

Article

Beyond Accuracy: A Multi-dimensional Cognitive Audit of Medical Large Vision–Language Models in Fundus Image Interpretation

Jingling Zhang ^{1,2} , Shuting Zheng ² , Xiangfei Liu ^{1,3}  and Jia Gu ^{1,*} ¹ Institute of Data Science, City University of Macau, Macau 999078, China; d24092100270@cityu.edu.mo (J.Z.); d24092100224@cityu.edu.mo (X.L.)² Institute of Artificial Intelligence, Putian University, Putian 351100, China; fgtear@ptu.edu.cn³ Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

* Correspondence: jiagu@cityu.edu.mo

Abstract

Reliance on standalone accuracy limits credible assessment of fundus-focused large vision–language models (LVLMs), as high scores often stem from linguistic shortcuts rather than real visual reasoning. This work develops the Cognitive Audit Framework (CAF), a four-module automated auditing pipeline that dissects model reasoning flaws: Visual–Linguistic Decoupling (textual dependency via modality ablation), Hierarchical Logical Consistency (lesion–diagnosis contradiction detection), Reasoning Fidelity Gap (chain-of-thought unfaithfulness scoring), and Contextual Robustness (positional bias under option permutation). Experiments on six 7B–31B LVLMs over FunBench reveal a notable gap between benchmark accuracy and reasoning quality: high accuracy coexists with measurable textual dependency, logical inconsistencies across diagnostic levels, limited chain-of-thought faithfulness, and non-trivial positional sensitivity. CAF serves as a reproducible complement to pure accuracy metrics for validating clinical competence of ophthalmic multimodal models.

Keywords: large vision–language models; visual reasoning; cognitive audit framework; clinical comprehension

1. Introduction

Driven by large-scale foundation models pre-trained on millions of retinal photographs [1,2] and ophthalmology-oriented multimodal architectures [3–6], large vision–language models (LVLMs) have rapidly advanced on fundus visual question answering (VQA) tasks spanning diabetic retinopathy (DR), age-related macular degeneration (AMD), glaucoma, and other retinal disorders [7–9]. Classification accuracy has accordingly become the dominant metric for quantifying ophthalmic multimodal model performance. However, relying solely on accuracy for performance validation has inherent limitations: this metric only captures label-matching probability under a fixed data distribution and cannot causally distinguish whether a model’s predictions originate from genuine fundus pathological features or from shortcut learning that exploits prompt wording and option arrangement [10]. Meanwhile, clinical ophthalmic diagnosis strictly follows hierarchical pathological criteria established by ICDR [11] and ETDRS [12] standards, where specific retinal lesions serve as necessary preconditions for corresponding disease grades—a layered constraint that a single accuracy score cannot quantify. The resulting disconnect between reported benchmark performance and authentic clinical reasoning capability underscores the need for richer evaluation tools.



Academic Editor: Christos Bouras

Received: 9 May 2026

Revised: 7 June 2026

Accepted: 10 June 2026

Published: 15 June 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](#)[Attribution \(CC BY\) license](#).

Two mainstream approaches currently exist for fine-grained medical multimodal evaluation. The first is manual expert auditing, which provides high clinical validity but incurs prohibitive cost and poor scalability for large-scale assessment [13,14]. The second employs external LLMs as automatic judges [15], enabling automation but introducing non-reproducible evaluation variance and circular assessment bias—particularly problematic for reproducible clinical benchmarking. Beyond these two paradigms, existing automated radiology report evaluation methods [16–18] are constrained by task-format discrepancies and cannot be directly transferred to fundus VQA scenarios. On the reasoning verification front, prior works on chain-of-thought faithfulness and logical consistency are predominantly designed for general-purpose benchmarks [19,20], lacking integration with ophthalmic-specific grading rules and thus poorly suited to the hierarchical diagnostic logic of fundus diseases. In summary, the field lacks an automated, reproducible, multi-dimensional evaluation tool that incorporates ophthalmic clinical standards. To address this research gap, we develop the Cognitive Audit Framework (CAF).

Built upon the FunBench dataset [21], which provides fine-grained hierarchical annotations suitable for multi-dimensional auditing, this work constructs a fully automated perturbation-based evaluation pipeline comprising four independent audit modules: Visual–Linguistic Decoupling (Audit A), Hierarchical Logical Consistency (Audit B), Reasoning Fidelity Gap (Audit C), and Contextual Robustness (Audit D). Among these, Audits A, C, and D leverage custom-built ophthalmic keyword dictionaries covering DR, AMD, glaucoma, and other common fundus disorders to enable quantitative scoring; Audit B distills 17 pathological contradiction rules from ICDR/ETDRS diagnostic criteria, enforcing lesion-to-grade logical constraint verification focused on diabetic retinopathy.

Our core contributions are summarized as follows:

- **Multi-dimensional Cognitive Audit Framework.** We propose CAF, a four-module automated auditing pipeline that moves beyond the single-accuracy paradigm to quantify clinical reasoning deficits from complementary dimensions—textual dependency, hierarchical logical coherence, reasoning faithfulness, and positional robustness.
- **Deterministic Rule-based Clinical Scoring.** We designed a CoT-Scorer and a Wrong-Reasoning Rate metric that quantify chain-of-thought quality through fixed clinical rules without invoking any external LLM judge (Audit C), ensuring full experimental reproducibility.
- **Multi-model Empirical Validation.** We conduct systematic experiments on six mainstream LVLMs spanning 7B–31B parameters, providing comprehensive evidence that high benchmark accuracy can coexist with measurable deficits in textual dependency, logical consistency, reasoning faithfulness, and positional robustness.

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 details the CAF design, Section 4 presents experiments and results, Section 5 discusses findings and limitations, and Section 6 concludes.

2. Related Work

2.1. Large Vision–Language Models in Medical Imaging

Medical AI has rapidly evolved from task-specific classifiers [22,23] to generalist large vision–language models (LVLMs) [24,25]. Foundational systems (e.g., LLaVA-Med [3], Med-PaLM M [4]) and ophthalmic-specific architectures (e.g., RETFound [1], VisionUnite [26], FundusExpert [6]) now achieve robust zero-shot diagnostic alignment and fine-grained visual grounding. Despite these advances, their clinical readiness is predominantly assessed through Visual Question Answering (VQA) benchmarks [9,21,27–29] that prioritize aggregate accuracy. This narrow, VQA-centric lens fundamentally overlooks the rigorous standards of traditional clinical AI validation, clinical safety paradigms, and automated

medical report evaluation [13,14,16–18,30,31]. By conflating superficial label-matching with genuine pathophysiological understanding and ignoring hierarchical consistency, current evaluations lack the mechanistic transparency and algorithmic vigilance crucial for clinical trustworthiness, directly inspiring our multi-dimensional cognitive audit scheme.

2.2. Evaluation Paradigms: Beyond the Accuracy Bottleneck

The dominant paradigm for evaluating medical LVLMs relies on VQA-driven benchmarks (e.g., VQA-RAD [28], PMC-VQA [32], FunBench [21]). Even recent reasoning-oriented datasets (e.g., GEMEX [33], DrVD-Bench [9]) structurally default to scalar accuracy, which conflates genuine visual grounding with linguistic shortcuts and ignores hierarchical causal constraints. To overcome this bottleneck, evaluation must transcend simple VQA accuracy and synthesize with traditional clinical AI validation and clinical safety paradigms (e.g., FDA guidelines [13], TRIPOD+AI [14], algorithmic vigilance [30]), which demand continuous failure mode characterization. Furthermore, inspiration must be drawn from automated medical report evaluation (e.g., RadGraph [16], GREEN [17], FineRadScore [18]), where clinical entity grounding and fine-grained penalty for diagnostic errors strictly outweigh simple label matching. By integrating these rigorous clinical validation standards into the VQA setting, our proposed framework bridges the gap between static benchmark accuracy and the trustworthy reasoning necessary for clinical deployment.

2.3. Shortcut Learning, Position Bias, and CoT Faithfulness

Beyond scalar accuracy, deep learning models frequently succumb to shortcut learning [10], which in medical VQA architectures manifests as position biases [34–37] and reliance on spurious visual cues [38,39]. Furthermore, while chain-of-thought (CoT) prompting aims to improve interpretability [40,41], LVLMs frequently exhibit unfaithful, post hoc rationalizations disconnected from their final predictions [19,42], alongside cross-query logical contradictions [20,43–45]. In clinical environments, such hallucinated reasoning paths pose safety risks [24,46] that remain invisible to standard accuracy-based scoring. Distinct from prior work that probes externally induced bias on general benchmarks [19] or verifies domain-agnostic propositional logic [20], our CAF enforces pathophysiological constraints inherent to ophthalmology—specifically, ICDR/ETDRS-derived causal hierarchies [11,12] between lesion findings and disease grades—through clinically grounded, deterministic auditing, ensuring that diagnostic outputs align with valid clinical causal chains.

3. Methodology: The Cognitive Audit Framework

To rigorously evaluate the clinical reasoning reliability of large vision–language models (LVLMs), we designed the Cognitive Audit Framework (CAF). Drawing on the interventional logic of causal inference [47], CAF adopts perturbation-based probing as its central design principle: rather than measuring aggregate accuracy over a fixed test set, each audit module applies a targeted intervention—to the model’s input content or its output behavior—and quantifies the causal impact of that informational change on predictions. We employ the FunBench dataset [21] as our evaluation corpus, leveraging its hierarchical task structure and fine-grained annotations to construct the pathological logic constraints required by Audit B. CAF comprises four complementary audit modules: Visual–Linguistic Decoupling (Audit A), Hierarchical Logical Consistency (Audit B), Reasoning Fidelity Gap (Audit C), and Contextual Robustness (Audit D). Figure 1 provides an overview of the framework architecture.

