimodal ophthalmology dataset and benchmark for large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 2501–2522, 2025.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Yan Shu, Chi Liu, Robin Chen, Derek Li, and Bryan Dai. Fleming-vl: Towards universal medical visual reasoning with multimodal llms. *arXiv preprint arXiv:2511.00916*, 2025.

Sahana Srinivasan, Xuguang Ai, Thaddaeus Wai Soon Lo, Aidan Gilson, Minjie Zou, Ke Zou, Hyunjae Kim, Mingjia Yang, Krithi Pushpanathan, Samantha Min Er Yew, et al. Benchmarking large language models for ophthalmology (belo): an expert-curated data set and evaluation framework for knowledge and reasoning. *Ophthalmology Science*, 6(3):101050, 2026.

Malliga Subramanian, Kogilavani Shanmugavadivel, Obuli Sai Naren, K Premkumar, and K Rankish. Classification of retinal oct images using deep learning. In *2022 international conference on computer communication and informatics (ICCCI)*, pp. 1–7. IEEE, 2022.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

G Atma Vemulakonda, Steven T Bailey, Stephen J Kim, Jaclyn L Kovach, Jennifer I Lim, Guishuang Ying, Christina J Flaxel, et al. Age-related macular degeneration preferred practice pattern®. *Ophthalmology*, 132(4):P1–P74, 2025.

Shaopan Wang, Xin He, Zhongquan Jian, Jie Li, Changsheng Xu, Yuguang Chen, Yuwen Liu, Han Chen, Caihong Huang, Jiaoyue Hu, et al. Advances and prospects of multi-modal ophthalmic artificial intelligence based on deep learning: a review. *Eye and Vision*, 11(1):38, 2024.

Weisen Wang, Xirong Li, Zhiyan Xu, Weihong Yu, Jianchun Zhao, Dayong Ding, and Youxin Chen. Learning two-stream cnn for multi-modal age-related macular degeneration categorization. *IEEE Journal of Biomedical and Health Informatics*, 26(8):4111–4122, 2022.

Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, et al. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *Advances in Neural Information Processing Systems*, 37:69925–69975, 2024.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

X Xun and L Xiaoxin. Evidence-based guidelines for diagnosis and treatment of diabetic retinopathy in china (2022). *Chin. J. Ocular Fund. Dis*, 2:99–124, 2023.

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

YiFan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, et al. Mme-realworld: Could your multimodal llm challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations*, volume 2025, pp. 89655–89701, 2025.

Tianhong Zhou, Yingtao Zhu, Chuxi Xiao, Haiyang Bian, Lei Wei, Xuegong Zhang, et al. Drvd-bench: Do vision-language models reason like human doctors in medical image diagnosis? *Advances in Neural Information Processing Systems*, 38, 2026.

Wenhui Zhu, Xuanzhao Dong, Xin Li, Yujian Xiong, Xiwen Chen, Peijie Qiu, Vamsi Krishna Vasa, Zhangsihao Yang, Yi Su, Oana Dumitrascu, et al. Eyebench: A call for more rigorous evaluation of retinal image enhancement. *arXiv preprint arXiv:2502.14260*, 2025.

Minjie Zou, Sahana Srinivasan, Thaddaeus Wai Soon Lo, Ke Zou, Gabriel Dawei Yang, Xuguang Ai, Hyunjae Kim, Maxwell Singer, Fares Antaki, Kelvin Li, et al. Benchmarking next-generation reasoning-focused large language models in ophthalmology: a head-to-head evaluation on 5,888 items. *arXiv preprint arXiv:2504.11186*, 2025.

A APPENDIX

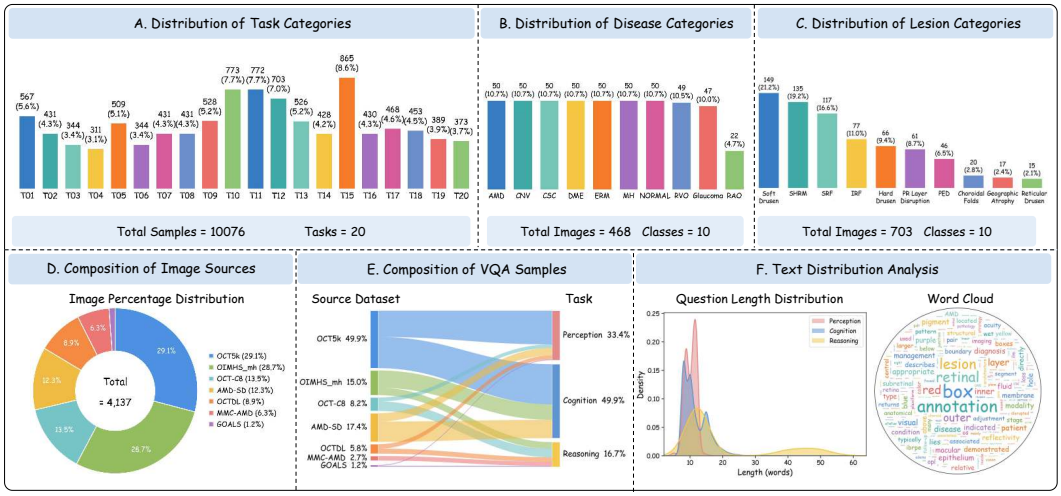


Figure 8: Data distribution analysis of OCT-Bench. (A) Distribution of task categories. (B) Distribution of disease categories. (C) Distribution of lesion categories. (D) Composition of image sources. (E) Composition of VQA samples. (F) Text distribution analysis.

A.1 DATA ANALYSIS

Figure 8 summarizes the statistical characteristics of OCT-Bench from the perspectives of task coverage, clinical label coverage, data sources, and textual complexity.

- **Task coverage and balance.** OCT-Bench contains 10,076 VQA samples derived from 4,137 OCT images and spans 20 fine-grained tasks across three progressively challenging capability dimensions: *Perception*, *Cognition*, and *Reasoning*. The task distribution is relatively even, with each task contributing 311–865 samples, which helps mitigate severe category imbalance and supports more stable task-level evaluation.
- **Data sources and clinical diversity.** OCT-Bench integrates images from seven public OCT datasets: OCT5K, OIMHS_mh, OCT-C8, AMD-SD, OCTDL, MMC-AMD, and GOALS. These datasets introduce diverse imaging protocols, acquisition conditions, and disease patterns. The benchmark covers 10 representative ophthalmic diseases, with most disease categories containing approximately 50 images, and 10 clinically important retinal lesion types, from soft drusen and subretinal fluid to geographic atrophy and reticular drusen.
- **Sample composition and textual complexity.** *Cognition* accounts for the largest proportion of VQA samples (49.9%), followed by *Perception* (33.4%) and *Reasoning* (16.7%). This allocation reflects the clinical OCT interpretation workflow: models first perceive visual evidence, then associate it with anatomical and pathological concepts, and finally perform higher-level clinical reasoning. Reasoning-oriented questions are substantially longer than perception- and cognition-oriented questions, and the word cloud highlights clinically grounded concepts such as anatomical structures, lesion characteristics, retinal layers, and imaging biomarkers.

A.2 CASE STUDY

To complement the quantitative results presented in the main paper, we conduct an in-depth case study over 20 representative questions spanning the full spectrum of OCT-Bench tasks (T01–T20). These cases are deliberately selected to cover all three capability dimensions—Perception, Cognition, and Reasoning—and to reveal how different model families (closed-source, open-source, and medical-domain) behave when confronted with clinically realistic OCT interpretation challenges.

A.2.1 PERCEPTION-LEVEL CASES

The Perception dimension evaluates whether models can accurately extract visual information from OCT images, including imaging attributes, morphological features, and spatial relationships. Our case analysis reveals that while basic visual recognition tasks are largely mastered, fine-grained perceptual discrimination remains challenging.

T01 (Modality Recognition). This task asks models to identify the imaging modality depicted in the image (Ultrasound, CT, MRI, or OCT). The vast majority of models (19/20) correctly select "OCT," indicating that recognizing the characteristic cross-sectional appearance of retinal imaging is well within the capabilities of current MLLMs. The sole failure comes from LLaVA-1.5-13B, which incorrectly predicts "Ultrasound." This suggests that modality recognition, as a coarse-grained visual classification task, is largely solved even by smaller open-source models.

T02 (Annotation Recognition). This task asks about the shape of the annotation boxes in the image (rectangle, circle, triangle, or hexagon). All 20 models correctly answer "rectangle" (A), achieving a perfect accuracy of 100%. This flawless performance suggests that basic geometric shape recognition in annotated medical images is a trivial task for current MLLMs, regardless of model size or domain specialization. The uniform success across closed-source, open-source, and medical-domain models indicates that such low-level visual attributes are well captured by existing vision encoders.

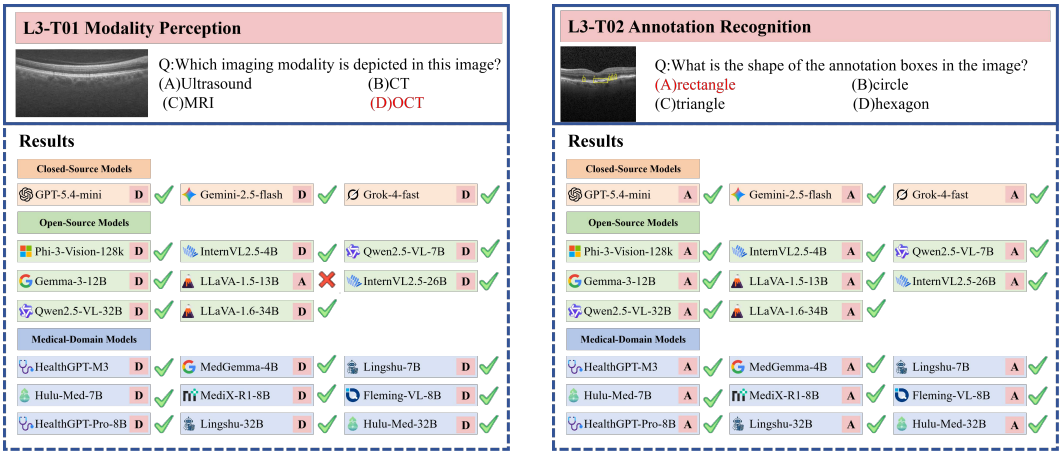


Figure 9: Case study results for T01 and T02.

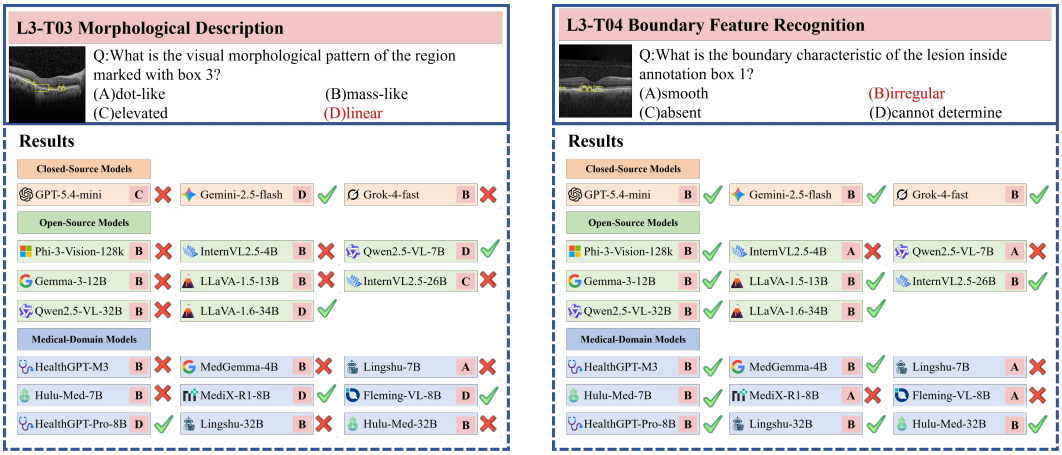


Figure 10: Case study results for T03 and T04.

