X-PCR: A Benchmark for Cross-modality Progressive Clinical Reasoning in Ophthalmic Diagnosis

Supplementary Material

1. Dataset Details

Complete Source List. Table 1 summarizes the 51 public ophthalmic imaging datasets integrated into X-PCR. These datasets were compiled from diverse geographic sources, including Asia (China, India, Bangladesh, Japan), Europe (Spain, Czech Republic), and North America (United States), capturing a broad spectrum of clinical settings and patient demographics.

Task Coverage and Annotation Richness. The source datasets span a wide range of supervision types, including image-level classification (presence/absence, severity grading such as DR stages, and multi-label diagnosis, *e.g.*, ODIR-5K with 8 classes, JSIEC Retinal39 with 39 classes, RFMiD2.0 with 45 labels), pixel-level segmentation of anatomical structures (optic disc/cup, vessels, retinal layers) and pathological lesions (hemorrhages, exudates, drusen, fluid), as well as detection and localization tasks for fovea, lesions, and image quality. Several datasets provide rich, multi-task annotations; for example, REFUGE and GAMMA jointly supply classification labels and optic disc/cup masks, while OIA-DDR combines DR grading with lesion segmentation and detection. This diversity of labels enables us to construct VQA pairs that probe recognition, localization, and reasoning in a unified framework.

Imaging Modalities and Disease Spectrum. The datasets cover six primary ophthalmic imaging modalities, namely *External Photography* (EP), *Color Fundus Photography* (CFP), *Fluorescein Angiography* (FFA), *Indocyanine Green Angiography* (ICGA), *Optical Coherence Tomography* (OCT), and *RetCam*. Collectively, they span eight major disease categories relevant to clinical practice. Diabetic retinopathy is the most represented condition (datasets 7–10, 25–26, 36–38, 45), reflecting its status as a leading cause of vision loss worldwide. Glaucoma datasets (1–6, 21–24, 34, 40–41, 44, 46) provide extensive coverage of this chronic optic neuropathy, from early screening to progression monitoring. Age-related macular degeneration and other macular diseases are captured by OCT-centric datasets (11, 16, 31, 43, 49, 51) and multi-disease collections (12, 14–15, 17–18, 42, 47). Additional datasets on retinopathy of prematurity (20, 28), pathological myopia (29), and toxoplasmosis (19) ensure that rarer but clinically important entities are also included.

Licensing and Accessibility. All datasets were selected for their public availability under open licenses or formal research agreements. Most are released under Creative Commons licenses (*e.g.*, CC BY, CC BY-NC) or comparably permissive terms that allow academic use and redistribution, enabling X-PCR to adhere to open-access and reproducibility principles. Overall, Table 1 provides the basis for understanding the breadth and depth of X-PCR's source data and illustrates our aim to comprehensively cover the ophthalmic imaging landscape.

2. Imaging Modality Specifications

X-PCR encompasses six distinct ophthalmic imaging modalities, each offering unique diagnostic capabilities and complementary perspectives on ocular pathology. This section provides detailed technical descriptions, clinical applications, and quality characteristics for each modality.

2.1. External Photography (EP)

Technical Principles. External photography captures the anterior segment and periocular structures using standard digital cameras or dedicated ophthalmic imaging systems. In contrast to fundus photography, which targets the posterior pole, EP focuses on external ocular anatomy, *e.g.*, eyelids, conjunctiva, cornea, iris, and anterior lens surface.

Clinical Applications. External photography is routinely used for cataract assessment (visualizing lens opacities, nuclear sclerosis, and cortical changes through the pupil), anterior segment pathology (corneal opacities, pterygium, pinguecula, conjunctival lesions), eyelid disorders (chalazion, hordeolum, blepharitis, ptosis), ocular surface disease (dry eye, related signs, limbal stem cell deficiency), and pupillary abnormalities (anisocoria, iris defects, posterior synechiae). In X-PCR, EP images primarily support cataract grading tasks, in which lens opacity severity is assessed from nuclear color and cortical spoke patterns.

Sample Characteristics in X-PCR. X-PCR curates 4,064 high-quality EP images from an initial collection of 5,335 in Dataset 47 (Table 1). Image resolutions range from 640×480 to 4288×2848 pixels, with most images acquired at $\geq$ 1920×1080. Representative examples are shown in Fig. 1, illustrating (a) a normal anterior segment, (b) a cataract with lens opacity visible through the pupil, and (c) conjunctival hyperemia consistent with "red eye."

2.2. Color Fundus Photography (CFP)

Technical Principles. Color fundus photography is the most widely used retinal imaging modality, employing

Table 1. Summary of 51 public ophthalmic datasets.

Index	Dataset	Size	Source	Description
1	ACRIMA [15]	705	FISABIO Oftalmologia Medica	Glaucoma classification
2	LAG [33]	11,760	Beijing Tongren Hospital	Glaucoma classification
3	BEH [25]	634	Bangladesh Eye Hospital	Glaucoma classification
4	AIROGS [10]	113,893	500 sites	Glaucoma classification
5	Harvard-GDP [1]	1,544	Harvard Medical School	Glaucoma classification
6	JustRAIGS [29]	111,183	EyePACS LLC, US	2 classes (RG, NRG) + 10
7	Messidor-2 [11]	1,748	Brest University Hospital	Diabetic retinopathy
8	Eyepacs [27]	35,126	Aravind, Sankara Nethralaya, Narayana Nethralaya	DR severity grading
9	DeepDRiD [37]	2,000	Shanghai Sixth People's Hospital	5 classes
10	BiDR [60]	2,838	Not reported	DR severity grading
11	OCT2017 [28]	108,312	Not reported	4 classes: CNV, DME, DRUSEN, NORMAL
12	ODIR-5K [34]	5,000	Shanggong Medical Technology	8 classes
13	SUSTech-SYSU [13]	712	Zhongshan Ophthalmic Center, Sun Yat-sen University	Multi-label: 3 categories $\times$ 5 types $\times$ 5 grades
14	JSIEC Retinal39 [5]	1,000	Joint Shantou International Eye Center	39-class classification
15	MuReD [47]	2,208	ARIA, STARE, RFMiD datasets	20-disease classification
16	Retinal OCT-C8 [28]	24,000	OCT2017 dataset	8 classes: AMD, CNV, CSR, DME, MH, Drusen, DR, Normal
17	RFMiD2.0 [44]	3,200	Shri Guru Gobind Singhji	45 labels
18	MedMNISTv2 [57]	109,309	12 2D datasets and 6 3D datasets	12 2D datasets
19	ToxoFundus [22]	412	Hospital de Clinicas Medical Center	4 classes
20	FARFUM RoP [40]	1,533	Farabi Eye Hospital	3 stages
21	REFUGE [43]	1,200	Not reported	Glaucoma classification and optic disc/cup segmentation
22	RIM-ONE DL [21]	485	HUC, HUMS, HCSC, Spain	Glaucoma classification and optic disc/cup segmentation
23	GAMMA [32]	300	Zhongshan Ophthalmic Center, Sun Yat-sen University	Glaucoma grading, fovea localization, optic disc/cup segmentation
24	Harvard-FairSeg [19]	10,000	Harvard University and large academic eye hospitals	Segmentation
25	DRAC22 [38]	1,103	Not reported	segmentation, 3-class(Non-DR, NPDR, PDR)
26	MAPLES-DR [8]	198	MESSIDOR dataset	12-class
27	HVDROPDB [2]	600	Desai Eye Hospital	Segmentation
28	RetinalROP [53]	6,004	University Hospital Ostrava	segmentation
29	PALM [20]	1,200	Not reported	classification, segmentation, localization
30	FIVES [26]	800	Second Affiliated Hospital of Zhejiang University	4-class task
31	AMD-SD [23]	3,048	Second Affiliated Hospital of Nanchang University	Segmentation: IRF, SRF, PED, SHRM, IS/OS disruption
32	OCTA-500 [46]	500	Jiangsu Province Hospital	—
33	OLIVES [48]	1,268	Retina Consultants of Texas	Segmentation
34	Harvard-GF [16]	3,300	Not reported	Glaucoma detection, segmentation
36	OIA-DDR [35]	13,673	147 hospitals across 23 provinces in China	DR classification, segmentation, detection
37	IDRiD [45]	1,032	India	Labels for DR severity and diabetic macular edema
38	MESSIDOR2 [**?**]	1,748	Brest University Hospital	Diabetic retinopathy (DR)
40	PAPILA [14]	488	Carlos III Health Institute	segmentation
41	Glaucoma Fundus [17]	1,544	Harvard database	—
42	Retina [54]	601	Not reported	4 classes
43	OCTID [39]	587	University of Waterloo	6-class classification task
44	REFUGE2 [52]	1,200	Not reported	segmentation, localization
45	APTOS-2019 [51]	5,590	Aravind Eye Hospital, India	5 DR grades
46	Harvard-GDP [18]	3,300	Asian, Black, and White	Includes visual field, demographics, and glaucoma labels
47	Eye Disease [24]	5,335	Bangladesh Eye Hospital	10-class classification
48	OIMHS [58]	3,859	Multiple centers in Asia	Segmentation of retina, macular hole, intraretinal cysts, choroid
49	OCT5k [61]	1,672	UCL Institute of Ophthalmology	Segmentation and detection tasks
50	EyePhotos [9]	30,000	Not reported	—
51	Retinal OCT [28]	84,495	Not reported	4 classes: NORMAL, CNV, DME, DRUSEN

