UniSurgBench: Benchmarking Surgical Video Understanding with Step-wise Reasoning and Workflow Aligned Evaluation

Zhenxin Lei[1,2], **Mengyang Li**[3], **Qian Li**[1], **Haozhe Cui**[3], **Zhiming Zhao** [3], **Hu Han**[1], **Shiguang Shan**[1], **Xilin Chen**[1]
[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]The First Medical Center, Chinese People's Liberation Army General Hospital

Abstract

Multimodal large language models have made substantial progress in medical visual understanding, but temporal reasoning and knowledge-guided reasoning over real surgical videos remain underexplored. Existing surgical benchmarks mainly focus on static frames, single procedures, or simple questions, which limits their ability to capture the complexity of real clinical workflows. To address this gap, we introduce **UniSurgBench**, a unified benchmark for open-ended surgical video reasoning, comprising 2,302 question-answer samples collected from real surgical videos spanning 45 procedures and 29 anatomical structures. Each sample is annotated with temporally grounded evidence and structured reasoning steps. We also propose **SurgExpEval**, a two-stage evaluation protocol that extracts structured reasoning components from free-form model outputs and jointly assesses answer accuracy, reasoning quality, and temporal grounding. Evaluations with 23 proprietary, open-source, and medical-related MLLMs show that MLLMs can produce clinically reasonable and causally coherent responses. However, they often fail to associate these responses with fine-grained visual evidence, highlighting evidence-level visual grounding as a key bottleneck for reliable surgical intelligence. Our code and data will be available at UniSurgBench.

1 Introduction

Multimodal large language models (MLLMs) have demonstrated rapid and wide-ranging advances in medical image analysis, achieving remarkable results in tasks such as medical report generation (Chen et al., 2025c; Baharoon et al., 2025; Zhang et al., 2025a), radiological diagnosis (Lee et al., 2025; Li et al., 2025; Pan et al., 2025; Jing et al., 2025), and pathology interpretation (Chen et al., 2025a; Liang et al., 2025; Sun et al., 2025; Yang et al., 2026). However, despite their strong performance on medical image analysis, the adoption of MLLMs in surgical scenarios for intraoperative assistance, risk prediction, and decision support remains limited (Zhu et al., 2026). This limitation primarily stems from the lack of comprehensive benchmarks and evaluation protocols that are well aligned with real-world clinical workflows. In contrast to static medical image understanding, surgical video understanding requires the integration of dynamic temporal information, fine-grained spatiotemporal perception, and high-level clinical reasoning (Wang et al., 2025a; Wei et al., 2025; Perez et al., 2026b; Li et al., 2026a). This complex reasoning paradigm in surgical scenario poses significant challenges for MLLMs in surgical video understanding.

Current benchmarks for surgical video understanding can be grouped into two primary categories: general-purpose medical benchmark (Royer et al., 2024; Hu et al., 2024b; Pengcheng et al., 2024; Zuo et al., 2025; Yue et al., 2024, 2025) and domain-specific surgical benchmark (Perez et al., 2026b; Qin et al., 2026; Cheng et al., 2026; Zhu et al., 2026; He et al., 2024; Hu et al., 2024a; Seenivasan et al., 2022; Zeng et al., 2025). General-purpose benchmarks cover a broad range of medical tasks but typically with limited surgical data per task, such as OmniMedVQA (Hu et al., 2024b), MultiMedEval (Royer et al., 2024), and MMMU (Yue et al., 2024). In contrast, domain-specific benchmarks focus on individual procedures or subspecialties, such as EndoBench (Liu et al., 2026a), Cholec80-VQA (Seenivasan et al., 2022), and GI-Bench (Zhu et al., 2026). Although these benchmarks have been useful in advancing surgical scene understanding to some degree, there are two critical limitations for real-world clinical applicability: i) they are largely confined to isolated procedures, subspecialties, or static images (He et al., 2024; Wang et al., 2025a; Seenivasan et al., 2022; Xu et al., 2025; Zeng et al., 2025), limiting the assessment of

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 35125–35161
October 24-29, 2026 ©2026 Association for Computational Linguistics