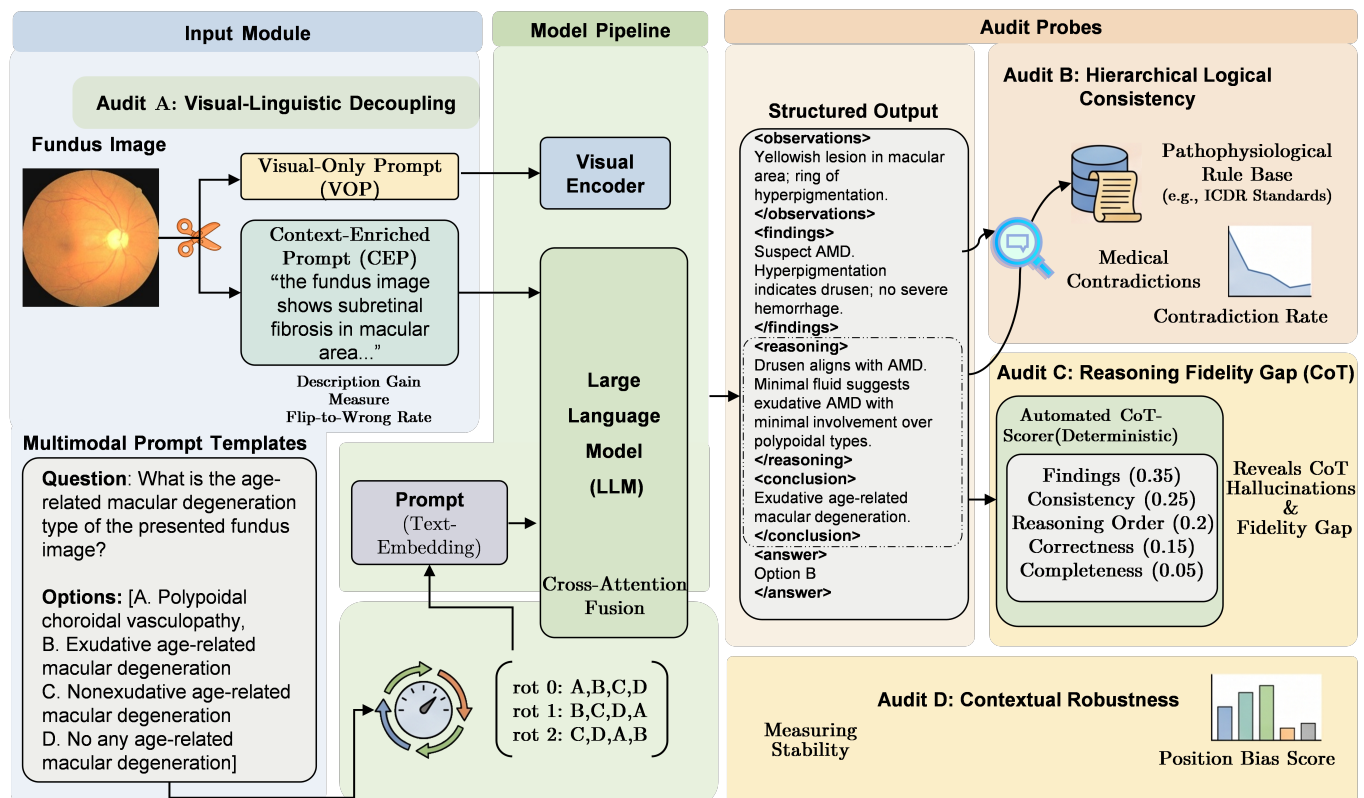


Figure 1. Overview of the Cognitive Audit Framework. The framework audits LVLM reliability through four complementary probes: Audit A (Visual–Linguistic Decoupling), Audit B (Hierarchical Logical Consistency), Audit C (Reasoning Fidelity Gap), and Audit D (Contextual Robustness under option-order perturbation). Dashed colored arrows indicate audit interventions, while solid arrows denote the standard inference flow.

3.1. Visual–Linguistic Decoupling (Audit A)

This module leverages the two input paradigms defined by FunBench [21]—CEP (Context-Enriched Prompt: fundus image paired with template-derived clinical text) and VOP (Visual-Only Prompt: fundus image without supplementary text)—to disentangle the respective contributions of visual features and auxiliary textual information to model predictions. By comparing model performance under both input conditions, we distinguish legitimate multimodal feature fusion from shortcut learning that relies on textual cues; simultaneously, we adopt the Macro-averaged Harmonic Accuracy (MHA) to eliminate evaluation bias introduced by class imbalance.

For image $i \in \mathcal{I}$, let $\hat{y}_i^{(\text{CEP})}$ and $\hat{y}_i^{(\text{VOP})}$ denote the predictions under CEP and VOP respectively, and y_i the ground-truth label. For each answer category c , we compute the harmonic mean of sensitivity (Sens_c) and specificity (Spec_c), then average across all C categories to obtain the Macro-averaged Harmonic Accuracy (MHA):

$$\text{MHA} = \frac{1}{C} \sum_{c=1}^C \frac{2 \text{Sens}_c \cdot \text{Spec}_c}{\text{Sens}_c + \text{Spec}_c} \quad (1)$$

Compared with the arithmetic mean, the harmonic mean penalizes models that sacrifice specificity for sensitivity (or vice versa), aligning with clinical practice where both missed diagnoses and false alarms carry significant risk. This avoids models that favor majority classes from inflating overall metrics, and ensures coverage of rare but critical categories (e.g., proliferative DR).

Based on MHA, we construct three metrics to quantify text dependence. The Description Gain Δ_{desc} measures the MHA difference between CEP and VOP:

$$\Delta_{\text{desc}} = \text{MHA}_{\text{CEP}} - \text{MHA}_{\text{VOP}} \quad (2)$$

A positive gain indicates that clinical text facilitates multimodal reasoning; an excessively large value suggests the model over-relies on textual cues rather than completing diagnosis from visual features alone. The Flip-to-Wrong rate counts samples where the image-only prediction is correct but adding clinical text flips the answer to incorrect, quantifying the negative interference of textual input on visual reasoning:

$$\text{Flip}_{\text{wrong}} = \frac{|\{i : \hat{y}_i^{(\text{CEP})} \neq y_i \wedge \hat{y}_i^{(\text{VOP})} = y_i\}|}{|\mathcal{I}|} \quad (3)$$

Visual Consistency captures the fraction of images on which both input modes yield identical predictions regardless of correctness, reflecting the stability of visual encoding:

$$\text{VisConsist} = \frac{|\{i \in \mathcal{I}_{\text{common}} : \hat{y}_i^{(\text{CEP})} = \hat{y}_i^{(\text{VOP})}\}|}{|\mathcal{I}_{\text{common}}|} \quad (4)$$

Together, the three metrics characterize model visual representation capability from precision loss, prediction reversal, and cross-input stability. A clinically reliable model should simultaneously exhibit low Description Gain, low Flip-to-Wrong rate, and high Visual Consistency.

Having quantified text dependence, Audit B further examines whether the model's diagnostic reasoning respects ophthalmological pathological hierarchies.

3.2. Hierarchical Logical Consistency (Audit B)

Ophthalmological diagnosis follows the strict hierarchical pathological constraints established by ICDR and ETDRS standards [11,12]: the presence of specific fundus lesions constitutes a necessary condition for corresponding diabetic retinopathy (DR) grading. Predictions that violate the causal relationship between lesion and disease grade are clinically incoherent, and such defects cannot be detected by conventional accuracy-based metrics. This module extracts the lesion-level prediction and disease-level grading for the same fundus image, verifies their joint consistency against deterministic clinical rules, and automates contradiction counting across the entire sample without requiring expert clinical review.

We derive 17 deterministic contradiction rules from the ICDR/ETDRS standards (the complete rule set is detailed in Appendix A). Each rule $r \in \mathcal{R}$ defines a bottom-up pathological constraint: if lesion ℓ^* is detected, the corresponding DR diagnosis must fall within the clinically permissible grade set $\mathcal{D}_r$:

$$r : \hat{\ell}_i = \ell^* \implies \hat{d}_i \in \mathcal{D}_r \quad (5)$$

When the model's lesion and grading predictions violate this constraint, we flag a binary contradiction indicator $\text{contra}(i, r) \in \{0, 1\}$. The Contradiction Rate quantifies the proportion of samples exhibiting at least one logical violation:

$$\text{ContradRate} = \frac{|\{i \in \mathcal{I} : \exists r \in \mathcal{R}, \text{contra}(i, r) = 1\}|}{|\mathcal{I}|} \quad (6)$$

The corresponding Consistency Score is defined as $1 - \text{ContradRate}$. High baseline accuracy coupled with a low consistency score reveals *isolated reasoning*: the model treats lesion detection and disease grading as independent classification tasks, lacking the capacity for causal pathological inference.

