EyePCR: A Comprehensive Benchmark for Fine-Grained Perception, Knowledge Comprehension and Clinical Reasoning in Ophthalmic Surgery

Anonymous CVPR submission

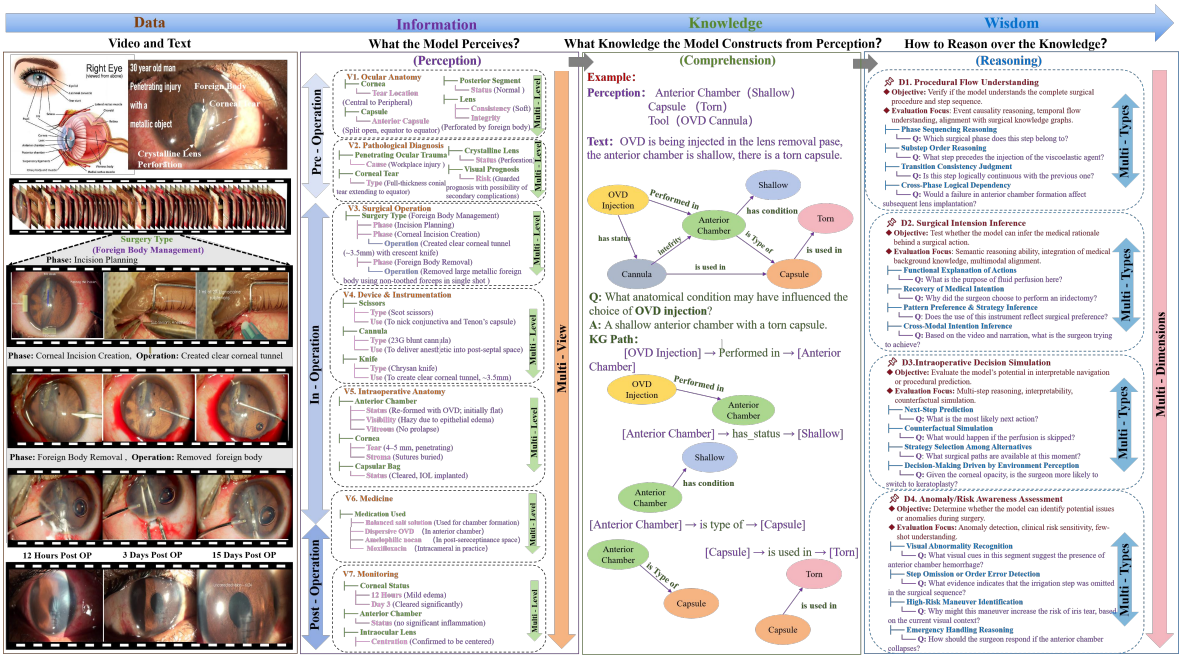

Figure 1. Overview of proposed EyePCR framework. It consists of a three stage cognitive hierarchy: fine-grained perception (1,048 attributes), scene graph comprehension (25K triplets), and multi-dimensional clinical reasoning(16 sub-tasks).

Abstract

*Multimodal Large Language Models (MLLMs) excel in general domains, but their reliability in high-stakes surgical scenarios is much less established. To bridge this gap, we develop **EyePCR**, a large-scale ophthalmic surgery benchmark grounded in structured clinical knowledge to evaluate surgical **Perception**, **Comprehension**, and **Reasoning**. The benchmark comprises 210k VQA pairs annotated with 1,048 fine-grained attributes for multi-view perception, a 25k-triplet knowledge graph for semantic comprehension, and 86k clinically grounded reasoning tasks across four categories. These structured annotations enable cognitive analysis that reflects surgical decision-making. We accordingly evaluate 18 state-of-the-art MLLMs, spanning commercial, open-source, and medical-specialized models, using expert clinicians as the reference standard. We then present **EyePCR-MLLM**, a surgical-domain model constructed by fine-tuning Qwen2.5-VL-7B on the EyePCR dataset. Our experiments reveal three key findings: 1) Commercial MLLMs outperform open-source and medical-specialized models overall, yet still lag behind human experts; 2) Supervised fine-tuning on medical data substantially improves performance; 3) **EyePCR-MLLM** markedly strengthens ophthalmic surgical **Perception**, **Comprehension**, and **Reasoning**, achieving performance comparable to GPT-5. This work sets a new public benchmark for ophthalmic surgical video understanding, highlighting the path toward expert-level clinical reasoning. The dataset and code will be released upon paper acceptance.*

1. Introduction

Multimodal large language models such as GPT-5 [42], Gemini [50] and LLaVA [33] have demonstrated significant capabilities in visual question answering (VQA), instruction following and cross-modal reasoning. However, their performance in scenarios that require domain-specific visual perception and structured semantic understanding, especially high-stakes medical scenarios, remains underexplored [23, 52]. Surgical video analysis demands fine-grained spatio-temporal reasoning to decipher subtle maneuvers in constrained anatomy, requiring interpretation of not only observable actions but also their intent, consequences, and subsequent steps [12, 36].

Previous surgical video analysis focuses on low-level tasks like tool segmentation [14, 58], phase recognition [16, 20], and temporal action detection [21, 45]. While valuable, these tasks do not probe models' capacity on multimodal cues fusion, domain knowledge alignment, or reasoning over surgical intent, procedural risks and intraoperative decision paths. Recent datasets have begun to focus on surgical VQAs [6, 16, 28]. However, they predominantly rely on LLMs to generate questions from semantic descriptions or captions, leading to issues of hallucination and poor surgical relevance [9, 24, 63]. Consequently, it cannot guarantee the factual accuracy or clinical validity. Ophthalmic surgery presents unique challenges, *e.g.*, fine-grained instrument handling, microanatomical variations, and high sensitivity to visual cues like anterior chamber depth, capsule tears, and corneal edema, all of which demand structured, knowledge-informed comprehension.

To bridge the gap, we develop **EyePCR**, a novel benchmark for evaluating surgical cognitive abilities under a **PCR** (**Perception** → **Comprehension** → **Reasoning**) paradigm shown in Fig. 1. For **Perception**, the model extracts *multi-view, multi-level attributes* from videos. The fine-grained annotations are curated across diverse semantic perspectives, *e.g.*, anatomy, tools, procedures, and physiological states, simulating how experts observe and interpret the operative field. For **Comprehension**, sensory observations are integrated into *Structured Scene Graphs* to align low-level cues with high-level surgical knowledge, each encoding intraoperative semantics using *entity–action–target* triplets grounded in surgical ontologies, *e.g.*, [`Forceps`] → `grasp` → [`Capsular Flap`]. Scene graphs are temporally aggregated into a *Semantic Memory Graph*, preserving procedural context and aligning observed events with clinical intent. Our structured graphs support a suite of reasoning tasks that are *knowledge grounded*, *clinically valid* and *interpretable*. For **Reasoning**, four categories of tasks are defined: *Procedural Flow Understanding* to verify the understanding over the complete surgical procedure; *Surgical Intent Inference* to assess medical reasoning inference for surgical actions; *Intraoperative Decision Simulation* to assess interpretable navigation/procedural prediction potential; and *Anomaly/Risk Awareness Assessment* to evaluate surgical anomaly identification ability. Beyond visual perception, these tasks interpret, predict, and reason over surgical behaviours using knowledge graphs.