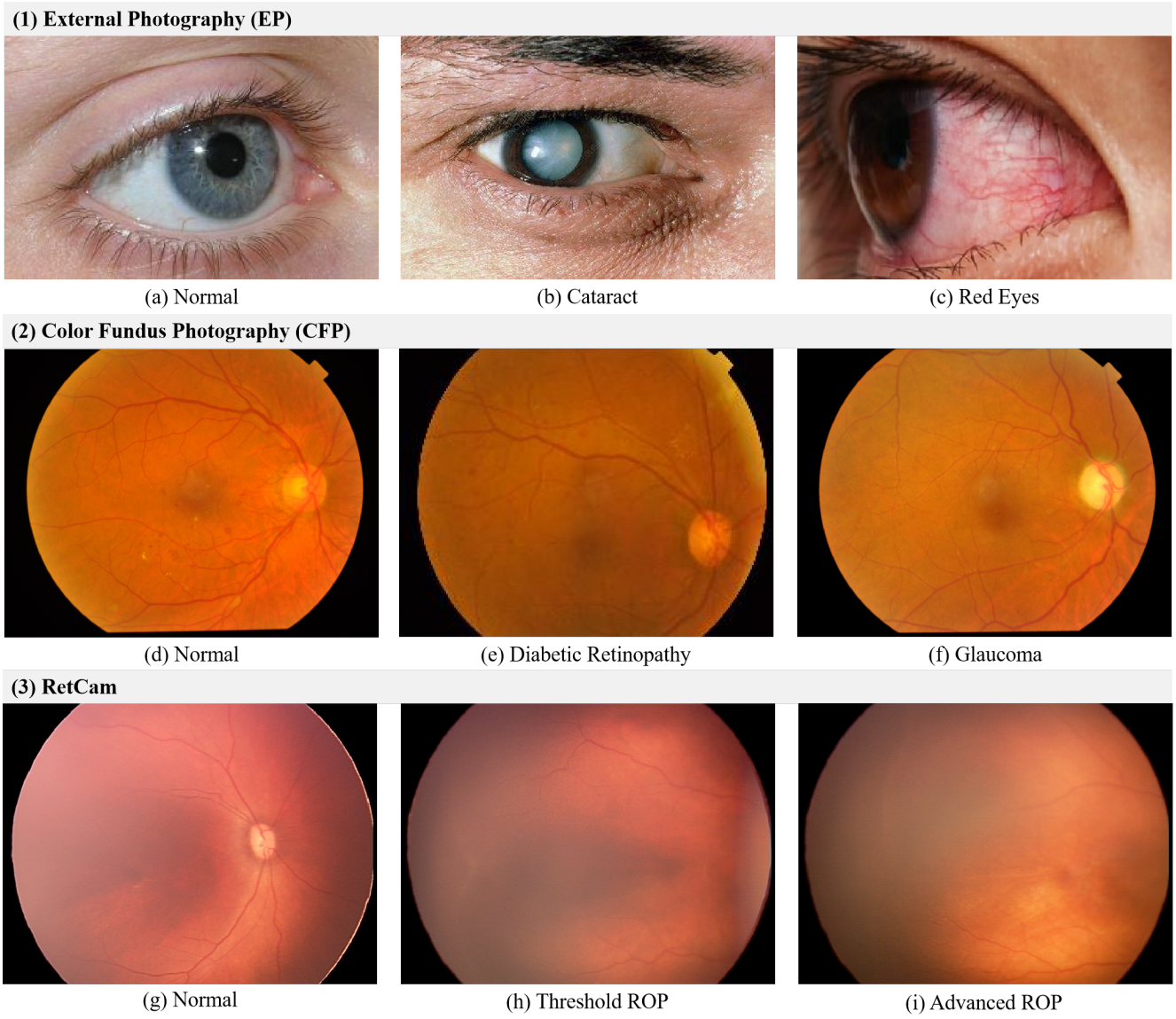

Figure 1. Representative ophthalmic images. (a–c) External photography (EP); (d–f) color fundus photography (CFP); (g–i) pediatric RetCam images.

white-light illumination to acquire two-dimensional color views of the posterior segment. Modern fundus cameras project light through the undilated or dilated pupil via a dedicated optical system that minimizes corneal reflections. Image contrast arises from the reflective properties of the retinal pigment epithelium and choroid, with melanin and hemoglobin absorption producing the characteristic red–orange fundus appearance.

Clinical Applications. Color fundus photography is a primary tool for screening and monitoring a broad spectrum of retinal diseases. In diabetic retinopathy, it enables the detection of microaneurysms, hemorrhages, exudates, cotton-wool spots, neovascularization, and retinal thickening. For glaucoma, CFP supports assessment of optic disc morphology, including cup-to-disc ratio, neuroretinal rim thinning, and disc hemorrhages, as well as retinal nerve fiber layer defects. In age related macular degeneration, it reveals drusen, geographic atrophy, pigmentary changes, and signs suggestive of choroidal neovascularization. CFP is also essential for grading hypertensive retinopathy (AV nicking, vessel tortuosity, flame hemorrhages, macular star), characterizing retinal vein occlusion (dilated tortuous veins, intraretinal hemorrhages, macular edema), and documenting pathological myopia (lacquer cracks, chorioretinal atrophy).

Sample Characteristics in X-PCR. In X-PCR, CFP constitutes the largest imaging subset, select 6,840 high-

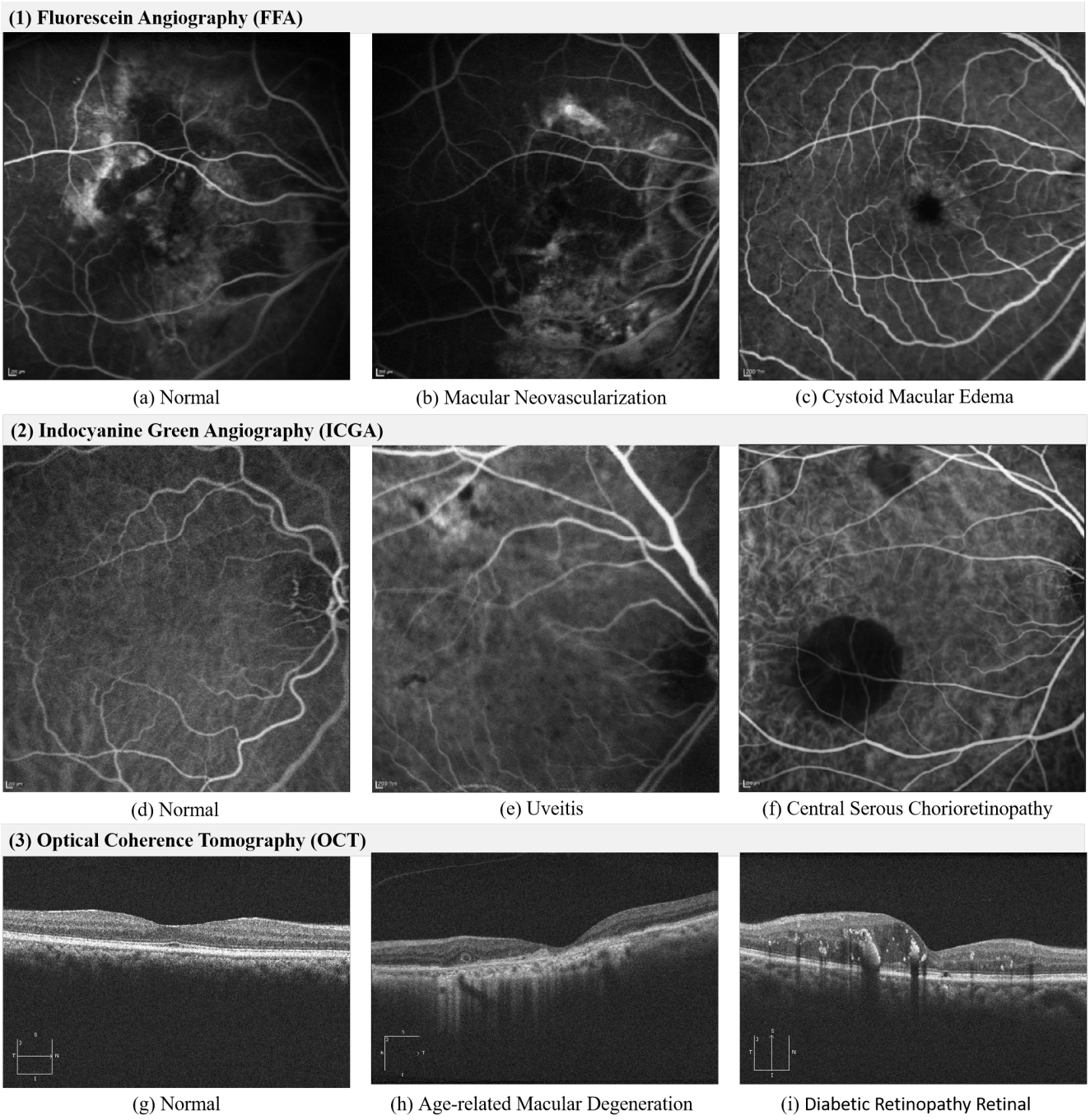

Figure 2. Representative ophthalmic images. (a–c) Fluorescein Angiography (FFA); (d–f) Indocyanine Green Angiography (ICGA); (g–i) Optical Coherence Tomography (OCT).

quality images from 8 datasets (including Eyepacs, Messidor-2, APTOS-2019, ODIR-5K, and others in Table 1). Fig. 1 shows representative cases: (d) a normal fundus with a well-defined optic disc and foveal reflex, (e) moderate non-proliferative diabetic retinopathy with microaneurysms and hard exudates, and (f) glaucomatous optic neuropathy with an enlarged cup-to-disc ratio and inferotemporal rim thinning.

2.3. RetCam

Technical Principles. RetCam (Natus Medical Inc.) is a wide-field digital fundus imaging system specifically en-

gineered for pediatric retinal examination, particularly in neonates and infants. Compared with conventional adult fundus cameras, RetCam uses a contact wide-angle lens and flexible optics that accommodate the steep corneal curvature, small palpebral fissure, and often undilated pupils of infant eyes, enabling panretinal views in this challenging population.

Clinical Applications. RetCam is now a core tool in pediatric ophthalmology, with retinopathy of prematurity (ROP) screening as its principal indication. It allows documentation of avascular peripheral retina, demarcation lines and ridges, plus disease (vascular tortuosity and dilation), and supports standardized staging (Stages 1–5) and zone classification (I–III). RetCam is also used to visualize peripheral vascular anomalies in familial exudative vitreoretinopathy and other infantile retinal vasculopathies.

Sample Characteristics in X-PCR. In X-PCR, the RetCam subset select 5,403 images curated from two ROP-focused datasets (datasets 20 and 28 in Table 1). Representative examples in Fig. 1 (RetCam panel) illustrate (g) a normal fully vascularized retina in a term infant, (h) threshold ROP with dense vitreoretinal proliferation obscuring posterior pole details, and (i) advanced ROP with diffuse media haze and poorly visualized retinal vasculature.

2.4. Fluorescein Angiography (FFA)

Technical Principles. Fluorescein angiography is a dynamic imaging technique that visualizes retinal and, to a lesser extent, choroidal circulation by recording the passage of intravenously injected sodium fluorescein. Blue excitation light (approximately 465–490 nm) stimulates fluorescein molecules, which emit yellow–green light (approximately 520–530 nm) that is captured through a barrier filter. Sequential images acquired at standardized time points sample the arterial, arteriovenous, and venous phases, enabling phase-resolved vascular assessment.

Clinical Applications. Fluorescein angiography provides critical diagnostic information for a range of retinal vascular disorders. In diabetic retinopathy, it delineates capillary non-perfusion, microaneurysm leakage, and neovascularization. In age-related macular degeneration, it characterizes choroidal neovascularization (classic, occult, mixed) and leakage patterns that guide treatment eligibility. For retinal vein occlusion, FFA quantifies ischemic areas, collateral vessel formation, and macular edema leakage. It is also central to diagnosing central serous chorioretinopathy (focal leakage points, smokestack patterns), retinal vasculitis (vessel wall staining and leakage), and selected inherited retinal diseases (*e.g.*, transmission defects in retinitis pigmentosa and pattern dystrophies).

Sample Characteristics in X-PCR. In X-PCR, we select 3,356 high-quality images from 6 source datasets, primarily specialized retinal disease cohorts. The collection spans early ($\sim$30%), mid ($\sim$45%), and late ($\sim$25%) phases, providing complementary views of dye filling, leakage, and staining. Image resolution typically ranges from 768×768 to 2048×2048 pixels. Representative examples in Fig. 2 illustrate three macular phenotypes: (a) a normal macular angiogram, (b) macular neovascularization with abnormal subfoveal vascular complexes, and (c) cystoid macular edema with petaloid leakage in the macular region.

2.5. Indocyanine Green Angiography (ICGA)

Technical Principles. Indocyanine green angiography complements FFA by specifically visualizing choroidal circulation using indocyanine green (ICG), which fluoresces in the near-infrared range (excitation $\sim$790–805 nm, emission $\sim$835 nm). In contrast to fluorescein, ICG is $\sim$98% protein-bound, producing minimal leakage from choriocapillaris fenestrations. The longer wavelength penetrates the retinal pigment epithelium and hemoglobin more effectively, yielding superior visualization of choroidal vessels and filling patterns. ICGA follows FFA-like early, mid, and late phases but with emphasis on choroidal perfusion and dye pooling.