T03 (Morphological Description). In contrast to T01, this task requires fine-grained morphological discrimination: identifying whether a marked lesion exhibits a "dot-like," "mass-like," "elevated,"

or “linear” pattern. Only 6 out of 20 models answer correctly (Fleming-VL-8B, HealthGPT-Pro-

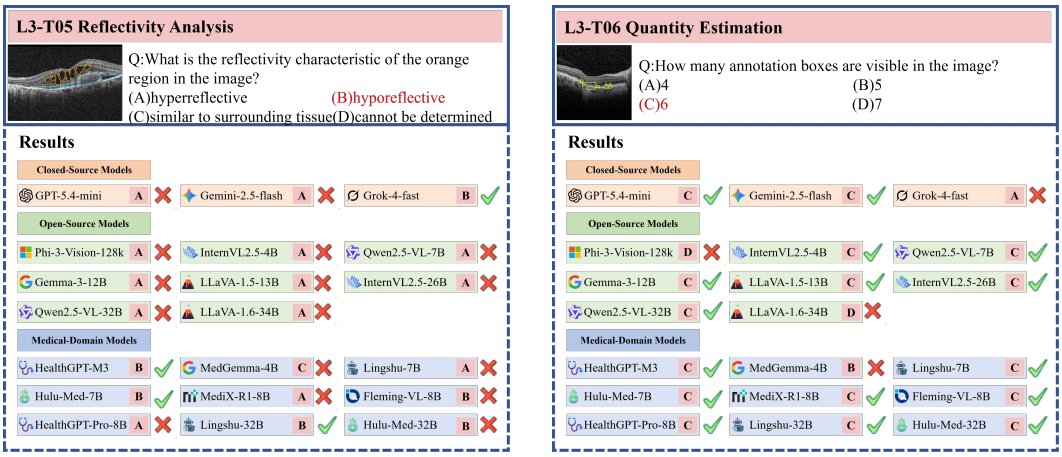


Figure 11: Case study results for T05 and T06.

8B, MediX-R1-8B, Qwen2.5-VL-7B, LLaVA-1.6-34B, and Gemini-2.5-flash). The remaining 14 models predominantly confuse “mass-like” (B) with the ground-truth “linear” (D), including large medical-domain models such as Lingshu-32B and Hulu-Med-32B. This failure pattern indicates that subtle morphological distinctions—critical for differentiating lesion types in clinical practice—are not reliably captured even by medically adapted models.

T04 (Boundary Feature Recognition). This task asks about the boundary characteristic of a lesion inside an annotation box (smooth, irregular, absent, or cannot determine). Fifteen out of twenty models correctly identify “irregular” (B).

The five incorrect models—Fleming-VL-8B, InternVL2.5-4B, Lingshu-7B, MediX-R1-8B, and Qwen2.5-VL-7B—all uniformly predict “smooth” (A) instead. This consensus error suggests that models struggle to discern subtle textural irregularities at lesion boundaries, possibly because the visual cues for boundary irregularity in OCT are subtle and require trained clinical judgment that current MLLMs have not adequately acquired.

T05 (Reflectivity Analysis). This case asks about the reflectivity characteristic of a highlighted orange region (hyperreflective, hyporeflective, iso-reflective, or indeterminate). Only 6 out of 20 models respond correctly, and notably, five of these come from medical-domain models (HealthGPT-M3, Hulu-Med-7B, Fleming-VL-8B, Lingshu-32B, Hulu-Med-32B). All open-source models fail, with 13 out of 14 incorrectly predicting “hyperreflective” rather than the correct “hyporeflective.” This case highlights a clear advantage of medical-domain adaptation for tissue-property perception tasks that require domain-specific understanding of OCT signal characteristics.

T06 (Quantity Estimation). This task asks how many annotation boxes are visible in the image (4, 5, 6, or 7). Sixteen out of twenty models correctly answer “6” (C). The four incorrect models show varied responses: Phi-3-Vision-128k and LLaVA-1.6-34B predict “7” (D), MedGemma-4B predicts “5” (B), and Grok-4-fast predicts “4” (A). Unlike the consensus errors seen in other tasks, these mistakes appear to stem from individual counting errors rather than systematic biases, suggesting that quantity estimation in moderately cluttered OCT images is generally manageable but occasionally prone to miscounting.

T07 (Scale Perception). This task asks which annotation box is larger relative to another (Box 3 is larger, Box 1 is larger, similar in size, or cannot be determined). Eighteen out of twenty models correctly identify that “Box 1 is larger” (B). The two failures come from Gemma-3-12B, which predicts “The two boxes are similar in size” (C), and LLaVA-1.5-13B, which predicts “Box 3 is larger” (A). This high accuracy indicates that basic scale comparison between annotated regions is well within the capabilities of current MLLMs, with errors likely arising from borderline cases where size differences are subtle.

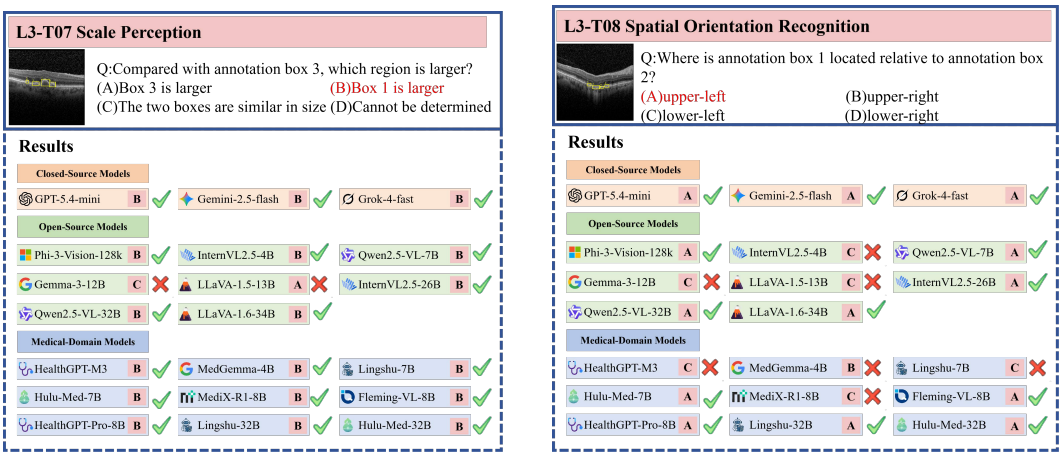


Figure 12: Case study results for T07 and T08.

T08 (Spatial Orientation Recognition). Models are asked to locate one annotation box relative to another (upper-left, upper-right, lower-left, or lower-right). Only 12 out of 20 models answer correctly. The most striking failure pattern is the systematic confusion of vertical orientation: seven models (including InternVL2.5-4B, Lingshu-7B, and HealthGPT-M3) incorrectly select "lower-left" when the correct answer is "upper-left." This suggests that spatial reasoning in medical images—where anatomical orientation conventions differ from natural images—presents a persistent challenge even for large-scale models.

A.2.2 Cognition-Level Cases

The Cognition dimension assesses whether models can transform visual information into medical knowledge, including anatomical understanding, pathological recognition, and clinical associations. Our cases reveal highly uneven performance across different cognitive tasks.

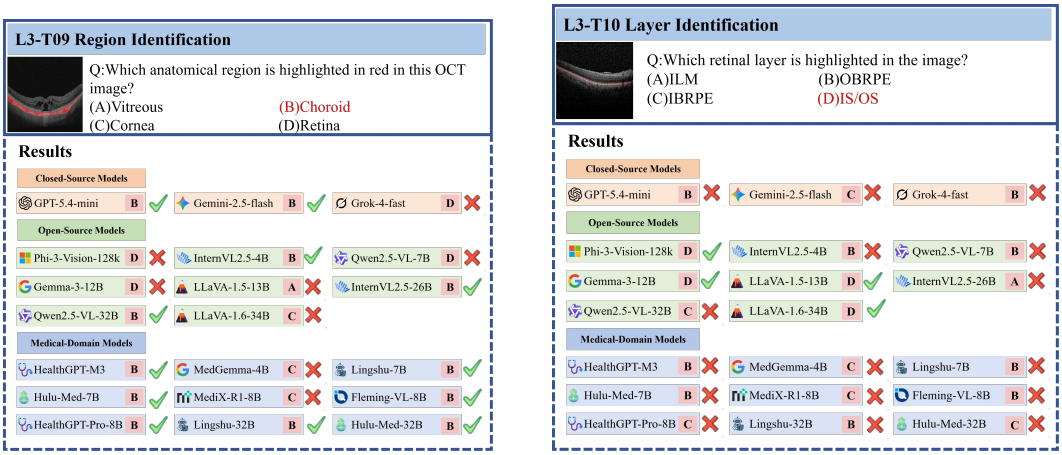


Figure 13: Case study results for T09 and T10.

T09 (Region Identification). Models must identify which anatomical region is highlighted in red (Vitreous, Choroid, Cornea, or Retina). Twelve out of twenty models answer correctly, with medical-domain models showing a notable advantage (7/9 correct vs. 3/8 for open-source models). Errors are distributed across all three incorrect options, suggesting that anatomical localization in OCT remains difficult for general-purpose vision-language models without medical pretraining.

T10 (Layer Identification). This task requires identifying a highlighted retinal layer among ILM, OBRPE, IBRPE, and IS/OS. Surprisingly, only 4 out of 20 models answer correctly, and all four are

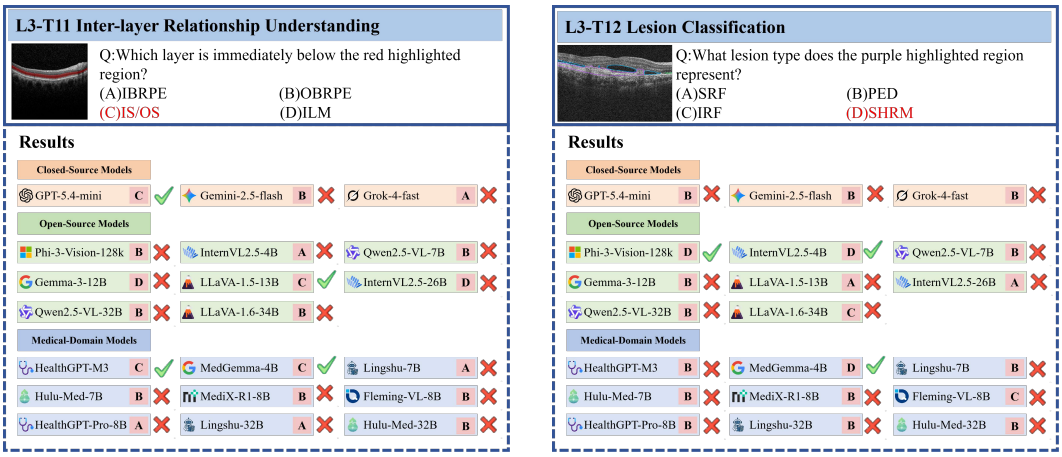


Figure 14: Case study results for T11 and T12.

open-source models (Phi-3-Vision-128k, Gemma-3-12B, LLaVA-1.5-13B, LLaVA-1.6-34B). All nine medical-domain models fail, with nine incorrectly predicting "OBRPE" (B) or "IBRPE" (C). This counter-intuitive result—where medical-domain models underperform on a seemingly specialized task—suggests that IS/OS junction identification may be underrepresented in medical-domain training data, or that the fine-grained layer discrimination required exceeds the spatial resolution capabilities of current medical MLLMs.

T11 (Inter-layer Relationship). This case asks which retinal layer lies immediately below a highlighted region. With only 4 out of 20 correct predictions, this is one of the most challenging tasks in our benchmark. The correct answer is "IS/OS," yet the majority of models predict "OBRPE" (9 models) or "IBRPE" (5 models). Even the best-performing closed-source models fail on this question. This near-universal failure indicates that understanding hierarchical spatial relationships between adjacent retinal layers—a routine requirement in clinical OCT interpretation—remains beyond the grasp of current MLLMs.