EyePCR is the first large-scale (210K) ophthalmic surgery VQA corpus grounded in structured clinical knowledge, with detailed annotations across all three cognitive stages: 1) 102K VQAs annotated with 1048 multi-view, multilevel attributes for fine-grained perception; 2) 24K VQAs supported by a scene graph abstraction of 25K triplets, providing structured knowledge for comprehension; 3) 86K VQAs spanning four reasoning tasks, each evaluating a distinct dimension of surgical intelligence.

Our main contributions are three-fold: 1) Our PCR framework offers a unified paradigm for evaluating surgical cognitive models by jointly integrating fine-grained visual perception, structured knowledge comprehension, and high-level clinical reasoning, supporting the construction of *knowledge-grounded*, *interpretable*, and *clinically reliable* surgical cognition models. 2) The **EyePCR** benchmark is designed to assess MLLMs on a progression of capabilities essential for real-world medical practice: from visual perception to clinical comprehension and decision-making. 3) Our PCR paradigm significantly enhances model cognition: **EyePCR-MLLM** surpasses open-source models and rivals GPT-5, with extensive benchmarking of 18 MLLMs revealing critical gaps toward expert-level reasoning.

2. Review of Surgical Video Analysis Datasets

While surgical video datasets have shifted the focus from visual perception [21, 58] to visual understanding/reasoning [6, 59], only EyePCR provides full-spectrum annotations for perception, comprehension and reasoning (Table 1).

Fine-Grained Perception. Existing surgical datasets use coarse labels like phase boundaries [23, 52] or action clips [26], restricting fine-grained visual understanding and semantic analysis. While datasets like Cataract [46] and OphNet [20] provide frame/segment-level annotations, they lack structured annotations on surgical tools, anatomy states, and procedural context, thereby limiting fine-grained reasoning. EyePCR is structured around a hierarchical taxonomy, systematically annotating 1,048 fine-grained attributes across seven dimensions (V1–V7) as defined in Sec. 3.2. While fine-grained video understanding has been widely explored in activity recognition [44], instructional video modeling [37], and object–action co-localization [5], surgical video analysis presents unique challenges in constrained perceptive environments, task-specific instruments, and high annotation costs. EyePCR bridges this gap with a large-scale, clinically validated dataset featuring structured annotations, establishing a foundational resource for advancing reasoning capabilities in surgical AI.

Table 1. Comparison of surgical video analysis datasets. Only **EyePCR** provides full-spectrum annotations, spanning multi-view perception (**V1–V7** defined in Sec. 3.2), knowledge graph (**KG**) comprehension and clinical reasoning (**D1–D4** defined in Sec. 3.4).

Dataset	Year	Domain	Size	Data	P							C	R			
					V1	V2	V3	V4	V5	V6	V7	KG	D1	D2	D3	D4
Cholecseg [19]	2020	Laparoscopic	8K	Object-Label	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗
SurVLP [58]	2023	Endoscope	25K	Video-Label	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗
OphNet [20]	2024	Ophthalmic	14K	Video-Label	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
Cataract [14]	2024	Ophthalmic	1K	Video-Label	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗
AVOS [15]	2024	Generalist	2K	Video-Label	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	✗	✗
MedVidQA [16]	2024	Aid	3.4K	Video-Label	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗
TEMSET [6]	2025	Endoscope	24K	Video-Label	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗
GenSurg+ [18]	2024	Laparoscopic	17K	Video-Caption	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
HecVL [60]	2024	Endoscope	37K	Video-Caption	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
OphVL [21]	2025	Ophthalmic	375K	Video-Caption	✗	✗	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗
GP-VLS [45]	2024	Generalist	120K	VQA	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗
SSG-VQA [59]	2024	Generalist	960K	VQA	✗	✗	✗	✓	✓	✗	✗	✓	✗	✓	✗	✗
Surg-QA [28]	2025	Generalist	102K	VQA	✗	✗	✓	✓	✗	✗	✗	✗	✓	✗	✓	✗
EyePCR(Ours)	**2025**	**Ophthalmic**	**210K**	**VQA**	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

KG-based Comprehension. While knowledge graphs (KGs) enable multi-hop reasoning and semantic alignment in natural image understanding [29, 61] and VQAs [51, 53], they remain unexplored for surgical video analysis. Existing surgical datasets like SurVLP [58], HecVL [60], and MedVidQA [16] lack entity–relation modeling and semantic scene representations, weakening the support for temporal reasoning, causal chaining, and inter-step coherence across surgical workflows. EyePCR constructs surgical scene graphs from fine-grained attributes to bridge multimodal cues with clinical reasoning. Unlike prior static KGs [22, 35], it aligns clinical knowledge with dynamic surgical environments through real-time visual observations, procedural flow, and task sequences.

Surgical Video Reasoning. Surgical video reasoning grows with multimodal VQA datasets like MedVidQA [16], Surg-QA [28], and SSG-VQA [59]. But these datasets emphasize general-purpose or retrieval-style shallow VQAs on tool identification, anatomy naming, and isolated action recognition, more towards visual recognition than surgical cognition. EyePCR explicitly aligns its QA structure with four surgical reasoning dimensions, mirroring surgeons' contextual and temporal decision-making. Compared to medical VQA datasets centred on diagnostic image interpretation like VQA-RAD [25], PathVQA [17], and SLAKE [31], EyePCR tackles surgical reasoning's core challenges, establishing the first benchmark combining multimodal QA with clinical reasoning.

3. EyePCR Dataset Construction

3.1. Data Collection and Curation

Video Collection and Preprocessing. We collected 2,968 surgical videos from public datasets like OphNet [20], LensID [13], and Cataract-1k [14], and from curated clinical repositories [55], selected institutional archives [54], Eyetube [11] and American Academy of Ophthalmology [2]. Personally identifiable information is removed from videos. All involved videos are publicly sourced. Each video underwent preliminary preprocessing to ensure procedural completeness, and the presence of narration or embedded subtitles to support textual information extraction.