We note that the current rule set covers only DR-related hierarchical constraints, as FunBench exclusively provides paired lesion and DR grading annotations for cross-task logical verification. After obtaining multi-level annotations for additional diseases, the framework can be extended to age-related macular degeneration, glaucoma, and other conditions.

3.3. Reasoning Fidelity Gap (Audit C)

Chain-of-thought (CoT) prompting [40] guides models to generate step-by-step diagnostic reasoning, theoretically enabling the reconstruction of a complete clinical inference chain from fundus image observations through lesion identification to disease grading—an important pathway for enhancing the interpretability of medical multimodal models. Although current mainstream multimodal large models have achieved excellent baseline accuracy on fundus diagnostic tasks, in certain scenarios a potential inconsistency remains between the model’s predicted results and the generated reasoning chains: models may achieve correct option selection by leveraging textual cues, while the generated reasoning descriptions exhibit limited alignment with fundus visual features. Relying solely on accuracy metrics makes it difficult to finely distinguish such reasoning deficiencies. To this end, this section constructs a deterministic, rule-driven automated CoT-Scorer quantification system grounded in clinical prior knowledge, without introducing third-party large models as judges. Paired with the Wrong-Reasoning Rate metric, it quantitatively characterizes the proportion of samples where “the diagnostic conclusion is correct, but the reasoning content provides insufficient support,” thereby complementing the automated evaluation of model clinical capability from the reasoning fidelity dimension.

To enable structured decomposition and fine-grained verification of reasoning content, we constrain models to output a complete reasoning chain in a fixed five-field XML format:

```
<observations>...</observations> <findings>...</findings> <reasoning>...</reasoning>
<conclusion>...</conclusion> <answer>...</answer>
```

Grounded in clinical keyword dictionaries, we design a deterministic, rule-based CoT-Scorer ($s_i \in [0, 1]$) that computes a weighted sum across five differentiable dimensions:

$$s_i = \sum_{d \in \mathcal{D}} w_d \cdot \text{score}_{i,d} \quad (7)$$

The dimensional weights are configured according to clinical diagnostic priority, with item-level scoring rules as follows:

1. Finding Coverage ($w = 0.35$): Queries the <findings> and <reasoning> fields against a curated clinical keyword dictionary. Matching any core lesion keyword yields a base score of 0.5; the remaining score is accumulated linearly based on the number of matched supporting modifier terms. The keyword dictionary integrates diabetic retinopathy, age-related macular degeneration, glaucoma, and other common fundus disease terminology, encompassing spelling variants and disease alias designations, totaling 126 mandatory keywords and 156 supporting descriptors.
2. Conclusion Consistency ($w = 0.25$): Applies clinical contradiction mapping rules to verify the <conclusion> text against the selected answer option. If the conclusion describes pathological changes yet the answer selects “Negative” (or vice versa), this dimension scores 0; when no logical conflict exists, full marks are awarded.

3. Reasoning Order ($w = 0.20$): Starting from 1.0 as baseline, a fixed deduction of 0.2 is applied for every missing or empty XML tag, constraining models to follow the clinical diagnostic logic of “image observation $\rightarrow$ lesion extraction $\rightarrow$ lesion reasoning $\rightarrow$ conclusion.”
4. Medical Correctness ($w = 0.15$): Uses regex-based expressions to identify whether the text explicitly denies the presence of certain lesions. If the reasoning text clearly states that a specific lesion is absent, yet the final diagnosis selects the corresponding positive option, this item scores 0, to prevent fundamental medical logic errors.
5. Chain Completeness ($w = 0.05$): Computed as the ratio of non-empty XML fields to the total field count, assessing only the completeness of format output without evaluating clinical content validity.

Building upon the CoT scoring system, we define the Wrong-Reasoning Rate (WrongReason) to quantify the phenomenon where model answers are correct but reasoning support is weak:

$$\text{WrongReason} = \frac{|\{i \in \mathcal{I} : c_i = 1 \wedge s_i < 0.5\}|}{|\{i \in \mathcal{I} : c_i = 1\}|} \quad (8)$$

Here $c_i = 1$ indicates that the sample’s final diagnostic option is correct; when this metric is numerically high, it suggests that the model achieves high accuracy partly through shortcut learning rather than grounding predictions in visual lesion reasoning. Beyond this, we additionally introduce two auxiliary indicators—CoT-Gain and Parse Failure Rate: CoT-Gain is defined as the MHA difference between CoT mode and direct answer mode ($\text{MHA}_{\text{CoT}} - \text{MHA}_{\text{direct}}$), used to quantify the gain in model diagnostic performance from enforced step-by-step reasoning; the Parse Failure Rate counts the proportion of samples that fail to comply with the specified XML format output, characterizing model adaptability to structured reasoning formats.

Unlike existing approaches that rely on third-party large models as judges [15]—which have proven effective in open-ended evaluation—this study adopts fixed clinical rules for scoring, motivated by three considerations: (1) reproducibility—identical inputs always yield identical scores; (2) transparency—every deduction traces to specific keyword matches or rule violations; and (3) independence—avoiding circular assessment variance inherent in one LLM judging another.

To mitigate the potential undercount of valid reasoning expressed through rare synonyms, we adopt several robustness measures: first, enriching a multi-variant clinical dictionary (formal names, abbreviations, synonyms, and notation variants) with graded rather than binary scoring; second, leveraging the structured XML format to constrain vocabulary diversity; and finally, employing context-aware regex to distinguish affirmative from negative semantics. Any residual systematic bias from rare terminology mismatch affects all models equally, preserving the validity of relative comparisons.

3.4. Contextual Robustness (Audit D)

A fundamental requirement of any reliable diagnostic system is *label permutation invariance*: the model’s diagnosis of a fundus image should not depend on the arbitrary ordering of the answer options. This invariance is a necessary (though not sufficient) condition for visual grounding—if a model’s prediction changes when options are relabeled but the image is unchanged, the prediction cannot be causally attributed to the visual content.

To evaluate this invariance robustly and mathematically without reliance on manual evaluation of reasoning strings, we employ purely structural cyclic permutations to the answer option set. This structural test modifies no substantive clinical data while fundamentally shifting the order architecture (generating two permuted variants alongside the

original option ordering). For each rotated permutation, we capture and decode the model's prediction mapping via deterministic inverse-logic maps. The result is three comparable predictions mathematically transformed into the canonical target label set: $\hat{y}_i^{(0)}, \hat{y}_i^{(1)}, \hat{y}_i^{(2)}$ for each image i . The Consistency Rate across permutations is:

$$\text{ConsistRate} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathbf{1}[\hat{y}_i^{(0)} = \hat{y}_i^{(1)} = \hat{y}_i^{(2)}] \quad (9)$$

The Position Bias Score is defined as $1 - \text{ConsistRate}$. Models with high Position Bias Score provide compelling evidence that their benchmark performance is partially a byproduct of positional heuristics—a form of shortcut learning that is entirely invisible to standard accuracy evaluation.

4. Experiments

4.1. Experimental Setup

4.1.1. Models

We adopt a layered evaluation design. Three primary models—HuatuoGPT-Vision-7B (medical-specialized), Qwen2.5-VL-7B-Instruct (general-purpose), and Gemma-4-31B-it (large-parameter frontier)—are evaluated across all four audits (A–D) in both direct-answer and CoT modes, comprehensively establishing the performance baselines and direct-vs.-CoT comparison patterns. Building on this established baseline, we introduce three additional reasoning-enhanced models (Qwen3.5-9B, Qwen3.5-Omni-Plus, Qwen3.6-Plus) with a specific diagnostic purpose: to test whether the decoupling between CoT Quality and diagnostic accuracy—observed in the primary models—persists even in architectures explicitly optimized for advanced reasoning. These models are evaluated exclusively on Audit C in CoT mode as a separate experiment (Table 5), since the direct-mode baseline is already fully characterized and repeating it would not yield new insights. Their role is to validate whether the superficial rationale generation phenomenon is a general structural limitation rather than a capacity bottleneck solvable by stronger reasoning ability. Table 1 summarizes the complete model $\times$ audit coverage.

Table 1. Model $\times$ audit coverage matrix. ✓ = systematically evaluated; — = not evaluated (baseline already established by the three primary models). [†] Models marked with [†] are reasoning-enhanced variants evaluated exclusively on Audit C to probe rationale quality.

Model	Audit A	Audit B	Audit C	Audit D
HuatuoGPT-Vision-7B	✓	✓	✓	✓
Qwen2.5-VL-7B-Instruct	✓	✓	✓	✓
Gemma-4-31B-it	✓	✓	✓	✓
Qwen3.5-9B [†]	—	—	✓	—
Qwen3.5-Omni-Plus [†]	—	—	✓	—
Qwen3.6-Plus [†]	—	—	✓	—

4.1.2. Dataset and Evaluation Modes

All experiments use the FunBench Benchmark [21], which aggregates multiple publicly available ophthalmic datasets covering fundus photography and optical coherence tomography (OCT): DDR [48], DeepDRiD [49], IDRiD [50], JSIEC [51], MMC-AMD [52], NEH [53], OCTDL [54], OCTID [55], OIA-ODIR [56], Retinal-Lesions [57], RETOUCH [58], RFMiD [59], TOP (Tsukazaki Ophthalmic Photography, a publicly available fundus dataset collected at Tsukazaki Hospital, Japan; available at: https://github.com/DateCazuki/Fundus_Diagnosis, accessed on 2 February 2026), and UCSD [23]. We evaluate results under two input modes natively provided by FunBench: CEP (Context-Enriched Prompt;

image + clinical description) and VOP (Visual-Only Prompt; image only), with clinical text derived from standardized templates based on source dataset annotations. For Audit B, each model is additionally prompted with a CoT instruction. For Audit D, options are cyclically permuted (original + two rotations) to probe positional sensitivity. All models use deterministic greedy decoding (temperature = 0) to ensure exact reproducibility. Full inference parameters and hardware details are reported in Appendix C.