Clinical Applications. Indocyanine green angiography is particularly valuable in diseases with predominant choroidal involvement. In polypoidal choroidal vasculopathy, it reveals polypoidal lesions and branching vascular networks that may be inconspicuous on FFA. In central serous chorioretinopathy, ICGA demonstrates choroidal vascular hyperpermeability and dilated choroidal vessels. It refines subtype classification in neovascular AMD by distinguishing occult CNV patterns and detecting retinal angiomatous proliferation. ICGA also aids in assessing choroidal tumors (melanoma, hemangioma) and inflammatory chorioretinopathies, such as multifocal choroiditis and birdshot chorioretinopathy, where hypofluorescent dark dots highlight areas of choroidal hypoperfusion or infiltration.

Sample Characteristics in X-PCR. In X-PCR, we select 3,356 high-quality images from 2 specialized macular disease datasets. Fig. 2 illustrates (d) a normal ICGA with regular choroidal filling and large-caliber vessels, (e) uveitis with vascular leakage and patchy hypofluorescent dark areas, and (f) central serous chorioretinopathy with regions of choroidal hyperpermeability in the macular area.

2.6. Optical Coherence Tomography (OCT)

Technical Principles. Optical coherence tomography uses low-coherence interferometry to acquire high-resolution cross-sectional images of retinal microstructure. Spectral-domain OCT (SD-OCT), the current clinical standard, records the interference pattern between light reflected from retinal layers in the sample arm and a reference beam using a spectrometer, enabling micrometer-scale axial resolution.

Clinical Applications. OCT has transformed the diagnosis and monitoring of macular and retinal disease. In diabetic macular edema, it quantifies central subfield thickness and reveals intraretinal cysts, subretinal fluid, and hard exudate deposits. In age-related macular degeneration, OCT visualizes drusen morphology, retinal pigment epithelium detachment, and subretinal or intraretinal fluid from choroidal neovascularization, as well as geographic atrophy. For glaucoma, it provides quantitative measurements of retinal nerve fiber layer thickness, ganglion cell complex integrity, and optic nerve head morphology. OCT is also indispensable for staging macular holes, characterizing epiretinal membranes and associated foveal contour distortion, assessing macular edema and inner retinal thickening in retinal vein occlusion, and evaluating subretinal fluid, RPE irregularities, and choroidal thickness in central serous chorioretinopathy.

Sample Characteristics in X-PCR. In X-PCR, OCT is the second-largest imaging modality, we select 8,199 B-scans and corresponding en face projections from 13 datasets (including OCT2017, Retinal OCT-C8, OLIVES, AMD-SD, OCTID, and OCT5k; see Table 1). Fig. 2 illustrates (g) a normal macular scan with well-delineated retinal layers and a central foveal depression, (h) age-related macular degeneration with drusen deposition and pigment epithelial detachment in the macular region, and (i) diabetic retinopathy with cystoid changes and macular thickening on OCT.

3. VQA Generation Details

X-PCR's VQA generation pipeline transforms heterogeneous clinical annotations into a unified question-answer framework aligned with the six-stage progressive clinical reasoning chain. This section details the systematic conversion strategies, template taxonomy, and quality assurance mechanisms.

3.1. Progressive Reasoning Aligned Question Taxonomy

Stage 1: Image Quality Assessment (IQA). Stage 1 targets pre-diagnostic image quality control, ensuring that subsequent clinical reasoning is not confounded by technically inadequate inputs. Question types are aligned with three core skills: (1) *Binary quality classification*, where the model judges whether a given modality-specific image is of sufficient quality for a specified clinical task (*e.g.*, DR screening, macular edema evaluation). (2) *Quality defect identification*, where the model enumerates technical artifacts (*e.g.*, blur, motion, under/overexposure, truncation, mirror or banding artifacts) and specifies which anatomical regions and diagnostic subtasks are affected.

Listing 1. Example template for Stage 1 IQA QA pair.

```
Q (Stage 1 - IQA, <MODALITY>):
  "Is this <MODALITY> image of sufficient
      quality for <CLINICAL_TASK>?"

A:
  "<YES/NO>. <JUSTIFICATION: affected
      regions (e.g., macula, optic disc,
      periphery).>"
```

Annotation Conversion. Source datasets with explicit image-quality flags are converted into binary IQA questions. Adversarial low-quality images (*e.g.*, low resolution, high blur, severe artifacts) serve as negative examples emphasizing defect recognition, while a curated high-quality subset provides positive examples focused on diagnostic adequacy.

Stage 2: Anatomical Localization (AL). Stage 2 establishes a spatial reference frame for lesion mapping and pathology interpretation. We design three AL-oriented question families: (1) *structure identification*, asking the model to localize key anatomy (*e.g.*, optic disc) using clock-hours, quadrants, or disc-diameter offsets; (2) *anatomical relationship mapping*, probing distances, directions, and preservation of normal spatial relationships; and (3) *modality-specific landmark detection*, requiring enumeration of canonical landmarks or layers (*e.g.*, retinal layers on OCT) under standard clinical conventions.

Annotation Conversion. Bounding-box annotations are turned into "locate [structure]" questions by mapping coordinates to anatomical descriptions (*e.g.*, nasal/temporal, superior/inferior, clock-hours). Segmentation masks support OCT layer-identification, while images without explicit labels use modality-standard landmarks as references.

Multi-Label Handling. For images with multiple annotated structures, we generate separate AL questions per landmark plus relational questions on pairwise spatial organization; for OCT with layer-wise segmentation, we also form composite questions that require listing all visible layers in anatomical order.

Listing 2. Example template for Stage 2 AL QA pair.

```
Q (Stage 2 - AL, <MODALITY>):
  "Identify and locate the <
      ANATOMICAL_STRUCTURE> in this image."

A:
  "<PRESENT/ABSENT>.
   <SPATIAL_DESCRIPTION using clock-hour
       notation, quadrants,
   or distances in disc diameters relative
       to the fovea/optic disc.>"
```

Stage 3: Lesion Characterization (LC). Stage 3 describes pathological findings with standardized clinical

terms as the basis for diagnosis. We define four LC question families: (1) *lesion detection*, where the model decides whether a lesion type is present and, if so, estimates its number and coarse distribution; (2) *morphological description*, where it summarizes lesion appearance (size, shape, color/reflectivity, borders, texture, associated features); (3) *distribution pattern analysis*, where it captures spatial patterns, anatomical zones, and severity gradients; and (4) *multi-lesion composite characterization*, where it lists all pathological findings in an image and anchors them to key landmarks (*e.g.*, arcades, macula, disc).

Annotation Conversion. Bounding boxes are turned into LC questions by mapping coordinates to anatomical regions (quadrants, clock-hours, disc–fovea distances) and generating brief location descriptions. Segmentation masks yield area and shape statistics that are expressed as size, shape, and location terms, while point annotations (*e.g.*, microaneurysms) support lesion counts and clustered vs. scattered patterns. Multi-class labels give rise to either per-lesion or composite "list all findings" questions, and numeric severity scores are converted to qualitative grades (mild, moderate, severe).

Multi-Label Handling. For images with multiple lesion types, we use three strategies: (1) *decomposition*, generating separate single-lesion questions for fine-grained evaluation; (2) *composite integration*, asking holistic questions that require listing and structuring all visible findings in complex cases (*e.g.*, advanced DR or AMD); and (3) *hierarchical prioritization*, focusing on vision-threatening and disease-defining lesions (*e.g.*, neovascularization, macular edema, drusen, glaucomatous cupping) while summarizing less critical findings (*e.g.*, scattered hemorrhages or exudates).

Listing 3. Example template for Stage 3 LC QA pair.

```
Q (Stage 3 - LC, <MODALITY>):
  "Describe the morphology and
     characteristics of <LESION_TYPE>
   in this image."

A:
  "Size: <SIZE or extent>.
   Shape: <SHAPE (e.g., round, irregular)>.
   Color / reflectivity: <COLOR or intensity
        pattern>.
   Borders: <BORDER definition (e.g., well/
       ill-defined)>.
   Texture: <TEXTURE (e.g., smooth, granular
       )>.
   Associated features: <RELATED FINDINGS
   (e.g., edema, hemorrhage, exudates, RPE
       changes)>."
```

Stage 4: Disease Diagnosis (DD). Stage 4 synthesizes anatomical (AL) and lesion-level (LC) evidence into explicit diagnostic conclusions. We define four DD question families: (1) *single-disease diagnosis*, where the model outputs the most likely diagnosis with brief supporting evidence; (2) *differential diagnosis ranking*, where it provides a ranked list of 2–3 candidate diseases with key discriminating features; (3) *multi-disease co-occurrence*, where it enumerates all concurrent conditions in the same image; and (4) *normal vs. abnormal assessment*, where it decides whether the image is normal and justifies this by confirming or excluding common pathologies.

Annotation Conversion. Single-disease labels are converted into "What is your diagnosis?" questions, with answers giving the primary diagnosis and a brief severity justification grounded in LC findings. Multi-class labels become questions that ask the model to identify all diseases present and, when needed, prioritize them by clinical urgency. Hierarchical disease codes (*e.g.*, ICD-10) are mapped to natural-language diagnoses, while "normal" or "no DR" labels become normality-check questions ("Is this image normal or abnormal?"). Low-confidence or ambiguous labels are used to construct differential-diagnosis questions requiring a ranked list of candidates and key discriminating features.

Consistency and Multi-Label Handling. DD answers are required to be consistent with earlier stages: automated checks verify that each diagnosis is supported by LC-described lesions, and inconsistencies are flagged for expert review. For multi-disease cases, the model should distinguish overlapping from independent features and briefly comment on the clinical impact of co-occurrence, keeping DD aligned with the structured evidence from AL and LC.

Listing 4. Example template for Stage 4 DD QA pair

```
Q (Stage 4 - DD):
  "Based on the imaging findings, what is
      the most likely diagnosis?"

A:
  Primary diagnosis: <PRIMARY DIAGNOSIS>
  Supporting evidence from LC: <EVIDENCE
      FROM LC FINDINGS>
  Differential considerations: <DIFFERENTIAL
       DIAGNOSES>
```

Stage 5: Severity Grading (SG). Stage 5 uses disease-specific scales to quantify progression and guide treatment. We define four SG question types: (1) *scale-based grading*, where the model applies a standard grading scale (*e.g.*, ICDR, ROP); (2) *component-based scoring*, evaluating specific disease features (*e.g.*, arteriolar tortuosity in ROP); (3) *progression assessment*, comparing current severity to prior images; and (4) *risk stratification*, assessing the likelihood of adverse outcomes (*e.g.*, glaucoma progression).

Annotation Conversion. Source annotations are

mapped to SG questions: ordinal labels are converted to scale grading, numeric scores to clinical terms, and multi-component scores into separate assessments. Longitudinal labels are used for progression, and missing severity labels are inferred from LC findings.

Listing 5. Example template for Stage 5 SG QA pair

```
Q (Stage 5 - SG):
  "Based on current findings, assess the
      risk of [adverse outcome]."

A:
  Risk category (low/moderate/high) +
      contributing factors + recommended
      monitoring
```

Stage 6: Clinical Decision-Making (CD).

Stage 6 applies diagnostic and severity assessments to create actionable management plans. It consists of four key question families: (1) Treatment recommendation, where the model suggests the primary intervention with rationale, alternatives, and contraindications; (2) Referral urgency, assessing how urgently a patient should be referred; (3) Monitoring strategy, recommending follow-up intervals, imaging, and clinical assessments; and (4) Contraindication assessment, evaluating risks and contraindications for the proposed treatment. Consistency Validation ensures that the CD recommendations align with the diagnosis and severity (DD-SG-CD coherence) and verifies that decisions follow clinical guidelines and include actionable details such as treatment dosing, follow-up, and escalation criteria.