Expert Screening. Three senior ophthalmic surgeons independently evaluated all 2,968 videos against four criteria: 1) Procedural completeness, covering essential surgical phases without critical omissions; 2) Maintaining educational clarity, unobstructed, and centered surgical field, minimizing instrument-induced occlusions; 3) Exhibiting clinical representativeness, depicting standard techniques across various ophthalmic procedures; 4) Presenting consistent and informative subtitles to accurately describe key surgical maneuvers and decision points. By expert consensus, 1,544 high-quality videos were retained, ensuring clinical authenticity and providing a robust foundation for subsequent annotation.

Textual Information Extraction. A dual-mode textual extraction strategy was employed to extract accurate procedural narration. 1) Automated speech recognition was performed using OpenAI's Whisper [38] to generate high-fidelity transcriptions of narrated surgical procedures. 2) Optical character recognition (OCR) was conducted using CnOCR [7], a multilingual OCR toolkit, to extract embedded subtitle texts from video frames. Domain-specific noise reduction heuristics [32] were applied to ensure textual consistency and fidelity. The generated text corpus serves as a reliable basis for structured knowledge extraction, scene graph construction, and VQA generation.

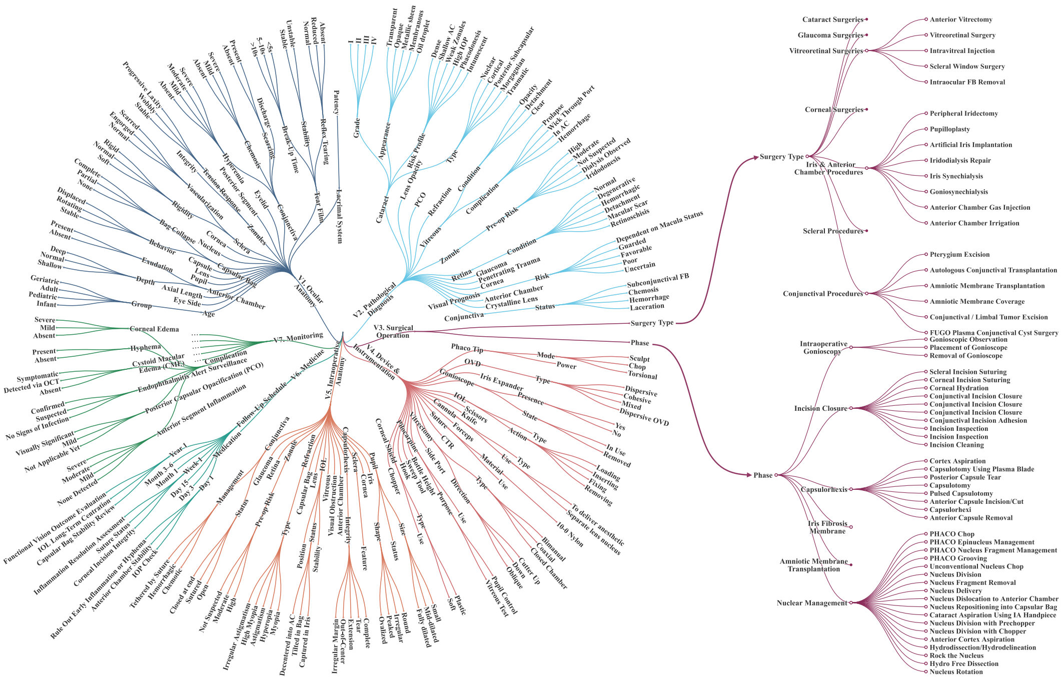

Figure 2. Hierarchy of multi-view multi-level fine-grained attributes on the full surgical continuum (Best viewed when zoomed in. Full version is provided in Appendix).

3.2. Perception Hierarchy of Multi-View Multi-Level Attributes

Multi-View Attributes. Fig. 2 depicts seven semantic views to capture distinct yet interrelated facets of surgical cognition. 1) *Ocular Anatomy*, encoding anatomical baselines essential for assessing the physiological state of the eye before intervention; 2) *Pathological Diagnosis*, documenting lesion characteristics, trauma indicators and pathological risks; 3) *Surgical Operation*, capturing procedural sequences and operational intent at different phases; 4) *Device & Instrumentation*, tracking presence, type and usage state of surgical tools; 5) *Intraoperative Anatomy*, reflecting real-time changes in ocular structures during manipulation; 6) *Medicine*, recording pharmacological management; 7) *Monitoring*, ensuring ongoing anatomical and functional surveillance post-surgery. These views collectively enable multidimensional surgical context reconstruction, facilitating visually grounded reasoning in clinical scenarios.

Multi-Level Attributes Surgical attributes are organized into three levels, reflecting how surgeon progressively process visual and procedural details during surgery. 1) *L1* outlines broad anatomical or procedural domains such as cornea, anterior chamber and surgical tools; 2) *L2* refines these into specific sub-features like corneal clarity or iris stability, capturing clinically relevant cues; 3) *L3* provides fine-grained descriptions like shape, location and dynamic state. The layered structure enables LLMs to perceive, organize, and reason with surgical data across multiple views and levels, extracting clinically meaningful insights with clarity, coherence, and purpose throughout the surgery.

Statistics of Annotated Attributes. Each video is segmented into temporal units and annotated with 1,048 fine-grained attributes across seven views. These are hierarchically organized into 76 coarse (*L1*), 240 subtype (*L2*), and 732 detailed (*L3*) attributes, collectively forming a multi-granular perceptual substrate for downstream reasoning.

3.3. Comprehension of Surgical Knowledge Graph

Structured Scene Graph To move beyond low-level perception, we transform visual attributes into coherent semantic graph representations, offering a structured foundation for clinically grounded comprehension. Specifically, for each video segment, we extract visual annotations in seven semantic views, and process video subtitles to align textual descriptions with visual observations. We then construct a *Structured Scene Graph* from these two sources of information. Each graph comprises a set of *entity–action-attribute* triplets that represent surgical interpretable events, *e.g.*, `[OVD]` $\rightarrow$ `injected into` $\rightarrow$ `anterior chamber`, allowing each predicted answer to be directly traced back to structured visual-semantic evidence grounded in the video.

Semantic Memory Graph To capture temporal and contextual continuity, we aggregate segment-level scene graphs across all videos into a *Semantic Memory Graph*, recording the evolution of the surgery by linking related enti-

ties and actions over time, modelling procedural flow, step dependencies, and recurrent structures, *e.g.*, repeated use of a device or transitions between phases. The semantic memory graph encodes spatial structure of the surgery in a clinically meaningful way, and encodes temporal continuity across video segments, enabling the model to perform cross-step reasoning and causal inference, such as understanding action preconditions or predicting subsequent maneuvers, thereby preserving the procedural integrity of surgical workflows.