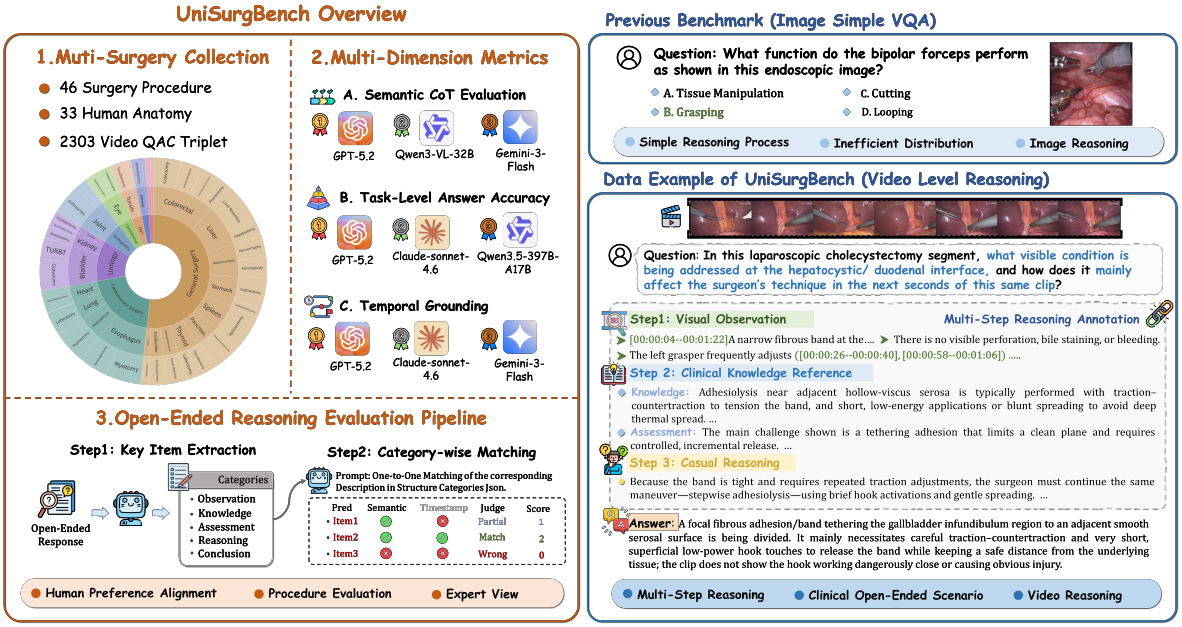


Figure 1: **Overview of UniSurgBench.** UniSurgBench is a unified testbed for open-ended surgical video reasoning, featuring 2,302 video question-answer-clue samples across 45 procedures and 29 anatomical structures. Each sample is organized as a question-answer-clue triplet with temporal evidence, structured rationales, and fine-grained visual clues. Together with SurgExpEval, UniSurgBench evaluates surgical MLLMs in answer correctness, reasoning quality, and temporal grounding.

model generalization across diverse anatomies, variable intraoperative workflows, and dynamic video scenarios; ii) existing benchmarks primarily rely on multiple-choice question (MCQ) accuracy (Liu et al., 2026a; Zhu et al., 2026). However, because of insufficiently rigorous distractor design and overly simplistic question settings, MLLMs may simply utilize language priors alone rather than using visual and textual evidence during reasoning, thereby undermining clinical trustworthiness.

To address these gaps, we introduce **UniSurgBench**, a unified open-ended benchmark for surgical video understanding (Fig. 1). UniSurgBench contains 2,302 video question-answer-clue (QAC) samples derived from real-world surgical videos, covering 29 anatomical structures and 45 surgical procedures. We construct the benchmark through a semi-automated pipeline consisting of structured video parsing, knowledge-guided question-answer generation, and human-in-the-loop quality control. Each sample is annotated with temporal evidence and structured reasoning steps. Inspired by clinical workflows conversation and surgeon expertise, we further defines 14 spatio-temporal reasoning tasks across four clinical dimensions, enabling systematic evaluation from low-level visual perception to high-level clinical reasoning.