Listing 6. Example template for Stage 6 CD QA pair

```
Q (Stage 6 - CD):
  "What treatment is recommended based on
      the diagnosis and severity?"

A:
  Primary intervention + rationale +
      alternative options +
      contraindications
```

3.2. Answer Generation and Validation Pipeline

The following outlines the detailed process for generating, validating, and ensuring quality control in the construction of diagnostic reasoning tasks in X-PCR. The process is divided into five key stages: Annotation Extraction, Template Selection, Answer Synthesis, Clinical Validation, and Quality Assurance. Each stage involves systematic tasks that ensure the accuracy, consistency, and relevance of the generated questions and answers.

Step 1: Annotation Extraction

– Parse source dataset labels (disease, severity, bounding boxes, etc.)
– Extract metadata (patient age, imaging device, field of view)
– Identify annotation gaps requiring inference

Step 2: Template Selection

– Map (annotation type + reasoning stage) → question template
– Prioritize templates based on annotation richness
– Generate question variations for data augmentation

Step 3: Answer Synthesis

– For explicitly labeled data:
  – Direct mapping: Label → structured answer
  – Enrichment: Add clinical context from guidelines
– For unlabeled aspects:
  – Expert-in-the-loop annotation (for critical cases)
  – GPT-4V-based answer generation with validation
– Multi-source reconciliation (for overlapping datasets)

Step 4: Clinical Validation

– Automated checks:
  – Anatomical plausibility (lesion locations)
  – Terminology consistency (standardized lexicon)
  – Cross-stage coherence (LC → DD → SG → CD chain)
  – Quantitative accuracy (measurements, counts)
– Expert review (board-certified ophthalmologists):
  – Sample 10% of answers per reasoning stage
  – Flag and correct errors
  – Refine answer quality guidelines
– Inter-annotator agreement:
  – Three experts independently review 500 VQA pairs
  – Cohen's kappa ¿0.85 required for release

Step 5: Quality Assurance

– Readability assessment (Flesch-Kincaid grade level)
– Answer length distribution analysis (avoid excessive verbosity)
– Template diversity verification (avoid repetitive phrasing)
– Adversarial filtering (remove ambiguous or misleading Q-A pairs)

The validation metrics for the X-PCR benchmark demonstrate high quality and consistency in the generated answers. The annotation consistency rate is 94.3%, with automated checks passing without the need for expert revision. Expert agreement, measured by Cohen's K, is strong, with values of 0.87 for disease diagnosis, 0.82 for severity grading, and 0.79 for clinical decisions. Additionally, 98.1% of answers address all question components, and 96.7% of clinical decision (CD) answers align with the AAO Preferred Practice Patterns or equivalent clinical guidelines, ensuring the benchmark's clinical relevance and reliability.

3.3. Distribution Across Diseases

Table 2 summarizes the disease taxonomy and per-class statistics in the X-PCR benchmark, which includes 52 distinct retinal and optic nerve conditions. The table catego-

Table 2. Disease taxonomy and per-class statistics in X-PCR.

Index	Category	Full name	Abbreviation	Images	QA Pairs
1	DR	No diabetic retinopathy	No-DR	1258	7128
2	DR	Mild non-proliferative diabetic retinopathy	Mild-NPDR	1048	5940
3	DR	Moderate non-proliferative diabetic retinopathy	Mod-NPDR	839	4752
4	DR	Severe non-proliferative diabetic retinopathy	Sev-NPDR	629	3564
5	DR	Proliferative diabetic retinopathy	PDR	629	3564
6	DR	Non–center-involving diabetic macular edema	Non-CI-DME	629	3564
7	DR	Center-involving diabetic macular edema	CI-DME	628	3564
8	DR	Clinically significant macular edema	CSME	419	2376
9	DR	Neovascularization elsewhere secondary to DR	NVE-DR	209	1187
10	DR	Neovascularization at the disc secondary to DR	NVD-DR	209	1187
11	DR	Fibrotic proliferative diabetic retinopathy	Fib-PDR	209	1187
12	DR	Diabetic tractional retinal detachment	DR-TRD	209	1187
13	GLA	Glaucoma suspect	GLA-Sus	981	5312
14	GLA	Primary open-angle glaucoma, early	POAG-Early	981	5312
15	GLA	Primary open-angle glaucoma, moderate	POAG-Mod	736	3984
16	GLA	Primary open-angle glaucoma, advanced	POAG-Adv	491	2656
17	GLA	Primary angle-closure glaucoma	PACG	490	2656
18	GLA	Normal-tension glaucoma	NTG	490	2656
19	GLA	Juvenile open-angle glaucoma	JOAG	245	1328
20	GLA	Ocular hypertension	OHT	736	3984
21	CAT	No cataract	No-CAT	198	1914
22	CAT	Nuclear sclerotic cataract	NS-CAT	398	3828
23	CAT	Cortical cataract	Cort-CAT	298	2871
24	CAT	Posterior subcapsular cataract	PSC-CAT	198	1914
25	AMD	Early age-related macular degeneration	Early-AMD	1187	8500
26	AMD	Intermediate age-related macular degeneration	Int-AMD	949	6800
27	AMD	Neovascular age-related macular degeneration	nAMD	712	5100
28	AMD	Geographic atrophy secondary to AMD	GA-AMD	474	3400
29	AMD	Retinal pigment epithelial detachment secondary to AMD	AMD-PED	474	3400
30	AMD	Polypoidal choroidal vasculopathy	PCV	237	1700
31	AMD	Retinal angiomatous proliferation secondary to AMD	RAP-AMD	237	1700
32	HTR	Mild-to-moderate hypertensive retinopathy	HTR-MildMod	936	8426
33	HTR	Severe hypertensive retinopathy	HTR-Sev	468	4213
34	HTR	Hypertensive maculopathy	HTR-Mac	233	2106
35	PM	Pathologic myopia without macular complication	PM-NoMac	727	6520
36	PM	Myopic macular degeneration	MMD	546	4890
37	PM	Myopic choroidal neovascularization	mCNV	363	3260
38	PM	Myopic traction maculopathy	MTM	181	1630
39	RVO	Non-ischemic branch retinal vein occlusion	BRVO-NI	977	7867
40	RVO	Ischemic branch retinal vein occlusion	BRVO-I	488	3933
41	RVO	Non-ischemic central retinal vein occlusion	CRVO-NI	733	5901
42	RVO	Ischemic central retinal vein occlusion	CRVO-I	488	3933
43	RVO	Venous stasis retinopathy	VSR	244	1966
44	Rare	Non-arteritic anterior ischemic optic neuropathy	NAION	474	2424
45	Rare	Bilateral acute macular neuroretinopathy	AMN	236	1212
46	Rare	Toxoplasmosis chorioretinitis	TOCR	356	1818
47	Rare	Tubercular multifocal choroiditis	TB-MC	356	1818
48	Rare	Multiple evanescent white dot syndrome	MEWDS	356	1818
49	Rare	Purtscher-like retinopathy	PLR	236	1211
50	Rare	Cytomegalovirus retinitis	CMV-Ret	236	1211
51	Rare	Susac syndrome	Susac	118	605
52	Rare	Terson syndrome	Terson	236	1211

DR: diabetic retinopathy; **AMD**: age-related macular degeneration; **GLA**: glaucoma; **RVO**: retinal vein occlusion; **PM**: pathological myopia; **HTR**: hypertensive retinopathy; **CAT**: cataract; **Rare**: rare retinal and optic nerve conditions.

rizes diseases into eight primary groups: diabetic retinopathy (DR), age-related macular degeneration (AMD), glaucoma (GLA), retinal vein occlusion (RVO), pathological myopia (PM), hypertensive retinopathy (HTR), cataract (CAT), and rare conditions. Each disease class is listed with the number of images and question-answer (QA) pairs as-

sociated with it. The dataset covers a wide range of severity levels, from early stages (*e.g.*, mild diabetic retinopathy) to advanced forms (*e.g.*, proliferative diabetic retinopathy, macular degeneration, and various rare retinal diseases), enabling comprehensive evaluation of pathological findings and clinical decision-making.

4. Metrics Details

4.1. Difficulty Aware Metrics

We implement a three-tier grading protocol that reflects progressive clinical expertise. *Resident-level (R-level)* items probe foundational knowledge and pattern perception capability expected during the first two years of ophthalmology training. *Attending-level (A-level)* items require integration of clinical context, formulation and prioritization of differential diagnoses, and selection of appropriate management consistent with standard guidelines. *Specialist-level (S-level)* items demand subspecialty proficiency (*e.g.*, medical retina), including recognition of uncommon subtypes, interpretation of complex multimodal patterns and evidence-based therapeutic decision-making. This stratification enables performance analyses aligned with real-world competency milestones, distinguishing surface-level pattern recognition from context-aware reasoning and subspecialty expertise.

Question difficulty was assigned via a controlled, multi-stage protocol with *small-sample, high-quality calibration*. First, on a stratified *gold set* (2% per disease–modality–stage stratum), fifteen board-certified ophthalmologists (five attending; ten specialists) independently labeled each case as *resident* (R), *attending* (A), or *specialist* (S), guided by: (i) knowledge requirements, (ii) diagnostic complexity, and (iii) decision-making nuance. Second, inter-rater reliability on the gold set was assessed using Fleiss' κ. Items with $\kappa < 0.60$ underwent structured adjudication and iterative re-annotation until $\kappa \geq 0.70$, ensuring consistent and transparent tier construction. Third, we validated the gold-set difficulty labels against stratified human performance. Residents ($n = 25$) were expected to achieve $> 70\%$ on R-level, 40–60% on A-level, and $< 40\%$ on S-level; attendings ($n = 18$) $> 70\%$ on A-level and 50–70% on S-level; subspecialty fellows ($n = 12$) $> 80\%$ on S-level. Items outside these envelopes were excluded. Fourth, calibrated by the gold set, the remaining large corpus received *provisional* difficulty labels via rule-based features, followed by periodic expert audits on 1–2% samples per stratum to control drift. Finally, because clinical utility extends beyond difficulty, each case received an *importance* score (1–5) reflecting diagnostic impact (risk of missing sight-threatening disease), treatment implications (likelihood of inappropriate management). The final weight was

$$\tilde{w}_i \;= \alpha \cdot \widehat{\mathrm{Diff}}_d + (1-\alpha) \cdot \widehat{\mathrm{Impact}}_d, \tag{1}$$

where $\widehat{\mathrm{Diff}}_d \in \{1,2,3\}$ (R=1, A=2, S=3), $\widehat{\mathrm{Impact}}_d \in [1,5]$ (importance), and $\alpha = 0.6$ reflects expert consensus prioritizing safety-critical scenarios. This scheme integrates pedagogic difficulty with clinical consequence, yielding evaluation scores aligned with real-world practice. Let N be the number of cases and $S{=}6$ be the number of reasoning stages. Denote by $a_{i,j} \in \{0,1\}$ the correctness of case j at stage i and by $\mathbb{1}[\cdot]$ the indicator function.

