SurgCoT: Advancing Spatiotemporal Reasoning in Surgical Videos through a Chain-of-Thought Benchmark

Anonymous CVPR submission

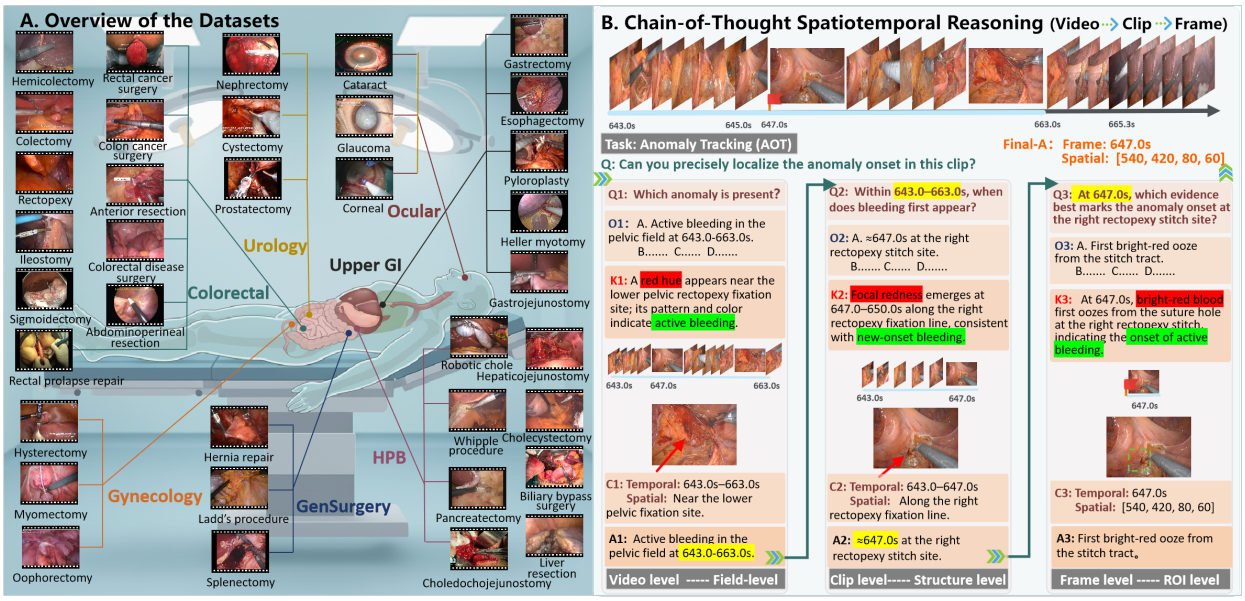

Figure 1. SurgCoT comprises **2,841 surgical videos** across **7 specialties** and **35 procedures**, with **19,345 main questions** and **59,177 sub-questions**. SurgCoT advances beyond frame-level tasks (*e.g.*, phase/tool recognition) by introducing a three-stage, five-tuple annotation protocol (`Question`→`Option`→`Knowledge`→`Clue`→`Answer`) to scaffold chain-of-thought reasoning. The framework's efficacy stems from its multi-stage reasoning structure and the synergistic interaction between the `Knowledge` field, which supplies contextual background, and the `Clue` field, which provides targeted spatiotemporal evidence, jointly enabling hierarchical reasoning.

Abstract

Fine-grained spatiotemporal reasoning on surgical videos is critical, yet the capabilities of Multi-modal Large Language Models (MLLMs) in this domain remain largely unexplored. To bridge this gap, we introduce ***SurgCoT****, a unified benchmark for evaluating chain-of-thought (CoT) reasoning in MLLMs across* ***7 surgical specialties*** *and* ***35 diverse procedures****. SurgCoT assesses five core reasoning dimensions: Causal Action Ordering, Cue–Action Alignment, Affordance Mapping, Micro-Transition Localization, and Anomaly Onset Tracking, through a structured CoT framework with an intensive annotation protocol (`Question`→`Option`→`Knowledge`→`Clue`→`Answer`), where the `Knowledge` field provides essential background context and `Clue` provides definitive spatiotemporal evidence. Evaluation of 10 leading MLLMs shows: 1) commercial models outperform open-source and medical-specialized variants; 2) significant gaps exist in surgical CoT reasoning; 3) SurgCoT enables effective evaluation and enhances progressive spatiotemporal reasoning. SurgCoT provides a reproducible testbed to narrow the gap between MLLM capabilities and clinical reasoning demands. Code and data will be released.*

1. Introduction

Surgical videos are a cornerstone of perioperative care and education, rich with dynamic anatomical and procedural

data [32, 52]. The growing volume and complexity of this data make fine-grained spatiotemporal understanding critical for clinical safety and efficiency [20]. In recent years, Multi-modal Large Language Models (MLLMs) have been adapted for surgical settings, enabling interactive analysis for tasks like surgical phase recognition [37, 44, 58], instrument recognition [9, 18], tissue detection [9, 18], and surgery understanding [7, 11, 13]. This rapid progress, however, creates a parallel need for robust evaluation frameworks to assess their true clinical utility.

Surgical benchmarks fall into two categories: General-purpose benchmarks [37, 55], covering broad scenarios but emphasize generic question answering (*e.g.*, phases, instruments) while specialized benchmarks [24, 54] focus on narrow procedural domains (*e.g.*, ophthalmic, endoscopic) but rely on frame/clip-level VQAs that treat videos as discrete segments, ignoring cross-temporal dependencies and failing to evaluate spatiotemporal or causal reasoning [53, 59]. In contrast, surgeons must track subtle, rapid spatiotemporal changes to support fine-grained inference and decision-making [30, 39]. A critical question remains: *Can MLLMs achieve expert-level progressive spatiotemporal reasoning and causal judgment across diverse surgical contexts?*

To tackle this, we introduce **SurgCoT** (Fig. 1), a unified surgical video benchmark for evaluating MLLMs under a chain-of-thought (CoT) reasoning framework. SurgCoT advances the field in two key aspects: 1) **Comprehensive Task Coverage**: Unlike prior benchmarks focused on narrow tasks [24, 45, 54] or generic queries [37, 53, 55], SurgCoT spans 35 procedures across 7 diverse specialties, enabling broad and realistic assessment of expert-level cognitive reasoning. 2) **In-depth Reasoning Dimensions**: SurgCoT introduces 5 clinical-relevant reasoning dimensions (*Causal Action Ordering*, *Cue-Action Alignment*, *Affordance Mapping*, *Micro-Transition Localization*, and *Anomaly Onset & Tracking*), reflecting critical cognitive processes in real world surgery. Collectively, these rich reasoning dimensions on diverse procedures enable a fine-grained evaluation of spatiotemporal understanding and clinical decision-making, closely mirroring actual surgical workflows.