We further propose **SurgExpEval**, a tailored evaluation framework designed to assess open-ended responses in UniSurgBench. Unlike prior methods that evaluate model responses solely based on semantic answer consistency (Zhu et al., 2026; Cheng et al., 2026), SurgExpEval first parses model outputs into structured reasoning steps and then applies category-aware semantic matching. This design enables us to disentangle the understanding ability of MLLMs and precisely determine their reasoning quality, answer correctness, and temporal grounding ability.

We evaluate 23 proprietary, open-source, and surgery-specialized MLLMs on UniSurgBench. The results show that current models are consistently stronger in knowledge-driven reasoning than in visually grounded perception. For example, GPT-5.4-Thinking achieves 74.6% F1 on causal reasoning but only 60.2% F1 on visual perception. Moreover, although many models can recognize basic surgical elements, they struggle with temporal understanding and specific scenario analysis tasks. These findings suggest that precise surgical perception remains a challenge for current MLLMs.

2 UniSurgBench

2.1 Task Definition

To construct a benchmark grounded in real clinical workflows, we conducted in-depth observations

Table 1: **Comparison with existing medical and surgical benchmarks. Scale** denotes the number of videos/images and QA pairs; **#Proc./Scope** denotes the surgery procedure or clinical/anatomical scope; **QA Fmt.** denotes the answer format, including MCQ, Open-End-Short/Open-End-Long, Classification, Sentence-level answer and Localization; **CoT Ann.** denotes multi-step reasoning annotations; **Temp. Clue** denotes temporal evidence annotations; **CoT Eval.** denotes explicit evaluation of reasoning steps; **Temp. Eval.** denotes explicit evaluation of temporal grounding.

						Annotation		Evaluation	
Benchmark	**Mod.**	**Domain**	**Scale**	**#Proc./Scope**	**QA Fmt.**	**CoT Ann.**	**Temp. Clue**	**CoT Eval.**	**Temp. Eval.**
Medical Video Benchmarks									
MedVidQA (Gupta et al., 2023)	Vid.	Medical education	899 / 3,010	-	Loc.	✗	✗	✗	✓
MedFrameQA (Yu et al., 2025)	Img.	Medical imaging	9,237 / 2,851	9 sys./43 org.	MCQ	✓	✗	✗	✗
M^3-Med (Liu et al., 2025)	Vid.	Medical education	3,748 / 12.7K	-	MCQ/Loc.	✗	✓	✗	✓
Surgical VQA Benchmarks									
EndoVis18-VQA (Seenivasan et al., 2022)	Img.	Robotic surgery	2K / 11.8K	1 proc. / 1 org.	Cls./Sent.	✗	✗	✗	✗
Cholec80-VQA (Seenivasan et al., 2022)	Img.	Laparoscopy	21.6K / 43.2K	1 proc. / 1 org.	Cls./Sent.	✗	✗	✗	✗
SSG-VQA (Yuan et al., 2024a)	Img.	Laparoscopy	25K / 960K	1 proc. / 1 org.	Cls.	✗	✗	✗	✗
PitVQA (He et al., 2024)	Img.	Pituitary surgery	109K / 884K	1 proc. / 1 org.	Cls.	✗	✗	✗	✗
EndoBench (Liu et al., 2026a)	Img.	Endoscopy	– / 6,832	4 scen. / -	MCQ	✗	✗	✗	✗
GI-Bench (Zhu et al., 2026)	Img.	GI endoscopy	947 / 4,735	2 proc. / 3 org.	MCQ/OE-S	✗	✗	✗	✗
Surgical Reasoning VQA Benchmarks									
CholeVidQA-32K (Li et al., 2026a)	Vid.	Cholecystectomy	3,855 / 32.3K	1 proc. / 1 org.	OE	✗	✗	✗	✗
SurgCoT (Wang et al., 2026)	Vid.	Multi-Surgery	2,841 / 19.3K	35 proc. / 7 org.	MCQ+SubQ	✓	✓	✓	✗
UniSurgBench	Vid.	General surgery	2,302 / 2,302	45 proc. / 29 org.	OE-L	✓	✓	✓	✓

of surgical procedures alongside three senior surgeons and systematically analyzed recurring information needs at each stage of surgical decision-making. This process revealed a clear cognitive hierarchy: surgeons first perceive and identify fundamental elements, then analyze interactions and dynamics, then engage in intent-driven planning, and finally perform high-level contextual reasoning. Motivated by this hierarchy, we define 14 spatio-temporal reasoning tasks across four levels:

Level 1: Surgical Element Identification. Short-term visual perception of fundamental surgical elements: instrument detection (ID), anatomy identification (AI), and anomaly recognition (AR).