4.2. Audit A: Visual–Linguistic Decoupling

This section employs CEP and VOP as a controlled experiment to quantify model dependence on textual cues. As shown in Table 2 and Figure 2, HuatuoGPT-Vision-7B and Qwen2.5-VL-7B-Instruct yield Description Gains of +0.205 and +0.206 respectively—nearly identical values indicating that roughly 20% of their diagnostic accuracy depends on the appended clinical text rather than intrinsic image feature extraction. Gemma-4-31B-it exhibits a reduced gain of +0.163, reflecting comparatively lower text dependency and stronger visual representation. All three gains pass 1000-resample bootstrap testing with 95% CI lower bounds above zero ($p \leq 0.002$), confirming that the accuracy improvement genuinely originates from textual information rather than random error.

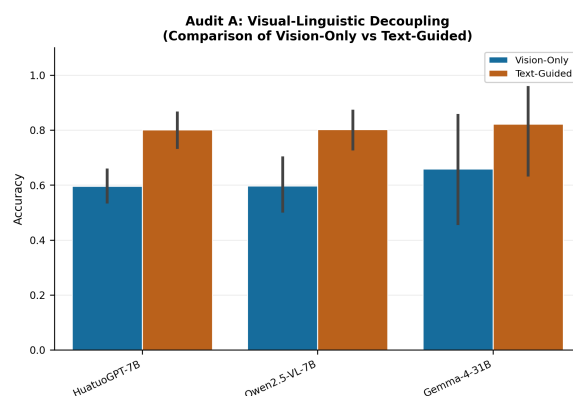


Figure 2. Audit A—Visual–Linguistic Decoupling (CEP vs. VOP). Grouped bars compare CEP and VOP accuracy per model. HuatuoGPT-Vision-7B and Qwen2.5-VL-7B-Instruct both show a substantial Description Gain (+0.205 and +0.206, respectively), indicating heavy reliance on textual cues rather than pure visual reasoning.

Visual consistency scores range from 0.651 to 0.723, meaning that more than 27% of samples undergo prediction changes when clinical text is removed. This demonstrates universal visual encoding instability across the tested models. Synthesizing all three metrics, current mainstream 7B–31B fundus multimodal models exhibit text-shortcut learning where high baseline accuracy does not equate to reliable visual pathological reasoning; the larger-parameter Gemma-4-31B-it holds a relative advantage in visual architectural design, yet still cannot entirely disengage from textual assistance.

Table 2. Audit A: Visual–Linguistic Decoupling (CEP vs. VOP); 95% CIs are bootstrapped (1000 resamples, seed = 42). *** $p < 0.001$, ** $p < 0.01$. H_0 : Desc. Gain ≤ 0 .

Model	CEP MHA	VOP MHA	Visual Consistency	Desc. Gain (95% CI)
HuatuoGPT-Vision-7B	0.801	0.596	0.651	0.205 (0.146 to 0.257) ***
Qwen2.5-VL-7B-Instruct	0.803	0.597	0.666	0.206 (0.129 to 0.281) ***
Gemma-4-31B-it	0.822	0.659	0.723	0.163 (0.053 to 0.291) **

4.3. Audit B: Hierarchical Logical Consistency

This section examines the alignment between lesion recognition and DR grading through 17 clinical rules derived from ICDR/ETDRS standards, using contradiction rate to quantify whether model diagnoses satisfy fundus-level diagnostic logic. The experiment is split into CEP (Image + Clinical Text) and VOP (Image Only) modes, with results reported in Table 3.

Under CEP, HuatuoGPT-Vision-7B exhibits the highest contradiction rate (0.513), Qwen2.5-VL-7B-Instruct follows at 0.459, and Gemma-4-31B-it achieves the lowest (0.352)—consistent with the visual capability ranking from Audit A. Pairwise model comparisons under CEP confirm: HuatuoGPT exceeds Gemma by 0.161 ($p < 0.001$), Qwen exceeds Gemma by 0.107 ($p < 0.05$), while the HuatuoGPT–Qwen difference is 0.054 at borderline significance ($p = 0.05$). All models maintain consistency scores below 0.72, meaning that even with overall diagnostic accuracy above the baseline, more than 28% of samples contain lesion-to-grade predictions that violate clinical hierarchical constraints.

Switching to VOP, contradiction rates drop consistently across all three models. The CEP–VOP differences are 0.140 for HuatuoGPT ($p < 0.001$), confirming that extraneous text clearly disrupts hierarchical reasoning; for Qwen2.5-VL the paired difference is only 0.061, not reaching significance at $p = 0.05$. Viewed across models in a layered perspective, HuatuoGPT is most affected by additional text, Gemma shows the best robustness against text-induced perturbation, and Qwen2.5-VL’s cross-modal fusion ability lies between the two. The graded nature of these results validates that the audit framework can effectively differentiate the inherent characteristics of different multimodal models, while also demonstrating that metric changes are difficult to attribute to significance testing alone.

Table 3. Audit B: Hierarchical Logical Consistency (micro-lesion → macro-disease contradiction rates). Consistency score = $1 - \text{Contrad Rate}$; 95% CIs are bootstrapped. *** $p < 0.001$, * $p < 0.05$, $^{\dagger} p = 0.05$. H_0 : difference = 0 (two-sided).

Model	Mode	Contrad Rate (95% CI)
HuatuoGPT-Vision-7B	CEP	0.513 (0.455 to 0.573)
HuatuoGPT-Vision-7B	VOP	0.373 (0.318 to 0.424)
Qwen2.5-VL-7B-Instruct	CEP	0.459 (0.398 to 0.513)
Qwen2.5-VL-7B-Instruct	VOP	0.398 (0.344 to 0.455)
Gemma-4-31B-it	CEP	0.352 (0.295 to 0.411)
Gemma-4-31B-it	VOP	0.281 (0.228 to 0.338)
Pairwise model comparisons (CEP mode):		
Δ Model: HuatuoGPT–Gemma		0.161 (0.078 to 0.244) ***
Δ Model: Qwen2.5-VL–Gemma		0.107 (0.025 to 0.189) *
Δ Model: HuatuoGPT–Qwen2.5-VL		0.054 (0.003 to 0.102) †

4.4. Audit C: Reasoning Fidelity Gap (CoT)

This experiment employs direct answer and chain-of-thought reasoning as two output formats under CEP and VOP input conditions as a controlled comparison. The experiment is divided into two groups: Table 4 evaluates three primary models under a full-condition Direct+CoT protocol, while Table 5 tests three reasoning-enhanced models exclusively under CoT mode, with results in Tables 4 and 5.

As shown in Table 4, after enabling chain reasoning, HuatuoGPT exhibits significantly negative CoT Gain under both CEP and VOP ($p < 0.001$), indicating that step-by-step reasoning noticeably degrades overall diagnostic accuracy. Meanwhile, its Wrong-Reasoning Rate reaches 0.813–0.983, confirming that the majority of correct answers lack reliable pathological reasoning support. Qwen2.5-VL shows significantly negative CoT Gain under CEP while VOP yields no statistically significant difference; combined with its relatively

low Wrong-Reasoning Rate, this model exhibits better vision-text fusion stability that makes it difficult for step-by-step reasoning to clearly interfere with prediction outcomes. Gemma’s CoT Gain under both input modes shows no significant deviation from baseline, with Wrong-Reasoning Rate below 0.01—indicating that its reasoning descriptions closely match diagnostic conclusions. The above results demonstrate that some models achieve superior accuracy through shortcuts, but their reasoning logic remains hollow; relying solely on accuracy cannot reveal deficiencies, embodying the necessity of CAF’s reasoning evaluation metrics.

Table 5 shows that Qwen3.6-Plus possesses top-tier reasoning organization ability with CoT Quality scores (0.854/0.850) superior to all other models, yet its diagnostic accuracy on fundus-specific data is notably low. This result reflects not a deficiency in general reasoning ability per se, but rather that general pre-training knowledge and ophthalmic domain-specific visual tasks have a domain adaptation gap—the model possesses general logical architecture capability but lacks specialized fundus lesion-domain visual training, resulting in the observed inconsistency between reasoning quality and diagnostic effectiveness. We term this phenomenon superficial rationale generation.

Synthesizing all data, the tested models can be divided into three categories: Gemma’s reasoning content closely matches diagnostic results with high coherence; Qwen3.6’s general reasoning architecture is mature and complete, yet lacks domain-specific data support leading to low diagnostic accuracy; HuatuoGPT readily derives answers by relying on textual shortcuts, with matched reasoning that lacks pathological reasoning support. The differentiated performance across the three models further demonstrates that CAF multi-dimensional evaluation can complement the limitations of accuracy-only assessment, as shown in Tables 4 and 5.