Stage-Wise Accuracy (SWA). The per-stage accuracy is

$$\mathrm{SWA}_i \;=\; \frac{1}{N}\sum_{j=1}^{N} a_{i,j}\,, \quad i \in \{1,\ldots,6\}. \tag{2}$$

Chain Completion Rate (CCR). The proportion of cases with all six stages correct is

$$\mathrm{CCR} \;=\; \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}\left[\bigwedge_{i=1}^{6}(a_{i,j}=1)\right]. \tag{3}$$

Expertise-Stratified Accuracy (ESA). For level $\ell \in \{R, A, S\}$,

$$\mathrm{ESA}_\ell \;=\; \frac{\text{Correct answers at level } \ell}{\text{Total questions at level } \ell}\,. \tag{4}$$

For each case, the model answers a six-stage chain of questions (one per stage). We evaluate under two regimes: Independent (IND), where each stage is answered in isolation with no prior-stage context; and Cascaded (CAS), where each stage consumes the model's own prediction from the preceding stage as context.

4.2. Uncertainty Aware Metrics

For each case, model outputs are mapped to four confidence-aware categories: *(CC) Correct–Confident*: correct with confidence ≥ 0.70; *(CU) Correct–Uncertain*: correct with confidence < 0.70; *(IU) Incorrect–Uncertain*: incorrect with confidence < 0.50 (model acknowledges limits); *(IC) Incorrect–Confident*: incorrect with confidence ≥ 0.50 (overconfident error).

We score each response with a confidence-aware scheme that rewards calibration. Let N be the number of items and s_i the score for item i:

$$s_i^{\text{base}} = \begin{cases} 1.0, & \text{CC} \\ 0.7, & \text{CU} \\ 0.3, & \text{IU} \\ -0.5, & \text{IC} \end{cases} \tag{5}$$

Final Weighted UAS

$$\mathrm{UAS} \;=\; \sum_{i=1}^{N} \tilde{w}_i \, s_i^{\text{base}} \tag{6}$$

Table 3. Evaluation of MLLMs across 6 stages on EP modality

	MLLM	IQA	AL	LC	DD	SG	CD
Commercial	GPT-5 [41]	99.90	89.07	81.96	73.95	62.36	55.34
	Gemini-2.5-Pro [12]	95.01	84.43	71.47	72.81	58.33	37.52
	GLM-4.5v [59]	96.10	86.74	69.31	75.07	53.55	49.43
	GPT-5-nano [42]	95.38	81.29	67.99	67.29	56.30	51.10
	Gemini-2.5-Flash [12]	93.90	85.90	65.78	78.19	51.59	37.99
	Claude-Haiku-4.5 [3]	86.91	85.01	70.33	70.82	53.46	49.09
Open-Source	Qwen2.5-VL-72B [4]	93.92	82.23	73.13	67.72	53.98	47.46
	Qwen3-VL-30B-A3B [55]	89.39	79.45	73.00	65.24	51.59	44.47
	InternVL3-32B [50]	94.35	83.25	74.17	71.01	54.82	43.22
	LLaVA-v1.5-13B [36]	56.93	55.55	49.99	42.13	32.73	37.58
	qwen3-VL-8B [55]	89.53	70.10	79.66	73.74	50.60	41.08
	Qwen2.5-VL-7B [4]	85.55	66.12	75.68	69.76	46.62	37.10
	LLaMA3-LLaVA-Next-8B [31]	69.99	61.24	79.44	47.60	44.86	38.27
	InternVL3-8B [50]	87.57	73.97	69.42	65.53	53.91	37.22
	ShareGPT4V-7B [7]	72.43	69.24	58.99	50.66	38.03	38.31
	LLaVA-v1.5-7B [36]	59.39	62.24	55.59	44.66	28.82	32.73
Medical	MedGemma-27B-IT [49]	92.72	74.60	74.44	66.19	48.34	46.67
	HuaTuoGPT-Vision-34B [6]	75.94	66.01	55.56	58.48	47.59	40.64
	Lingshu-7B [56]	77.01	63.33	70.19	64.49	47.49	31.39
	LLaVA-Med-7B [30]	49.36	36.03	53.84	44.74	20.00	26.28
	HuaTuoGPT-Vision-7B [6]	60.22	54.69	46.63	57.48	42.88	35.02

Table 4. Evaluation of MLLMs across 6 stages on CFP modality

	MLLM	IQA	AL	LC	DD	SG	CD
Commercial	GPT-5 [41]	99.80	98.65	90.78	81.91	69.07	61.30
	Gemini-2.5-Pro [12]	99.96	97.53	82.56	84.11	67.39	43.34
	GLM-4.5v [59]	99.91	97.19	77.66	84.12	60.01	55.38
	GPT-5-nano [42]	99.97	98.13	82.06	81.22	67.95	61.68
	Gemini-2.5-Flash [12]	99.82	100.00	78.21	92.97	61.34	45.17
	Claude-Haiku-4.5 [3]	94.45	92.39	76.43	76.97	58.10	53.35
Open-Source	Qwen2.5-VL-72B [4]	99.78	93.13	82.82	76.69	61.13	53.75
	Qwen3-VL-30B-A3B [55]	99.88	93.09	85.53	76.44	60.44	52.10
	InternVL3-32B [50]	94.35	83.25	74.17	71.01	54.82	43.22
	LLaVA-v1.5-13B [36]	60.02	58.57	52.71	44.42	34.51	39.62
	qwen3-VL-8B [55]	99.70	86.20	97.88	90.65	61.54	50.81
	Qwen2.5-VL-7B [4]	99.86	80.82	92.50	85.27	56.98	45.35
	LLaMA3-LLaVA-Next-8B [31]	61.06	53.43	69.30	41.53	39.14	33.39
	InternVL3-8B [50]	87.57	73.97	69.42	65.53	53.91	37.22
	ShareGPT4V-7B [7]	65.34	62.47	53.22	45.71	34.31	34.56
	LLaVA-v1.5-7B [36]	51.79	54.27	48.47	38.94	25.13	28.54
Medical	MedGemma-27B-IT [49]	99.89	89.37	89.17	79.29	57.91	55.91
	HuaTuoGPT-Vision-34B [6]	75.94	66.01	55.56	58.48	47.59	40.64
	Lingshu-7B [56]	99.87	82.93	91.92	84.45	62.19	41.10
	LLaVA-Med-7B [30]	78.61	57.39	85.75	71.26	31.80	41.85
	HuaTuoGPT-Vision-7B [6]	74.86	69.33	61.27	72.12	57.52	49.66

This formulation incentivizes appropriate calibration: models should express high confidence when correct (especially on R-level items) and low confidence when uncertain (particularly on ambiguous S-level items).

We assess calibration using three complementary measures in a unified formulation. The *Expected Calibration Error (ECE)* aggregates bin-wise confidence accuracy gaps: letting $\{B_m\}_{m=1}^{M}$ be confidence bins (*e.g.*, $[0, 0.25), \ldots, [0.75, 1.0]$) with $\mathrm{acc}(B_m)$ and $\mathrm{conf}(B_m)$ denoting average accuracy and confidence, respectively,

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|. \qquad (7)$$

For each response, we require the model to report its confidence ("Please rate your confidence from 0 to 100%"). We parse this as a percentage and normalize it to the $[0, 1]$ range, clipping out-of-range values to that interval. The resulting confidences are then used in the uncertainty-aware scoring and calibration metrics described in Eq. (7).

5. Detailed Experimental Results

In the main manuscript, we have presented the evaluation results across six modalities combined. To further supplement these findings, we provide individual modality results across six distinct stages in subsequent subsections.

5.1. Results on EP Modality

Tab. 3 exhibits a step-wise degradation on the EP modality: most models achieve high accuracy on Image Quality Assessment (IQA) and basic Anatomical Localization (AL), indicating that "seeing clearly and localizing the region of interest" is largely solved using external photography; from from Lesion Characterization (LC) to Disease Diagnosis (DD), performance drops substantially and gaps between commercial, open-source, and medical-specialized models become more pronounced, showing that inferring a concrete diagnosis from EP alone remains unstable. At the Severity Grading (SG) and Clinical Decision (CD) stages, accuracies fall further into the mid-to-low range, with MedGemma-27B-IT [49] and HuaTuoGPT [6] offering only modest gains. Overall, current MLLMs can provide reliable basic readability and structural localization using external photography, but they still struggle with higher-level clinical reasoning that requires integrating severity, disease course, and management strategies, functioning more as visual aids than autonomous decision makers.

5.2. Results on CFP Modality

Tab. 4 reports six-stage performance on the CFP modality and reveals a clear high-to-low pattern. Most models are nearly saturated on image quality assessment and basic anatomical localization, indicating that tasks of "seeing clearly and finding the right structure" are no longer challenging for CFP. However, accuracies drop markedly from lesion characterization to disease diagnosis, and the gaps between models become more pronounced, suggesting that the transition from merely spotting lesions to assigning a specific diagnosis forms the first major bottleneck. In the severity grading and clinical decision stages, performance further declines to roughly the 40–60% range, with medical-specialized models only slightly outperforming general MLLMs. This pattern shows that while current MLLMs can perform reliable visual recognition on CFP, they still struggle to produce guideline-consistent grading and downstream management decisions when relying on this single modality.

Table 5. Evaluation of MLLMs across 6 stages on RetCam modality.

	MLLM	IQA	AL	LC	DD	SG	CD
Commercial	GPT-5 [41]	97.65	86.94	80.00	72.19	60.87	54.02
	Gemini-2.5-Pro [12]	89.30	79.36	67.18	68.44	54.83	35.27
	GLM-4.5v [59]	95.44	86.14	68.84	74.56	53.19	49.09
	GPT-5-nano [42]	87.83	74.86	62.61	61.96	51.84	47.06
	Gemini-2.5-Flash [12]	72.51	66.34	50.79	60.38	39.84	29.33
	Claude-Haiku-4.5 [3]	57.41	56.16	46.46	46.78	35.32	32.43
Open-Source	Qwen2.5-VL-72B [4]	92.95	81.38	72.37	67.02	53.42	46.97
	Qwen3-VL-30B-A3B [55]	96.09	85.41	78.48	70.13	55.46	47.80
	InternVL3-32B [50]	94.35	83.25	74.17	71.01	54.82	43.22
	LLaVA-v1.5-13B [36]	80.93	78.98	71.07	59.90	46.53	53.43
	qwen3-VL-8B [55]	86.17	65.60	75.72	69.45	46.10	36.02
	Qwen2.5-VL-7B [4]	90.60	70.03	80.15	73.88	49.37	39.29
	LLaMA3-LLaVA-Next-8B [31]	72.09	63.08	81.83	49.03	46.21	39.42
	InternVL3-8B [50]	87.57	73.97	69.42	65.53	53.91	37.22
	ShareGPT4V-7B [7]	78.26	74.82	63.74	54.75	41.10	41.39
	LLaVA-v1.5-7B [36]	80.24	84.09	75.10	60.34	38.94	44.22
Medical	MedGemma-27B-IT [49]	88.92	71.55	71.39	63.48	46.36	44.76
	HuaTuoGPT-Vision-34B [6]	75.94	66.01	55.56	58.48	47.59	40.64
	Lingshu-7B [56]	85.53	70.34	77.96	71.62	52.74	34.86
	LLaVA-Med-7B [30]	45.06	32.90	49.15	40.85	20.00	23.99
	HuaTuoGPT-Vision-7B [6]	70.50	64.97	56.91	67.76	53.16	45.30

5.3. Results on RetCam Modality

On the RetCam modality, Tab. 5 shows the same six-stage "high-to-low" gradient: most models remain strong on IQA and AL (commercial systems often above 90–95%), indicating that assessing image readability and coarse structural localization on wide-field pediatric fundus images is largely solved; however, performance consistently drops from LC to DD, SG, and CD. Even the best models (GPT-5 [41], Qwen2.5-VL-72B [4], MedGemma-27B-IT [49]) only reach about 60–70% on DD, with SG/CD typically falling into the 40–60% band, and while some open-source or medical models achieve relatively high LC scores, their advantages shrink at staging and decision levels. Overall, RetCam is clearly more challenging than CFP/EP: current MLLMs can detect and localize lesions reasonably well, but still struggle to deliver stable, guideline-aligned diagnosis, grading, and management decisions, behaving more like pattern spotters than reliable agents for structured ROP-style clinical reasoning.