Level 2: Surgical Operation Analysis. Spatiotemporal understanding of surgical interactions: action recognition (AcR), triplet extraction (TE), and interaction dynamics analysis (IDA).

Level 3: Surgical Phase Understanding Evidence-grounded reasoning for mid-term planning: phase recognition (PR), intent inference (II), safety assessment (SA), and next-step prediction (NSP).

Level 4: Surgical Procedure Reasoning High-level reasoning in surgical contexts: pathology inference (PI), anatomical variation inference (AVI), surgical summarization (SS), and difficulty-specific assessment (DSA).

All 14 tasks in UniSurgBench are formulated as open-ended question-answering-clue tasks, which can better reflect the free-form interaction that occurs in real-world scenes.

2.2 Data Collection

We curate professionally narrated surgical videos from WebSurg (WebSurg, 2026) and medical YouTube channels (YouTube, 2026). Videos are initially screened for visual quality, duration, narration density, procedural completeness, and temporal clarity, yielding 2,574 candidates across laparoscopic, open, and minimally invasive procedures. To ensure clinical relevance and safety, two senior surgeons independently review the candidates and exclude irrelevant, low quality, unsafe, or high-risk content, resulting in 254 full high-quality surgery videos. All videos are de-identified to satisfy ethical and privacy requirements, and their audio tracks are transcribed into timestamp-aligned text using WhisperX (Bain et al., 2023) .

2.3 Data Annotation Pipeline

2.3.1 Structured Surgical Video Parsing

We primarily parse each video into a structured multi-scale representation from time-aligned transcripts (Perez et al., 2026a). GPT-5 (OpenAI, 2025a) is prompted to decompose long surgical videos into three temporal levels, namely *Phase*, *Step*, and *Task* (Korndorffer Jr et al., 2010; Perez et al., 2026a). It also normalizes informal transcript expressions and generates a procedure-level summary as global context. To enable visually grounded QA generation, each segment is enriched with hierarchy-aware visual captions. For *Task*-level clips, uniformly sampled frames and transcript captions are used to generate detailed descriptions. For coarser *Step*- and *Phase*-level clips, child-clip captions and transcripts provide contextual priors