To enable CoT reasoning in surgical video understanding, as shown in Fig. 1, we introduce a three-stage progressive reasoning framework (`Q1`→`Q2`→`Q3`) with a comprehensive five-tuple annotation protocol (`Question`→`Option`→`Knowledge`→`Clue`→`Answer`). The novelty is two-fold: 1) It establishes a hierarchical reasoning pathway progressing from video-level comprehension through clip-level analysis to frame-level localization. As shown in Fig. 1, all five annotation tuples evolve through these stages, with each stage's `Answer` serving as contextual evidence for the subsequent stage, forming a continuous reasoning chain. 2) It employs synergistic annotation strategies where the `Knowledge` field provides clinically relevant background context based on the `Question` and `Option`, while the `Clue` supplies targeted spatiotemporal evidence extracted directly from the surgical video content. This integrated design progressively narrows both temporal and spatial scope, effectively guiding MLLMs through structured hierarchical reasoning.

To validate its effectiveness as a unified spatiotemporal reasoning benchmark, we systematically evaluate 12 state-of-the-art MLLMs across diverse surgical procedures within a CoT framework, including open-source [23, 31, 46, 50], medically specialized [5, 38, 51], and commercial variants [2, 10, 33]. Results show that while commercial models lead their counterparts, all models exhibit significant limitations in fine-grained spatiotemporal understanding. Notably, by utilizing the three-stage progressive reasoning framework with five-tuple annotation protocol, considerable accuracy improvements could be achieved, indicating a viable path toward enhanced reasoning capability. SurgCoT provides a reproducible foundation for methodological development and equitable evaluation in whole-body surgical video analysis, underscoring both the current progress and the considerable gap remaining to clinical-level reasoning.

Our contributions are three-fold: 1) We introduce SurgCoT, the first surgical video reasoning benchmark offering cross-specialty procedural coverage and a five-dimensional evaluation framework for fine-grained assessment of spatiotemporal understanding and clinical decision-making aligned with surgical workflows. 2) We propose a novel three-stage reasoning framework with a five-tuple annotation protocol, combining hierarchical video-to-frame localization with synergistic `Knowledge`/`Clue` annotations to enable structured CoT reasoning. 3) Our evaluation of 12 leading MLLMs reveals that significant gaps persist in surgical CoT reasoning, while SurgCoT effectively enables progressive spatiotemporal reasoning evaluation.

2. Related Work

Multi-Modal Large Language Models. MLLMs have achieved notable performance on complex vision-language tasks through large-scale pretraining and instruction tuning. Initial joint-encoder models such as BLIP [25, 26] and Flamingo [1] connected visual encoders with cross-attention layers for multimodal fusion. Subsequent autoregressive MLLMs like LLaVA [28] tokenize images as visual prompts for LLM conditioning, showing strong few-shot generalization. Recent models such as Qwen-VL [46] and InternVL [8] scale visual backbones and alignment methods, approaching the performance of proprietary systems like GPT-4o [19]. Video-language models such as Video-LLaVA [27] further improve temporal modeling, offering a suitable substrate for surgical video analysis.

Table 1. Comparison of surgical benchmarks. Our SurgCoT uniquely spans 7 surgical specialties with multi-level annotations (video/-clip/frame), supporting hierarchical spatiotemporal reasoning with localization supervision and clinician-derived reference standards.

Benchmark	Domain	#Spe.	#Pro.	Scale	Unit	ST	Pro.	Loc.	Clin.
Surgical-VQA [36](MICCAI'21)	Endoscope	∼	3	11.8K	Frame	✗	✗	✓	✓
EndoVis-VQLA [3](ICRA'23)	Endoscope	∼	1	9.5K	Frame	✗	✗	✓	✓
Endo-FM [48](MICCAI'23)	Endoscope	∼	2	33K+50K	Clip/Frame	✗	✗	✓	✓
CoPESD [42](arXiv'24)	Endoscope	∼	1	17.7K	Frame	✗	✓	✓	✓
SurVLP [54](MedIA'25)	Endoscope	∼	11	25.5K	Frame	✗	✗	✗	✗
Surg-3M [4](arXiv'25)	Endoscope	∼	35	30K	Frame	✗	✗	✗	✗
Surg-396K [43](MIA'26)	Endoscope	3	3	396K	Frame	✗	✗	✓	✓
Cholec80-VQA [36](MICCAI'22)	Laparoscopic	1	1	43K	Frame	✗	✗	✗	✓
GSViT [35](arXiv'24)	Laparoscopic	1	28	7000K	Frame	✗	✗	✗	✗
GenSurg+ [15](arXiv'24)	Laparoscopic	2	28	17K	Clip	✗	✗	✗	✗
MedVidQA [12](BioNLP@ACL'24)	Instructional	∼	∼	1.2K	Clip	✗	✗	✗	✓
Surg-QA [24](arXiv'24)	Instructional	∼	∼	10.2K	Video	✗	✗	✗	✗
M ^{3}-Med [29](arXiv'25)	Instructional	∼	∼	3.7K	Clip	✗	✗	✗	✓
OphNet [16](ECCV'24)	Ophthalmic	1	∼	14K	Video	✗	✓	✓	✓
OphVL [17](arXiv'24)	Ophthalmic	1	∼	375K	Clip	✗	✗	✗	✓
EyePCR [45]((arXiv'25)	Ophthalmic	1	∼	210K	Clip/Frame	✗	✓	✗	✓
PSI-AVA-VQA [37](MICCAI'23)	Generalist	1	1	10.3K	Frame	✗	✗	✓	✓
SSG-VQA [55](IJCARS'24)	Generalist	1	1	960K	Frame	✗	✗	✓	✓
SurgVLM-Bench [56](arXiv'25)	Generalist	4	16	7798.4K	Frame	✗	✓	✓	✗
MedFrameQA [53](arXiv'25)	Generalist	9	∼	3.4K	Multi-Frame	✓	✗	✗	✓
SurgBench [49](arXiv'25)	Generalist	11	22	23K+5300K	Clip/Frame	✗	✗	✓	✓
SurgCoT(Ours)	**Generalist**	**7**	**35**	2.8K+19K+19K	Video/Clip/Frame	✓	✓	✓	✓

#Spe.: #Specialties; **#Pro.**:#Procedures; **ST**: Spatiotemporal modeling; **Pro.**: Progressive reasoning protocol; **Loc.**: Localization supervision (caption/bbox/mask/timestamps/tracks.); **Clin.**: Clinicians as a reference standard. ✓: yes, ✗: no, ∼: unknown.

In the medical domain, MLLM adaptation has progressed through specialized pretraining and instruction tuning, *e.g.*, LLaVA-Med [21] leverage biomedical corpora to enhance visual diagnostic capabilities, while HuatuoGPT-Vision [5] introduces large-scale medical VQA datasets for knowledge injection. Retrieval-augmented approaches such as MedDr [14] further strengthen factual grounding through external knowledge bases. Surgical-specific MLLMs including SurgicalGPT [37] and Surgical-LVLM [41] have demonstrated promising results in operative tasks like phase recognition and report generation. However, current evaluations remain confined to narrow tasks and short video segments, leaving fine-grained spatiotemporal reasoning largely unexamined [53, 59]. This limitation underscores the need for unified benchmarks capable of assessing expert-level reasoning in realistic surgical workflows.