Reasoning Paths We extract *Reasoning Paths* from the semantic memory graph to support high-level reasoning tasks. Each path is a multi-hop subgraph capturing clinically meaningful event sequences, *e.g.*, `incision` → `viscoelastic injection` → `capsulorhexis`. These paths mirror the procedural logic and step dependencies used in surgical reasoning and serve as templates for reasoning focused VQAs. For example, a path ending in `capsulorhexis` may yield the question "What steps are typically required before capsulorhexis?" or "Why is viscoelastic injected at this stage?" This ensures that each question is grounded in verifiable surgical logic, enabling interpretable and clinically aligned model evaluation. These reasoning paths not only serve as templates for constructing clinically meaningful VQAs but also constrain model outputs to medically valid and contextually grounded event sequences, mitigating hallucinations and enabling interpretable, logic driven surgical reasoning.

Statistics of Knowledge Graph The constructed semantic memory graph comprises 25,567 triplets spanning 16,034 entities and 3,507 relations, from which we extract 11,462 logical paths (9,499 two-hop and 1,963 multi-hop).

3.4. Surgical Video Reasoning

Perception and comprehension extract and structure visual semantics, while reasoning requires interpreting, inferring, and simulating surgical cognition. We evaluate this through four clinically grounded dimensions. 1) **Procedural Flow Understanding** evaluates the model's ability to grasp the temporal causal structure of surgical workflows, *e.g.*, ordering sub-steps within a phase, recognizing cause effect relations between adjacent or distant actions, and verifying the clinical validity of phase transitions. These tasks require reasoning beyond step recognition to grasp procedural flow and inter-step dependencies. 2) **Surgical Intent Inference** evaluates the ability of correctly interpreting the clinical rationale behind a given action, *e.g.*, attributing actions to goals, recognizing patterns in strategy selection, and aligning procedural behavior with the physiological or pathological state in the operative field. Rather than identifying tools or anatomy in isolation, the model must infer why a specific tool is used at a given stage and how its use supports the surgical objective. 3) **Intraoperative Decision Simulation** simulates real time surgical planning and adaptive decision making, *e.g.*, forecasting the next step in an ongoing procedure, planning contingencies for complications such as anterior chamber collapse, proposing viable alternatives under uncertainty, and adjusting decisions to observed scene conditions. These tasks test the model's ability to move beyond static recognition and reason under dynamic constraints. 4) **Anomaly/Risk Awareness Assessment** focuses on detecting deviations from expected patterns and inferring their clinical implications, *e.g.*, identifying subtle pathological changes, predicting complications from anatomical or procedural abnormalities, and estimating the consequences of incorrect actions. These tasks require integrating visual cues with latent medical knowledge to approximate expert-level risk sensitivity and safety judgment. Across all four dimensions, we cover multiple high-level clinical inference types, from temporal logic modeling and surgical goal attribution to strategic forecasting and risk simulation, instantiated as VQAs grounded in structured scene graphs and domain ontologies, enabling rigorous and interpretable assessment of surgical cognition.

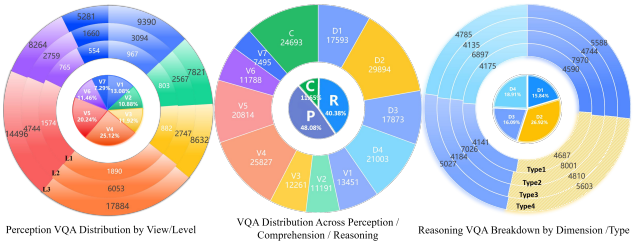

Figure 3. Distribution of 210K VQAs across **Perception**, **Comprehension**, and **Reasoning** stages.

3.5. Construction of VQAs

VQA on Attributes We design a comprehensive set of VQA templates targeting directly observable surgical scene properties. Using `Gemini-2.5-Pro` [10], we automatically generate structured VQA pairs for each video segment, including MCQs (e.g., *"Which ocular structure is being manipulated in this segment?"* with options *"Cornea"*, *"Conjunctiva"*, *"Sclera"*, *"Iris"*) and OEQs (e.g., *"Describe the baseline anatomical condition of the cornea before any surgical intervention."*). As shown in the left panel of Fig. 3, we obtain **64,426** MCQs and **38,401** OEQs from **47,336** video segments under the perception stage, spanning seven semantic views and three attribute levels. This design ensures both breadth and depth of perceptual VQAs and evaluates the model's ability to extract diverse clinical attributes directly from surgical videos.

VQA on KG-based Comprehension. VQAs at the comprehension stage assess whether models move beyond perception to interpret structured surgical semantics

grounded in knowledge graphs. We generate **24,693** OEQs from **16,556** video segments using *entity–action–attribute* triplets from scene graphs, capturing intraoperative actions, object interactions, and contextual relations. For example, the question *"Describe the path of the anchoring suture"* expects an answer such as *"Passes into the sclera near the limbus and out through the edge of the conjunctival graft"*, grounded in triplets `[Suture]` → `passes into` → `Sclera` and `[Suture]` → `related to` → `Conjunctival Graft`. As shown in the middle panel of Fig. 3, this stage constitutes ∼11.5% of all VQAs and bridges low-level visual parsing with high-level surgical reasoning, emphasizing intra-segment comprehension.

VQA on Surgical Reasoning. We derive **86,363** OEQs from **19,001** video segments with explicit reasoning paths, targeting four key reasoning dimensions: (1) Procedural Flow Comprehension (e.g., "Which surgical phase does this action belong to?"), probing stepwise dependencies; (2) Surgical Intent Inference (e.g., "Why did the surgeon perform an iridectomy?"), inferring action purpose from context; (3) Intraoperative Decision Simulation (e.g., "What would happen if perfusion is skipped?"), predicting clinical consequences; (4) Anomaly/Risk Awareness (e.g., "What potential complications could result from this iris manipulation?"), assessing risk-aware reasoning. Open-ended questions better capture surgical reasoning complexity by avoiding fixed options and allowing multiple valid solutions.

Statistics of VQAs We curate a total of 82,893 video segments from 1,544 full-length ophthalmic surgery videos, generating 213,883 high-quality VQAs, *i.e.*, 64,425 MCQs and 149,458 OEQs, categorized into three cognitive stages: **102,827** for **Perception**, **24,693** for **Comprehension**, and **86,363** for **Reasoning**. The VQA distribution (Fig. 3) reflects our emphasis on progressing from visual perception to comprehension and clinical reasoning.

3.6. Evaluation Protocols

Dataset Split. The EyePCR dataset is split at the patient level, avoiding any overlap of patient identities or videos between training and test sets. It contains 82,893 annotated video segments (77,133 for training, 5,760 for testing) and 213,883 VQAs (198,630 for training, 15,253 for testing). The training set contains 95,246, 23,072, and 80,312 VQAs for perception, comprehension, and reasoning, respectively, while the test set includes 7,581, 1,621, and 6,051 VQAs for these categories. This structured split enables fine-grained evaluation of perception, comprehension, and reasoning, and ensures fair, reproducible performance comparison.