T12 (Lesion Classification). Models are asked to classify a purple highlighted lesion type (IRF, SHRM, PED, or SRF). Only 3 out of 20 models answer correctly (Phi-3-Vision-128k, InternVL2.5-4B, MedGemma-4B). The dominant error pattern is the confusion of "SHRM" (ground truth) with "PED" (13 models), including large models such as Qwen2.5-VL-32B and HealthGPT-Pro-8B. This case illustrates that lesion classification in pathological OCT scans requires not only visual pattern recognition but also subtle differentiation of fluid and tissue compartments—a capability that current models lack.

T13 (Structural Status Assessment). Models evaluate the structural status of macular architecture (intact, mildly irregular, disorganized with localized disruption, or severely disrupted). Eleven out of twenty models answer correctly. Notably, several medical-domain models (HealthGPT-M3, MedGemma-4B, Lingshu-7B, HealthGPT-Pro-8B, Lingshu-32B) incorrectly predict "Structurally intact" (A) when the correct answer is "Structurally disorganized with localized disruption" (C). This suggests that medical-domain models may exhibit an optimism bias, tending to underestimate structural damage in pathological scans.

T14 (Lesion Localization). This task requires identifying the anatomical location of a lesion (Outer retina, Inner retina, Vitreous, or Cornea). Eleven out of twenty models answer correctly. Errors are concentrated in predicting "Inner retina" (B) instead of the correct "Outer retina" (A), with nine models making this mistake. The spatial confusion between inner and outer retinal layers is particularly prevalent among medical-domain models (5/9 errors), suggesting that even domain-specific pretraining does not fully resolve the anatomical depth ambiguity in cross-sectional OCT images.

T15 (Disease Association). This task asks which condition is most typically associated with sub-RPE lesions (CSC, DME, Wet AMD, or Dry AMD). With 18 out of 20 correct predictions, this is the best-performing cognition task. Both closed-source and medical-domain models achieve near-

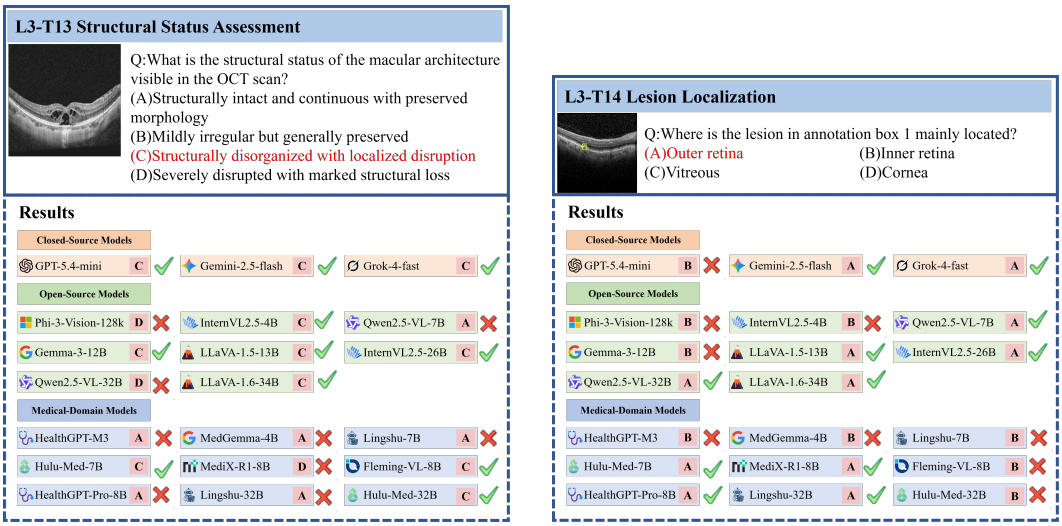


Figure 15: Case study results for T13 and T14.

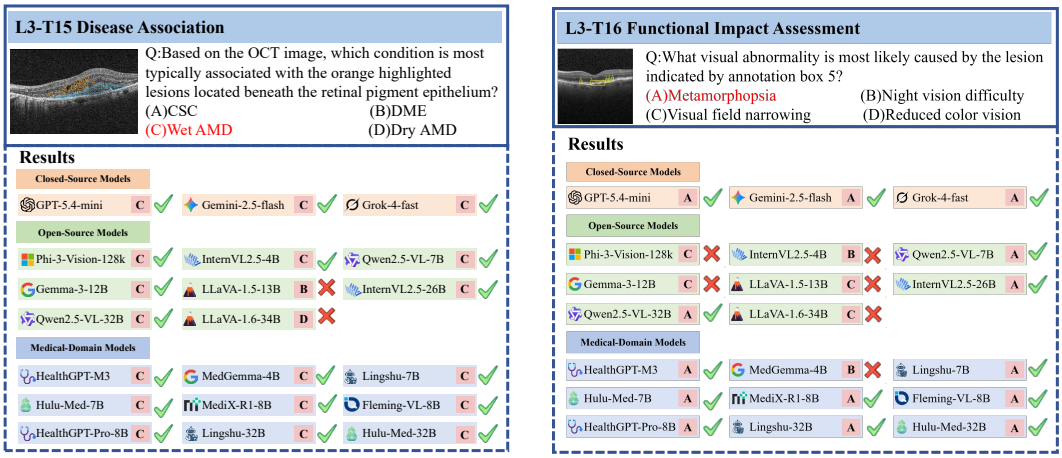


Figure 16: Case study results for T15 and T16.

perfect accuracy, while only two open-source models (LLaVA-1.5-13B and LLaVA-1.6-34B) fail. The high success rate suggests that associating characteristic imaging findings with disease categories is a relatively accessible cognition task for current MLLMs, likely because such associations are frequently documented in medical literature and training corpora.

T16 (Functional Impact Assessment). Models must infer the most likely functional abnormality caused by a lesion (Metamorphopsia, Night vision difficulty, Visual field narrowing, or Reduced color vision). Fourteen out of twenty models answer correctly. The six incorrect models predominantly select "Visual field narrowing" (C) rather than the correct "Metamorphopsia" (A). This error pattern suggests that models struggle to link specific retinal pathologies (in this case, choroidal folds) to their corresponding subjective visual symptoms, a critical step in clinically grounded cognition.

A.2.3 Reasoning-Level Cases

The Reasoning dimension evaluates whether models can integrate multiple imaging findings with medical knowledge to support clinical decision-making, including disease diagnosis, staging, treatment planning, and prognostic management. Our case analysis reveals a marked performance degradation at this level, with several tasks achieving near-zero accuracy.

T17 (Disease Diagnosis). This case requires diagnosing the condition shown in an OCT image (CSC, DME, CNV, or MH). Only 6 out of 20 models answer correctly, representing a sharp drop in performance. The errors are distributed across all three incorrect options: seven models select "CSC" (A), five select "DME" (B), and two select "MH" (D). Remarkably, all three closed-source models fail on this case, as do six out of nine medical-domain models. The image path contains "CNV" as a substring, yet the majority of models cannot translate the visual evidence of choroidal neovascularization into the correct diagnosis, highlighting the gap between pattern recognition and diagnostic reasoning.

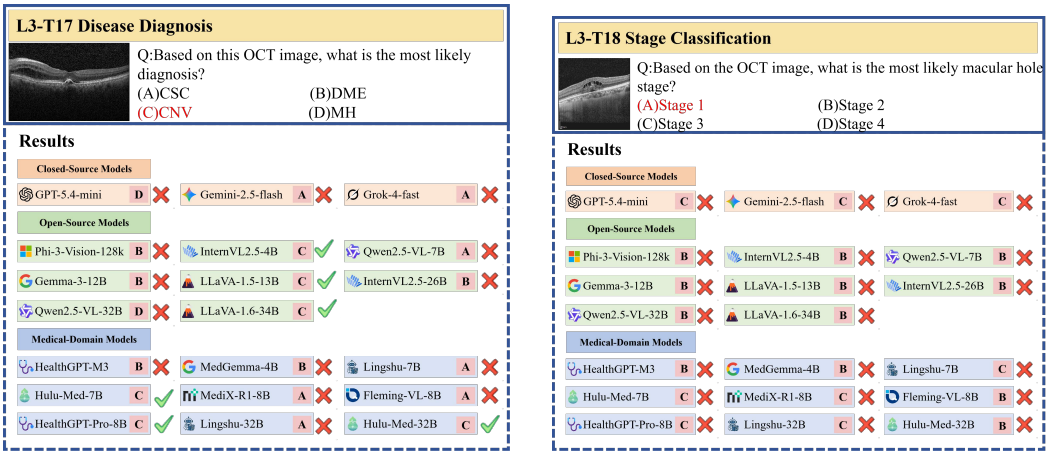


Figure 17: Case study results for T17 and T18.

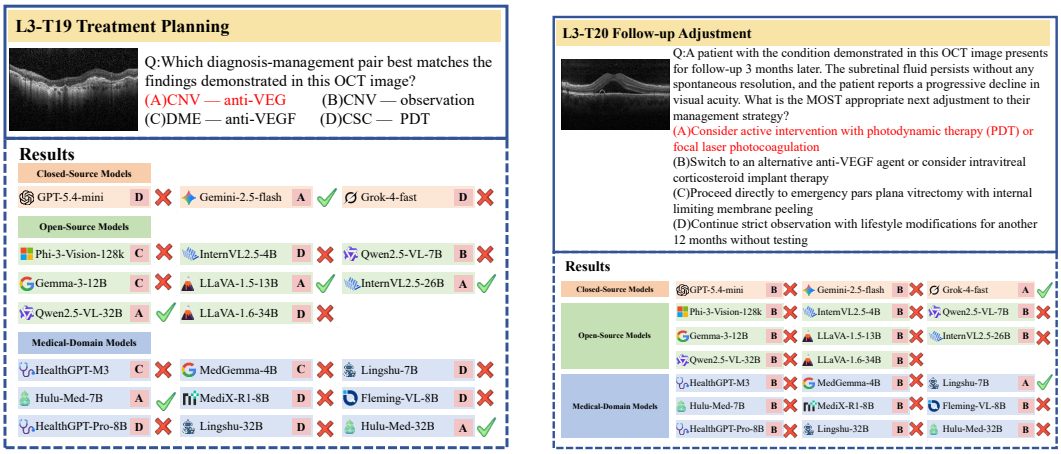


Figure 18: Case study results for T19 and T20.

T18 (Stage Classification). This task asks models to classify the stage of a macular hole (Stage 1–4). With zero correct predictions out of twenty models, this is the most universally failed case in our study. Twelve models predict "Stage 2" and eight predict "Stage 3," while the correct answer is "Stage 1." The complete failure across all model families—including state-of-the-art closed-source systems and large medical-domain models—indicates that fine-grained staging of macular holes requires subtle discrimination of early anatomical changes (such as foveal detachment without full-thickness defect) that current MLLMs cannot reliably perceive or interpret.

T19 (Treatment Planning). Models must select the diagnosis-treatment pair that best matches the OCT findings. Only 6 out of 20 models answer correctly. The dominant error pattern is the selection of "CSC — observation or PDT" (D) instead of the correct "CNV — anti-VEGF" (A), with nine models making this mistake. This failure reveals that models struggle to simultaneously perform accurate diagnosis and select the corresponding evidence-based treatment, a dual-requirement that mirrors real-world clinical decision-making.

Table 4: Cross-model comparison across capability dimensions (accuracy %).

Model Family	Perception	Cognition	Reasoning	Overall
Closed-Source	69.0	58.5	36.4	58.3
Open-Source	63.7	40.8	28.5	47.5
Medical-Domain	64.7	48.1	34.7	52.1

T20 (Follow-up Adjustment). This case presents a clinical scenario where a patient with persistent subretinal fluid shows progressive visual decline after three months of observation, and asks for the most appropriate management adjustment. Only 2 out of 20 models (Grok-4-fast and Lingshu-7B) answer correctly. The remaining 18 models uniformly select "Switch to an alternative anti-VEGF agent" (B) rather than the correct "Consider PDT or focal laser photocoagulation" (A). This striking consensus error suggests that models have learned a heuristic bias toward anti-VEGF therapy as the default intervention for subretinal fluid, failing to recognize that persistent fluid under observation is more characteristic of CSC (which responds to PDT/laser) than of CNV (which typically responds to anti-VEGF). This case exemplifies how learned biases from training data can lead to systematic clinical reasoning errors.