Benchmarks for Surgical MLLMs. Table 1 summarizes the landscape of surgical benchmarks. Early works including SurgicalVQA and Cholec80-VQA [36] establish foundational protocols, while subsequent efforts like SSG-VQA [55] extend to complex laparoscopic scenarios with enhanced geometric and procedural analysis. Specialized benchmarks (*e.g.*, EndoVis-VQLA [3] and CoPESD [42]) focus on tool analysis and localization within narrow procedural contexts, but with limited generalization. To address single-specialty constraints, cross-specialty benchmarks have emerged: SurgVLM-Bench [56] covers 16 procedures across 4 specialties with hierarchical knowledge organization, and SurgBench [49] spans 22 procedures over 11 specialties. MedFrameQA [53] pioneers spatiotemporal modeling via cross-frame reasoning but lacks scale and fine-grained localization. SurgCoT bridges this gap by unifying spatiotemporal reasoning, hierarchical knowledge, and localization supervision in a clinically validated framework, establishing a comprehensive evaluation baseline for surgical MLLMs under CoT reasoning.

3. Construction of SurgCoT Benchmark

Fig. 2 outlines the benchmark construction pipeline with expert oversight at each stage: 1) **Data Processing** with multi-source video curation, hierarchical segmentation, and evidence mining; 2) **Three-Stage Progressive Reasoning** with five-tuple annotation protocol; 3) **VQA Generation** producing 78,522 QA pairs via structured task templates and ontology-driven design; 4) **Quality Control** via dual-pass validation and multi-criteria verification.

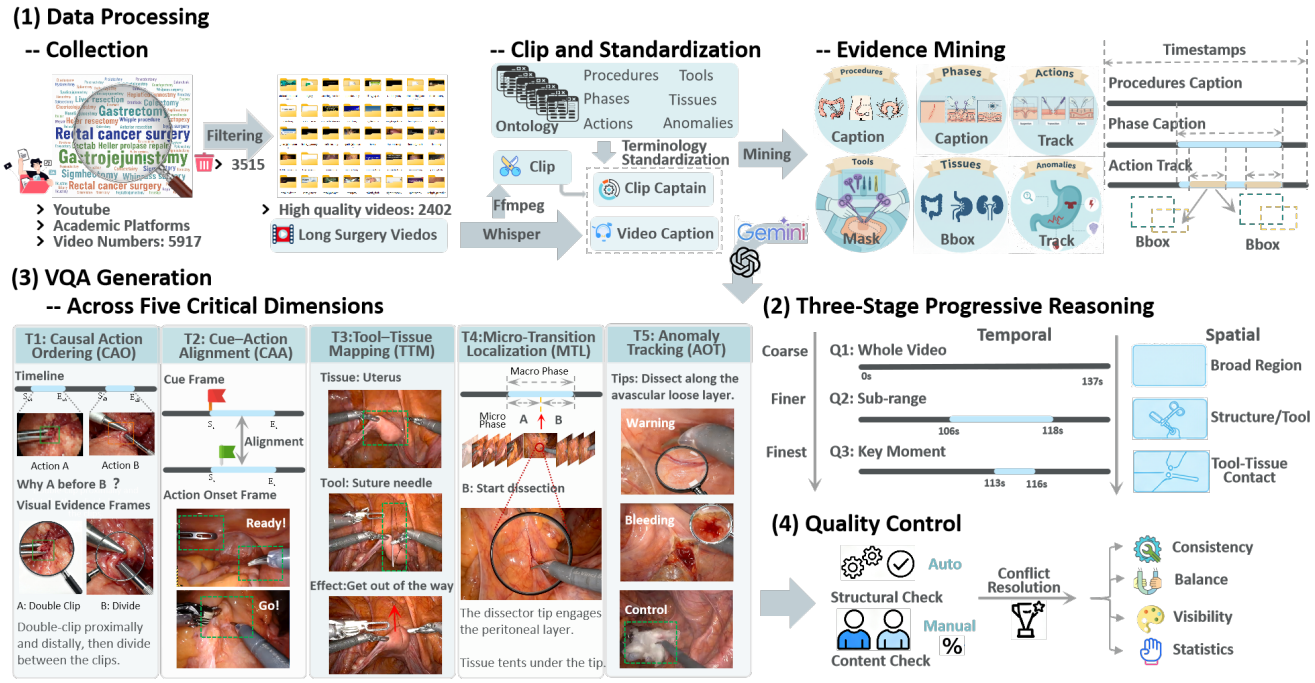

Figure 2. Construction pipeline of SurgCoT benchmark.

3.1. Data Processing

Data Collection. Our benchmark is constructed from a multi-source surgical video collection systematically organized across seven clinical specialties: Colorectal, Urological, Upper Gastrointestinal (Upper GI), Ocular, Gynecologic, General Surgery (GenSurgery) and Hepatobiliary-Pancreatic (HPB). The dataset integrates public platform content (*e.g.*, YouTube, ASVIDE), ten open-source repositories, and proprietary clinical archives, yielding **8,917** initial surgical cases. Following standardized filtration prioritizing procedural completeness, clinical validity, and bilingual narration for temporal alignment, **2,841** high-quality videos (31.9% of original corpus) were retained. All data were rigorously de-identified in compliance with ethical standards, with extended analyses available in Appendix.

Clip and Standardization. Surgical videos are processed into standardized clips to enable fine-grained spatiotemporal reasoning: 1) **Multi-source segmentation** uses hierarchical cue fusion (visual scenes, tool/tissue transitions, and ASR anchors) to create semantically coherent segments; 2) **ASR alignment** produces millisecond-precise, temporally consistent captions without altering medical content; 3) **Ontology-driven normalization** maps surface terms to canonical entities using a controlled vocabulary, ensuring terminological consistency while preserving original captions. This workflow guarantees temporally aligned and semantically standardized clips.

Evidence Mining. We retain public dataset annotations and introduce an end-to-end evidence mining paradigm to convert unannotated clips into spatiotemporally-grounded supervision units. 1) The pipeline utilizes ASR captions as semantic anchors, performs ontology-driven terminology normalization to create lexical-canonical mappings, and annotates `Procedures` and `Phases` by merging semantically consistent sentence windows. 2) **Spatial evidence** is constructed through frame-wise `Tissue` detection by YOLOv10 [40] and `Tool` segmentation by SAM2 [34], with cross-frame tracking via ByteTrack [57]. 3) **Temporal evidence** is extracted by detecting `Action` onsets from appearance-change indicators, using minimal visual cues as frame-level anchors. `Anomalies` are annotated with tracks specifying onset time and minimal ROIs. All evidence is bidirectionally aligned with ASR timestamps, with caption windows split at event boundaries while preserving original timing. The generated annotations integrate `Procedure`/`Phase` captions, `Tool` masks, `Tissue` bounding boxes, and `Action`/`Anomaly` tracks, enabling progressive spatiotemporal reasoning through three-stage window narrowing and region-to-interface grounding for comprehensive VQA evaluation.

3.2. Three-Stage Progressive Reasoning with Five-Tuple Annotation Protocol

To enable structured, verifiable CoT reasoning in surgical video understanding, we introduce a three-stage progressive reasoning framework (`Q1`→`Q2`→`Q3`)

with a comprehensive five-tuple annotation protocol (`Question`→`Option`→`Knowledge`→`Clue`→`Answer`).