5.4. Results on FFA Modality

In Tab. 6, the FFA modality exhibits a similar "high-to-low" six-stage staircase: leading commercial models achieve very high scores on IQA and AL (mostly above 80–95%), suggesting that overall image quality assessment and coarse anatomical localization on FFA are largely solved; performance then drops into the 60–70% band for LC and DD, especially for open-source and medical-specialized models, indicating that stably recognizing leakage patterns and non-perfusion across phases and mapping them to correct diagnoses remains challenging; by the SG and CD stages, most models fall to roughly 40–60%, with medical models only modestly better, highlighting that current MLLMs can "see where is bright or dark" but still struggle to translate multi-phase perfusion patterns and lesion extent into guideline-consistent staging and management, making the full chain from lesion detection to hemodynamic understanding and decision-making particularly weak on FFA.

Table 6. Evaluation of MLLMs across 6 stages on FFA modality.

	MLLM	IQA	AL	LC	DD	SG	CD
Commercial	GPT-5 [41]	96.71	82.54	75.95	68.54	57.79	51.29
	Gemini-2.5-Pro [12]	97.08	86.27	73.03	74.39	59.60	38.34
	GLM-4.5v [59]	86.74	78.29	62.56	67.77	48.34	44.62
	GPT-5-nano [42]	82.68	70.48	58.94	58.33	48.80	44.30
	Gemini-2.5-Flash [12]	95.37	87.25	66.81	79.42	52.40	38.58
	Claude-Haiku-4.5 [3]	75.02	73.39	60.71	61.14	46.15	42.38
Open-Source	Qwen2.5-VL-72B [4]	91.36	79.99	71.13	65.88	52.50	46.17
	Qwen3-VL-30B-A3B [55]	82.46	73.29	67.34	60.18	47.59	41.02
	InternVL3-32B [50]	94.35	83.25	74.17	71.01	54.82	43.22
	LLaVA-v1.5-13B [36]	47.59	46.44	41.79	35.22	27.36	31.41
	qwen3-VL-8B [55]	84.39	65.73	74.91	69.23	46.99	37.84
	Qwen2.5-VL-7B [4]	82.20	63.54	72.72	67.04	44.80	35.65
	LLaMA3-LLaVA-Next-8B [31]	65.41	57.24	74.25	44.49	41.93	35.77
	InternVL3-8B [50]	87.57	73.97	69.42	65.53	53.91	37.22
	ShareGPT4V-7B [7]	58.74	56.16	47.84	41.09	30.84	31.07
	LLaVA-v1.5-7B [36]	51.85	54.34	48.53	38.99	25.16	28.58
Medical	MedGemma-27B-IT [49]	79.34	63.84	63.70	56.64	41.37	39.94
	HuaTuoGPT-Vision-34B [6]	75.94	66.01	55.56	58.48	47.59	40.64
	Lingshu-7B [56]	75.27	61.90	68.60	63.03	46.42	30.68
	LLaVA-Med-7B [30]	40.78	29.77	44.48	36.97	20.00	21.71
	HuaTuoGPT-Vision-7B [6]	51.51	45.98	37.92	48.77	34.17	26.31

Table 7. Evaluation of MLLMs across 6 stages on ICGA modality.

	MLLM	IQA	AL	LC	DD	SG	CD
Commercial	GPT-5 [41]	98.71	82.54	75.95	68.54	57.79	51.29
	Gemini-2.5-Pro [12]	97.08	86.27	73.03	74.39	59.60	38.34
	GLM-4.5v [59]	86.74	78.29	62.56	67.77	48.34	44.62
	GPT-5-nano [42]	82.68	70.48	58.94	58.33	48.80	44.30
	Gemini-2.5-Flash [12]	95.37	87.25	66.81	79.42	52.40	38.58
	Claude-Haiku-4.5 [3]	75.02	73.39	60.71	61.14	46.15	42.38
Open-Source	Qwen2.5-VL-72B [4]	89.99	78.79	70.07	64.89	51.72	45.47
	Qwen3-VL-30B-A3B [55]	79.85	70.97	65.21	58.28	46.08	39.72
	InternVL3-32B [50]	94.35	83.25	74.17	71.01	54.82	43.22
	LLaVA-v1.5-13B [36]	48.93	47.75	42.96	36.21	28.13	32.30
	qwen3-VL-8B [55]	83.44	65.09	74.12	68.53	46.67	37.68
	Qwen2.5-VL-7B [4]	80.79	62.44	71.47	65.88	44.02	35.03
	LLaMA3-LLaVA-Next-8B [31]	62.45	54.64	70.88	42.47	40.03	34.15
	InternVL3-8B [50]	87.57	73.97	69.42	65.53	53.91	37.22
	ShareGPT4V-7B [7]	58.78	56.19	47.88	41.12	30.87	31.09
	LLaVA-v1.5-7B [36]	51.70	54.18	48.39	38.88	25.09	28.49
Medical	MedGemma-27B-IT [49]	80.21	64.54	64.40	57.26	41.82	40.38
	HuaTuoGPT-Vision-34B [6]	75.94	66.01	55.56	58.48	47.59	40.64
	Lingshu-7B [56]	74.68	61.42	68.07	62.54	46.05	30.44
	LLaVA-Med-7B [30]	40.75	29.75	44.45	36.94	20.00	21.70
	HuaTuoGPT-Vision-7B [6]	51.52	45.99	37.93	48.78	34.18	26.32

5.5. Results on ICGA Modality

Tab. 7 shows a similar six-stage "degrading" pattern on ICGA: most models maintain high performance on IQA and AL (commercial models often in the 80–95% range), indicating that assessing angiogram quality and roughly localizing abnormal regions is largely solved; however, accuracy drops steadily from LC to DD and further to SG and CD, with staging/decision performance frequently falling into the 40–60% band or below, and open-source or medical-specialized models offering only modest gains over strong

Table 8. Evaluation of MLLMs across 6 stages on OCT modality.

	MLLM	IQA	AL	LC	DD	SG	CD
Commercial	GPT-5 [41]	97.81	63.30	58.25	52.56	40.32	30.33
	Gemini-2.5-Pro [12]	89.52	61.78	52.30	53.28	32.68	27.45
	GLM-4.5v [59]	51.34	46.34	37.03	40.11	28.61	26.41
	GPT-5-nano [42]	67.67	57.68	48.24	47.74	39.94	36.26
	Gemini-2.5-Flash [12]	65.36	59.79	45.78	54.43	35.91	26.44
	Claude-Haiku-4.5 [3]	59.53	58.23	48.18	48.51	36.62	33.63
Open-Source	Qwen2.5-VL-72B [4]	60.55	53.01	47.14	43.66	34.80	30.60
	Qwen3-VL-30B-A3B [55]	61.00	54.22	49.82	44.52	35.20	30.34
	InternVL3-32B [50]	84.35	53.25	54.17	41.01	38.82	23.22
	LLaVA-v1.5-13B [36]	33.75	32.93	29.63	24.98	20.00	22.28
	qwen3-VL-8B [55]	56.10	43.36	49.63	45.75	30.57	24.33
	Qwen2.5-VL-7B [4]	56.10	43.36	49.63	45.75	30.57	24.33
	LLaMA3-LLaVA-Next-8B [31]	46.40	40.60	52.66	31.55	29.74	25.37
	InternVL3-8B [50]	87.57	73.97	69.42	65.53	53.91	37.22
	ShareGPT4V-7B [7]	40.87	39.08	33.29	28.59	21.46	21.62
	LLaVA-v1.5-7B [36]	39.60	41.50	37.07	29.78	20.00	21.83
Medical	MedGemma-27B-IT [49]	60.59	48.75	48.65	43.26	31.59	30.50
	HuaTuoGPT-Vision-34B [6]	75.94	66.01	55.56	58.48	47.59	40.64
	Lingshu-7B [56]	51.18	42.09	46.65	42.86	31.56	20.86
	LLaVA-Med-7B [30]	28.25	20.62	30.81	25.61	20.00	20.00
	HuaTuoGPT-Vision-7B [6]	48.07	42.54	34.48	45.33	30.73	22.87

commercial baselines. Given that ICGA is crucial for differentiating entities such as PCV and AMD subtypes, this "can see and localize, but cannot reliably subtype or manage" behavior suggests that current MLLMs can exploit ICGA for lesion recognition and coarse diagnosis, but still struggle to integrate temporal phase patterns and subtle choroidal vascular signatures into guideline-consistent staging and treatment planning.

5.6. Results on OCT Modality

Tab. 8 indicates that OCT is arguably the toughest modality for all models: while top commercial systems (*e.g.*, GPT-5 [41]) still achieve near-ceiling performance on IQA, accuracy drops sharply from AL onward, with LC and DD typically hovering around the 40–50% range and SG/CD further collapsing into the low 30–40% band; most open-source and even medical-specialized models perform at or below this level. Unlike the smoother "high-to-low" pattern seen on CFP/EP, OCT exhibits a consistently low profile across all six stages, suggesting that current MLLMs remain weak at modeling layered retinal microstructures, subtle reflectivity changes, and quasi-3D context: they can roughly "see" the B-scan, but struggle to complete the full chain from structural parsing and lesion characterization to diagnosis, staging, and management, making OCT-centric diseases such as DME or nAMD far from being reliably handled in an end-to-end fashion by present models.

5.7. Analysis of Single-Modality Findings

A comprehensive analysis of model performance across six ophthalmic imaging modalities (EP, CFP, RetCam, OCT, FFA, ICGA) reveals a consistent and pronounced performance degradation as task complexity increases from basic image interpretation to advanced clinical reasoning. The following key findings emerge:

Consistent Performance Decline from IQA to DD Spanning Six Modalities. All modalities exhibit a "high-to-low" performance gradient across the six-stage clinical reasoning chain. Models achieve near-saturation performance on IQA and AL, indicating that foundational visual recognition capabilities are largely solved. However, a significant and consistent accuracy drop occurs from LC to DD, which represents the first major bottleneck. Performance further declines at SG and CD, with accuracies often falling into the 40–60% range, demonstrating a fundamental limitation in higher-order clinical reasoning.

Limitations of Monomodal Diagnosis. The steep decline in performance underscores the inherent limitations of single-modality analysis. While models function effectively as "pattern spotters" for basic tasks, they struggle to integrate the contextual, temporal, and structural subtleties required for guideline consistent diagnosis and management. This gap is most acute for complex modalities like FFA, ICGA, and OCT, where interpreting dynamic sequences (FFA, ICGA) or 3D structural relationships (OCT) remains a substantial challenge. Consequently, current MLLMs act more as visual aids than autonomous diagnostic agents when limited to a single data source.