Evaluation Metrics. We use accuracy for MCQs and the following for OEQs: 1) linguistic metrics, including BLEU-1–4 and ROUGE-1 [43], ROUGE-2, and ROUGE-L [30], to measure lexical overlap with reference answers; 2) GPT-4.1-based rubric scoring [34, 41], assessing factual correctness, specificity, and completeness on a 5-point Likert scale via standardized grading prompts.

4. Experimental Results

4.1. Experimental Setup

Evaluation Models. The proposed EyePCR benchmark is evaluated on a broad spectrum of MLLMs, encompassing commercial, open-source, and medical-specialized across varying capability tiers. **Commercial models** include GPT-5 [42], GPT-4.1 [40], GPT-o3 [39], Claude 3.7 (Sonnet) [3], as well as Gemini-2.0-Flash, Gemini-2.5-Flash, and Gemini-2.5-Pro [10], representing the state of the art in general purpose multimodal reasoning. **Open-source models** include Qwen2.5-VL-7B, Qwen2.5-VL-32B [4], Qwen3-VL-30B-A3B [56], InternVL-14B [48], LLaVA-Video-7B [64], and VideoLLaMA3-7B [62]. **Med-specialized models** include MedGemma-4B [47], Lingshu-7B [57], LLaVA-Med-7B [27], HuatuoGPT-Vision-7B [8].

Fine-Tune Models. We fine-tune the open-source Qwen2.5-VL-7B model [4] on the EyePCR training set to obtain **EyePCR-MLLM**, a domain-specialized model for surgical video understanding and clinical VQA. Fine-tuning spans all three cognitive levels, enabling the model to encode structured surgical knowledge and strengthen multimodal alignment. Training details are provided in the Appendix.

Human Study. We incorporate human experts to contextualize model performance. In total, **66 ophthalmology professionals** (47 trainees and 19 surgeons) participated in the evaluation. Each expert was randomly assigned one of five questionnaires comprising 50 MCQs and 20 OEQs, covering all three cognitive stages, Perception, Comprehension, and Reasoning, across diverse semantic views. This human benchmark supports assessment of both perceptual accuracy and clinical reasoning, and highlights the performance gap between domain-informed experts and general-purpose MLLMs.

4.2. Evaluation of OEQs on Perception, Comprehension and Reasoning

Tab. 2 summarizes the performance of various MLLMs on OEQs across the three cognitive stages. We report BLEU-4, ROUGE-L scores, and GPT-4.1–based rubric evaluations. Several key findings emerge:

EyePCR Are Challenging Yet Valid. All models perform markedly worse on OEQs; *e.g.*, Gemini 2.5 Pro achieves the highest BLEU-4 of 0.0731, while most open-source models fall below 0.04, highlighting the increased complexity of generative surgical understanding. Interestingly, human experts also score poorly under lexical metrics, *e.g.*, Ophthalmic Surgeon obtains a BLEU-4 of 0.0116, not due to limited expertise but because expert answers tend

Table 2. **OEQs** performance in BLEU-4 (B-4) and ROUGE-L (R-L) across **Perception (P)**, **Comprehension (C)**, and **Reasoning (D1–D4)**, alongside average and ChatGPT-based rubric scores. **Blue**,**Green**, **Red**, **Yellow**: highest among **commercial models**, **Med-Specialized models**, **open-source models** and **Expert**; **Bold**: best overall.

	Participants	P		C		R								Avg		ChatGPT
		B-4	R-L	B-4	R-L	D1 B-4	D1 R-L	D2 B-4	D2 R-L	D3 B-4	D3 R-L	D4 B-4	D4 R-L	B-4	R-L	(0→5)
Commercial	Gemini 2.0 Flash [10]	0.0624	0.3783	0.0603	0.2926	0.0329	0.1956	0.0513	0.2499	0.0338	0.2105	0.0392	0.2214	0.0467	0.2581	3.2417
	Claude 3.7 [3]	0.0497	**0.4207**	0.0518	0.2888	0.0375	0.2371	0.0652	0.2850	0.0506	0.2818	0.0499	0.2736	0.0508	0.2978	3.5261
	Gemini 2.5 Flash [10]	0.0568	0.3993	0.0593	0.2871	0.0448	0.2567	0.0649	0.2790	0.0576	0.2772	0.0482	0.2615	0.0553	0.2935	3.5048
	GPT-o3 [39]	0.0449	0.3757	0.0475	0.2820	0.0469	0.2722	0.0470	0.2606	0.0484	0.2798	0.0446	0.2686	0.0466	0.2898	3.9349
	GPT-4.1 [40]	0.0398	0.3069	0.0513	0.2696	0.0365	0.2297	0.0507	0.2636	0.0445	0.2451	0.0411	0.2351	0.0440	0.2583	3.8405
	GPT-5 [42]	0.0586	0.4164	0.0548	0.3002	0.0506	**0.2920**	0.0568	0.2855	0.0502	0.2817	0.0453	0.2727	0.0527	0.3080	3.8907
	Gemini 2.5 Pro [10]	0.0685	0.4060	**0.0697**	**0.3200**	**0.0573**	0.2877	**0.0893**	**0.3213**	**0.0829**	**0.3245**	**0.0709**	**0.3163**	**0.0731**	**0.3293**	3.6949
Open-Source	VideoLLaMA3-7B [62]	0.0174	0.1351	0.0211	0.1546	0.0189	0.1398	0.0174	0.1469	0.0198	0.1364	0.0157	0.1235	0.0183	0.1393	2.3605
	LLaVA-Video-7B [64]	0.0060	0.0598	0.0087	0.0808	0.0077	0.0683	0.0091	0.0965	0.0080	0.0754	0.0081	0.0766	0.0079	0.0762	1.8765
	Qwen2.5-VL-7B [4]	0.0049	0.0465	0.0062	0.0597	0.0076	0.0653	0.0072	0.0685	0.0060	0.0563	0.0056	0.0527	0.0063	0.0582	2.1659
	Qwen2.5-VL-32B [4]	0.0416	0.2995	0.0495	0.2631	0.0371	0.2124	0.0398	0.2278	0.0315	0.1938	0.0288	0.1902	0.0381	0.2311	2.7163
	Qwen3-VL-30B-A3B [56]	0.0468	0.3582	0.0510	0.2820	0.0419	0.2303	0.0503	0.2593	0.0414	0.2261	0.0364	0.2184	0.0446	0.2624	3.3280
	InternVL-14B [48]	0.0127	0.1182	0.0158	0.1287	0.0143	0.1155	0.0171	0.1408	0.0128	0.1128	0.0137	0.1120	0.0144	0.1213	2.5446
Med-Specialized	MedGemma-27B [47]	0.0044	0.0512	0.0067	0.0614	0.0063	0.0541	0.0073	0.0629	0.0062	0.0553	0.0054	0.0512	0.0060	0.0560	2.6959
	Lingshu-7B [57]	0.0113	0.0957	0.0125	0.1074	0.0121	0.0981	0.0172	0.1350	0.0119	0.1016	0.0142	0.1165	0.0132	0.1090	2.5766
	LLaVA-Med-7B [27]	0.0167	0.1375	0.0193	0.1420	0.0235	0.1592	0.0164	0.1451	0.0207	0.1458	0.0154	0.1245	0.1867	0.1424	2.3153
	HuatuoGPT-Vision-7B [8]	0.0110	0.1053	0.0124	0.1049	0.0128	0.1009	0.0139	0.1200	0.0107	0.0967	0.0121	0.1051	0.0122	0.1055	2.4097
	EyePCR-MLLM (7B)(Ours)	**0.1099**	0.3795	0.0513	0.2513	0.0510	0.2355	0.0703	0.2792	0.0645	0.2638	0.0597	0.2612	0.0678	0.2784	3.0062
Expert	Ophthalmology Trainee	0.0092	0.0808	0.0057	0.0927	0.0010	0.0105	0.0075	0.0782	0.0012	0.1253	0.0142	0.1581	0.0065	0.0909	3.1697
	Ophthalmic Surgeon	0.0108	0.1431	0.0085	0.1245	0.0054	0.0633	0.0080	0.1028	0.0141	0.1617	0.0230	0.2769	0.0116	0.1454	**4.1583**