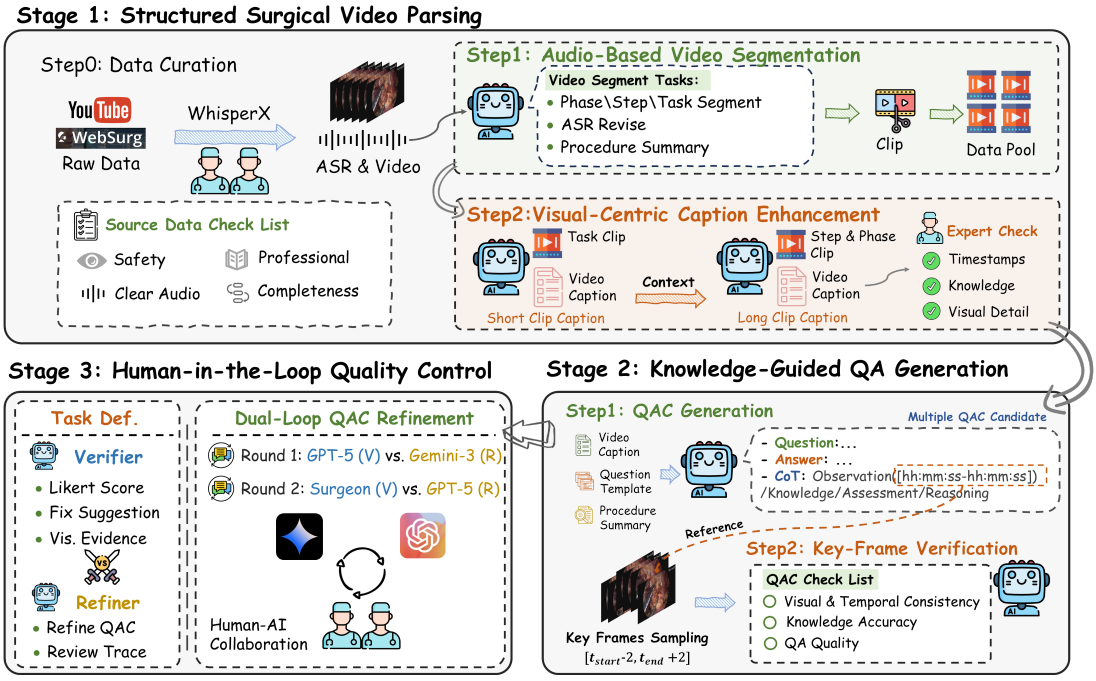


Figure 2: **Annotation pipeline overview.** We integrate expert-narrated clinical knowledge with visual cues from the videos to construct video descriptions and question-answer pairs. Through a human-in-the-loop process with multiple rounds of validation and discussion, we select high-quality samples with appropriate difficulty and clear clues to build UniSurgBench.

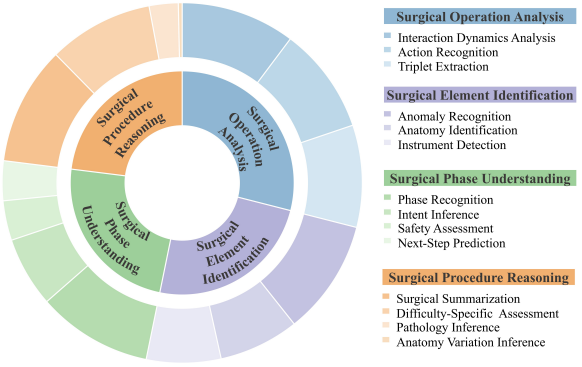


Figure 3: Task distribution of UniSurgBench across 14 spatio-temporal tasks.

Table 2: Dataset Annotation Statistics for UniSurgBench

Metric	Value
Number of videos	2,302
Maximum video duration (s)	1,621.0
Average video duration (s)	95.0
Average words per question	39.4
Average words per answer	51.4
Average words per reasoning step	224.7
Number of clues	12,768
Average clues per QAC	6.0
Average clue duration (s)	19.19

for caption generation. This hierarchical captioning strategy reduces evidence misalignment from sparse long-video sampling while preserving fine-grained visual details. Finally, we uniformly sampled 3,574 initial candidate samples from the video clip candidate pool according to duration, information, and category, and invited a surgeon to review the candidate videos' descriptions and segmentations.

2.3.2 Knowledge-Guided QA Generation

Given the parsed video-caption pairs, we prompt GPT-5 (OpenAI, 2025a) with predefined QA templates (Section 2.1), clip-level descriptions, and procedure summaries to generate video QAC pairs with structured reasoning chains. Each chain contains four components: *Observation*, *Knowledge*, *Assessment*, and *Reasoning*, which reflect expert clinical reasoning process. To match question complexity with temporal context, *Task*-level clips address perception and action understanding, while *Step*- and *Phase*-level clips support more integrative, knowledge-intensive reasoning. To ensure visual grounding, each timestamp in the *Observation* category is linked to corresponding keyframes, and Gemini-3-Pro (Doshi, 2025) is used to review and filter the samples with incorrect visual evidence, temporal localization, reasoning, and question design. We further screen for language-only leakage by removing samples that can be answered from the question text without visual or contextual evidence