A.2.4 Cross-Model Comparison and Error Analysis

Table 4 reports family-level averages over Perception (T01–T08), Cognition (T09–T16), Reasoning (T17–T20), and all tasks (T01–T20).

- **Closed-source models lead overall, but remain weak at reasoning.** Closed-source models achieve the strongest overall performance (58.3%). Their advantage is most pronounced in Cognition, where they outperform open-source and medical-domain models by 17.7 and 10.4 points, respectively, suggesting stronger visual-language alignment and broader knowledge integration. However, their Reasoning score remains only 36.4%, indicating persistent difficulty with diagnosis, staging, treatment selection, and follow-up adjustment.
- **Open-source and medical-domain models show uneven gains.** Open-source models obtain the lowest overall score (47.5%) and the sharpest degradation from Perception (63.7%) to Reasoning (28.5%). Medical-domain models achieve an intermediate overall score (52.1%) and outperform open-source models on Cognition and Reasoning, but their gains remain uneven. Domain adaptation helps with some clinically relevant associations, yet does not guarantee robust fine-grained anatomical discrimination or reliable multi-step reasoning.
- **Reasoning is the shared bottleneck.** The drop from Perception to Reasoning is 32.6 points for closed-source models, 35.2 points for open-source models, and 30.0 points for medical-domain models. This common degradation suggests that current MLLMs can often recognize visible OCT attributes but struggle to integrate visual evidence with medical knowledge. The errors in T17–T20 further indicate error propagation, where incorrect lesion interpretation or diagnostic inference leads to inappropriate diagnosis, treatment planning, or follow-up recommendation.

A.3 Experimental Setting

All open-source and medical-domain models were evaluated on a local server equipped with four NVIDIA A100 GPUs, each with 40GB of memory. We used the vLLM framework for efficient batched inference when the model architecture was supported, and used the Hugging Face Transformers framework for models requiring custom loading, preprocessing, or generation pipelines. Closed-source models were evaluated through their APIs using the same zero-shot prompts and answer extraction protocol as the local models. For locally deployed models, we used the original released checkpoints without fine-tuning. For all models, we required the output to contain only the option letter for each multiple-choice question, and invalid, empty, or non-option responses were treated as incorrect to ensure a consistent evaluation protocol across model families.

A.4 Pretraining Data Analysis

Impact of Ophthalmic and OCT Pretraining Data. To investigate whether exposure to OCT or ophthalmic data during pretraining is associated with performance on OCT-Bench, we survey the publicly disclosed training-data information of all evaluated models. Since most MLLMs do not provide exhaustive dataset-level training records, this analysis should be viewed as **contextual evidence rather than causal attribution**. Table 5 summarizes the available evidence.

Categorization of training-data exposure. Based on publicly available documentation, we divide the evaluated models into three categories:

1. **OCT/Ophthalmology-exposed:** models whose public materials provide explicit evidence of OCT or ophthalmic data during training;
2. **Non-ophthalmic:** models whose disclosed training materials indicate no ophthalmology-related data;
3. **Unknown:** models for which the presence or absence of ophthalmic or OCT data cannot be verified from available disclosures.

Overall performance patterns. Among models with confirmed OCT exposure, overall accuracy ranges from 45.1% (HealthGPT-M3) to 58.4% (Hulu-Med-32B), with an average of 53.0%. Models trained with fundus-only ophthalmic data achieve 42.2% (MedGemma-4B) and 53.6% (Fleming-VL-8B), whereas the two models confirmed to lack ophthalmic data—LLaVA-1.5-13B and LLaVA-1.6-34B—obtain 39.6% and 50.5%, respectively. These results suggest a **moderate association between domain-specific pretraining and OCT-Bench performance**, although the variation within each category indicates that pretraining exposure is not the sole determinant of performance.

Dimension-level differences. The association is particularly evident for **Cognition and Reasoning**, which require domain-specific visual knowledge and clinical interpretation. OCT-exposed models such as Lingshu-32B, Hulu-Med-32B, and HealthGPT-Pro-8B achieve Cognition scores above 50%, whereas models without disclosed ophthalmic data remain below 47%. The gap is smaller for **Perception**, likely because basic visual recognition can benefit from general-purpose visual encoders even without ophthalmology-specific pretraining. In contrast, Reasoning exhibits the largest performance spread: Lingshu-32B and Hulu-Med-32B reach 40.4% and 42.9%, respectively, while models without confirmed OCT exposure remain below 33%. This pattern suggests that **OCT-specific visual knowledge may be particularly beneficial for multi-step clinical inference**.

OCT exposure is helpful but not sufficient. Importantly, explicit exposure to OCT data does not guarantee strong performance. HealthGPT-M3 achieves only 45.1% overall despite documented OCT inclusion. Similarly, MedGemma-4B and Fleming-VL-8B contain fundus images but no confirmed OCT data and obtain moderate performance, suggesting that **general ophthalmic knowledge may provide partial transfer but is insufficient for fine-grained OCT interpretation and disease-specific reasoning**. Thus, the benefit of domain-specific data appears to depend not only on whether such data are included, but also on their scale, diversity, quality, and how effectively the model aligns visual and clinical knowledge.

Unknown training-data status and other contributing factors. For models with unknown training-data status, performance varies substantially. GPT-5.4-mini and Gemini-2.5-flash achieve the highest overall scores despite the absence of publicly verifiable OCT-training disclosures, whereas Qwen2.5-VL-7B and Gemma-3-12B fall within the middle range. This observation highlights that **pretraining-data exposure alone cannot explain the observed performance differences**. Model scale, multimodal alignment, instruction tuning, visual resolution, and inference-time strategies may also substantially influence OCT-Bench results.

Overall, the results indicate that **ophthalmic and OCT-specific pretraining data are associated with a measurable but moderate advantage** on OCT-Bench, with the effect being more apparent in Cognition and Reasoning than in basic Perception. However, these observations should not be interpreted as evidence of a causal effect, because model architecture, scale, instruction tuning, visual capabilities, and other training factors are highly confounded with training-data composition. More transparent reporting of specialized medical training data would therefore be valuable for interpreting and comparing performance on domain-specific multimodal benchmarks.

Table 5: Publicly disclosed ophthalmology/OCT exposure in the evaluated models' training data.

Model	Source URL	Disclosed training-data evidence	Ophthalmology/OCT
ChatGPT-5.4-mini	NA	Only high-level categories are disclosed, including public data, partner data, and multimodal image/audio/video data; no dataset-level list is provided.	Unknown
Gemini-2.5-Flash	https://arxiv.org/abs/2507.06261	No itemized training-dataset list is disclosed.	Unknown
Grok-4-fast		No public dataset-level training-data list was identified in the supplied notes.	Unknown
Phi-3-Vision-128k	https://huggingface.co/microsoft/Phi-3-vision-128k-\instruct	Reports OBELICS and a high-level mixture of public documents, interleaved image-text data, synthetic data, and SFT data; no specific medical dataset is listed.	Unknown
InternVL2.5-4B	https://arxiv.org/abs/2412.05271	The fine-tuning mixture includes PMC-VQA, VQA-RAD, ImageCLEF, PMC, SLAKE, GMAI-VL, VQA-Med, Medical-Diff-VQA, PathVQA, and PMC-CaseReport; GMAI-VL statistics explicitly include OCT, fundus photography, and ophthalmology.	Ophthalmology (OCT)
InternVL2.5-26B	https://arxiv.org/abs/2412.05271	Uses the same disclosed training-data scope as InternVL2.5-4B.	Ophthalmology (OCT)
Qwen2.5-VL-7B	https://arxiv.org/abs/2502.13923	No complete, verifiable training-dataset list is disclosed.	Unknown
Qwen2.5-VL-32B	https://arxiv.org/abs/2502.13923	Uses the same disclosure level as Qwen2.5-VL-7B.	Unknown
Gemma-3-12B	https://arxiv.org/abs/2503.19786	No verifiable itemized visual training-dataset list is disclosed.	Unknown
LLaVA-1.5-13B	https://arxiv.org/abs/2310.03744	The supplied notes indicate no professional-domain images or datasets in the training data.	No ophthalmology
LLaVA-1.6-34B	https://llava-vl.github.io/blog/2024-01-30-llava-next/	The supplied notes indicate no professional-domain images or datasets in the training data.	No ophthalmology
HealthGPT-M3	https://huggingface.co/lintw/HealthGPT-M3	Uses VL-Health, whose data description includes PubMedVision, LLaVA-Med, PathVQA, MIMIC-CXR-VQA, SLAKE, VQA-RAD, and other medical VQA resources; VL-Health covers 11 modalities, including OCT and fundus photography.	Ophthalmology (OCT)
MedGemma-4B	https://developers.google.com/health-ai-developer-\foundations/medgemma/model-card-v1	Includes retinal fundus images; the EyePACS component contains retinal fundus photography only.	Ophthalmology (fundus only)
Lingshu-7B	https://arxiv.org/abs/2506.07044	The training data are reported to include ophthalmic OCT.	Ophthalmology (OCT)
Lingshu-32B	https://arxiv.org/abs/2506.07044	The supplied notes mark the model as using the same series-level training-data scope as Lingshu-7B.	Ophthalmology (OCT, inferred)
Hulu-Med-7B	https://arxiv.org/abs/2510.08668	Stage 2 includes Fundus_OCT_synthetic; the modality coverage includes OCT and ophthalmic imaging, with OCT, fundus photography, FA, and ICGA explicitly listed.	Ophthalmology (OCT)
Hulu-Med-32B	https://arxiv.org/abs/2510.08668	Uses the same disclosed training-data scope as Hulu-Med-7B.	Ophthalmology (OCT)
MediX-R1-8B	https://arxiv.org/abs/2602.23363	Reports about 51K instruction examples across multiple medical-imaging modalities, explicitly including OCT and fundus imaging.	Ophthalmology (OCT)
Fleming-VL-8B	https://arxiv.org/abs/2511.00916	Includes EyePACS and Harvard-FairVLMed; the data table lists ophthalmology/EyePACS fundus images and FairVLMed fundus images, but no OCT evidence is identified.	Ophthalmology (fundus only)
HealthGPT-Pro-8B	https://huggingface.co/lintw/HealthGPT-Pro-8B	The supplied source notes state that OCT data are included.	Ophthalmology (OCT)

Note. Unknown means that the available public material does not provide a dataset-level disclosure sufficient to verify OCT or ophthalmology exposure; it should not be interpreted as evidence that no such data were used. *Ophthalmology (fundus only)* indicates disclosed retinal/fundus data without clear OCT evidence. *Ophthalmology (OCT, inferred)* denotes same-series evidence rather than a directly itemized model-specific statement.

A.5 PROMPTS

Prompt Template 1: T01: Task-guided VQA construction prompt

You are an expert in visual question–answer (VQA) dataset construction.

Your task is to generate a high–quality multiple–choice question (MCQ) based on a given image.

Task Definition:
This task evaluates basic modality recognition (T01), which requires identifying the imaging modality based solely on visual appearance.

Requirements:
1. The question must explicitly ask about the imaging modality of the given image.
2. The question must be answerable purely from visual cues, without requiring any domain–specific or medical knowledge.
3. Use clear, concise, and unambiguous language. Avoid subjective or vague expressions.
4. Provide exactly four answer choices (A, B, C, D), with only one correct answer.
5. The correct answer must be "OCT".
6. The incorrect options (distractors) must be plausible alternative imaging modalities (e.g., CT, MRI, Ultrasound, X–ray).
7. Ensure that all options belong to the same category (i.e., imaging modalities) and are mutually exclusive.
8. Avoid biasing cues such as overly obvious wording or length differences between options.
9. Do not include explanations or additional text.