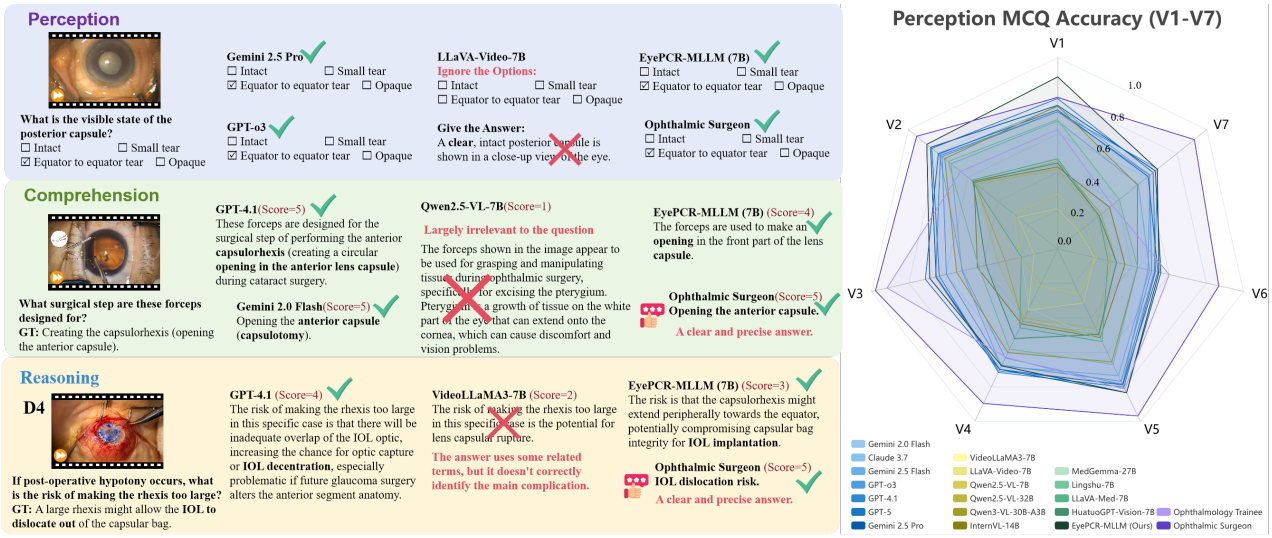

Figure 4. **Left:** Qualitative comparisons across Perception, Comprehension, and Reasoning tasks in EyePCR. **Right: MCQs** performance across seven perceptual views (V1–V7) in the **Perception** stage.

to be concise and semantically precise, with low n-gram overlap. To address this, we adopt a rubric based evaluation of factuality, specificity, and completeness. Under this rubric, EyePCR's OEQs are clinically validated, with the Ophthalmic Surgeon achieving the highest score of 4.1583.

EyePCR Comprehensively and Accurately Assesses MLLMs. EyePCR covers Perception (P), Comprehension (C), and four Reasoning tasks (D1–D4), spanning capabilities from low-level perception to high-level inference. It evaluates with BLEU-4, ROUGE-L, and a Chat-

GPT rubric, which reliably separate model families and competency tiers. Gemini 2.5 Pro achieves the highest B-4/R-L across many cells among commercial models, while EyePCR-MLLM (7B) leads multiple P/R entries within the medical-specialized group; open-source models are markedly lower (e.g., Qwen3-VL-30B-A3B, ChatGPT = 3.3280), indicating demands for fine-grained visual/semantic alignment beyond generic pretraining. Human experts (Ophthalmic Surgeon) score 4.1583, exceeding all models (commercial best GPT-o3 = 3.9349), validating clinical soundness and revealing headroom. By combining multi-stage tasks with multi-metric evaluation and anchoring validity in expert performance, EyePCR accurately profiles model strengths/weaknesses in ophthalmic settings and highlights where further improvements are needed.

Rubric Scores Better Assess Semantic Understanding. The disparity between lexical metrics and rubric evaluations underscores the limitations of BLEU and ROUGE in clinical QA. These metrics rely on n-gram overlap and tend to reward surface level lexical similarity [1, 49], penalizing semantically concise, expert-style answers. In contrast, rubric scoring evaluates factuality, specificity, and completeness, giving credit to clinically correct responses even when lexically dissimilar to the reference. Consequently, although the Ophthalmic Surgeon attains low BLEU-4, it achieves the highest rubric score, surpassing all models, indicating that rubric evaluation better aligns with domain expertise and the demands of safety-critical medical QA. Integrating semantic aware evaluation is thus crucial for measuring generative quality in high stakes medical applications.

Fig. 4 (left) summarizes qualitative findings across **Perception**, **Comprehension**, and **Reasoning**. Gemini 2.5 Pro, GPT-o3, and **EyePCR-MLLM** correctly identified posterior capsule status, indicating robust low-level parsing; LLaVA-Video-7B failed to follow the required answer format. GPT-4.1, Gemini 2.0 Flash, and **EyePCR-MLLM** correctly linked forceps to anterior capsulorhexis, whereas Qwen2.5-VL-7B misattributed the tool's use, revealing contextual misunderstanding. GPT-4.1 and **EyePCR-MLLM** converged on the risk of IOL dislocation; only the latter earned full credit by explicitly naming the key complication. Using GPT-4.1 rubric scores, these cases distinguished clinically precise answers from partial or diffuse ones, highlighting **EyePCR-MLLM**'s consistent expert-level responses and exposing the grounding and reasoning gaps of general-purpose or insufficiently adapted models.