Table 4. Audit C, experiment 1: Direct vs. CoT under a unified protocol. CoT Quality ($\bar{s}_i$) is the mean deterministic CoT-Scorer (0–1 scale); 95% CIs are bootstrapped (CoT Gain) or Normal Approximation (Wrong-Reas. Rate). *** $p < 0.001$. H_0 : CoT Gain = 0 (two-sided).

Model	Mode	Direct MHA	CoT MHA	CoT Quality	Wrong-Reas. Rate (95% CI)	CoT Gain (95% CI)
HuatuoGPT-7B	CEP	0.801	0.279	0.411	0.983 (0.979 to 0.987)	−0.500 (−0.598 to −0.428) ***
HuatuoGPT-7B	VOP	0.472	0.126	0.427	0.813 (0.802 to 0.824)	−0.340 (−0.384 to −0.191) ***
Qwen2.5-VL-7B	CEP	0.803	0.591	0.746	0.026 (0.023 to 0.029)	−0.155 (−0.433 to −0.071) ***
Qwen2.5-VL-7B	VOP	0.419	0.410	0.726	0.197 (0.189 to 0.205)	−0.062 (−0.184 to 0.118)
Gemma-4-31B	CEP	0.822	0.873	0.743	0.003 (0.002 to 0.004)	0.051 (−0.071 to 0.271)
Gemma-4-31B	VOP	0.733	0.691	0.776	0.004 (0.002 to 0.006)	−0.009 (−0.226 to 0.138)

Table 5. Audit C, experiment 2: reasoning-enhanced architectures (CoT-only). Probes whether advanced reasoning resolves the quality–accuracy decoupling; 95% CIs via Normal Approximation. *** $p < 0.001$. H_0 : CoT Quality difference = 0 vs. Gemma-4-31B (two-sided).

Model	Mode	CoT MHA	Wrong-Reas. Rate (95% CI)	CoT Quality (95% CI)
Qwen3.5-9B	CEP	0.196	0.725 (0.713 to 0.737)	0.479 (0.465 to 0.493)
Qwen3.5-9B	VOP	0.338	0.615 (0.602 to 0.628)	0.516 (0.502 to 0.530)
Qwen3.5-Omni-Plus	CEP	0.380	0.689 (0.677 to 0.701)	0.493 (0.479 to 0.507)
Qwen3.5-Omni-Plus	VOP	0.384	0.732 (0.720 to 0.744)	0.508 (0.494 to 0.522)
Qwen3.6-Plus	CEP	0.133	0.002 (0.001 to 0.003)	0.854 (0.845 to 0.863) ***
Qwen3.6-Plus	VOP	0.047	0.001 (0.000 to 0.002)	0.850 (0.841 to 0.859) ***

4.5. Audit D: Contextual Robustness (Option-Order/Position Bias)

This module fixes the image and question stem while cyclically rotating option orders as a perturbation experiment, employing consistency rate and position bias scores to quantify the degree to which answer arrangement interferes with model predictions. Results

are summarized in Table 6. The overall mean consistency across three models is 0.867, indicating that approximately 12.5% of samples change their output due to option-order variation, which demonstrates that positional bias is a shared vulnerability across models.

Results show that HuatuoGPT-Vision-7B exhibits the highest position bias and lowest consistency (0.825), Gemma-4-31B-it achieves the best robustness with the highest consistency (0.910), and Qwen2.5-VL-7B-Instruct falls between the two on all metrics. The difference between HuatuoGPT and Gemma is statistically significant ($p = 0.007$), whereas pairwise comparisons of Qwen2.5 against either group do not reach significance. Under unified evaluation samples, this phenomenon arises because Qwen2.5's positional sensitivity lies between the high and low groups, and the metric difference itself is small, making it difficult to pass significance testing. All models' consistency coefficients are significantly below 1 ($p < 0.001$), proving that every model suffers from position-dependent behavior to varying degrees—a class of defect that cannot be discovered through conventional accuracy metrics alone. The stratified experimental results validate the effectiveness of CAF's perturbation testing, which can uncover positional-bias defects masked by accuracy, complementing the limitations of traditional evaluation approaches.

Table 6. Audit D: option-order robustness across key evaluation models; 95% CIs computed via bootstrap (Gemma)/Normal Approximation ($N = 222$). *** $p < 0.001$. H_0 : Consistency Rate = 1 (no position bias).

Model	Position Bias	Consistency Rate (95% CI)
HuatuoGPT-Vision-7B	0.175	0.825 (0.775 to 0.875) ***
Qwen2.5-VL-7B-Instruct	0.136	0.864 (0.819 to 0.909) ***
Gemma-4-31B-it	0.090	0.910 (0.874 to 0.946) ***

4.6. Qualitative Analysis and Cognitive Trace Examples

This section presents a case study drawn from the MMC-AMD [52] dataset, dissecting the complete CoT output of Qwen2.5-VL-7B-Instruct to intuitively demonstrate how the model performs clinical reasoning in practice (Figure 3). The gold-standard label for this sample is exudative age-related macular degeneration; the textual description notes the presence of subretinal fibrosis, RPE hemorrhage, and choroidal neovascular membrane in the macular area. However, the imagery generated in the <observations> segment does not correspond to the pathological features stated in the original text—failing to identify hemorrhage and fibrotic lesions. Even when the visual observations contain inaccuracies, the model ultimately outputs the correct answer in the conclusion by relying on key textual information.

This case directly illustrates a high-level logical pattern of current models: arriving at the correct conclusion without genuinely relying on visual image features, instead depending on textual keywords to complete option selection. It also provides case-level corroboration for the quantitative audit statistics reported in Audits A/B/C/D above, explaining the co-existence of high accuracy and reasoning deficits—the “Illusion of Clinical Comprehension.” Quantitative metrics combined with case-level analysis jointly demonstrate that relying solely on accuracy to evaluate model clinical competence has clear limitations.

Qualitative Case Study: AMD Type Classification (MMC-AMD)

Sample ID: MMC-AMD/ImageData/cfp-448x448/f-w-43949233.jpg
 Ground Truth: A (Exudative age-related macular degeneration)

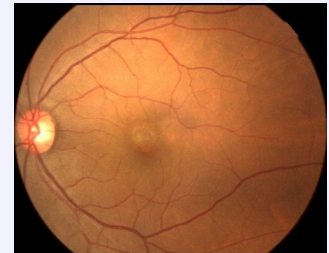
Prompt:

Please choose the most suitable option based on the image, description and the question. Answer with the option's letter directly.

Description: the fundus image shows subretinal fibrosis in macular area, subretinal pigment epithelium hemorrhage in macular area and choroidal neovascular membrane in macular area.

The question is: What type of age-related macular degeneration is depicted in this fundus image?

The options are: A. Exudative age-related macular degeneration, B. Not any age-related macular degeneration, C. Polypoidal choroidal vasculopathy, D. Nonexudative age-related macular degeneration.

**Qwen2.5-VL-7B-Instruct Output:**

<observations>

- The image shows a fundus view with a yellowish-white lesion located at the fovea.
- The surrounding retina appears relatively normal without significant hemorrhages or exudates.
- The retinal vessels are visible and appear to be within normal limits.
- There is no evidence of subretinal fluid or hemorrhage.

</observations>

<findings>

- The presence of a yellowish-white lesion at the fovea is indicative of a choroidal neovascular membrane, which is characteristic of exudative age-related macular degeneration.
- The absence of subretinal fluid or hemorrhage suggests that it is not nonexudative age-related macular degeneration.
- There is no evidence of polypoidal choroidal vasculopathy, as this condition typically presents with multiple polyps and often has a different appearance.

</findings>

<reasoning>

- Exudative age-related macular degeneration is characterized by the formation of choroidal neovascular membranes, which can appear as yellowish-white lesions at the fovea.
- Nonexudative age-related macular degeneration, also known as dry AMD, typically does not present with such a lesion but rather with drusen and atrophy.
- Polypoidal choroidal vasculopathy is a distinct condition that presents with multiple polyps and is usually associated with a different clinical picture.
- Based on the observation of a single yellowish-white lesion at the fovea, the most likely diagnosis is exudative age-related macular degeneration.

<conclusion>

The fundus image depicts exudative age-related macular degeneration.

</conclusion>

<answer>

A. Exudative age-related macular degeneration

</answer>

Figure 3. Cognitive Trace Example (MMC-AMD). A single-case audit sample showing the original prompts, ground-truth label, and the model-generated chain-of-thought trace for f-w-43949233.jpg.