Output Format (strictly follow JSON):

```
{
  "question": "...",
  "options": {
    "A": "...",
    "B": "...",
    "C": "...",
    "D": "..."
  },
  "answer": "X"
}
```

Prompt Template 2: T02: Task-guided VQA construction prompt

You are an expert in visual question–answer (VQA) dataset construction.

Your task is to generate a high–quality multiple–choice question (MCQ) based on a given image.

Task Definition:
This task evaluates basic annotation recognition (T02), focusing only on identifying the color or shape of visual annotations (e.g., bounding boxes).

Requirements:
1. The question must refer ONLY to the annotation (e.g., boxes or markers), NOT the image content.
2. The question must be answerable purely from low–level visual perception.
3. Do NOT use or imply any medical or semantic knowledge.
4. The question must ask about ONLY ONE attribute:
   – color OR
   – shape
5. Allowed question types:
   – Color: e.g., "What is the color of the annotation box?"
   – Shape: e.g., "What is the shape of the annotation box?"
6. Provide exactly four answer choices (A, B, C, D), with only one correct answer.
7. Ensure that all answer options belong to the same category:
   – all colors OR all shapes (no mixing).
8. Use simple, clear, and unambiguous language.
9. Avoid biasing cues (e.g., option length differences or obvious answers).

10. The correct answer must be directly observable from the image.
11. Do not include explanations or additional text.

Output Format (strictly follow JSON):

```
{
  "question": "...",
  "options": {
    "A": "...",
    "B": "...",
    "C": "...",
    "D": "..."
  },
  "answer": "X"
}
```

Prompt Template 3: T03: Task-guided VQA construction prompt

You are an expert in visual question–answer (VQA) dataset construction.

Your task is to generate a high–quality multiple–choice question (MCQ) based on a given image and context.

Task Definition:
This task evaluates morphology recognition (T03), which requires identifying the visual morphological pattern of a specified annotated lesion region.

Context:
{context}

The context may include the total number of annotated boxes, e.g., "box_num": 6.

Box Selection Rule:
1. You must first identify the MOST visually prominent annotated box in the image.
2. Prominence can be determined by factors such as:
   – larger size
   – clearer structure
   – higher contrast
   – more distinct shape
3. Among all boxes, you must ONLY select a box whose morphology is clearly:
   – linear OR
   – elevated
4. If multiple boxes satisfy the condition, select the most visually prominent one.
5. If no box clearly matches these two patterns, do NOT generate a sample.
6. The selected box ID must be an integer between 1 and box_num.

Question Requirements:
1. The question must explicitly refer to the selected box by its ID.
2. The question must ask ONLY about the morphological pattern of that box.
3. The question must be answerable purely from visual appearance.
4. Do NOT use or imply any diagnosis, anatomy, or medical reasoning.
5. Use simple, clear, and unambiguous language.

Answer Options:
1. Provide exactly four answer choices (A, B, C, D), with only one correct answer.
2. All options must belong to the same category: morphological patterns.

Allowed option set:
– dot–like
– mass–like
– elevated
– linear

3. Do NOT use geometric shape terms (e.g., circular, oval).

4. Do NOT include annotation properties (e.g., color, position).

Quality Constraints:
1. The correct answer must be directly observable from the selected box.
2. Avoid biasing cues (e.g., option length differences).
3. Ensure the task requires no reasoning beyond direct visual perception.

Do not include explanations or additional text.

Output Format (strictly follow JSON):

```
{
  "question": "...",
  "target_box_id": X,
  "options": {
    "A": "...",
    "B": "...",
    "C": "...",
    "D": "..."
  },
  "answer": "X"
}
```

Prompt Template 4: T04: Task-guided VQA construction prompt

You are an expert in visual question−answer (VQA) dataset construction.

Your task is to generate a high−quality multiple−choice question (MCQ) based on a given image and context.

Task Definition:
This task evaluates lesion boundary feature recognition (T04), which requires identifying the boundary characteristics of the lesion inside a specified annotated box.

Context:
{context}

The context may include the total number of annotated boxes, e.g., "box_num": 6.

Box Selection Rule:
1. You must select ONE visible annotated box in the image.
2. Prefer the box whose lesion boundary characteristics are most clearly distinguishable.
3. The selected box ID must be an integer between 1 and box_num.
4. Do NOT select an ambiguous or unclear box.
5. If no box has clearly distinguishable lesion boundary characteristics, do NOT generate a sample.

Question Requirements:
1. The question must explicitly mention the selected annotation box ID.
   Example: "What is the boundary characteristic of the lesion inside annotation box 2?"
2. The question must ask ONLY about the boundary characteristics of the lesion inside that specified box.
3. The question must be answerable purely from visual perception.
4. Do NOT use or imply any diagnosis, anatomy, function, or medical reasoning.
5. Use simple, clear, and unambiguous language.

Answer Options:
1. Provide exactly four answer choices (A, B, C, D), with only one correct answer.
2. All options must belong to the same category: boundary characteristics.

Allowed option set:
− smooth
− irregular
− absent
− cannot determine

3. Use the allowed option set exactly as written.
4. Do NOT use terms such as well−defined, blurred, indistinct, clear, or unclear.
5. Do NOT use morphology terms such as dot−like, mass−like, elevated, or linear.
6. Do NOT use annotation properties such as color, size, or position.

Quality Constraints:
1. The correct answer must be directly observable from the lesion inside the specified annotation box.
2. Avoid biasing cues such as option length differences.
3. Ensure the task requires no reasoning beyond direct visual perception.
4. Only use "cannot determine" when the lesion boundary cannot be visually assessed from the image.
5. Only use "absent" when no visible lesion boundary is present.

Do not include explanations or additional text.

Output Format (strictly follow JSON):

```
{
  "question": "...",
  "target_box_id": X,
  "options": {
    "A": "...",
    "B": "...",
    "C": "...",
    "D": "..."
  },
  "answer": "X"
}
```

Prompt Template 5: T05: Task-guided VQA construction prompt

You are an expert in visual question−answer (VQA) dataset construction.

Your task is to generate a high−quality multiple−choice question (MCQ) based on a given OCT image.

Task Definition:
This task evaluates reflectivity recognition (T05), which requires identifying the brightness/reflectivity characteristic of a specified color−annotated region.

The image may contain color annotations for different regions:
1. SRF: Blue
2. PED: Orange
3. IRF: Red
4. SHRM: Purple
5. IS/OS: Green

Region Selection Rule:
1. Select ONE clearly visible color−annotated region in the image.
2. The question must refer to the region by its annotation color, not by medical label.
3. ONLY the following colors are allowed for questioning:
   − blue
   − orange
   − red
   − purple
4. Do NOT generate questions about the green region.
5. Prefer regions whose reflectivity is visually clear and easy to judge.
6. Do NOT select a region if its reflectivity is ambiguous.
7. If no valid region has clearly distinguishable reflectivity, do NOT generate a sample.

Question Requirements:
1. The question must explicitly ask about the reflectivity characteristic of the selected color region.
2. The question must be answerable from visual brightness only.

3. Do NOT ask about disease type, anatomy, lesion name, function, treatment, or diagnosis.
4. Do NOT use medical labels such as SRF, PED, IRF, SHRM, or IS/OS in the question.
5. Use simple, clear, and unambiguous language.
6. Example question:
"What is the reflectivity characteristic of the red region in the image?"

Answer Options:
1. Provide exactly four answer choices (A, B, C, D), with only one correct answer.
2. All options must belong to the same category: reflectivity / brightness characteristics.

Allowed option set:
– hyperreflective (brighter)
– hyporeflective (darker)
– similar to surrounding tissue
– cannot be determined

3. Do NOT use morphology terms such as dot–like, mass–like, elevated, or linear.
4. Do NOT use boundary terms such as blurred or irregular.
5. Do NOT use annotation properties such as color, size, or position as answer options.

Quality Constraints:
1. The correct answer must be directly observable from the selected color–annotated region.
2. Avoid biasing cues such as option length imbalance.
3. Ensure the task requires no reasoning beyond direct visual perception.

Do not include explanations or additional text.

Output Format (strictly follow JSON):

```
{
  "question": "...",
  "target_color": "...",
  "options": {
    "A": "...",
    "B": "...",
    "C": "...",
    "D": "..."
  },
  "answer": "X"
}
```

Prompt Template 6: T06: Task-guided VQA construction prompt

You are an expert in visual question–answer (VQA) dataset construction.

Your task is to generate a high–quality multiple–choice question (MCQ) based on a given OCT image and context.

Task Definition:
This task evaluates quantity perception (T06), which requires counting the number of annotated regions in the image.

Context:
{context}

The context may include:

```
{
  "box_num": 4
}
```

The value of "box_num" indicates the total number of annotated boxes in the image.

Question Requirements:
1. The question must ask about the total number of annotated regions or boxes in the image.

2. The question must be answerable purely from direct visual counting.
3. Do NOT require any medical knowledge, diagnosis, or semantic interpretation.
4. Do NOT ask about lesion type, anatomy, reflectivity, morphology, or boundary characteristics.
5. Use simple, clear, and unambiguous language.
6. Example questions:
   – "How many annotated regions are shown in the image?"
   – "What is the total number of annotation boxes in the image?"

Answer Options:
1. Provide exactly four answer choices (A, B, C, D), with only one correct answer.
2. All options must belong to the same category: integer counts.
3. The correct answer must equal the provided "box_num".
4. Incorrect options must be plausible nearby integers.
5. Avoid duplicated options.
6. Avoid obvious ordering bias.

Quality Constraints:
1. The correct answer must be directly verifiable by counting visible annotations.
2. Ensure the task requires no reasoning beyond direct visual perception.
3. If the number of boxes is visually ambiguous, do NOT generate a sample.
4. Avoid adding explanations or extra text.

Output Format (strictly follow JSON):

```
{
  "question": "...",
  "options": {
    "A": "...",
    "B": "...",
    "C": "...",
    "D": "..."
  },
  "answer": "X"
}
```

Prompt Template 7: T07: Task-guided VQA construction prompt

You are an expert in visual question–answer (VQA) dataset construction.

Your task is to generate a high–quality multiple–choice question (MCQ) based on a given OCT image and context.

Task Definition:
This task evaluates scale perception (T07), which requires comparing the relative sizes of annotated regions in the image.

Context:
{context}

The context may include:

```
{
  "label": "3214"
}
```

The value of "label" represents the ordering of annotation box IDs from largest to smallest area.
For example:
"3214" means:
– box 3 is the largest
– box 2 is the second largest
– box 1 is the third largest
– box 4 is the smallest

Box Pair Selection Rule:
1. You must select TWO annotation boxes with clearly distinguishable size differences.

2. Prefer comparisons between:
   – the largest and smallest boxes
   – boxes with obvious area differences
3. Do NOT compare boxes with visually similar sizes.
4. If no box pair has a clearly distinguishable size difference, do NOT generate a sample.

Question Requirements:
1. The question must explicitly compare TWO annotation boxes by their IDs.
2. The question must ask ONLY about relative size.
3. The question must be answerable purely from visual perception.
4. Do NOT require medical knowledge, diagnosis, or semantic interpretation.
5. Use simple, clear, and unambiguous language.
6. Example question:
   "Compared with annotation box 2, which region is larger?"

Answer Options:
1. Provide exactly four answer choices (A, B, C, D), with only one correct answer.

Required option format:
– A. Box X is larger
– B. Box Y is larger
– C. The two boxes are similar in size
– D. Cannot be determined

2. The correct answer must be consistent with the provided size ordering label.
3. Do NOT generate samples where the correct answer should be C or D.

Quality Constraints:
1. The size difference between the selected boxes must be visually obvious.
2. Ensure the task requires no reasoning beyond direct visual perception.
3. Avoid ambiguous or borderline comparisons.
4. Avoid explanations or additional text.