4.3. Evaluation of Fine-Grained Perception (MCQ)

Fig. 4 (right radar chart) summarizes the MCQ comparisons for fine-grained perception. We can observe the following.

Ophthalmic Surgery Remains Challenging for MLLMs Ophthalmic surgeons achieve high accuracy (0.9090), confirming the benchmark's clinical validity, while leading models (GPT-5: 0.6958, Gemini-2.5 Pro: 0.7138) and open-source alternatives (< 0.60) lag significantly. Though EyePCR-MLLM and Gemini-2.5 Pro surpass trainee performance, they fall short of expert levels, revealing MLLMs' limitations in interpreting nuanced surgical cues despite recognizing common patterns. This human–AI gap highlights that clinically reliable perception requires deeper integration of surgical semantics beyond data scale alone.

Domain-Specific Fine-Tuning Greatly Improves Performance. Fine-tuning on domain-specific medical data yields substantial and parameter-efficient gains across open-source MLLMs. Even models of similar size (e.g., 7B) can outperform much larger general counterparts when guided by clinical priors and curated supervision. Such targeted adaptation enhances perceptual accuracy and cross-view consistency, confirming that surgical semantics and domain alignment contribute more than sheer model scale in improving clinical understanding.

EyePCR-MLLM Markedly Strengthens Ophthalmic Surgical. Built upon Qwen2.5-VL-7B, the domain-tuned **EyePCR-MLLM** achieves a remarkable gain in average MCQ accuracy from 0.3866 to 0.7412. It consistently surpasses larger open-source models (Qwen2.5-VL-32B, InternVL-14B) and even commercial systems like GPT-5 and Gemini-2.5 Pro, despite having far fewer parameters. When compared to human participants, EyePCR-MLLM outperforms the Ophthalmology Trainee and approaches the Ophthalmic Surgeon, especially in visual categories such as **V1** and **V5**, highlighting the efficacy of the EyePCR dataset in closing the clinical perception gap.

5. Conclusion and Future Work

We present **EyePCR**, the first comprehensive benchmark for assessing cognitive intelligence in ophthalmic surgery, integrating multi-view fine-grained attributes, structured knowledge graphs, and clinically grounded reasoning tasks within a unified **Perception**, **Comprehension**, **Reasoning** framework. The benchmark comprises 210k VQAs evaluating fine-grained attribute perception, procedural semantic alignment, and expert level clinical inference. Across 18 MLLMs, we observe a clear gap: commercial models show strong perceptual ability, but their reasoning remains shallow without domain adaptation. Our domain-adapted **EyePCR-MLLM** achieves SOTA performance among open-source models and is competitive with leading commercial systems, highlighting the value of structured surgical representations and task-aware fine-tuning.

Looking ahead, EyePCR provides a reference benchmark for surgical AI and lays the groundwork for clinically reliable, interpretable multimodal foundation models. Beyond advancing MLLM research, we aim to catalyze broader efforts toward open, transparent, and generalizable systems for surgical video understanding.

References

[1] Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial Intelligence in Medicine*, 155:102938, 2024. 8

[2] American Academy of Ophthalmology. Aao education – clinical surgical videos, 2024. https://www.aao.org/education. 3

[3] Anthropic. Claude 3.7 sonnet. https://www.anthropic.com/index/claude, 2024. Accessed: May 2025. 6, 7

[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. 6, 7

[5] Soufiane Belharbi, Shakeeb Murtaza, Marco Pedersoli, Ismail Ben Ayed, Luke McCaffrey, and Eric Granger. Colocam: Class activation mapping for object co-localization in weakly-labeled unconstrained videos. *Pattern Recognition*, 144:109924, 2023. 2

[6] Muhammad Bilal, Mahmood Alam, Deepa Bapu, Stephan Korsgen, Neeraj Lal, Simon Bach, Amir M Hajivanand, Muhammed Ali, Kamran Soomro, Iqbal Qasim, et al. Temset-24k: Densely annotated dataset for indexing multipart endoscopic videos using surgical timeline segmentation. *arXiv preprint arXiv:2502.06708*, 2025. 2, 3

[7] BreezeWhite. Cnocr: Paddleocr-based toolkit for chinese text recognition. https://github.com/breezedeus/CnOCR, 2023. 3

[8] 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. 6, 7

[9] Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, and Joyce Chai. Multi-object hallucination in vision language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:44393–44418, 2024. 2

[10] Google DeepMind. Gemini 1.5: Scaling up multimodal reasoning, 2024. https://deepmind.google/technologies/gemini/. 5, 6, 7

[11] Eyetube. Eyetube – surgical video platform, 2024. https://eyetube.net/. 3

[12] Zhuo Gao, Liu Jiang, Tianyi Shao, and Guang-Zhong Yang. Trans-svnet: Hybrid transformer network for surgical video analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 208–218. Springer, 2022. 2

[13] Negin Ghamsarian, Mario Taschwer, Doris Putzgruber-Adamitsch, Stephanie Sarny, Yosuf El-Shabrawi, and Klaus Schoeffmann. Lensid: a cnn-rnn-based framework towards lens irregularity detection in cataract surgery videos. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 76–86. Springer, 2021. 3

[14] Negin Ghamsarian, Yosuf El-Shabrawi, Sahar Nasirihaghighi, Doris Putzgruber-Adamitsch, Martin Zinkernagel, Sebastian Wolf, Klaus Schoeffmann, and Raphael Sznitman. Cataract-1k dataset for deep-learning-assisted analysis of cataract surgery videos. *Scientific Data*, 11(1):373, 2024. 2, 3

[15] Emmett D. Goodman, Krishna K. Patel, Yilun Zhang, William Locke, Chris J. Kennedy, Rohan Mehrotra, Stephen Ren, Melody Guan, Orr Zohar, Maren Downing, Hao Wei Chen, Jevin Z. Clark, Margaret T. Berrigan, Gabriel A. Brat, and Serena Yeung-Levy. Analyzing surgical technique in diverse open surgical videos with multitask machine learning. *JAMA Surgery*, 159(2):185–192, 2024. 3

[16] Deepak Gupta, Kush Attal, and Dina Demner-Fushman. A dataset for medical instructional video classification and question answering. *Scientific Data*, 10(1):158, 2023. 2, 3

[17] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. 3

[18] Mohammadmahdi Honarmand, Muhammad Abdullah Jamal, and Omid Mohareri. Vidlpro: A video-language pretraining framework for robotic and laparoscopic surgery. In *Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond*, 2024. 3

[19] W-Y Hong, C-L Kao, Y-H Kuo, J-R Wang, W-L Chang, and C-S Shih. Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. *arXiv preprint arXiv:2012.12453*, 2020. 3