Five-Tuple Annotation Protocol. For each stage, we annotate five fields to collectively enable explicit, verifiable reasoning. `Question`: A clinically meaningful query aligned with surgical workflow; `Option`: Curated, mutually exclusive candidates designed to disambiguate similar phenomena (*e.g.*, tool reflections vs. true bleeding) and constrain the hypothesis space; `Knowledge`: Provides domain priors (*e.g.*, color/flow patterns, typical anatomy, device behavior) that explain clinical plausibility independent of video content; `Clue`: Supplies video-grounded evidence (*e.g.*, temporal windows, spatial ROIs, landmarks) that directs attention to case-specific diagnostic anchors; `Answer`: the adjudicated target. Critically, `Answer` is carried forward as conditioning context for the next stage, enforcing causal dependency and spatiotemporal narrowing. Serializing `Knowledge` and `Clue` before `Answer` encourages a "*reason-then-decide*" pattern: `Knowledge` supplies clinical "*why*," while `Clue` anchors "*where/when*," yielding a transparent, video-grounded chain of thought.

Three Stages of Progressive Reasoning. The framework operationalizes clinical reasoning by decomposing complex spatiotemporal diagnostics into three hierarchically interdependent stages, progressively focusing subproblems while conditioning each stage on validated evidence from its predecessor. **Q1. Video-Level Comprehension**: identifies high-level clinical events (*e.g.*, "*active bleeding present?*") across the surgical video, establishing a global hypothesis space by determining whether a target phenomenon occurs. Its output constrains the reasoning trajectory and provides the primary conditioning signal for subsequent stages. **Q2. Video-Level Comprehension**: operates under Q1's validated output, performing spatiotemporal localization by determining when a target event (*e.g.*, *active bleeding*) first emerges within a bounded temporal window and where it occurs at region-of-interest granularity. This process prunes the hypothesis space from video-level to a focused spatiotemporal segment, providing temporal anchors and spatial regions-of-interest for the next stage. **Q3. Frame-/Patch-Level Localization**: strictly conditioned on Q2's spatiotemporal bounds, requiring precise, fine-grained localization of exact onset frames and anatomical sites (*e.g.*, suture hole vs. adjacent tissue) with pixel-/bbox-level precision. Its output enables detailed auditability and provides supervision for downstream tasks. The cascading framework ensures each reasoning step is grounded in validated prior evidence, enforcing logical coherence and traceability throughout the diagnostic workflow.

Progressive Conditioning Reasoning Chain. We establish a rigorous context-carry mechanism, which enforces explicit logical dependencies across three refined reasoning stages, with each stage building strictly upon its predecessor's validated constraints:

$$\underbrace{(\texttt{Q1},\ \texttt{O1},\ \texttt{K1},\ \texttt{C1})}_{\text{Global Video Comprehension}} \Rightarrow \texttt{A1}$$
$$\rightarrow \underbrace{(\texttt{Q2},\ \texttt{O2},\ \texttt{K2},\ \texttt{C2},\ \texttt{A1})}_{\text{Conditioned Clip Analysis}} \Rightarrow \texttt{A2}$$
$$\rightarrow \underbrace{(\texttt{Q3},\ \texttt{O3},\ \texttt{K3},\ \texttt{C3},\ \texttt{A2})}_{\text{Fine-grained Frame Localization}} \Rightarrow \texttt{A3}.$$

The chain operates through three critical dependency mechanisms: **1) Semantic Constraint Propagation**: Each stage's output forms an immutable foundation for the next. For instance, if `Q1` confirms "*active bleeding present*," `Q2` must proceed within this context without reconsideration, ensuring logical consistency and narrative coherence. **2) Spatiotemporal Scope Refinement**: The analysis scope narrows progressively. `Q1` processes the full video (minutes), `Q2` focuses on relevant clips (seconds-minutes), and `Q3` zooms into critical frames (sub-second) and anatomical regions (pixel/bbox level), mirroring the clinical workflow of identifying and then localizing abnormalities. **3) Evidence Accumulation and Validation**: `Knowledge` evolves from general anatomy (`K1`) to lesion-specific details (`K3`), while clues progress from temporal landmarks (`C1`) to spatial regions (`C2`) and pixel-level evidence (`C3`), reflecting increasing diagnostic certainty.

This design enforces a strict chain of dependencies for diagnostic decisions. For instance, a surgical recommendation (`A3`) must be grounded in a confirmed lesion location (`A2`), which itself requires established pathology (`A1`). This creates an auditable trail that mirrors clinical reasoning, ensuring treatment is justified by diagnostic evidence. The process guarantees progressively decreasing uncertainty and evaluates both diagnostic accuracy and adherence to the required logical progression.

3.3. VQA Generation: Five Critical Dimensions

With clinical experts, we define five spatiotemporal reasoning tasks to investigate deep cognitive processes. Each task uses the three-stage framework (Sec. 3.2) to generate ontology-driven VQA pairs containing spatial, temporal, and semantic distractors. This yields the SurgCoT benchmark, comprising 19,345 main and 59,177 sub-questions, for evaluating hierarchical reasoning in surgical videos.

Causal Action Ordering (CAO). This task determines the causal sequence of surgical micro-actions by using minimal visual evidence and fine-grained spatiotemporal cues (*e.g.*, action onset, interface-level ROIs) to enable verifiable rationale inference. Example: ***Q**: In cholecystectomy, which occurs first: clip placement or scissor division?* ***A**: Clip placement precedes and enables subsequent division.*

Cue-Action Alignment (CAA). CAA aligns pre-action cues with micro-actions to pinpoint action initiation in time

and space, thus disambiguating intent from execution and providing verifiable onset anchors, a critical capability for surgical video reasoning. Example: ***Q**: When does visible execution first begin?* ***A**: Around $t = 425$s, initiating at the hook–peritoneum interface over Calot's triangle.*

Affordance Mapping (AM). AM formalizes tool–tissue interactions by grounding them with spatiotemporal evidence and modeling their relationships, enabling progressive reasoning from coarse recognition to fine-grained localization. Example: ***Q**: Identify the first affordance event that improves exposure.* ***A**: uterine suspension at the anterior serosal needle entry (∼485s).*

Micro-Transition Localization (MTL). MTL identifies frame-level boundaries between micro-phases using minimal visual evidence, providing verifiable spatiotemporal evidence of transitions to enable fine-grained temporal reasoning. Example: ***Q**: At what moment does the micro-phase shift from positioning to dissection?* ***A**: At ≈312s, when the hook–peritoneum interface over Calot's triangle first incises and tents the peritoneal layer.*

Anomaly Onset Tracking (AOT). AOT identifies the onset and early trajectory of surgical anomalies, providing precise spatiotemporal localization to enable evidence-based detection and assessment of workflow deviations. Example: ***Q**: Can you precisely localize the anomaly onset in this clip?* ***A**: At 647.0s, Spatial: [540, 420, 80, 60], where bleeding first appears at the cystic artery stump before tracking its spread until hemostasis.*

These five dimensions form a cohesive framework for surgical reasoning assessment: **CAO** establishes causal logic; **CAA** anchors intent to execution; **AM** grounds interactions physically; **MLT** tracks micro-scale progressions; and **AOT** monitors pathway deviations. Together, they enable holistic evaluation from normative workflow reasoning (**CAO**, **CAA**, **AM**, **MLT**) to abnormal scenario handling (**AOT**) within a unified paradigm.