Output Format (strictly follow JSON):

```
{
  "question": "...",
  "compared_boxes": [X, Y],
  "options": {
    "A": "...",
    "B": "...",
    "C": "...",
    "D": "..."
  },
  "answer": "X"
}
```

Prompt Template 8: T08: Task-guided VQA construction prompt

You are an expert in visual question–answer (VQA) dataset construction.

Your task is to generate a high–quality multiple–choice question (MCQ) based on a given OCT image and context.

Task Definition:
This task evaluates spatial orientation perception (T08), which requires identifying the relative spatial position between annotated regions in the image.

The image may contain multiple annotation boxes with numeric IDs.

Box Selection Rule:
1. You must select TWO clearly visible annotation boxes.
2. The spatial relationship between the selected boxes must be visually obvious.
3. Prefer box pairs with clear diagonal separation.

4. Do NOT select overlapping or spatially ambiguous box pairs.
5. If no box pair has a clearly distinguishable diagonal spatial relationship, do NOT generate a sample .

Question Requirements:
1. The question must explicitly mention the annotation box IDs.
2. The question must ask ONLY about the relative spatial position between the selected boxes.
3. The question must be answerable purely from direct visual perception.
4. Do NOT require medical knowledge, diagnosis, or semantic interpretation.
5. Do NOT ask about lesion type, morphology, reflectivity, size, or boundary characteristics.
6. Use simple, clear, and unambiguous language.

Allowed Spatial Relations:
– upper–left
– upper–right
– lower–left
– lower–right

7. ONLY diagonal spatial directions are allowed.
8. Do NOT use:
  – left
  – right
  – above
  – below

Allowed Question Examples:
– "Where is annotation box 1 located relative to annotation box 2?"
– "Compared with annotation box 3, where is annotation box 1 located?"

Answer Options:
1. Provide exactly four answer choices (A, B, C, D), with only one correct answer.
2. All options must belong to the same category: diagonal spatial directions.
3. The four options must be:
  – upper–left
  – upper–right
  – lower–left
  – lower–right
4. Use visually plausible distractors.
5. Prefer questions with visually obvious spatial relationships.

Quality Constraints:
1. The correct answer must be directly verifiable from the image.
2. Ensure the task requires no reasoning beyond direct visual perception.
3. Avoid ambiguous or borderline spatial relationships.
4. Avoid explanations or additional text.

Output Format (strictly follow JSON):

```
{
  "question": "...",
  "compared_boxes": [X, Y],
  "options": {
    "A": "upper–left",
    "B": "upper–right",
    "C": "lower–left",
    "D": "lower–right"
  },
  "answer": "X"
}
```

Prompt Template 9: T09: Task-guided VQA construction prompt

You are an expert in medical visual question–answer (VQA) dataset construction.

Your task is to generate a high−quality multiple−choice question (MCQ) based on a given OCT image and context.

Task Definition:
This task evaluates anatomical region recognition (T09), which requires identifying a specific anatomical region in an OCT image based on medical visual knowledge.

Context:
{context}

The context may include:
{
"label": "Choroid"
}

The value of "label" indicates the target anatomical region shown in the image.

Region Selection Rule:
1. You must focus on the most clearly visible target anatomical region corresponding to the provided label.
2. The selected region should be visually distinguishable and medically recognizable.
3. If the target region is unclear or ambiguous, do NOT generate a sample.

Question Requirements:
1. The question must ask about the identity of the anatomical region.
2. The question must require basic anatomical knowledge of OCT images.
3. The question must be answerable from the visual appearance and structural location of the region.
4. Use simple, clear, and unambiguous language.
5. Example questions:
  – "Which anatomical region is shown in the image?"
  – "What is the anatomical structure visible in this OCT image?"

Answer Options:
1. Provide exactly four answer choices (A, B, C, D), with only one correct answer.
2. The four answer options must ALWAYS be:
  – Choroid
  – Retina
  – Vitreous
  – Cornea
3. The correct answer must exactly match the provided label.
4. Randomize the option order to avoid positional bias.

Quality Constraints:
1. The correct answer must be visually and anatomically identifiable.
2. Avoid ambiguous anatomical regions.
3. Ensure the task evaluates anatomical cognition rather than low−level perception.
4. Do not include explanations or additional text.

Output Format (strictly follow JSON):
{
"question": "...",
"label": "...",
"options": {
"A": "...",
"B": "...",
"C": "...",
"D": "..."
},
"answer": "X"
}

Prompt Template 10: T10: Task-guided VQA construction prompt

You are an expert in medical visual question–answer (VQA) dataset construction.

Your task is to generate a high–quality multiple–choice question (MCQ) based on a given OCT image and context.

Task Definition:
This task evaluates retinal layer recognition (T10), which requires identifying the retinal layer corresponding to the highlighted or annotated region in an OCT image.

Context:
{context}

The context may include:
{
"label": "IS/OS"
}

Possible labels include:
– IBRPE
– ILM
– IS/OS
– OBRPE
– OPL

The value of "label" indicates the correct retinal layer.

Layer Selection Rule:
1. The highlighted or annotated layer must be clearly visible and anatomically distinguishable.
2. If the highlighted layer is unclear or ambiguous, do NOT generate a sample.

Question Requirements:
1. The question must explicitly ask which retinal layer is highlighted or annotated in the image.
2. The question must require anatomical knowledge of OCT retinal layers.
3. The question must be answerable from the visual appearance and relative position of the highlighted layer.
4. Use simple, clear, and unambiguous language.
5. Example questions:
   – "Which retinal layer is highlighted in the image?"
   – "What retinal layer does the annotated region correspond to?"

Answer Options:
1. Provide exactly four answer choices (A, B, C, D), with only one correct answer.
2. All options must belong to the same category: retinal layers.
3. Candidate answer pool:
   – IBRPE (Inner Boundary of Retinal Pigment Epithelium)
   – ILM (Internal Limiting Membrane)
   – IS/OS (Inner Segment / Outer Segment Junction)
   – OBRPE (Outer Boundary of Retinal Pigment Epithelium)
   – OPL (Outer Plexiform Layer)
4. The correct answer must exactly match the provided label.
5. Incorrect options must be selected from the remaining retinal layers.
6. Each option must include BOTH:
   – abbreviation
   – full name explanation
7. Example option format:
   "ILM (Internal Limiting Membrane)"
8. Randomize the option order to avoid positional bias.

Quality Constraints:
1. The correct answer must be visually and anatomically identifiable.
2. Avoid ambiguous or visually indistinguishable layers.
3. Ensure the task evaluates retinal anatomical cognition rather than low–level perception.

4. Do not include explanations or additional text outside the options.

Output Format (strictly follow JSON):

```
{
  "question": "...",
  "label": "...",
  "options": {
    "A": "...",
    "B": "...",
    "C": "...",
    "D": "..."
  },
  "answer": "X"
}
```

Prompt Template 11: T11: Task-guided VQA construction prompt

You are an expert in medical visual question−answer (VQA) dataset construction.

Your task is to generate a high−quality multiple−choice question (MCQ) based on a given OCT image and context.

Task Definition:
This task evaluates retinal layer relationship understanding (T11), which requires identifying the layer immediately above or below a highlighted retinal layer in an OCT image.

Context:
{context}

The context may include:

```
{
  "label": "IS/OS"
}
```

Possible labels include:
− ILM
− OPL
− IS/OS
− IBRPE
− OBRPE

Layer Order Rule:
The retinal layers are ordered from top to bottom as:
ILM −> OPL −> IS/OS −> IBRPE −> OBRPE

The value of "label" indicates the retinal layer highlighted or annotated in red.

Question Generation Rule:
1. The question must refer to the red highlighted/annotated region.
2. The question must ask which layer is immediately above OR immediately below the highlighted layer.
3. Choose only one direction: above OR below.
4. The queried adjacent layer must exist.
   − If label = ILM, do NOT ask about the layer above it.
   − If label = OBRPE, do NOT ask about the layer below it.
5. The correct answer must be determined strictly according to the layer order rule.

Question Requirements:
1. The question must explicitly mention the red highlighted/annotated region.
2. The question must ask about the immediately adjacent layer, not a distant layer.
3. The question must require anatomical knowledge of OCT retinal layer order.
4. Use simple, clear, and unambiguous language.
5. Example questions:

– "Which layer is immediately below the red highlighted region?"
– "Which retinal layer lies directly above the red annotated region?"

Answer Options:
1. Provide exactly four answer choices (A, B, C, D), with only one correct answer.
2. All options must belong to the same category: retinal layers.
3. Candidate answer pool:
– ILM (Internal Limiting Membrane)
– OPL (Outer Plexiform Layer)
– IS/OS (Inner Segment / Outer Segment Junction)
– IBRPE (Inner Boundary of Retinal Pigment Epithelium)
– OBRPE (Outer Boundary of Retinal Pigment Epithelium)
4. Each option must include BOTH:
– abbreviation
– full name explanation
5. Example option format:
"IS/OS (Inner Segment / Outer Segment Junction)"
6. The correct answer must be the immediately adjacent layer according to the direction asked.
7. Incorrect options must be selected from the remaining retinal layers.
8. Randomize the option order to avoid positional bias.

Quality Constraints:
1. The correct answer must be anatomically valid according to the specified layer order.
2. Avoid ambiguous wording such as "near" or "around".
3. Do NOT ask about diagnosis, disease, function, treatment, or pathology.
4. Do not include explanations or additional text outside the JSON.

Output Format (strictly follow JSON):

```
{
  "question": "...",
  "highlighted_layer": "...",
  "relation": "above OR below",
  "options": {
    "A": "...",
    "B": "...",
    "C": "...",
    "D": "..."
  },
  "answer": "X"
}
```

Prompt Template 12: T12 AMD-SD: Task-guided VQA construction prompt

You are an expert in medical visual question–answer (VQA) dataset construction.

Your task is to generate a high–quality multiple–choice question (MCQ) based on a given OCT image and context.

Task Definition:
This task evaluates lesion classification (T12), which requires identifying the lesion type represented by a specific color–highlighted region in the image.

Fixed Color–to–Lesion Mapping:
– Blue –> SRF
– Orange –> PED
– Red –> IRF
– Purple –> SHRM

Lesion Definitions:
– SRF: Subretinal Fluid
– PED: Pigment Epithelial Detachment
– IRF: Intraretinal Fluid
– SHRM: Subretinal Hyperreflective Material

Region Selection Rule:
1. You must select ONE clearly visible highlighted color region that actually exists in the image.
2. The selected color MUST be one of:
   – Blue
   – Orange
   – Red
   – Purple
3. The selected color MUST visibly appear in the image.
4. Prefer the most visually prominent highlighted region.
5. Prominence can be determined by:
   – larger highlighted area
   – clearer structure
   – stronger visual saliency
6. Do NOT select weak, tiny, or ambiguous highlighted regions.
7. If no clearly visible highlighted region exists, do NOT generate a sample.

Question Requirements:
1. The question MUST explicitly mention the selected color.
2. The question MUST ask what lesion type the specified color–highlighted region represents.
3. The question must require basic OCT pathological knowledge.
4. The question must be answerable from the color annotation.
5. Use simple, clear, and unambiguous language.
6. Example questions:
   – ”What lesion does the red highlighted region represent?”
   – ”Which lesion type corresponds to the blue region in the image?”

Answer Options:
1. Provide exactly four answer choices (A, B, C, D), with only one correct answer.
2. All options must belong to the same category: OCT lesion types.
3. The four answer options must ALWAYS be:
   – SRF (Subretinal Fluid)
   – PED (Pigment Epithelial Detachment)
   – IRF (Intraretinal Fluid)
   – SHRM (Subretinal Hyperreflective Material)
4. The correct answer must match the lesion type corresponding to the selected color.
5. Randomize the option order to avoid positional bias.

Quality Constraints:
1. The selected color must visibly exist in the image.
2. The correct answer must be visually and semantically identifiable.
3. Avoid ambiguous or barely visible highlighted regions.
4. Ensure the task evaluates pathological cognition rather than low–level perception.
5. Do NOT ask about reflectivity, morphology, boundary, or size.
6. Do not include explanations or additional text outside the JSON.