5. Discussion

This study leverages CAF's multi-dimensional assessment to confirm that current fundus LVLs universally exhibit an Illusion of Clinical Comprehension: high test-set accuracy cannot be equated with genuine clinical reasoning ability. Audit A results show that all evaluated models exhibit textual dependence—nearly 20% of diagnostic accuracy for HuatuoGPT and Qwen2.5-VL relies on supplementary clinical text, and even Gemma-4-31B, despite its larger parameter count, cannot fully circumvent textual shortcuts, with over 27% of samples changing predictions after auxiliary text removal, reflecting deficiencies in visual lesion extraction. Audit B reveals that all models frequently produce logical conflicts between lesion identification and DR grading that violate ICDR/ETDRS diagnostic standards, with additional textual input further exacerbating hierarchical reasoning disorder, demonstrating that models lack the capacity for learning causal relationships in ophthalmic pathology. A particularly noteworthy counter-intuitive finding is that the Vision-Only mode consistently yields lower contradiction rates than the Text-Guided mode across all three models (e.g., HuatuoGPT: 0.373 vs. 0.513, $p < 0.001$). This indicates that supplementary clinical descriptions, rather than reinforcing correct diagnostic logic, introduce spurious linguistic shortcuts that cause models to anchor on surface-level textual cues instead of adhering to strict lesion-to-disease causal constraints. In effect, multi-

modal text guidance can inadvertently pollute the model's hierarchical clinical reasoning pathway—weakening logical consistency rather than solidifying it—highlighting a critical safety concern for deployment scenarios that integrate auxiliary clinical notes alongside fundus images. Audit C identifies two typical defects: some models arrive at correct answers via textual cues while their accompanying reasoning diverges from the actual image-based pathology; Qwen3.6-Plus possesses superior general chain-of-thought reasoning capabilities yet performs poorly on fundus diagnosis, confirming a domain gap between general logical modeling and ophthalmology-specific visual diagnosis. Audit D verifies that all models are subject to option-position bias, with nearly 12.5% of samples changing their diagnostic output following option reordering—fixed option layouts inflate baseline test scores, undermining reliability in real open-ended clinical scenarios. These shortcomings carry practical medical risks: text-dependent models suffer sharp performance degradation in field screening scenarios lacking accompanying clinical notes, and unfaithful reasoning can mislead primary-care clinical decisions. This study employs 7B–31B lightweight open-source models—precisely the architecture class most likely to be deployed in community-level fundus screening—making these conclusions directly relevant to real-world deployment.

This study also has several limitations. First, the entire CAF framework is currently developed exclusively for fundus imaging tasks and has not yet been extended to radiology, pathology, or other medical imaging domains. Second, constrained by FunBench dataset annotations, the hierarchical logic verification rules are designed only for diabetic retinopathy; macular degeneration, glaucoma, and other common fundus diseases have not been incorporated into the rule system. Third, experiments employ only six open-source lightweight models and do not include ultra-large closed-source commercial models, limiting the generalizability of conclusions. Finally, the clinical keyword dictionary used for CoT evaluation cannot exhaustively cover rare lesion aliases and uncommon subtype expressions; a small number of specialized medical terms may introduce minor scoring discrepancies. Future work will progressively address these limitations.

6. Conclusions

This paper proposes the automated multi-dimensional Cognitive Audit Framework (CAF), which compensates for the deficiencies of single-metric accuracy evaluation from four complementary dimensions: Visual–Linguistic Decoupling, hierarchical pathological logic, reasoning fidelity, and option robustness. Controlled experiments on six mainstream multimodal large models confirm that fundus LVLs universally suffer from inflated scores—behind strong baseline accuracy commonly lurk textual shortcut learning, pathological logic contradictions, ineffective reasoning, and option-position bias, four classes of defect that are difficult to detect through traditional metrics. CAF achieves automated scoring via fixed clinical rules without requiring a third-party large model as an evaluator, offering excellent reproducibility and the ability to efficiently screen for latent reasoning defects in models, making it well-suited for iterative quality control of medical multimodal models. Future work will expand multi-disease clinical constraint rules and professional dictionaries, extend CAF to multiple medical imaging modalities, and help transition medical large model evaluation systems from single-accuracy assessment toward clinical reasoning authenticity evaluation.

Author Contributions: Conceptualization, J.Z. and J.G.; methodology, J.Z. and S.Z.; validation, S.Z. and X.L.; formal analysis, J.Z. and X.L.; investigation, J.Z. and S.Z.; data curation, S.Z. and X.L.; writing—original draft preparation, J.Z.; writing—review and editing, J.Z., J.G., S.Z. and X.L.; visualization, J.Z. and X.L.; supervision, J.G.; project administration, J.G.; funding acquisition, J.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Science and Technology Development Fund of Macao under Grant No. 0002/2024/RIA1, and by Fujian Province Young and Middle-aged Teachers’ Educational Science Research Projects (Science and Technology Category) under Grant No. JAT232014.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available at <https://github.com/ruc-aimc-lab/FunBench> (accessed on 2 February 2026).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

LVLM	Large Vision–Language Model
CAF	Cognitive Audit Framework
CoT	Chain-of-Thought
MHA	Macro-averaged Harmonic Accuracy
VQA	Visual Question Answering
DR	Diabetic Retinopathy
PDR	Proliferative Diabetic Retinopathy
AMD	Age-related Macular Degeneration
ICDR	International Clinical Diabetic Retinopathy
ETDRS	Early Treatment Diabetic Retinopathy Study
CEP	Context-Enriched Prompt
VOP	Vision-Only Prompt
OCT	Optical Coherence Tomography

Appendix A. Complete Clinical Contradiction Rule Set

Table A1 presents the complete set of 17 cross-task contradiction rules used in Audit B. The rules are organized into two clinically motivated categories:

- Category 1 (R1–R8): Any DR-indicative lesion predicted as present is incompatible with a “No DR” diagnosis. Under ICDR standards, the presence of any retinal lesion characteristic of diabetic retinopathy definitionally excludes a normal (No DR) classification.
- Category 2 (R9–R17): Lesions pathognomonic of severe or proliferative disease are incompatible with low-grade DR diagnoses. Neovascularization, fibrous proliferation, vitreous hemorrhage, and preretinal hemorrhage are hallmarks of advanced DR stages; their co-occurrence with Mild or Moderate NPDR constitutes a hierarchical logic violation.

These 17 rules are applied to all fundus images that have both lesion recognition and DR grading annotations in FunBench.

Table A1. Complete cross-task contradiction rule set for Audit B (17 rules in two categories). Each rule specifies the lesion finding and the DR grade that constitutes a medical contradiction when co-occurring.

Rule	Lesion Type	Contradictory DR Grade	Clinical Rationale
Category 1: Lesion present + No DR (any DR lesion excludes normal status)			
R1	Microaneurysm	No DR	Earliest DR lesion; definitionally excludes No DR
R2	Hard Exudate	No DR	Vascular leakage marker; incompatible with No DR
R3	Soft Exudate	No DR	Nerve fiber ischemia sign; indicates DR presence
R4	Neovascularization	No DR	Pathognomonic of proliferative disease
R5	Fibrous Proliferation	No DR	Late-stage proliferative sequela
R6	Vitreous Hemorrhage	No DR	Indicates severe vascular compromise
R7	Preretinal Hemorrhage	No DR	Hallmark of advanced DR
R8	Intraretinal Hemorrhage	No DR	Characteristic DR finding at any grade
Category 2: Severe/proliferative lesion + low-grade DR (hierarchical severity violation)			
R9	Neovascularization	Mild NPDR	Neovascularization mandates PDR diagnosis
R10	Neovascularization	Moderate NPDR	Neovascularization mandates PDR diagnosis
R11	Neovascularization	Severe NPDR	Neovascularization mandates PDR diagnosis
R12	Fibrous Proliferation	Mild NPDR	Fibrous tissue indicates prior proliferative stage
R13	Fibrous Proliferation	Moderate NPDR	Fibrous tissue indicates prior proliferative stage
R14	Vitreous Hemorrhage	Mild NPDR	Vitreous bleeding requires severe/proliferative DR
R15	Vitreous Hemorrhage	Moderate NPDR	Vitreous bleeding requires severe/proliferative DR
R16	Preretinal Hemorrhage	Mild NPDR	Preretinal bleeding indicates advanced disease
R17	Preretinal Hemorrhage	Moderate NPDR	Preretinal bleeding indicates advanced disease

Rule application scope.

The 17 rules are applied to all images in the FunBench dataset that have both lesion recognition and DR grading annotations. For each image, the system extracts the model's predicted answer for the relevant lesion recognition task and the DR grading task, then evaluates whether the joint prediction violates any rule. An image is flagged as contradictory if any rule is violated. The Contradiction Rate (Equation (5)) reports the fraction of images with at least one violation.

Scope limitation.

The current rule set is limited exclusively to the DR diagnostic hierarchy, as FunBench only provides paired lesion and DR grading annotations for cross-task logical verification. Although our CoT-Scorer in Audit C already contains keyword dictionaries for AMD and multi-condition diagnosis, FunBench lacks the required multi-level paired annotations to extend the hierarchical consistency audit to AMD and glaucoma. This remains a valuable direction for future research.

Appendix B. Complete Clinical Keyword Dictionary

Table A2 present the complete clinical keyword dictionary used by the CoT-Scorer's Finding Coverage dimension. Each clinical concept maps to a set of required keywords (at least one must appear for a non-zero score) and supporting keywords (each additional hit increases the score proportionally). For grading tasks, forbidden keywords trigger a Medical Correctness penalty when present in the reasoning chain for an incompatible grade. All matching is performed after text normalization (lowercasing and whitespace collapsing).

Table A2. Lesion keyword dictionary (12 categories, 43 required terms, 66 supporting variants).