3.4. Quality Control Protocol

The protocol integrates *expert validation* with *multi-criteria verification* via a dual-pass human-in-the-loop process. Automated structural validation ensures schema integrity and logical consistency (*e.g.*, temporal/spatial progression, no answer leakage), while dual-annotator content validation with expert adjudication verifies task dependency, rationale verifiability, and ontology compliance. The protocol addresses four quality dimensions: *Consistency* (schema/logic checks), *Balance* (class-wise quotas), *Visibility* (obstruction tagging for difficulty filtering), and *Statistical Reliability* (inter-rater agreement metrics and acceptance thresholds). Disagreements are resolved by majority voting, with inconsistent samples discarded to ensure robust annotation.

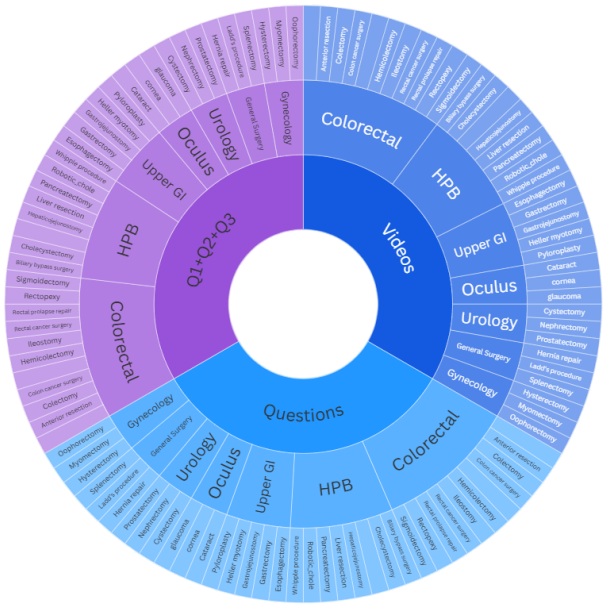

Figure 3. Statistics of SurgCoT: 2,841 videos, 19,345 questions, and 59,177 sub-questions across 35 procedures and 7 specialties.

3.5. Dataset Statistics

SurgCoT contains 2,841 surgical videos across 35 procedures in 7 specialties (Fig. 3), covering abdominal, pelvic, and ophthalmic surgeries. The benchmark includes 19,345 main questions and 59,177 sub-questions, averaging 7 main and 21 sub-questions per video. Question distribution aligns with specialty coverage, ensuring balanced supervision density while maintaining a long-tail of rare procedures to evaluate both common and uncommon workflows.

4. Experimental Results

4.1. Experimental Setup

Evaluated Models. We evaluate 10 MLLMs: **Commercial models** include GPT-5 [33], Claude-Sonnet-4.5 [2], and Gemini-2.5-Pro [10], serving as leading multimodal baselines with strong performance on diverse vision-language tasks; **Open-source models** include Qwen2.5-VL-7B, Qwen3-VL-8B [50] and Intern3.5-VL-8B [47]; **Med-specialized models** include MedGemma-27B-IT [38], Lingshu-7B [51], LLaVA-Med-7B [22], and HuatuoGPT-Vision-7B [6]. We use accuracy as the primary evaluation metric.

Evaluation Protocol. We design a three-stage protocol to systematically assess MLLMs' CoT reasoning capabilities: **1) Baseline: (BL)**: Models receive only the surgical video and main question, evaluating end-to-end reasoning without guidance. **2) Knowledge-Enhanced (KE)**: Adds clinical background knowledge; models answer both main

Table 2. Evaluation of 10 MLLMs across five clinical reasoning tasks under progressive settings (**BL→KE→FC**) shows: 1) commercial models outperform open-source and medical-specialized counterparts, and 2) the five-tuple annotation protocol improves reasoning accuracy under both (**KE**) and (**FC**) settings. Best results in **bold**, second-best underlined.

	Model	CAO. (%) ↑			CAA. (%) ↑			AM. (%) ↑			MTL.(%) ↑			AOT. (%) ↑			Avg. (%) ↑		
		BL	KE	FC	BL	KE	FC	BL	KE	FC	BL	KE	FC	BL	KE	FC	BL	KE	FC
Com.	GPT-5 [33]	85.07	91.56	94.79	80.99	85.29	90.64	**78.97**	**77.67**	82.34	69.14	81.63	86.71	68.95	66.53	78.23	**76.62**	80.54	**87.58**
	Gemini-2.5-Pro [10]	89.66	93.99	96.92	77.97	76.75	84.62	60.31	73.20	86.55	69.82	82.73	84.05	52.32	76.48	80.85	70.02	**81.83**	87.20
	claude-sonnect-4.5 [2]	**92.81**	**97.99**	98.33	**83.66**	86.46	89.12	59.80	68.10	**87.51**	**79.49**	72.45	86.39	54.74	67.35	75.34	74.10	78.87	87.54
Medical	MedGemma-27B-IT [38]	83.67	94.99	96.99	74.02	81.28	86.49	73.39	70.83	79.26	72.69	75.54	85.60	51.02	59.27	80.52	70.96	76.37	86.37
	Lingshu-7B [51]	86.68	97.24	97.50	65.83	64.57	72.27	49.06	53.98	63.50	65.12	68.99	73.84	50.46	62.46	67.64	63.43	69.45	74.95
	LLaVA-Med-7B [22]	89.95	96.93	97.50	66.10	72.34	71.52	62.29	74.68	76.16	71.43	81.84	**88.67**	50.97	47.25	74.81	68.15	75.22	81.73
	HuatuoGPT-Vision-7B [6]	78.47	80.53	88.74	68.28	65.91	73.24	67.24	68.63	81.37	72.53	72.69	87.37	63.89	68.21	77.95	70.08	71.19	83.73
Open.	InternVL-8B [8]	90.72	94.17	97.53	70.08	73.88	82.75	59.12	59.45	68.46	68.35	69.52	84.85	51.46	65.90	76.03	67.95	73.58	82.32
	Qwen2.5-VL-7B [46]	84.42	87.17	**98.98**	65.64	68.77	83.56	60.95	66.78	68.41	76.59	**83.40**	85.02	56.64	49.97	61.30	68.85	71.22	79.45
	Qwen3-VL-8B [50]	87.96	92.69	96.77	75.51	**87.42**	**91.57**	60.67	66.75	73.30	74.07	76.10	83.30	**79.00**	**77.46**	**86.68**	75.44	81.48	86.92

and sub-questions (`Q1`, `Q2`, `Q3`), evaluate if `Knowledge` aids progressive reasoning. **3) Full-Context (FC)**: Provides full annotation support (Video, `Knowledge` and `Clue`) evaluating the comprehensive diagnostic capability with complete reasoning scaffolds.