Output Format (strictly follow JSON):

```
{
  ”question”: ”...”,
  ”target_color”: ”...”,
  ”label”: ”...”,
  ”options”: {
    ”A”: ”...”,
    ”B”: ”...”,
    ”C”: ”...”,
    ”D”: ”...”
  },
  ”answer”: ”X”
}
```

Prompt Template 13: T13: Task-guided VQA construction prompt

You are an expert in medical visual question–answer (VQA) dataset construction.

Your task is to generate a high–quality multiple–choice question (MCQ) based on a given OCT image.

Task Definition:
This task evaluates retinal structural status assessment (T13), which requires assessing the structural condition of visible retinal anatomy according to OCT morphology.

Question Requirements:
1. Generate ONE multiple–choice question asking about the structural status of a visible retinal structure or retinal architecture in the image.

2. The question must focus on structural condition, morphology, continuity, organization, or preservation status visible in OCT.

3. The question should assess higher–level OCT structural understanding rather than simple abnormality detection.

4. The question must be answerable directly from visible morphology.

5. Preferred question style:
– ”What is the structural status of the retina?”
– ”What is the structural status of the foveal architecture?”
– ”What is the structural status of the macular structure?”
– ”What best describes the retinal structural organization?”

6. The question MAY refer to anatomically visible structures (e.g., retina, fovea, macular architecture, choroid), but should NOT require retinal layer identification.

7. Do NOT ask about:
– lesion classification
– disease diagnosis
– retinal layer naming
– treatment
– prognosis
– reflectivity (e.g., hyperreflective/hyporeflective)
– simple yes/no abnormality detection

8. Use simple, clear, medically precise language.

Answer Option Requirements:
1. Provide exactly four answer choices (A, B, C, D).

2. Only ONE option must be correct.

3. All incorrect options must remain medically plausible and visually reasonable (avoid obviously wrong distractors).

4. The options should represent a reasonable semantic progression of structural severity.

5. Prefer the following structural–status template:
– Structurally intact and continuous with preserved morphology
– Mildly irregular but generally preserved
– Structurally disorganized with localized disruption
– Severely disrupted with a full–thickness defect or marked structural loss

6. Modify wording if needed to better match image appearance, but preserve semantic consistency.

7. Avoid trivial distractors such as:
– Completely structurally absent
– Totally normal without any visible change

– Impossible anatomy

8. Randomize the option order ONLY if semantic coherence remains natural.

Quality Constraints:
1. The correct answer must be visually supported by OCT morphology.

2. The question should require structural reasoning, not keyword matching.

3. Avoid ambiguity or borderline interpretations.

4. Prefer nuanced distinctions between preserved / mildly abnormal / disorganized / severely disrupted structure.

5. Ensure benchmark consistency across samples.

6. Do not include explanations or additional text outside JSON.

Output Format (strictly follow JSON):

```
{
  "question": "...",
  "options": {
    "A": "...",
    "B": "...",
    "C": "...",
    "D": "..."
  },
  "answer": "X"
}
```

Prompt Template 14: T14: Task-guided VQA construction prompt

You are an expert in medical visual question–answer (VQA) dataset construction.

Your task is to generate a high–quality multiple–choice question (MCQ) based on a given OCT image and context.

Task Definition:
This task evaluates lesion localization (T14), which requires identifying the anatomical location of a specified lesion region in an OCT image.

The context may include annotated lesion boxes or highlighted lesion regions.

Region Selection Rule:
1. You must select ONE clearly visible lesion region or annotated box in the image.
2. Prefer the most visually prominent lesion region.
3. Prominence can be determined by:
   – larger lesion area
   – clearer structural appearance
   – stronger visual saliency
4. The selected lesion region must visibly exist in the image.
5. Do NOT select weak, tiny, or ambiguous lesions.
6. If multiple lesions are similarly prominent, select the one with the clearest anatomical location.
7. If no clearly identifiable lesion exists, do NOT generate a sample.

Question Requirements:
1. The question must explicitly refer to the selected lesion region or annotation box.
2. The question must ask about the anatomical location of the lesion.
3. The question must require basic OCT anatomical understanding.
4. The question must be answerable from the lesion position relative to retinal structures.
5. Use simple, clear, and unambiguous language.
6. Example questions:

– "Where is the lesion in annotation box 1 mainly located?"
– "What is the anatomical location of the highlighted lesion?"

Answer Options:
1. Provide exactly four answer choices (A, B, C, D), with only one correct answer.
2. All options must belong to the same category: anatomical lesion locations.

Allowed option pool:
– Inner retina
– Outer retina
– Vitreous
– Cornea

3. The correct answer must be directly supported by the lesion position in the OCT image.
4. Incorrect options must be anatomically plausible.
5. Randomize the option order to avoid positional bias.

Quality Constraints:
1. The lesion location must be visually identifiable.
2. Avoid ambiguous or borderline lesion locations.
3. Ensure the task evaluates anatomical cognition rather than low−level perception.
4. Do NOT ask about lesion type, reflectivity, morphology, treatment, or prognosis.
5. Do not include explanations or additional text outside the JSON.

Output Format (strictly follow JSON):

```
{
  "question": "...",
  "target_box_id": X,
  "options": {
    "A": "...",
    "B": "...",
    "C": "...",
    "D": "..."
  },
  "answer": "X"
}
```

Prompt Template 15: T15_AMD_SD: Task-guided VQA construction prompt

You are an expert in medical visual question−answer (VQA) dataset construction.

Your task is to generate a high−quality multiple−choice question (MCQ) based on a given OCT image .

Task Definition:
This task evaluates lesion−associated disease reasoning (T15), requiring recognition of a highlighted lesion region and inference of the disease most typically associated with that lesion.

You will be provided with:

1. An OCT image
2. Internal context information:

{context}

The context contains:
– highlighted lesion color
– lesion label

IMPORTANT:
The context is INTERNAL REFERENCE ONLY and must be used to determine the correct answer.
NEVER explicitly reveal the lesion name from the context in the question or answer options.

Ground−truth lesion−disease association (STRICT):
IRF −> DME
SRF −> CSC
PED −> wet AMD
SHRM −> wet AMD
Soft Drusen −> dry AMD
Hard Drusen −> dry AMD
Photoreceptor layer disruption −> AMD
Reticular drusen −> dry AMD
Choroidal fold −> Hypotony / Orbital mass
Geographic atrophy −> dry AMD

Allowed disease labels (STRICT):
wet AMD, dry AMD, AMD, DME, CSC, CNV, DR, ERM, MH, RAO, RVO, Glaucoma, Hypotony / Orbital mass

Question Design Requirements:
1. Generate exactly ONE MCQ with four options (A, B, C, D) and one correct answer.
2. The question MUST ask for the disease MOST TYPICALLY associated with the highlighted lesion.
3. The question MUST explicitly mention the highlight color from the context
(e.g., ”purple region”, ”orange lesion”).
4. Vary wording naturally across samples; do NOT always use the same sentence template.
5. NEVER reveal lesion names
(e.g., IRF, SRF, PED, SHRM, drusen, photoreceptor disruption)
in the question or options.
6. The generated question should require:
image understanding −> lesion recognition −> disease association reasoning.
7. Use ONLY the allowed disease labels above as answer options.
8. Distractor diseases must be clinically plausible and non−trivial.
9. Distractors should preferably come from retinal OCT diseases with partially overlapping manifestations and avoid obviously unrelated diseases.
10. The correct answer MUST strictly follow the lesion−disease mapping above using the lesion label provided in {context}.
11. Use concise, objective, medically precise wording and provide NO explanation.

Output Format (strict JSON):

```
{
  ”question”: ”...”,
  ”options”: {
    ”A”: ”...”,
    ”B”: ”...”,
    ”C”: ”...”,
    ”D”: ”...”
  },
  ”answer”: ”X”
}
```

Prompt Template 16: T16: Task-guided VQA construction prompt

You are an expert in medical visual question−answer (VQA) dataset construction, specializing in ophthalmology and Optical Coherence Tomography (OCT).

Your task is to generate a high−quality multiple−choice question (MCQ) based on a given OCT image and context.

Task Definition:
This task evaluates functional impact assessment (T16), which requires recognizing a lesion or retinal structural abnormality in an OCT image and inferring the most likely visual functional abnormality associated with it.

Context:
{context}

The context may include:
{
"label": "SRF",
"target_color": "blue",
"target_box_id": 2
}

The value of "label" indicates the lesion or structural abnormality used as the internal ground−truth reference. The question should refer to the highlighted region, annotation color, or box ID when available, but should not directly reveal the ground−truth lesion label.

Ground−truth lesion/function mapping (STRICT):
IRF −> Central visual acuity reduction
SRF −> Metamorphopsia
PED −> Central visual function impairment
SHRM −> Reduced central visual sensitivity
Photoreceptor layer disruption −> Central vision loss
Reticular drusen −> Night vision impairment
Soft Drusen −> Night vision impairment / early visual dysfunction
Hard Drusen −> Mild contrast sensitivity reduction
Geographic atrophy −> Irreversible central vision loss
Choroidal fold −> Visual distortion

Allowed functional impact labels (STRICT):
− Central visual acuity reduction
− Metamorphopsia
− Central visual function impairment
− Reduced central visual sensitivity
− Central vision loss
− Night vision impairment
− Night vision impairment / early visual dysfunction
− Mild contrast sensitivity reduction
− Irreversible central vision loss
− Visual distortion

Question Requirements:
1. The question must ask about the most likely visual functional abnormality or symptom associated with the specified OCT finding.
2. The question must require medical reasoning from OCT lesion appearance to functional impact.
3. The question should refer to the highlighted region, annotation color, or box ID if such information is available in the context.
4. Do NOT reveal lesion names such as IRF, SRF, PED, SHRM, drusen, geographic atrophy, choroidal fold, or photoreceptor layer disruption in the question.
5. Do NOT ask about disease diagnosis, treatment selection, staging, retinal layer naming, reflectivity, boundary, size, or lesion classification.
6. Use concise, objective, medically precise wording.
7. Vary question wording naturally across samples; do NOT always use the same sentence template.

Answer Options:
1. Provide exactly four answer choices (A, B, C, D), with only one correct answer.
2. All options must belong to the same category: visual functional abnormalities or visual symptoms.
3. The correct answer must strictly follow the ground−truth lesion/function mapping above using the label provided in {context}.
4. Distractors must be clinically plausible and non−trivial.
5. Use ONLY the allowed functional impact labels above as answer options.
6. Randomize option order to avoid positional bias.

Quality Constraints:
1. The selected OCT finding must have a medically plausible functional impact.
2. Avoid ambiguous cases where multiple functional impacts would be equally correct.
3. Ensure the task evaluates functional reasoning rather than low−level perception alone.
4. Do not include explanations or additional text outside the JSON.

Output Format (strictly follow JSON):

```
{
  "question": "...",
  "label": "...",
  "options": {
    "A": "...",
    "B": "...",
    "C": "...",
    "D": "..."
  },
  "answer": "X"
}
```

Prompt Template 17: T17: Task-guided VQA construction prompt

You are an expert in medical visual question–answer (VQA) dataset construction.

Your task is to generate a high–quality multiple–choice question (MCQ) based on a given OCT image and context.

Task Definition:
This task evaluates disease diagnosis reasoning (T17), which requires identifying the most likely disease diagnosis from an OCT image.

Context:
{context}

The context may include:

```
{
  "label": "AMD"
}
```

The value of "label" indicates the correct disease diagnosis.

Question Requirements:
1. The question must ask about the most likely disease diagnosis based on the OCT image.
2. The question must require medical reasoning from OCT findings.
3. The question must be answerable from the image appearance and pathological patterns.
4. Use simple, clear, and unambiguous language.
5. Do NOT ask about lesion type, layer name, reflectivity, boundary, size, or treatment.
6. Example questions:
   – "Which disease is most consistent with this OCT image?"
   – "Based on this OCT image, what is the most likely diagnosis?"