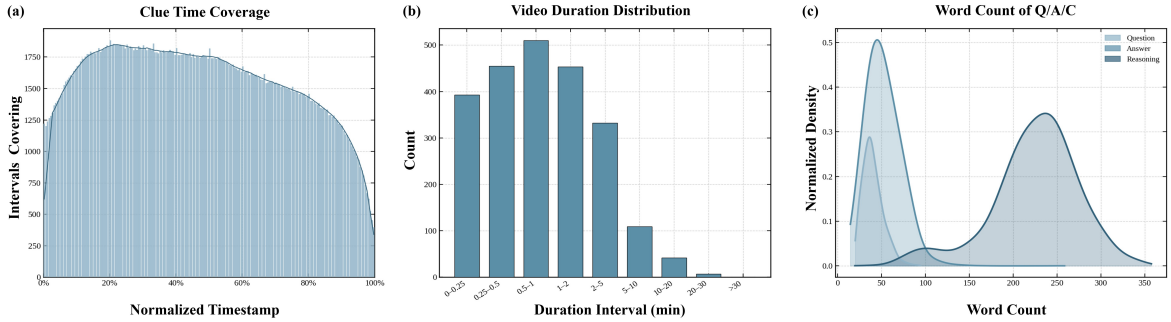


Figure 4: **Data analysis of UniSurgBench.** (a) Temporal coverage of time intervals over normalized video timestamps; (b) Distribution of video durations; (c) Word-count distributions of data samples.

(Zuo et al., 2025). Finally, we filtered out 701 and 324 video QAC pairs through keyframe verification and language-only leakage test, respectively.

2.3.3 Human-in-the-Loop Quality Control

We adopt a two-stage iterative verification-and-refinement process for the remaining 2,594 samples to further align model-generated annotations with expert clinical judgment. At each stage, we use a verifier to score each video QA pair on a Likert scale using both the video frames and contextual captions, record issues requiring correction, and pass the feedback to a refiner for revision. Then, the refiner will revise the corresponding issue and record the revision history for further verification.

In the first round, GPT-5 (OpenAI, 2025a) serves as the verifier and Gemini-3-Pro (Doshi, 2025) as the refiner. They perform multiple rounds of automatic evaluation and revision, after which each sample is labeled as *reject*, *revise*, or *accept* according to the final scores. Only samples labeled *revise* or *accept* proceed to the second stage.

In the second round, two senior surgeons review the outputs from the first stage using the same Likert-scale protocol and manually correct ambiguous references, logical inconsistencies, and clinically inappropriate terminology. GPT-5 then revises the annotations according to this structured expert feedback. The two-stage design with iterative expert review progressively aligns model-generated annotations with expert clinical judgment, simultaneously reducing expert burden and expanding question coverage without compromising quality. For more details on the annotation pipeline, please refer to Appendix Sec C.

2.4 Statistics

UniSurgBench provides broad procedural coverage and dense temporal evidence annotations for surgical video reasoning. After annotation and filtering, the benchmark contains 2,302 high-quality video QAC pairs spanning 45 surgical procedures and 29 anatomical structures, with video durations ranging from under 1 minute to over 27 minutes (Figure 1). The 14 task types are distributed across different task dimensions, surgical scenes, procedural stages, and information densities, as shown in Figure 3. In addition, UniSurgBench includes 12,768 annotated clues and 12,359 time intervals, with an average interval duration of 19.2s. Figure 4(a) shows that these clue intervals are broadly distributed over normalized video timestamps, indicating that the annotated evidence covers diverse stages of the surgical workflow. Additional dataset statistics are provided in Appendix Sec E.

3 Evaluation

We introduce SurgExpEval, a two-stage evaluation framework for open-ended surgical QA (Figure 1(c)). The evaluation pipeline and metrics are detailed in the following subsections.