Implementation Details. We adopt unified prompts and decoding settings: a zero-shot template with fixed decoding parameters, including temperature $= 0.0$, top_p $= 1.0$, max_new_tokens $= 4096$, and repetition_penalty $= 1.0$; sampling retries are disabled. API-hosted models are invoked with these parameters, while local open-source models are run with `Torch 2.9.0` and `Transformers 4.57.1` on `CUDA 12.4`, using `bf16` inference on NVIDIA A100 80 GB $\times$ 8 GPUs.

4.2. Evaluation of MLLMs

Table 2 summarizes the evaluation of 10 leading MLLMs, revealing four key observations.

Commercial Models Lead Peers. As demonstrated in Table 2, commercial models consistently achieve superior performance compared to both open-source and medically-specialized counterparts across all five clinical reasoning dimensions. Leading commercial systems including GPT-5 [33], claude-sonnect-4.5 [2], and Gemini-2.5-Pro [10] demonstrate particularly strong capabilities in progressive reasoning scenarios (**BL** $\rightarrow$ **KE** $\rightarrow$ **FC**), *i.e.*, there is a stable performance progression as additional knowledge elements (**KE**) and spatiotemporal constraints (**FC**) are introduced. GPT-5 [33], for instance, shows consistent performance gains across all task categories (on average +10.96%), while Claude-Sonnet-4.5 [2] demonstrates notable improvements in fine-grained decision-making tasks such as AOT (+20.60%) and AM (+27.71%). Conversely, open-source and medically specialized models display significant limitations in tasks requiring precise temporal and spatial reasoning, particularly struggling with the integration of multi-modal information sources.

Five-Tuple Annotation Protocol Boosts Accuracy. Medical-specific models such as LLaVA-Med-7B [22] exhibit substantial performance gains when transitioning from **BL** to **KE**, with an average accuracy improvement of nearly 7% across the five reasoning dimensions. This pronounced improvement indicates that explicit knowledge augmentation effectively compensates for domain-specific limitations. In contrast, commercial models like GPT-5 [33] show more moderate improvements (nearly 4%), suggesting their stronger inherent language reasoning capabilities allow for more seamless knowledge integration.

Table 3. Sub-question accuracy under progressive reasoning settings (**BL→KE→FC**). Evaluation reveals: 1) MLLMs show chain-of-thought reasoning gaps, with performance dropping at intermediate steps; 2) SurgCoT's structured framework supports progressive reasoning, yielding gains under **KE/FC** settings. Best results in **bold**, second-best underlined.

	Model	BL	KE				FC			
		Q	Q	Q1	Q2	Q3	Q	Q1	Q2	Q3
Com.	GPT-5 [33]	**76.62**	80.54	54.71	55.85	47.60	**87.58**	56.12	56.15	50.51
	Gemini-2.5-Pro [10]	70.02	**81.83**	34.15	44.04	43.06	87.20	35.85	34.02	52.32
	Claude-Sonnect-4.5 [2]	74.10	78.87	49.38	41.42	36.50	87.54	53.04	34.85	30.44
Medical	MedGemma-27B-IT [38]	70.96	76.37	54.82	56.75	**76.25**	86.37	61.29	**71.25**	70.42
	Lingshu-7B [51]	63.43	69.45	36.06	47.93	43.53	74.95	48.36	58.58	56.52
	LLaVA-Med-7B [22]	68.15	75.22	31.71	36.65	36.46	81.73	34.40	39.75	37.57
	HuatuoGPT-Vision-7B [6]	70.08	71.19	69.97	**67.69**	55.62	83.73	72.13	70.31	69.00
Open	Qwen3-VL-8B [50]	75.44	81.48	**72.90**	51.00	73.04	86.92	**74.75**	67.49	**70.98**
	Qwen2.5-VL-7B [46]	68.85	71.22	31.85	27.02	25.33	79.45	38.14	29.87	36.80
	InternVL-8B [8]	67.95	73.58	67.80	65.79	48.86	82.32	64.89	55.28	51.63

The transition from **KE** to **FC** further demonstrates the critical importance of spatiotemporal grounding, *e.g.*, Qwen2.5-VL-7B [50] showed a notable 8.23% increase, highlighting its enhanced ability to utilize temporal and spatial clues for refined reasoning. Similarly, Claude-Sonnet-4.5 [2], despite strong baseline performance (BL = 74.10%), improved by approximately 13.44% with spatiotemporal clues, underscoring its advanced capacity for multimodal fusion. These observations collectively validate that the five-tuple annotation framework enables progressive spatiotemporal reasoning through structured cue enhancement across three reasoning settings: **BL**, **KE**, and **FC**.

MLLMs Demonstrate Limitations in CoT Reasoning To systematically evaluate CoT's reasoning capabilities, we analyze MLLM's performance across the three progressive

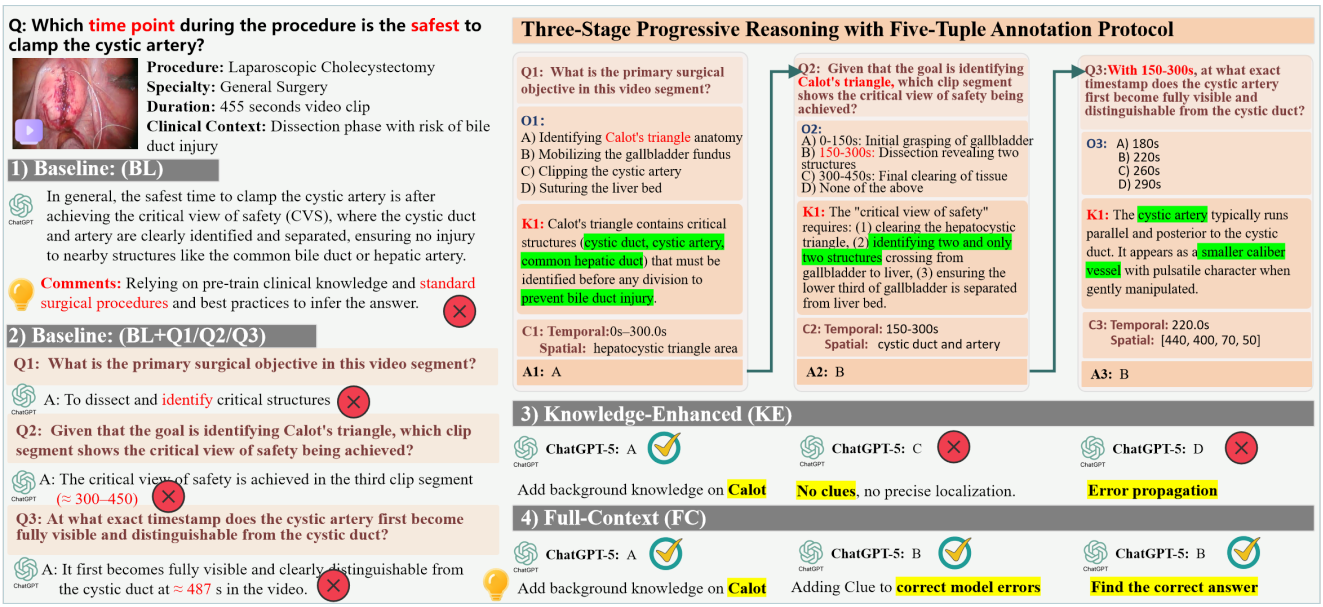

Figure 4. SurgCoT constructs a diagnostic chain-of-thought by progressively decomposing a flawed baseline (**BL**) into clinical sub-questions, correcting semantics with knowledge (**KE**), and refining evidence with spatiotemporal clues (**FC**).

reasoning stages (Q1, Q2, Q3) that form the foundational substructure of each main question (Q). As quantified in Table 3, a pronounced performance dissociation emerges between main question accuracy and sub-question proficiency, revealing fundamental limitations in progressive reasoning.