Lesion Category	Required Keywords	Supporting Keywords
Fibrous Proliferation	fibrous, proliferation, fibrosis, fibrovascular	membrane, traction, white tissue, scar, epiretinal
Hemorrhage	hemorrhage, hemorrhage, bleeding, blood	flame, blot, dot-blot, dark red, intraretinal, subretinal, preretinal
Hard Exudate	hard exudate, exudate, lipid exudate	yellow, bright, lipid, waxy, well-defined, crystalline
Intraretinal Fluid	intraretinal fluid, IRE, cystoid, cyst	hyporefective, space, edema, swelling, retinal thickening
Intraretinal Hemorrhage	intraretinal hemorrhage, intraretinal hemorrhage, dot hemorrhage, blot hemorrhage	dark, round, oval, deep retina, inner nuclear
Microaneurysm	microaneurysm, micro-aneurysm	small red dot, dot hemorrhage, punctate, capillary, round red, tiny red

Table A2. *Cont.*

Lesion Category	Required Keywords	Supporting Keywords
Neovascularization	neovascularization, new vessel, NVD, NVE, new blood vessel	proliferative, fibrovascular, frond, disc vessel, retinal vessel
PED	pigment epithelial detachment, PED, RPE detachment	dome, elevation, subretinal, RPE, pigment epithelium
Preretinal Hemorrhage	preretinal hemorrhage, preretinal hemorrhage, boat-shaped, subhyaloid	flat top, horizontal level, vitreous, anterior, large
Soft Exudate	soft exudate, cotton wool, cotton-wool spot, CWS	white, fluffy, nerve fiber, ischemia, superficial, feathery
Subretinal Fluid	subretinal fluid, SRF, subretinal fluid	hyporeflective, beneath retina, detachment, neurosensory, accumulation
Vitreous Hemorrhage	vitreous hemorrhage, vitreous hemorrhage, vitreous bleeding	vitreous, haze, obscure, red reflex, floater, dense

Appendix C. Reproducibility Details

Prompt templates.

The exact prompt templates used for each evaluation mode are:

- Direct mode (CEP): “Please choose the most suitable option based on the image, description and the question. Answer with the option’s letter directly. Description: [TEXT]. The question is: [QUESTION]. The options are: [OPTIONS].”
- Direct mode (VOP): “Please choose the most suitable option based on the image and the question. Answer with the option’s letter directly. The question is: [QUESTION]. The options are: [OPTIONS].”
- CoT mode: Appends to the above: “Please analyze the image step by step. Structure your response using the following XML tags: <observations>describe what you see</observations> <findings>identify clinical findings</findings> <reasoning>explain your diagnostic reasoning</reasoning> <conclusion>state your conclusion</conclusion> <answer>give your final answer letter</answer>”

Inference parameters.

All models: temperature = 0, top_p = 1.0. No system prompt is used. Each image is provided as a single-turn input with the prompt text. All experiments are conducted on two NVIDIA RTX 5090 GPUs. Deterministic greedy decoding guarantees that identical inputs produce bit-identical outputs, rendering multi-run repetition uninformative (inference variance is exactly zero) and ensuring that all reported metrics are fully reproducible. The bootstrap confidence intervals reported in the main experimental tables address a complementary concern: sample-level stability of aggregate metrics across the finite evaluation population.

Random seed control.

For open-weight models served via vLLM, we fix all random seeds (Python 3.10.20, NumPy 2.2.6, PyTorch 2.7.0+cu128, CUDA 12.8) to ensure deterministic behavior at the framework level. Combined with temperature = 0 greedy decoding at the generation level, this dual-layer determinism guarantees exact reproducibility of all predictions across independent runs.

Model versions and access dates.

- HuatuoGPT-Vision-7B: ModelScope, FreedomIntelligence/HuatuoGPT-Vision-7B;
- Qwen2.5-VL-7B-Instruct: ModelScope, Qwen/Qwen2.5-VL-7B-Instruct;
- Qwen3.5-9B: ModelScope, Qwen/Qwen3.5-9B;
- Gemma-4-31B-it: HuggingFace, google/gemma-4-31B-it;
- Qwen3.5-Omni-Plus: Alibaba Cloud Bailian API, qwen3.5-omni-plus-2026-03-15;
- Qwen3.6-Plus: Alibaba Cloud Bailian API, qwen3.5-plus-2026-04-20.

References

1. Zhou, Y.; Chia, M.A.; Wagner, S.K.; Ayhan, M.S.; Williamson, D.J.; Struyven, R.R.; Liu, T.; Xu, M.; Lozano, M.G.; Woodward-Court, P.; et al. A foundation model for generalizable disease detection from retinal images. *Nature* **2023**, *622*, 156–163. [\[CrossRef\]](#)
2. Shi, D.; Zhang, W.; Yang, J.; Huang, S.; Chen, X.; Yusufu, M.; Jin, K.; Lin, S.; Liu, S.; Zhang, Q.; et al. Eyeclip: A visual-language foundation model for multi-modal ophthalmic image analysis. *arXiv* **2024**, arXiv:2409.06644.
3. Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; Gao, J. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 28541–28564.
4. Tu, T.; Azizi, S.; Driess, D.; Schaekermann, M.; Amin, M.; Chang, P.C.; Carroll, A.; Lau, C.; Tanno, R.; Ktena, I.; et al. Towards generalist biomedical AI. *NEJM AI* **2024**, *1*, A10a2300138. [\[CrossRef\]](#)
5. Chen, Y.; Xu, D.; Huang, Y.; Zhan, S.; Wang, H.; Chen, D.; Wang, X.; Qiu, M.; Li, H. MIMO: A medical vision language model with visual referring multimodal input and pixel grounding multimodal output. In *Proceedings of the Computer Vision and Pattern Recognition Conference*; IEEE: New York, NY, USA, 2025; pp. 24732–24741.
6. Liu, X.; Song, D. Constructing ophthalmic mllm for positioning-diagnosis collaboration through clinical cognitive chain reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2025; pp. 21547–21556.
7. Wang, M.; Lin, T.; Lin, A.; Yu, K.; Peng, Y.; Wang, L.; Chen, C.; Zou, K.; Liang, H.; Chen, M.; et al. Enhancing diagnostic accuracy in rare and common fundus diseases with a knowledge-rich vision-language model. *Nat. Commun.* **2025**, *16*, 5528. [\[CrossRef\]](#)
8. Li, S.; Lin, T.; Lin, L.; Zhang, W.; Liu, J.; Yang, X.; Li, J.; He, Y.; Song, X.; Xiao, J.; et al. Eyecaregpt: Boosting comprehensive ophthalmology understanding with tailored dataset, benchmark and model. In *Proceedings of the 33rd ACM International Conference on Multimedia*; Association for Computing Machinery: New York, NY, USA, 2025; pp. 3893–3902.
9. Zhou, T.; Xu, Y.; Zhu, Y.; Xiao, C.; Bian, H.; Wei, L.; Zhang, X. DrVD-Bench: Do Vision-Language Models Reason Like Human Doctors in Medical Image Diagnosis? In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2026; Volume 38.
10. Geirhos, R.; Jacobsen, J.H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; Wichmann, F.A. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2020**, *2*, 665–673. [\[CrossRef\]](#)
11. Wilkinson, C.P.; Ferris, F.L., III; Klein, R.E.; Lee, P.P.; Agardh, C.D.; Davis, M.; Dills, D.; Kampik, A.; Pararajasegaram, R.; Verdager, J.T.; et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology* **2003**, *110*, 1677–1682. [\[CrossRef\]](#)
12. Early Treatment Diabetic Retinopathy Study Research Group. Grading diabetic retinopathy from stereoscopic color fundus photographs—An extension of the modified Airlie House classification: ETDRS report number 10. *Ophthalmology* **1991**, *98*, 786–806. [\[CrossRef\]](#)
13. U.S. Food and Drug Administration (FDA). *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan*; U.S. Food and Drug Administration: Silver Spring, MD, USA, 2021.
14. Collins, G.S.; Moons, K.G.; Dhiman, P.; Riley, R.D.; Beam, A.L.; Van Calster, B.; Ghassemi, M.; Liu, X.; Reitsma, J.B.; Van Smeden, M.; et al. TRIPOD+ AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* **2024**, *385*, e078378. [\[CrossRef\]](#)
15. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 46595–46623.
16. Jain, S.; Agrawal, A.; Saporta, A.; Truong, S.Q.; Duong, D.N.; Bui, T.; Chambon, P.; Zhang, Y.; Lungren, M.P.; Ng, A.Y.; et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv* **2021**, arXiv:2106.14463. [\[CrossRef\]](#)
17. Ostmeier, S.; Xu, J.; Chen, Z.; Varma, M.; Blankemeier, L.; Bluethgen, C.; Md, A.E.M.; Moseley, M.; Langlotz, C.; Chaudhari, A.S.; et al. Green: Generative radiology report evaluation and error notation. *Proc. Find. Assoc. Comput. Linguist. EMNLP* **2024**, *2024*, 374–390.
18. Huang, A.; Banerjee, O.; Wu, K.; Reis, E.P.; Rajpurkar, P. Fineradscore: A radiology report line-by-line evaluation technique generating corrections with severity scores. *arXiv* **2024**, arXiv:2405.20613.
19. Turpin, M.; Michael, J.; Perez, E.; Bowman, S. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 74952–74965.
20. Ghosh, B.; Hasan, S.; Arafat, N.A.; Khan, A. Logical consistency of large language models in fact-checking. *arXiv* **2024**, arXiv:2412.16100.
21. Wei, Q.; Qian, K.; Li, X. FunBench: Benchmarking fundus reading skills of MLLMs. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2025; pp. 278–288.
22. Topol, E.J. High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **2019**, *25*, 44–56. [\[CrossRef\]](#)
23. Kermany, D.S.; Goldbaum, M.; Cai, W.; Valentim, C.C.; Liang, H.; Baxter, S.L.; McKeown, A.; Yang, G.; Wu, X.; Yan, F.; et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **2018**, *172*, 1122–1131. [\[CrossRef\]](#)