[20] Ming Hu, Peng Xia, Lin Wang, Siyuan Yan, Feilong Tang, Zhongxing Xu, Yimin Luo, Kaimin Song, Jurgen Leitner, Xuelian Cheng, et al. Ophnet: A large-scale video benchmark for ophthalmic surgical workflow understanding. In *European Conference on Computer Vision (ECCV)*, pages 481–500. Springer, 2024. 2, 3

[21] Ming Hu, Kun Yuan, Yaling Shen, Feilong Tang, Xiaohao Xu, Lin Zhou, Wei Li, Ying Chen, Zhongxing Xu, Zelin Peng, et al. Ophclip: Hierarchical retrieval-augmented learning for ophthalmic surgical video-language pretraining. *arXiv preprint arXiv:2411.15421*, 2024. 2, 3

[22] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021. 3

[23] Yuxing Jin, Qi Dou, Hao Chen, Lequan Yu, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. Multi-task recurrent model for surgical workflow and phase recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5744–5752, 2020. 2

[24] Junho Kim, Hyunjun Kim, Kim Yeonju, and Yong Man Ro. Code: Contrasting self-generated description to combat hallucination in large multi-modal models. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:133571–133599, 2024. 2

[25] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):1–10, 2018. 3

[26] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[27] 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. 6, 7

[28] Jiajie Li, Garrett Skinner, Gene Yang, Brian R Quaranto, Steven D Schwaitzberg, Peter CW Kim, and Jinjun Xiong. Llava-surg: towards multimodal surgical assistant via structured surgical video learning. *arXiv preprint arXiv:2408.07981*, 2024. 2, 3

[29] Xinyue Li, Chen Zhang, Yen-Chun Lin, and et al. Boosting image-grounded dialogue with grounded scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 964–965, 2020. 3

[30] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, 2004. 6

[31] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021. 3

[32] Feifan Liu, Shubo Shen, and Hongfang Liu. Syntactic and semantic pre-processing for clinical text mining. In *AMIA Annual Symposium Proceedings*, pages 750–759. American Medical Informatics Association, 2021. 3

[33] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:34892–34916, 2023. 2

[34] Xiaoxuan Liu, Rui Zhang, and et al. Truthfulqa-rubric: Measuring clinical factuality with gpt-4-based rubric scoring. *arXiv preprint arXiv:2306.05337*, 2023. 6

[35] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409, 2022. 3

[36] Chensu Ma, Yukun Zhang, Lequan Yu, Qi Chen, and Pheng-Ann Heng. Self-supervised surgical phase recognition with consistent temporal synchronization. *Medical Image Analysis*, 72:102117, 2021. 2

[37] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2630–2640, 2019. 2

[38] OpenAI. Whisper: Open-source automatic speech recognition. https://openai.com/index/whisper/, 2022. 3

[39] OpenAI. Gpt-3.5 overview. https://platform.openai.com/docs/models/gpt-3-5, 2024. Accessed: 2025-05-13. 6, 7

[40] OpenAI. Gpt-4.1. https://openai.com/chatgpt, 2024. Accessed: 2025-05-13. 6, 7

[41] OpenAI. Gpt-4o technical report. https://openai.com/research/gpt-4o, 2024. Accessed: 2025-05-13. 6

[42] OpenAI. Chatgpt-5. https://chat.openai.com/, 2025. Large language model developed by OpenAI, accessed 2025-10-19. 2, 6, 7

[43] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6

[44] AJ Piergiovanni and Michael S Ryoo. Fine-grained activity recognition in baseball videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1805–1814, 2020. 2

[45] Samuel Schmidgall, Joseph Cho, Cyril Zakka, and William Hiesinger. Gp-vls: A general-purpose vision language model for surgery. *arXiv preprint arXiv:2407.19305*, 2024. 2, 3

[46] Klaus Schoeffmann, Mario Taschwer, Stephanie Sarny, Bernd Münzer, Manfred Jürgen Primus, and Doris Putzgruber. Cataract-101: video dataset of 101 cataract surgeries. In *Proceedings of the 9th ACM Multimedia Systems Conference (ACM MM)*, pages 421–425, 2018. 2

[47] 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. 6, 7

[48] Shanghai AI Laboratory. Internvl-14b: Vision-language model by shanghai ai lab. https://huggingface.co/OpenGVLab/InternVL-14B-224px, 2024. Accessed: 2025-05-13. 6, 7

[49] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*, pages 1–8, 2025. 8

[50] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2

[51] Damien Teney, Lingqiao Liu, and Anton van den Hengel. Graph-structured representations for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3233–3241, 2017. 3

[52] Apurva Twinanda, Sara Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. En-

donet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 36 (1):86–97, 2016. 2

[53] Yufei Wang, Yaodong Yang, Peng Gao, and et al. Vqa-gnn: Reasoning with multimodal knowledge via graph neural networks for visual question answering. *arXiv preprint arXiv:2205.11501*, 2022. 3

[54] Wenzhou Medical University Eye Hospital. Wenzhou eye surgical education repository, 2024. A curated set of ophthalmic surgical training videos reviewed and released for clinical education and research by Wenzhou Medical University Eye Hospital. 3

[55] Wills Eye Hospital. Wills eye surgical video repository, 2024. A curated repository of expert-labeled ophthalmic surgical videos for clinical training and research, maintained by Wills Eye Hospital. 3

[56] 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. 6, 7

[57] 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. 6, 7

[58] Kun Yuan, Vinkle Srivastav, Tong Yu, Joel L Lavanchy, Pietro Mascagni, Nassir Navab, and Nicolas Padoy. Learning multi-modal representations by watching hundreds of surgical video lectures. *arXiv preprint arXiv:2307.15220*, 2023. 2, 3

[59] Kun Yuan, Manasi Kattel, Joël L Lavanchy, Nassir Navab, Vinkle Srivastav, and Nicolas Padoy. Advancing surgical vqa with scene graph knowledge. *International Journal of Computer Assisted Radiology and Surgery*, 19(7):1409–1417, 2024. 2, 3

[60] Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. Hecvl: Hierarchical video-language pretraining for zero-shot surgical phase recognition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 306–316. Springer, 2024. 3

[61] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5831–5840, 2018. 3

[62] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 6, 7

[63] Mengxi Zhang, Wenhao Wu, Yu Lu, Yuxin Song, Kang Rong, Huanjin Yao, Jianbo Zhao, Fanglong Liu, Haocheng Feng, Jingdong Wang, et al. Automated multi-level preference for mllms. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:26171–26194, 2024. 2

[64] Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model. https://llava-vl.github.io/blog/2024-04-30-llava-next-video/, 2024. Blog post. Accessed: 2025-05-13. 6, 7