Notably, GPT-5 [33] achieves 76.62% accuracy on main questions (Q) but exhibits a substantial performance degradation in sub-questions, dropping to 47.60% in Q3. This pattern is consistently observed across other leading commercial models, including Gemini-2.5-Pro [10] and Claude-Sonnet-4.5 [2]. Open-source models such as Qwen2.5-VL-7B [50] demonstrate even more pronounced deficiencies, particularly in the initial (Q1) and final (Q3) sub-questions, underscoring their limited capacity for coherent multi-step inference. ***These observations collectively highlight a critical gap in current MLLMs' ability to sustain logically consistent reasoning chains***, a capability essential for reliable clinical decision-making where each diagnostic step must follow rigorously from preceding evidence.

SurgCoT Enables CoT Reasoning SurgCoT facilitates surgical CoT reasoning by decomposing tasks into clinical sub-questions (Q1→Q2→Q3) and progressively integrating knowledge (**KE**) and spatiotemporal clues (**FC**) to guide progressive, transparent inference. As shown in Table 3, the introduction of Knowledge Enhancement (**KE**) yields substantial improvements in sub-question performance across diverse model architectures. For instance, LLaVA-Med-7B [22] demonstrates a notable accuracy increase in main Q from 68.15% to 75.22% when transitioning from **BL** to **KE**. The subsequent integration of spatiotemporal clues (**FC**) further elevates performance, particularly in stages requiring fine-grained temporal and spatial analysis (Q2, Q3), where models must ground their reasoning in concrete visual evidence.

Fig. 4 illustrates a typical case when `Knowledge` and `Clue` are gradually introduced, the model refines its reasoning process, correcting initial mistakes and arriving at the correct answer. In Fig. 4(1), the model's response is based solely on clinical knowledge from pretraining, without effectively integrating the video content. Fig. 4(2) expands the model's reasoning process, but significant errors and deviations remain. Fig. 4(3) shows that after adding `Knowledge`, the model adjusts its reasoning to produce a more accurate answer. Finally, Fig. 4(4) shows that added `Clues` further improve performance, highlighting the benefits of SurgCoT for enhancing progressive reasoning.

5. Conclusion

SurgCoT introduces a unified benchmark for evaluating surgical reasoning in MLLMs via a chain-of-thought framework, combining multi-specialty coverage with fine-grained reasoning tasks through a three-stage, five-tuple annotation protocol. Evaluation of 12 leading MLLMs shows commercial models lead overall but all exhibit significant gaps in progressive reasoning, *i.e.*, often answering correctly while failing intermediate reasoning steps. SurgCoT not only exposes these limitations but also provides a scaffold for improvement through structured reasoning and spatiotemporal grounding, establishing a foundation for clinically-validated MLLMs.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 2

[2] Anthropic. Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5, 2025. Accessed: 2025-11-09. 2, 6, 7, 8

[3] Long Bai, Mobarakol Islam, Lalithkumar Seenivasan, and Hongliang Ren. Surgical-vqla: Transformer with gated vision-language embedding for visual question localized-answering in robotic surgery. *arXiv preprint arXiv:2305.11692*, 2023. 3

[4] Chengan Che, Chao Wang, Tom Vercauteren, Sophia Tsoka, and Luis C Garcia-Peraza-Herrera. Surg-3m: A dataset and foundation model for perception in surgical settings. *arXiv preprint arXiv:2503.19740*, 2025. 3

[5] 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. 2, 3

[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. 6, 7

[7] Zhen Chen, Xingjian Luo, Jinlin Wu, Danny Chan, Zhen Lei, Jinqiao Wang, Sebastien Ourselin, and Hongbin Liu. Vs-assistant: versatile surgery assistant on the demand of surgeons. *arXiv preprint arXiv:2405.08272*, 2024. 2

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23904–23915, 2024. 2, 7

[9] Zhen Chen, Xingjian Luo, Kun Yuan, Jinlin Wu, Danny Chan, Nassir Navab, Hongbin Liu, Zhen Lei, and Jiebo Luo. Surgllm: A versatile large multimodal model with spatial focus and temporal awareness for surgical video understanding. *arXiv preprint arXiv:2509.00357*, 2025. 2

[10] Google DeepMind. Gemini 1.5: Scaling up multimodal reasoning, 2024. https://deepmind.google/technologies/gemini/. 2, 6, 7, 8

[11] Xuan Gong, Balu Harshavardan Koduru, Yuanhao Zhai, Shun Liu, Nan Xi, Xi Tang, Yuan Zhang, Tenzin Lhakpa, Yunjie Tian, Yuxuan Sun, et al. Endoassistant: A large-scale vision-language dataset for endoscopic surgery understanding from open-source videos. 2

[12] Deepak Gupta and Dina Demner-Fushman. Overview of the medvidqa 2022 shared task on medical video question-answering. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 264–274, 2022. 3

[13] Pengfei Hao, Shuaibo Li, Hongqiu Wang, Zhizhuo Kou, Junhang Zhang, Guang Yang, and Lei Zhu. Surgery-r1: Advancing surgical-vqla with reasoning multimodal large language model via reinforcement learning. *arXiv preprint arXiv:2506.19469*, 2025. 2

[14] Sunan He, Yuxiang Nie, Zhixuan Chen, Zhiyuan Cai, Hongmei Wang, Shu Yang, and Hao Chen. Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. *CoRR*, 2024. 3

[15] Mohammadmahdi Honarmand, Muhammad Abdullah Jamal, and Omid Mohareri. Vidlpro: A video-language pretraining framework for robotic and laparoscopic surgery. *arXiv preprint arXiv:2409.04732*, 2024. 3

[16] 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*, pages 481–500. Springer, 2024. 3

[17] 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. 3

[18] Wenguo Huang, Jin Yi, Ziyao Liu, and Wenjing Zhou. Mllms for versatile scene understanding: Towards embodied intelligent surgical robots. In *2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA)*, pages 1300–1303. IEEE, 2025. 2

[19] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2

[20] Kyle Lam, Junhong Chen, Zeyu Wang, Fahad M Iqbal, Ara Darzi, Benny Lo, Sanjay Purkayastha, and James M Kinross. Machine learning for technical skill assessment in surgery: a systematic review. *NPJ Digital Medicine*, 5(1):24, 2022. 2