3.1 Evaluation Pipeline

SurgExpEval consists of two stages: structured prediction extraction and category-aware semantic matching. In the first stage, inspired by (Qi et al., 2025; Jing et al., 2025), we use LLM-judge model with a customized prompt to standardize each open-ended response into five categories consistent with the expert data review workflow: *Observation* O, *Knowledge* K, *Assessment* A, *Reasoning* R, and *Conclusion* C. *Observation* contains timestamped visual evidence, while the other categories capture referenced clinical knowledge, surgical scene

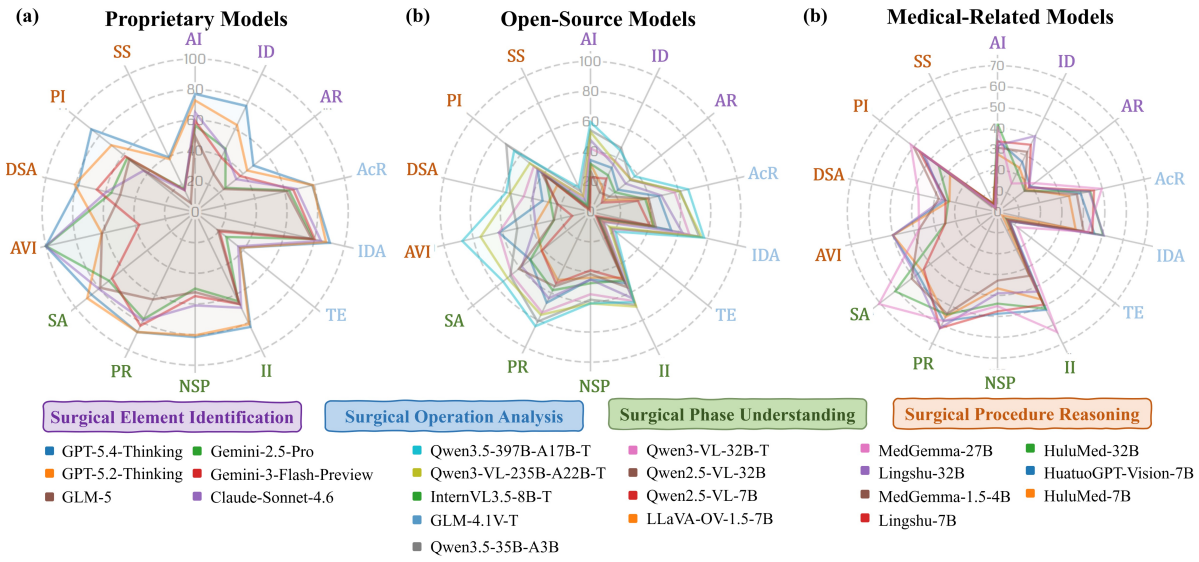


Figure 5: **Answer accuracy performance comparison** across 14 tasks for proprietary (Left Panel), open-source (Middle Panel), and medical-related models (Right Panel).

assessment, inference steps, and the final answer. Each category may contain multiple items, and each item is treated as an individual matching unit. In the second stage, given the extracted prediction $\hat{z}_i$ and ground truth z_i, SurgExpEval performs semantic matching within each category $u \in \{O, K, A, R, C\}$. For each ground-truth item g and predicted item p, Judge model assigns a semantic score $s_i^u(g,p) \in \{0,1,2\}$, where 0, 1, and 2 denote mismatch, partial match, and exact match. Each GT item is matched to at most one prediction, namely the most relevant one, and the score is normalized to $[0,1]$ by $w(\ell) = \ell/2$. The resulting matched triplets (g,p,ℓ), where $\ell = s_i^u(g,p)$, are used for metric computation.

3.2 Evaluation Metrics

Semantic Reasoning Evaluation. We evaluate each structured component after semantic alignment, treating each extracted item as one reasoning step. For sample i and component u, we use recall to measure the completeness of key reasoning steps and precision to measure the validity of generated reasoning steps. The metrics are defined as follows:

$$\begin{aligned} \text{Recall}_i^u &= \frac{\sum_{(g,p,\ell)} w(\ell)}{|z_i^u|}, \\ \text{Precision}_i^u &= \frac{\sum_{(g,p,\ell)} w(\ell)}{|\hat{z}_i^u|}. \end{aligned} \tag{1}$$

We macro-average recall and precision over valid samples for each component and compute F1 from the macro-averaged values. We further group *Knowledge* and *Assessment* as Medical Knowledge (MK), *Reasoning* as Causal Reasoning (CR), and *Observation* as Visual Perception (VP).