24. Thirunavukarasu, A.J.; Ting, D.S.J.; Elangovan, K.; Gutierrez, L.; Tan, T.F.; Ting, D.S.W. Large language models in medicine. *Nat. Med.* **2023**, *29*, 1930–1940. [\[CrossRef\]](#)
25. Liu, C.; Pan, J.; Shen, W.; Bai, W.; Rueckert, D.; Arcucci, R. How Far Have Medical Vision-Language Models Come? A Comprehensive Benchmarking Study. *arXiv* **2025**, arXiv:2507.11200. [\[CrossRef\]](#)
26. Li, Z.; Song, D.; Yang, Z.; Wang, D.; Li, F.; Zhang, X.; Kinahan, P.E.; Qiao, Y. Visionunite: A vision-language foundation model for ophthalmology enhanced with clinical knowledge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 11848–11862. [\[CrossRef\]](#)
27. Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. Mmbench: Is your multi-modal model an all-around player? In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2024; pp. 216–233.
28. Lau, J.J.; Gayen, S.; Ben Abacha, A.; Demner-Fushman, D. A dataset of clinically generated visual questions and answers about radiology images. *Sci. Data* **2018**, *5*, 180251. [\[CrossRef\]](#)
29. Liu, B.; Zhan, L.M.; Xu, L.; Ma, L.; Yang, Y.; Wu, X.M. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *Proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*; IEEE: New York, NY, USA, 2021; pp. 1650–1654.
30. Park, S.H.; Han, K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* **2018**, *286*, 800–809. [\[CrossRef\]](#)
31. Smit, A.; Jain, S.; Rajpurkar, P.; Pareek, A.; Ng, A.Y.; Lungren, M. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics (ACL): Kerrville, TX, USA, 2020; pp. 1500–1519.
32. Zhang, X.; Wu, C.; Zhao, Z.; Lin, W.; Zhang, Y.; Wang, Y.; Xie, W. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv* **2023**, arXiv:2305.10415.
33. Liu, B.; Zou, K.; Zhan, L.M.; Lu, Z.; Dong, X.; Chen, Y.; Xie, C.; Cao, J.; Wu, X.M.; Fu, H. Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2025; pp. 21310–21320.
34. Ko, M.; Lee, J.; Kim, H.; Kim, G.; Kang, J. Look at the first sentence: Position bias in question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; Association for Computational Linguistics (ACL): Kerrville, TX, USA, 2020; pp. 1109–1121.
35. Pezeshkpour, P.; Hruschka, E. Large language models sensitivity to the order of options in multiple-choice questions. *Proc. Find. Assoc. Comput. Linguist. NAACL* **2024**, *2024*, 2006–2017.
36. Zheng, C.; Zhou, H.; Meng, F.; Zhou, J.; Huang, M. Large language models are not robust multiple choice selectors. *Proc. Int. Conf. Learn. Represent.* **2024**, *2024*, 19426–19454.
37. Schilcher, P.; Karasin, D.; Schöpf, M.; Saleh, H.; Tommasel, A.; Schedl, M. Characterizing Positional Bias in Large Language Models: A Multi-Model Evaluation of Prompt Order Effects. *Proc. Find. Assoc. Comput. Linguist. EMNLP* **2025**, *2025*, 20643–20664.
38. Kang, S.; Kim, J.; Kim, J.; Hwang, S.J. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*; IEEE: New York, NY, USA, 2025; pp. 9339–9350.
39. Xu, Y.; Zhu, L.; Yang, Y. Mc-bench: A benchmark for multi-context visual grounding in the era of mllms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: New York, NY, USA, 2025; pp. 17675–17687.
40. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837.
41. Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; Zhou, D. Self-consistency improves chain of thought reasoning in language models. *arXiv* **2022**, arXiv:2203.11171.
42. Lin, G.Y.; Sun, C.E.; Weng, T.W. FoCUS: Improving Faithfulness in Chain-of-Thoughts by Training on Structured Reasoning Data. In *Proceedings of the 5th Workshop on Mathematical Reasoning and AI at NeurIPS*, San Diego, CA, USA, 6 December 2025.
43. Calanzone, D.; Teso, S.; Vergari, A. Logically consistent language models via neuro-symbolic integration. *Proc. Int. Conf. Learn. Represent.* **2025**, *2025*, 54495–54522.
44. Liu, Y.; Guo, Z.; Liang, T.; Shareghi, E.; Vulić, I.; Collier, N. Aligning with logic: Measuring, evaluating and improving logical preference consistency in large language models. *arXiv* **2024**, arXiv:2410.02205. [\[CrossRef\]](#)
45. Bansal, A.; Gangwani, I. Logical Consistency Under Pressure: Probing and Repairing Cross-Query Contradictions in LLMs. In *Proceedings of the ICLR 2026 Workshop on Logical Reasoning of Large Language Models*, Rio de Janeiro, Brazil, 26–27 April 2026.
46. Sun, Z.; Wang, Q.; Wang, H.; Zhang, X.; Xu, J. Detection and mitigation of hallucination in large reasoning models: A mechanistic perspective. *arXiv* **2025**, arXiv:2505.12886. [\[CrossRef\]](#)
47. Didelez, V.; Pigeot, I. Causality: Models, reasoning, and inference. *Politische Vierteljahresschr. (PVS)* **2001**, *42*, 313–315. [\[CrossRef\]](#)
48. Li, T.; Gao, Y.; Wang, K.; Guo, S.; Liu, H.; Kang, H. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Inf. Sci.* **2019**, *501*, 511–522. [\[CrossRef\]](#)

49. Liu, R.; Wang, X.; Wu, Q.; Dai, L.; Fang, X.; Yan, T.; Son, J.; Tang, S.; Li, J.; Gao, Z.; et al. DeepDRiD: Diabetic retinopathy—Grading and image quality estimation challenge. *Patterns* **2022**, *3*, 100512. [[CrossRef](#)]
50. Porwal, P.; Pachade, S.; Kamble, R.; Kokare, M.; Deshmukh, G.; Sahasrabuddhe, V.; Meriaudeau, F. Indian diabetic retinopathy image dataset (IDRiD): A database for diabetic retinopathy screening research. *Data* **2018**, *3*, 25. [[CrossRef](#)]
51. Cen, L.P.; Ji, J.; Lin, J.W.; Ju, S.T.; Lin, H.J.; Li, T.P.; Wang, Y.; Yang, J.F.; Liu, Y.F.; Tan, S.; et al. Automatic detection of 39 fundus diseases and conditions in retinal photographs using deep neural networks. *Nat. Commun.* **2021**, *12*, 4828. [[CrossRef](#)] [[PubMed](#)]
52. Wang, W.; Li, X.; Xu, Z.; Yu, W.; Zhao, J.; Ding, D.; Chen, Y. Learning two-stream CNN for multi-modal age-related macular degeneration categorization. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 4111–4122. [[CrossRef](#)] [[PubMed](#)]
53. Sotoudeh-Paima, S.; Jodeiri, A.; Hajizadeh, F.; Soltanian-Zadeh, H. Multi-scale convolutional neural network for automated AMD classification using retinal OCT images. *Comput. Biol. Med.* **2022**, *144*, 105368. [[CrossRef](#)] [[PubMed](#)]
54. Kulyabin, M.; Zhdanov, A.; Nikiforova, A.; Stepichev, A.; Kuznetsova, A.; Ronkin, M.; Borisov, V.; Bogachev, A.; Korotkich, S.; Constable, P.A.; et al. Octdl: Optical coherence tomography dataset for image-based deep learning methods. *Sci. Data* **2024**, *11*, 365. [[CrossRef](#)]
55. Gholami, P.; Roy, P.; Parthasarathy, M.K.; Lakshminarayanan, V. OCTID: Optical coherence tomography image database. *Comput. Electr. Eng.* **2020**, *81*, 106532. [[CrossRef](#)]
56. Li, N.; Li, T.; Hu, C.; Wang, K.; Kang, H. A benchmark of ocular disease intelligent recognition: One shot for multi-disease detection. In *Proceedings of the International Symposium on Benchmarking, Measuring and Optimization*; Springer: Cham, Switzerland, 2020; pp. 177–193.
57. Wei, Q.; Li, X.; Yu, W.; Zhang, X.; Zhang, Y.; Hu, B.; Mo, B.; Gong, D.; Chen, N.; Ding, D.; et al. Learn to segment retinal lesions and beyond. In *Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR)*; IEEE: New York, NY, USA, 2021; pp. 7403–7410.
58. Bogunović, H.; Venhuizen, F.; Klimscha, S.; Apostolopoulos, S.; Bab-Hadiashar, A.; Bagci, U.; Beg, M.F.; Bekalo, L.; Chen, Q.; Ciller, C.; et al. RETOUCH: The retinal OCT fluid detection and segmentation benchmark and challenge. *IEEE Trans. Med. Imaging* **2019**, *38*, 1858–1874. [[CrossRef](#)]
59. Pachade, S.; Porwal, P.; Thulkar, D.; Kokare, M.; Deshmukh, G.; Sahasrabuddhe, V.; Giancardo, L.; Quellec, G.; Mériaudeau, F. Retinal fundus multi-disease image dataset (rfmid): A dataset for multi-disease detection research. *Data* **2021**, *6*, 14. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.