[21] 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. 3

[22] 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, 8

[23] 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. 2

[24] Jiajie Li, Garrett Skinner, Brian R Quaranto, Gene Yang, Steven D Schwaitzberg, Peter CW Kim, and Jinjun Xiong. Llava-surg: Towards multimodal surgical assistant via structured lecture learning. 2, 3

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023. 2

[27] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 2

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

[29] Shenxi Liu, Kan Li, Mingyang Zhao, Yuhang Tian, Bin Li, Shoujun Zhou, Hongliang Li, and Fuxia Yang. M 3-med: A benchmark for multi-lingual, multi-modal, and multi-hop reasoning in medical instructional video understanding. *arXiv preprint arXiv:2507.04289*, 2025. 3

[30] Yang Liu, Maxence Boels, Luis C Garcia-Peraza-Herrera, Tom Vercauteren, Prokar Dasgupta, Alejandro Granados, and Sebastien Ourselin. Lovit: Long video transformer for surgical phase recognition. *Medical Image Analysis*, 99: 103366, 2025. 2

[31] LLaVA Team. Llava-next: A strong zero-shot video understanding model, 2024. Accessed: 2025-10-13. 2

[32] Lena Maier-Hein, Matthias Eisenmann, Duygu Sarikaya, Keno März, Toby Collins, Anand Malpani, Johannes Fallert, Hubertus Feussner, Stamatia Giannarou, Pietro Mascagni, et al. Surgical data science–from concepts toward clinical translation. *Medical Image Analysis*, 76:102306, 2022. 2

[33] OpenAI. Gpt-5: Multimodal large language models for understanding complex visual inputs. *OpenAI Technical Report*, 2025. 2, 6, 7, 8

[34] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 4

[35] Samuel Schmidgall, Ji Woong Kim, Jeffrey Jopling, and Axel Krieger. General surgery vision transformer: A video pre-trained foundation model for general surgery. *arXiv preprint arXiv:2403.05949*, 2024. 3

[36] Lalithkumar Seenivasan, Mobarakol Islam, Adithya K Krishna, and Hongliang Ren. Surgical-vqa: Visual question answering in surgical scenes using transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 33–43. Springer, 2022. 3

[37] Lalithkumar Seenivasan, Mobarakol Islam, Gokul Kannan, and Hongliang Ren. Surgicalgpt: end-to-end language-vision gpt for visual question answering in surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 281–290. Springer, 2023. 2, 3

[38] 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. 2, 6, 7

[39] Gabriel Szydlo Shein, Elisa Bannone, Silvia Seidlitz, Mohamed Hassouna, Luca Baratelli, Arturo Pardo, Sylvain Lecler, Frédéric Triponez, Manish Chand, Sylvain Gioux, et al. Surgical optomics: a new science towards surgical precision. *NPJ Gut and Liver*, 2(1):1–14, 2025. 2

[40] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, et al. Yolov10: Real-time end-to-end object detection. *Advances in Neural Information Processing Systems*, 37:107984–108011, 2024. 4

[41] Guankun Wang, Long Bai, Wan Jun Nah, Jie Wang, Zhaoxi Zhang, Zhen Chen, Jinlin Wu, Mobarakol Islam, Hongbin Liu, and Hongliang Ren. Surgical-lvlm: Learning to adapt large vision-language model for grounded visual question answering in robotic surgery. *arXiv preprint arXiv:2405.10948*, 2024. 3

[42] Guankun Wang, Han Xiao, Huxin Gao, Renrui Zhang, Long Bai, Xiaoxiao Yang, Zhen Li, Hongsheng Li, and Hongliang Ren. Copesd: A multi-level surgical motion dataset for training large vision-language models to co-pilot endoscopic submucosal dissection. *arXiv preprint arXiv:2410.07540*, 2024. 3

[43] Guankun Wang, Long Bai, Junyi Wang, Kun Yuan, Zhen Li, Tianxu Jiang, Xiting He, Jinlin Wu, Zhen Chen, Zhen Lei, et al. Endochat: Grounded multimodal large language model for endoscopic surgery. *arXiv preprint arXiv:2501.11347*, 2025. 3

[44] Guankun Wang, Long Bai, Junyi Wang, Kun Yuan, Zhen Li, Tianxu Jiang, Xiting He, Jinlin Wu, Zhen Chen, Zhen Lei, et al. Endochat: Grounded multimodal large language model for endoscopic surgery. *arXiv preprint arXiv:2501.11347*, 2025. 2

[45] Gui Wang, Yang Wennuo, Xusen Ma, Zehao Zhong, Zhuoru Wu, Ende Wu, Rong Qu, Wooi Ping Cheah, Jianfeng Ren, and Linlin Shen. Eyepcr: A comprehensive benchmark for fine-grained perception, knowledge comprehension and clinical reasoning in ophthalmic surgery. *arXiv preprint arXiv:2509.15596*, 2025. 2, 3

[46] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 7

[47] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 6

[48] Zhao Wang, Chang Liu, Shaoting Zhang, and Qi Dou. Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 101–111. Springer, 2023. 3

[49] Jianhui Wei, Zikai Xiao, Danyu Sun, Luqi Gong, Zongxin Yang, Zuozhu Liu, and Jian Wu. Surgbench: A unified large-scale benchmark for surgical video analysis. *arXiv preprint arXiv:2506.07603*, 2025. 3

[50] 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. 2, 6, 7, 8

[51] 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. 2, 6, 7

[52] Michael Yip, Septimiu Salcudean, Ken Goldberg, Kaspar Althoefer, Arianna Menciassi, Justin D Opfermann, Axel Krieger, Krithika Swaminathan, Conor J Walsh, He Huang, et al. Artificial intelligence meets medical robotics. *Science*, 381(6654):141–146, 2023. 2

[53] Suhao Yu, Haojin Wang, Juncheng Wu, Cihang Xie, and Yuyin Zhou. Medframeqa: A multi-image medical vqa benchmark for clinical reasoning. *arXiv preprint arXiv:2505.16964*, 2025. 2, 3

[54] 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

[55] 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

[56] Zhitao Zeng, Zhu Zhuo, Xiaojun Jia, Erli Zhang, Junde Wu, Jiaan Zhang, Yuxuan Wang, Chang Han Low, Jian Jiang, Zilong Zheng, et al. Surgvlm: A large vision-language model and systematic evaluation benchmark for surgical intelligence. *arXiv preprint arXiv:2506.02555*, 2025. 3

[57] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, pages 1–21. Springer, 2022. 4

[58] Yiyi Zhang, Yuchen Yuan, Ying Zheng, Jialun Pei, Jinpeng Li, Zheng Li, and Pheng-Ann Heng. Mejo: Mllm-engaged surgical triplet recognition via inter-and intra-task joint optimization. *arXiv preprint arXiv:2509.12893*, 2025. 2

[59] Junze Zhu, Dasen Gu, Mengwei Sha, Yang Peng, Haotian Yang, and Bin Li. Sfn-esvqa: Spatio-temporal frame-aware network for endoscopic surgery visual question answering. 2025. 2, 3