Modality-Specific Difficulty Hierarchy. A clear hierarchy of modality difficulty is observed. EP and CFP are relatively easier, with models maintaining reasonable performance through LC. RetCam introduces wider-field complexity, while FFA and ICGA present challenges in interpreting temporal and hemodynamic patterns. OCT is unequivocally the most challenging modality, with performance collapsing early at AL and remaining low across all subsequent stages, highlighting a significant weakness in modeling layered retinal microstructures and quasi-3D context.

Marginal Gains from Medical Specialization. The performance advantage of medical-specialized models over strong generalist MLLMs is often modest and diminishes at higher reasoning stages (SG, CD). While these models sometimes show improvements in intermediate tasks like LC, they fail to demonstrate a decisive edge in complex diagnostic and decision-making tasks, indicating that current medical fine-tuning approaches are insufficient for instilling deep clinical reasoning capabilities.

6. Case Studies

6.1. Case Study of Six-Stage Clinical Reasoning Chain

Fig. 3 presents a case study of our six-stage clinical reasoning chain on an eye photograph with severe central lens opacity. The chain mirrors real clinical workflow: the model first evaluates image quality (Stage 1, IQA), local-

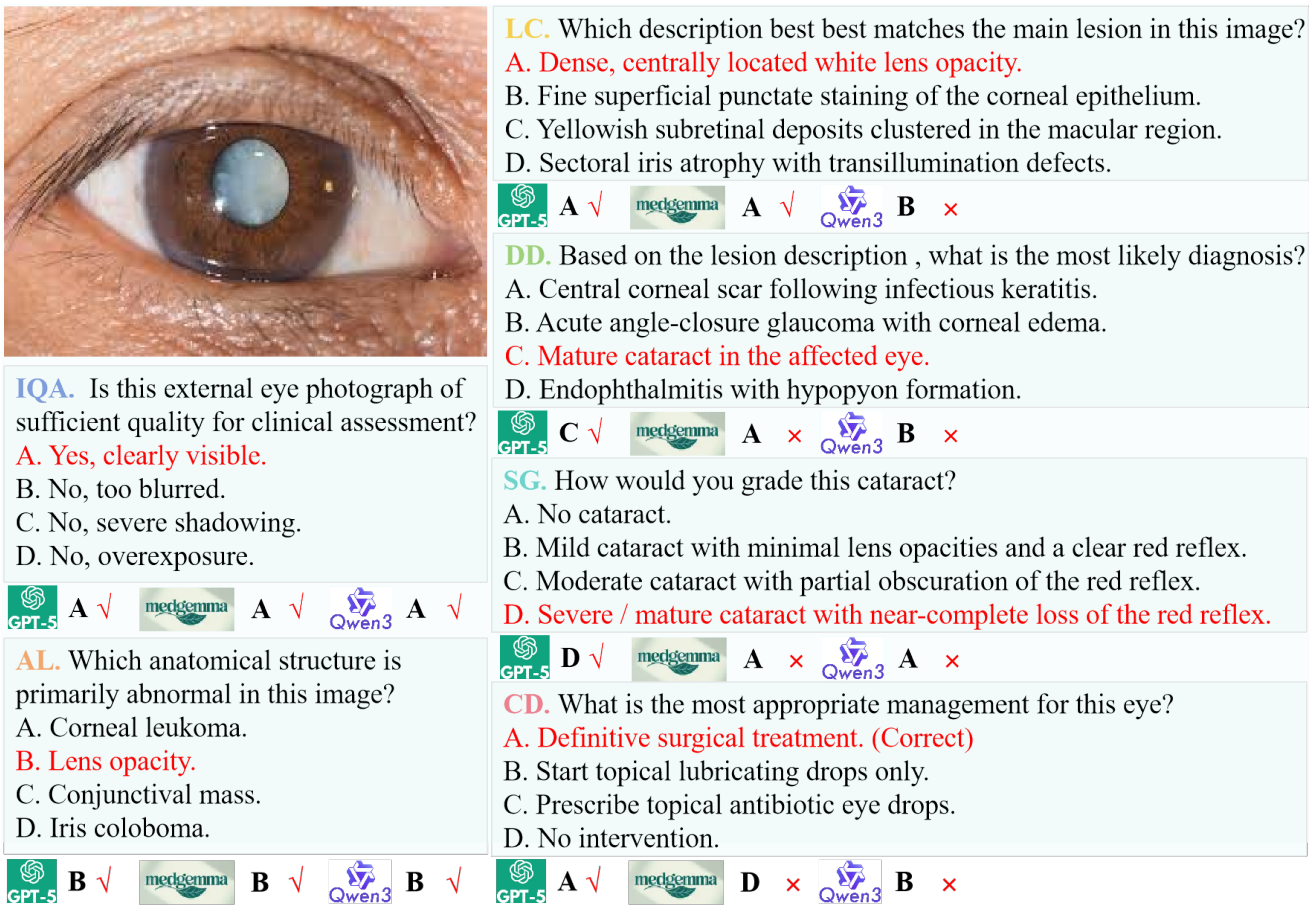

Figure 3. Case study of six-stage clinical reasoning chain. X-PCR serves as a unified benchmark for bridging task-specific performance and comprehensive diagnostic reasoning in MLLMs, via aligned multi-modal data across six stages.

izes the abnormality to the lens (Stage 2, AL), describes it as a dense central white opacity (Stage 3, LC), integrates these findings into a mature cataract diagnosis (Stage 4, DD), grades it as severe with near-complete loss of red reflex (Stage 5, SG), and finally recommends definitive surgical treatment (Stage 6, CD). GPT-5 completes all six stages, while Qwen3 performs well only on the earlier perception and localization stages (IQA, AL) but fails on higher-level diagnostic and management stages (LC–CD). MedGemma and Qwen3 both break down once diagnostic synthesis and treatment planning are required. This pattern, consistent performance across models in early stages and divergence in later stages, shows that our six-stage design faithfully captures clinical reasoning and exposes where different models fail, rather than obscuring these weaknesses with a single end-point accuracy metric.

6.2. Case Study of Cross-Modality Clinical Reasoning

Fig. 4 presents a case study of cross-modality clinical reasoning in a patient with polypoidal choroidal vasculopathy (PCV) in the right eye. The color fundus photograph (CFP) of the affected eye shows an orange-red subfoveal lesion with surrounding hard exudates and macular edema, while the fellow eye appears relatively normal. On fluorescein angiography (FFA), the corresponding lesion manifests as a temporal focal hyperfluorescent spot with late diffuse leakage and serous detachment, and OCT reveals subretinal hyper-reflective material with overlying intraretinal and subretinal fluid.

We cast this case into three cross-modal reasoning tasks. Correspondence Identification (Q1) asks the model to match the macular lesion on CFP with its FFA counterpart, testing precise spatial and semantic alignment. Diagnostic Integration (Q2) requires jointly interpreting CFP, FFA, and clinical context to favor PCV over mimics such as center-involving DME or neovascular AMD. Modality Selection (Q3) evaluates value-of-information reasoning by asking which further test is most useful; the correct choice is ICGA, the gold standard for confirming polypoidal lesions and branching vascular networks.

Despite their strong overall performance, all three mod-

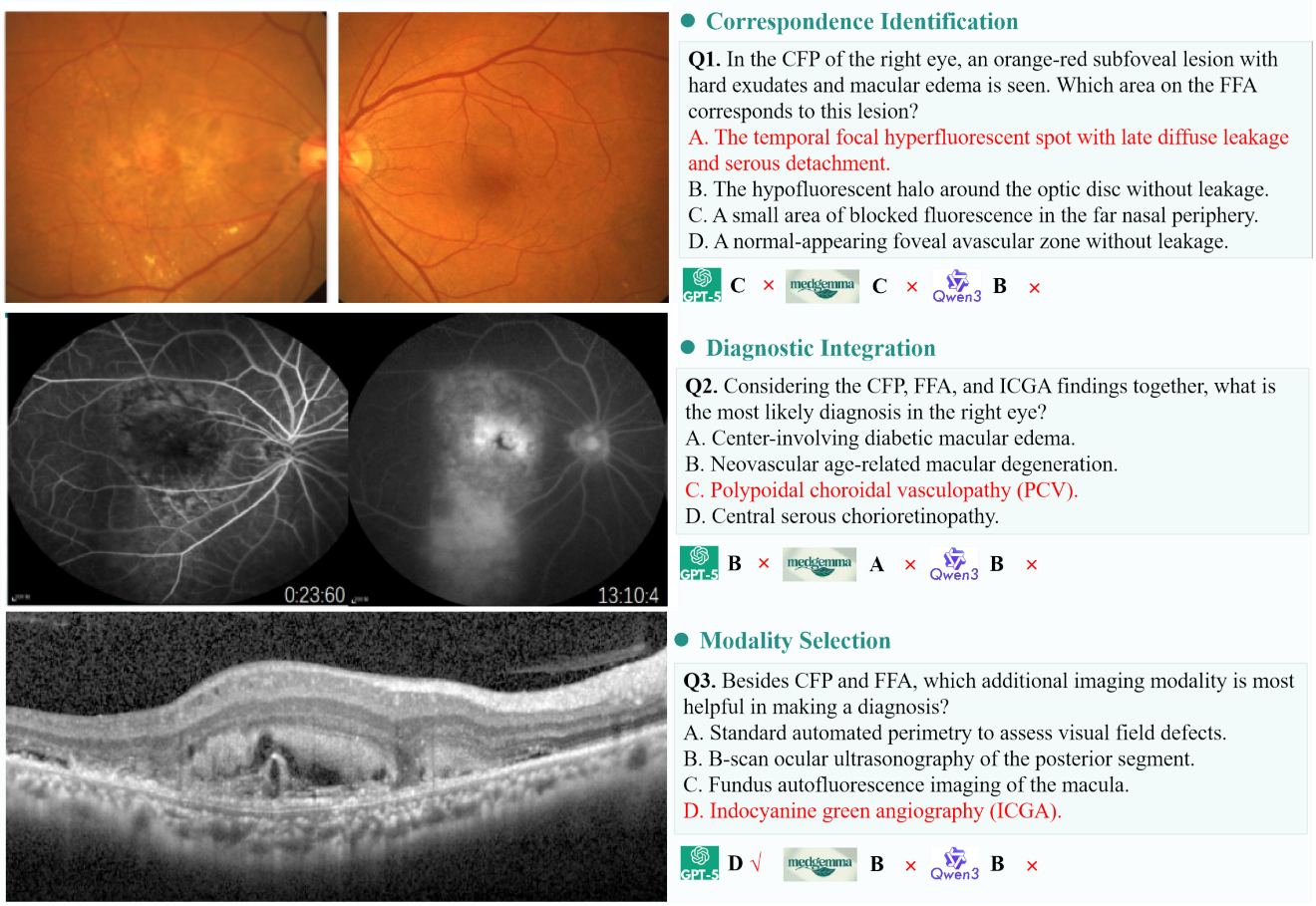

Figure 4. Case study of cross-modality clinical reasoning. We cast this case into three cross-modal reasoning tasks: matching CFP-FFA lesions through Correspondence Identification (Q1), integrating multi-modal data to differentiate PCV from mimics through Diagnostic Integration (Q2), and selecting ICGA as the definitive next test through Modality Selection (Q3).

els fail on Q1 and Q2, misaligning CFP and FFA findings and converging on incorrect diagnoses, while only GPT-5 correctly selects ICGA in Q3. This case study shows that our cross-modality tasks expose failures that would be invisible in single-modality or single-step diagnosis settings: models can describe individual images reasonably well, yet break down when required to align evidence across modalities and use it to drive diagnosis and imaging decisions. Consequently, this design demonstrates the necessity and utility of structured cross-modal reasoning benchmarks like X-PCR for realistically assessing and improving ophthalmic MLLMs.