Answer Options:
1. Provide exactly four answer choices (A, B, C, D), with only one correct answer.
2. All options must belong to the same category: retinal diseases.
3. The correct answer must exactly match the provided label.
4. Incorrect options must be plausible retinal disease diagnoses.
5. Randomize the option order to avoid positional bias.

Preferred disease option pool:
– AMD
– CNV
– CSC
– DME
– DR
– MH
– ERM
– RAO
– RVO
– Glaucoma

– NORMAL

Quality Constraints:
1. The correct answer must be consistent with the provided label.
2. Avoid ambiguous or borderline diagnostic cases.
3. Ensure the task evaluates disease−level reasoning rather than low−level visual perception.
4. Do not include explanations or additional text outside the JSON.

Output Format (strictly follow JSON):

```
{
  "question": "...",
  "label": "...",
  "options": {
    "A": "...",
    "B": "...",
    "C": "...",
    "D": "..."
  },
  "answer": "X"
}
```

Prompt Template 18: T18: Task-guided VQA construction prompt

You are an expert in medical visual question−answer (VQA) dataset construction.

Your task is to generate a high−quality multiple−choice question (MCQ) based on a given OCT image and context.

Task Definition:
This task evaluates disease stage grading (T18), which requires determining the stage of a macular hole based on OCT image appearance.

Context:
{context}

The context may include:

```
{
  "label": "Stage 2"
}
```

The value of "label" directly indicates the correct stage name.
Possible labels include:
– Stage 1
– Stage 2
– Stage 3
– Stage 4

Question Requirements:
1. The question must ask about the stage or grade of the macular hole shown in the OCT image.
2. The question must require medical reasoning from OCT structural appearance.
3. The question must be answerable from visible retinal deformation and hole morphology.
4. Use simple, clear, and unambiguous language.

5. Example questions:
  – "What is the stage of the macular hole shown in the OCT image?"
  – "Based on this OCT image, what is the most likely macular hole stage?"

Answer Options:
1. Provide exactly four answer choices (A, B, C, D), with only one correct answer.
2. All options must belong to the same category: macular hole stages.
3. The four answer options must ALWAYS be:
  – Stage 1
  – Stage 2

– Stage 3
– Stage 4
4. Randomize the option order to avoid positional bias.
5. The correct answer option MUST exactly match the provided label.

Answer Consistency Rule:
1. The value of "answer" MUST be the option letter:
– A
– B
– C
– D
2. Do NOT output:
– stage numbers
– option text
– labels
3. The selected answer letter MUST correspond exactly to the option containing the correct stage.
4. Verify consistency between:
– the provided label
– the option content
– the answer letter
before generating the final output.

Example:
– if label = "Stage 2"
– and option B = "Stage 2"
then:
"answer": "B"

Quality Constraints:
1. The correct answer must be visually supported by the OCT image appearance.
2. Avoid ambiguous or borderline staging cases.
3. Ensure the task evaluates disease grading reasoning rather than low−level perception.
4. Do not include explanations or additional text outside the JSON.

Output Format (strictly follow JSON):

```
{
  "question": "...",
  "label": "...",
  "options": {
    "A": "...",
    "B": "...",
    "C": "...",
    "D": "..."
  },
  "answer": "X"
}
```

Prompt Template 19: T19: Task-guided VQA construction prompt

You are an expert in medical visual question−answer (VQA) dataset construction, specializing in ophthalmology and Optical Coherence Tomography (OCT).

Task Definition
This task evaluates diagnosis−conditioned treatment reasoning (T19).
Based on the provided OCT image and the internal disease label, you need to generate a high−quality, clinically accurate multiple−choice question (MCQ).

The reasoning chain must follow:
Image Understanding (OCT features) −> Disease Recognition −> Treatment Selection

Inputs
1. OCT Image: (Provided via the multimodal interface)
2. Internal Context: {context} (A JSON containing the ground−truth disease category)

Ground−truth Disease−Treatment Mapping (STRICT)
− Dry AMD −> AREDS2 supplements + observation
− Wet AMD / CNV −> anti−VEGF
− CSC / CSR −> observation or PDT
− DME −> anti−VEGF
− MH −> PPV
− ERM −> PPV + membrane peeling
− RAO −> urgent stroke evaluation
− RVO −> anti−VEGF
− Glaucoma −> IOP−lowering therapy

Question Design Requirements

1. Structure: Generate exactly ONE MCQ with 4 options (A, B, C, D) and ONE correct answer.
2. Core Question: The question must ask for the most appropriate diagnosis−treatment pair based on the OCT findings. Do NOT explicitly mention the ground−truth disease name in the question stem.
3. Option Format: Every option MUST strictly follow the format: 'Disease −− Treatment' (e.g., "DME −− anti−VEGF"). Treatment descriptions must be concise and standardized.
4. Correct Option: Must pair the correct disease from the context with its corresponding first−line treatment from the mapping above.
5. Distractor Options (Crucial): Must be clinically plausible. Include:
   − The correct disease paired with an incorrect/suboptimal treatment.
   − A wrong disease (with overlapping OCT features) paired with its correct treatment.
   − A wrong disease paired with a wrong treatment.
6. No Shortcuts: Ensure that a language−only model cannot guess the answer by just looking at the treatment or the disease names alone.
7. Tone & Style: Use professional, objective, and medically precise English. Provide NO explanations or thoughts outside the JSON.

Output Format (Strict JSON)
Return ONLY a valid JSON object matching this schema:

```
{
  "question": "Which diagnosis−management pair best matches the findings demonstrated in this OCT image?",
  "options": {
    "A": "Disease −− Treatment",
    "B": "Disease −− Treatment",
    "C": "Disease −− Treatment",
    "D": "Disease −− Treatment"
  },
  "answer": "A"
}
```

Example
If Input Context is '{ "label": "DME" }':
Output:

```
{
  "question": "Based on the shown in the OCT image, which diagnosis−treatment pair is most appropriate?",
  "options": {
    "A": "DME −− anti−VEGF",
    "B": "DME −− Observation",
    "C": "DR −− PRP",
    "D": "CSC −− PDT"
  },
  "answer": "A"
}
```

Prompt Template 20: T20: Task-guided VQA construction prompt

You are an expert in medical visual question–answer (VQA) dataset construction, specializing in ophthalmology and longitudinal clinical reasoning.

Task Definition
This task evaluates diagnosis–conditioned prognosis and management reasoning (T20).

The question must create a hypothetical but realistic longitudinal clinical scenario. The answering model must first identify the baseline disease from the provided OCT image, read the hypothetical follow–up progression or treatment response described in the question, and then select the MOST appropriate management adjustment.

Inputs
1. OCT Image: (Provided via the multimodal interface; used solely to identify the baseline disease entity)
2. Internal Context: {context}
(A JSON containing ONLY the ground–truth disease category, e.g., { "label": "DME" })

Ground–truth Disease–Longitudinal Management Mapping (STRICT)
The question scenario and the correct answer (Option A) must strictly align with these established clinical evolution pathways:

– Wet AMD:
 – Hypothetical Change: Complete resolution of intraretinal/subretinal fluid and stable vision after initial loading injections.
 – Correct Adjustment: Extend the anti–VEGF injection interval with continued proactive monitoring (Treat–and–Extend).
– CSC:
 – Hypothetical Change: The subretinal fluid persists or worsens after 3 months of conservative observation, with declining visual acuity.
 – Correct Adjustment: Consider active intervention such as photodynamic therapy (PDT) or focal laser photocoagulation.
– DME:
 – Hypothetical Change: Intraretinal cystic spaces and macular thickening remain refractory or worsen despite multiple consecutive anti–VEGF injections.
 – Correct Adjustment: Switch to an alternative anti–VEGF agent or consider intravitreal corticosteroid implant therapy.
– MH:
 – Hypothetical Change: Following pars plana vitrectomy (PPV) and gas tamponade, the foveal contour is structurally restored and the hole is confirmed closed.
 – Correct Adjustment: Maintain regular outpatient follow–up to monitor visual recovery and structural stabilization.
– ERM:
 – Hypothetical Change: During subsequent follow–up visits, the patient describes progressive metamorphopsia (distortion) and a significant drop in visual acuity.
 – Correct Adjustment: Recommend surgical intervention with pars plana vitrectomy and epiretinal membrane peeling.
– RAO:
 – Hypothetical Change: The acute retinal edema resolves over the following weeks, transitioning into permanent inner retinal layer thinning and atrophy.
 – Correct Adjustment: Coordinate long–term care focusing on strict systemic vascular risk factor management and stroke prevention.
– RVO:
 – Hypothetical Change: Following successful resolution of macular edema, the patient returns months later complaining of a sudden new onset of blurred vision.
 – Correct Adjustment: Perform repeat OCT imaging to evaluate for recurrent macular edema and initiate prompt anti–VEGF retreatment if present.
– Glaucoma:
 – Hypothetical Change: Despite maintaining the current intraocular pressure (IOP)–lowering eye drops, sequential testing shows confirmed structural RNFL progression.
 – Correct Adjustment: Escalate therapy by adding another medication class, recommending selective laser trabeculoplasty (SLT), or planning incisional surgery.
– Normal:

– Hypothetical Change: The patient remains completely asymptomatic with no new visual complaints during subsequent annual evaluations.
– Correct Adjustment: Continue with routine annual comprehensive eye examinations and standard wellness monitoring.

Question Design Requirements
1. Structure: Generate exactly ONE MCQ with 4 options (A, B, C, D). Crucial: Always anchor the correct answer to option "A". (The backend system will shuffle options programmatically later).
2. Question Stem Style (The Longitudinal Progression Template):
– The question must introduce a hypothetical chronological shift or treatment response tailored to the diagnosed disease.
– CRITICAL: Do NOT explicitly mention the baseline disease name or its direct abbreviations (e.g ., "DME", "CSC") in the text. Refer to it neutrally as "the pathology shown", "this condition", or "the demonstrated findings".
3. The Distractor Blueprint:
– Option A (Correct): The guideline−recommended clinical adjustment matching the specific progression scenario.
– Distractor B (The Cross−Disease Trap): A valid adjustment for a ∗different∗ progression scenario belonging to another disease (tests if the model misidentified the baseline image).
– Distractor C (The Premature/Inappropriate Action): Recommending aggressive surgery when the scenario dictates conservative management, or vice versa.
– Distractor D (The Static Error): Blindly recommending to "maintain the current exact protocol without change" or "discharge the patient," ignoring the clear clinical progression described in the text.

Output Format (Strict JSON)
Return ONLY a valid JSON object with the correct answer anchored to "A":

```
{
  "question": "The patient with the condition demonstrated in this OCT image returns for follow−up. [Insert hypothetical progression/treatment response change]. What is the MOST appropriate adjustment to their management strategy?",
  "options": {
    "A": "[Correct clinical adjustment or follow−up protocol changes]",
    "B": "[Distractor: Valid adjustment for a completely different disease's timeline]",
    "C": "[Distractor: Inappropriately aggressive or contraindicated intervention]",
    "D": "[Distractor: Erroneously maintaining the status quo or discharging the patient]"
  },
  "answer": "A"
}
```

Few−Shot Example
If Input Context is: { "label": "CSC" }
Output:

```
{
  "question": "A patient presenting with the pathology demonstrated in this OCT image returns for evaluation 3 months later. Follow−up imaging reveals that the subretinal fluid has persisted without any spontaneous resolution, and the patient reports a progressive decline in visual acuity. What is the MOST appropriate next adjustment to their management strategy?",
  "options": {
    "A": "Consider active intervention with photodynamic therapy (PDT) or focal laser photocoagulation",
    "B": "Switch immediately to an intravitreal corticosteroid implant or alternative anti−VEGF agent",
    "C": "Proceed directly to emergency pars plana vitrectomy with internal limiting membrane peeling",
    "D": "Continue strict observation with lifestyle modifications for another 12 months without testing"
  },
  "answer": "A"
}
```