References

[1] S. O. Afolabi, D. K. Hwang, F. A. Medeiros, et al. Equity-enhanced glaucoma progression prediction from multimodal clinical data using a fairness-aware deep learning framework. *npj Digital Medicine*, 8, 2025. Introduces and uses the Harvard-GDP glaucoma progression dataset. 2

[2] Ankita Agrawal, Priyanka Kokil, Rahul Phalak, and et al. Hvdropdb: A multi-structure segmentation dataset for retinopathy of prematurity. *Scientific Data*, 10:595, 2023. 2

[3] Anthropic. Introducing claude haiku 4.5, 2025. Model announcement and overview. 11, 12, 13

[4] 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. *arXiv preprint arXiv:2502.13923*, 2025. 11, 12, 13

[5] Joint Shantou International Eye Center. 1000 fundus images with 39 categories (jsiec retinal-39). https://www.kaggle.com/datasets/jr2ngb/retinal-disease-classification, 2021. Accessed 2025-11-15. 2

[6] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, et al. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal

llms at scale. *arXiv preprint arXiv:2406.19280*, 2024. 11, 12, 13

[7] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision (ECCV)*, 2024. 11, 12, 13

[8] Xin Chen, Jing Wang, Qiang Li, and et al. Maples-dr: A multi-attribute pathology lesion segmentation dataset for diabetic retinopathy. *arXiv preprint*, arXiv:2403.XXXXX, 2024. 2

[9] Kaggle Community. Eyepacs/eyephotos fundus image collection. https://www.kaggle.com/datasets/jeanpat/eyepacs-eye-photos, 2019. Accessed 2025-11-15. 2

[10] C. De Vente, M. Wang, L. Räber, J. Schouten, B. van Ginneken, et al. Airogs: Artificial intelligence for robust glaucoma screening challenge. *IEEE Transactions on Medical Imaging*, 43(1):542–554, 2024. 2

[11] Etienne Decencière, Xiwei Zhang, Guy Cazuguel, Bruno Lay, Béatrice Cochener, Caroline Trone, Philippe Gain, Richard Ordonez, Pascale Massin, Ali Erginay, Béatrice Charton, and Jean-Claude Klein. Feedback on a publicly distributed image database: The messidor database. *Image Analysis & Stereology*, 33(3):231–234, 2014. 2

[12] Google DeepMind. Gemini 1.5: Scaling up multimodal reasoning, 2024. https://deepmind.google/technologies/gemini/. 11, 12, 13

[13] Lijie Deng, Junyan Lyu, Haixiang Huang, and et al. The sustech-sysu dataset for automatically segmenting and classifying corneal ulcers. *Scientific Data*, 7:23, 2020. 2

[14] Pablo Díaz, Laura Marcos, Antonio Anton, and et al. Papila: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. *Scientific Data*, 6:190083, 2019. 2

[15] A. Diaz-Pinto, S. Morales, V. Naranjo, T. Koehler, J. M. Mossi, and A. Navea. Cnns for automatic glaucoma assessment using fundus images: an extensive validation. *BioMedical Engineering OnLine*, 18(1):29, 2019. 2

[16] Zekun Dong, Xiang Luo, Tobias Elze, and et al. A large-scale functional dataset for glaucoma visual field prediction. *Scientific Data*, 10:597, 2023. 2

[17] Zekun Dong, Xiang Luo, Mina Huh, and et al. Harvard glaucoma detection and progression dataset. *Scientific Data*, 10: 600, 2023. 2

[18] Zekun Dong, Xiang Luo, Mina Huh, and et al. A dataset for glaucoma detection and progression with visual fields and fundus images. *Scientific Data*, 10:601, 2023. 2

[19] Zekun Dong, Xiang Luo, Mina Huh, and et al. Harvard-fairseg: A fair and diverse medical image segmentation dataset. *arXiv preprint*, arXiv:2404.XXXXX, 2024. 2

[20] Huazhu Fu, Fang Li, Qin Hu, and et al. Palm: Pathologic myopia challenge and dataset. In *MICCAI 2019 OMIA Workshop*, 2019. 2

[21] Francisco Fumero, José Sigut, Severiano Alayón, and et al. RIM-ONE DL: A glaucoma fundus image database for optic nerve head segmentation. *Scientific Data*, 7:1–10, 2020. 2

[22] Asim Ghosh, Prithwijit Das, Avik Choudhury, and et al. Toxofundus: A retinal fundus image dataset for the diagnosis of ocular toxoplasmosis. *Scientific Reports*, 11:14059, 2021. 2

[23] Yu Hu, Ying Gao, Wei Gao, and et al. Amd-sd: An optical coherence tomography image dataset for wet amd lesions segmentation. *Scientific Data*, 11:1014, 2024. 2

[24] Mohammad Nurul Islam and colleagues. Eye disease image dataset. https://www.kaggle.com/datasets/andrewmvd/eye-disease-classification, 2020. Accessed 2025-11-15. 2

[25] S. Islam, M. Hasan, M. Rahman, et al. Novel deep learning model for glaucoma detection using fusion of fundus and optical coherence tomography images. *Sensors*, 25(14):4337, 2025. Uses a private clinical dataset from Bangladesh Eye Hospital. 2

[26] Qin Jin, Zaiwang Chen, Yong Xu, and et al. Fives: A fundus image dataset for vessel segmentation. *Scientific Data*, 9:94, 2022. 2

[27] Kaggle and EyePACS. Diabetic retinopathy detection. https://www.kaggle.com/c/diabetic-retinopathy-detection, 2015. Kaggle competition dataset from EyePACS. 2

[28] Daniel S. Kermany, Michael Goldbaum, Wenjia Cai, and et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018. 2

[29] H. Lemij, C. de Vente, et al. Justified referral in ai glaucoma screening (justraigs) challenge. Zenodo, 2024. 2

[30] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023. 11, 12, 13

[31] Fuchen Li, Yuanhan Zhang, Sheng Shen, Yong Jae Lee, et al. Llava-next-interleave: Tackling multi-image, video, and multi-view in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 11, 12, 13

[32] Jianyong Li, Huazhu Fu, Xun Sun, and et al. Gamma: A large-scale multi-modal dataset for glaucoma assessment. *Scientific Data*, 9:17, 2022. 2

[33] Liu Li, Mai Xu, Xiaofei Wang, Lai Jiang, and Hanruo Liu. Attention based glaucoma detection: A large-scale database and cnn model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10571–10580, 2019. 2

[34] Ning Li, Ting Li, Huazhu Chen, and et al. A benchmark of ocular disease intelligent recognition: One shot for multi-disease detection. *arXiv preprint*, arXiv:2102.07978, 2021. 2

[35] Tao Li, Yingqi Gao, Kai Wang, Song Guo, Hanruo Liu, and Hong Kang. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, 501:511–522, 2019. OIA-DDR dataset. 2

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Llava-1.5: Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03761*, 2023. 11, 12, 13

[37] Risheng Liu, Chen Ma, Yutong Wang, et al. Deepdrid: Diabetic retinopathy–grading and image quality estimation challenge. *Patterns*, 3(7):100512, 2022. 2

[38] Yuxuan Lu, Han Zhang, Yong Xu, and et al. Drac 2022: Diabetic retinopathy analysis challenge on ultra-widefield octa. In *MICCAI 2022 Ophthalmic Medical Image Analysis Workshop*, 2022. 2

[39] Nasser Mohammed, Arash Tashk, and et al. OCTID: Optical coherence tomography image dataset. *Mendeley Data*, 3, 2016. 2

[40] Vahid Norouzi, Xiaoyu Guo, Manoj P. Nallabothula, and et al. Farfum: A multi-center fundus image dataset for retinopathy of prematurity screening. *Scientific Data*, 11: XXX, 2024. FARFUM RoP dataset. 2

[41] OpenAI. Gpt-5: Multimodal large language models for understanding complex visual inputs. *OpenAI Technical Report*, 2025. 11, 12, 13

[42] OpenAI. Gpt-5 nano, 2025. Model documentation. 11, 12, 13

[43] José Ignacio Orlando, Huazhu Fu, Joana Barbosa Breda, and et al. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical Image Analysis*, 59:101570, 2020. 2

[44] Saurabh Pachade, Deepayan De, Manesh Kokare, and et al. Rfmid: A retinal fundus multi-disease image dataset. *Scientific Data*, 8:249, 2021. 2

[45] Prasoon Porwal, Saurabh Pachade, Rahul Kamble, and et al. IDRiD: Diabetic retinopathy—segmentation and grading challenge. *Data*, 3(3):25, 2018. 2

[46] Zhen Qin, Lu Fang, Zhihong Ding, and et al. OCTA-500: A retinal oct angiography dataset for age-related macular degeneration and diabetic retinopathy. *IEEE Transactions on Medical Imaging*, 39(12):4038–4050, 2020. 2

[47] Diego Rodríguez, Ana Suárez, Jorge Novo, and et al. Mured: A multi-label retinal fundus image dataset for multi-disease classification. *Data in Brief*, 43:108357, 2022. 2

[48] Mohammed H. Sarhan, Jun Wu, Jian Zhang, and et al. Olives: Ophthalmic labels for investigating visual semantic segmentation. *Scientific Data*, 9:274, 2022. 2

[49] 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. 11, 12, 13

[50] Shanghai AI Laboratory. Internvl-14b: Vision-language model by shanghai ai lab. https://huggingface.co/OpenGVLab/InternVL-14B-224px, 2024. Accessed: 2025-05-13. 11, 12, 13

[51] Asia Pacific Tele-Ophthalmology Society. Aptos 2019 blindness detection. https://www.kaggle.com/c/aptos2019-blindness-detection, 2019. Accessed 2025-11-15. 2

[52] Maria Stegiou, José Ignacio Orlando, Huazhu Fu, and et al. Refuge2 challenge: Multi-domain glaucoma assessment from fundus photographs. In *MICCAI 2022 OMIA Workshop*, 2022. 2

[53] Jiri Timkovic, Radim Kolar, Martin Sramek, and et al. RetinalROP: A retinal fundus image dataset for retinopathy of prematurity screening. *Scientific Data*, 11:XXX, 2024. 2

[54] Unknown. Retinal disease classification dataset (4 classes). https://www.kaggle.com/datasets/andrewmvd/retinal-disease-classification, 2019. Accessed 2025-11-15. 2

[55] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. 11, 12, 13

[56] 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. 11, 12, 13

[57] Jun Ma Yang, Rui Shi, Ziyuan Wang, and et al. Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10:41, 2023. 2

[58] Xiang Ye, Shuai Zhang, Hao Li, and et al. OIMHS: An oct image and mask dataset for retinal macular hole and cyst segmentation. *Scientific Data*, 10:XXX, 2023. 2

[59] Z.ai. Glm-4.5v, 2025. Developer documentation. 11, 12, 13

[60] Jiacheng Zhang, Rui Li, Cheng Liu, and Xiang Ji. Improving domain transfer with consistency-regularized joint distribution alignment for medical image classification. *Symmetry*, 17(4):515, 2025. Introduces the BiDR (Binary Diabetic Retinopathy) dataset. 2

[61] Yong Zhao, Fan Yang, Xin Li, and et al. OCT5K: A large-scale optical coherence tomography dataset with multi-disease annotations. *Scientific Data*, 9:XXX, 2022. 2