Answer Evaluation. We evaluate final-answer correctness using semantic matching on the *Conclusion* component. We report Strict Accuracy, which counts only exact semantic matches, and Lenient Accuracy, which also counts partial matches.

Temporal Grounding Evaluation. We evaluate temporal grounding on timestamped *Observation* items. Rather than scoring temporal overlap alone, SurgExpEval adopts a verify-then-locate protocol that first checks semantic alignment and then measures localization quality. This design reduces over-generation bias and better reflects whether a model can identify relevant video evidence. For each semantically aligned reference-prediction pair with intervals I_{gt} and I_{pred}, temporal overlap is defined as:

$$\text{IoU} = \begin{cases} 1, & |t_{\text{gt}} - t_{\text{pred}}| \le 1\,\text{s}, \\ \dfrac{|I_{\text{gt}} \cap I_{\text{pred}}|}{|I_{\text{gt}} \cup I_{\text{pred}}|}, & \text{otherwise}. \end{cases} \tag{2}$$

Point-like cues are treated as observations with a single timestamp. After semantic alignment, each reference observation receives one localization score, and unmatched observations receive zero overlap. We report mIoU over all reference observations, together with IoU@0.3 and IoU@0.5.

Table 3: **Performance comparison on semantic reasoning evaluation and temporal grounding.** VP, MK, and CR denote visual perception, medical knowledge, and causal reasoning, respectively. R, P, and F1 denote recall, precision, and F1. Top-1 and Top-2 values in each metric column are highlighted.

Model	Semantic Reasoning							Temporal Grounding		
	VP@R	VP@P	MK@R	MK@P	CR@R	CR@P	Overall@F1	mIoU	IoU@0.3	IoU@0.5
Proprietary Models										
GPT-5.4 (OpenAI, 2026)	77.42	43.58	86.24	57.23	82.72	70.59	67.35	23.70	34.73	21.77
GPT-5.2 (OpenAI, 2025b)	72.83	51.32	79.21	57.57	79.27	70.53	67.33	23.03	33.05	21.66
Gemini-2.5-pro (Google DeepMind, 2025)	66.26	40.94	77.09	46.67	70.69	58.07	57.76	18.64	26.23	17.21
Gemini-3-Flash (Google, 2026)	65.29	48.48	71.09	55.72	70.01	64.56	61.86	21.97	31.38	20.76
Claude-Sonnet-4.6 (Anthropic, 2026)	71.83	36.33	79.32	40.37	71.78	55.17	55.22	22.95	33.17	19.97
GLM-5 (Z.AI, 2026)	58.54	41.77	71.52	40.27	63.56	51.32	52.65	20.88	30.77	18.72
Open-source Models										
GLM4.1V-9B-T (GLM-V Team et al., 2026)	11.28	52.61	41.74	59.37	51.37	63.08	44.12	3.90	4.98	3.52
LLaVA-OV-7B (Li et al., 2024a)	22.92	37.46	51.09	45.36	45.37	48.67	41.91	4.01	5.00	3.77
InternVL3.5-8B-T (Wang et al., 2025b)	23.98	50.42	57.26	52.97	55.49	56.44	49.31	8.49	11.23	7.76
Qwen2.5-VL-7B-I (Bai et al., 2025b)	28.65	44.99	56.94	48.81	51.19	50.76	46.77	9.92	13.42	9.26
Qwen3-VL-8B-T (Bai et al., 2025a)	47.38	59.60	70.92	59.94	65.25	62.62	60.93	18.97	26.21	18.30
Qwen2.5-VL-32B-I (Bai et al., 2025b)	27.25	37.95	63.58	45.66	54.36	52.76	47.37	5.31	6.96	4.94
Qwen3-VL-32B-T (Bai et al., 2025a)	52.61	60.35	75.13	60.71	68.54	64.88	63.60	20.08	27.50	19.46
Qwen3.5-35B-A3B (Qwen Team, 2026)	60.70	49.24	77.30	50.82	70.43	60.54	60.44	15.46	22.12	14.18
Qwen3-VL-235B-A22B-T (Bai et al., 2025a)	59.01	50.41	76.38	51.14	70.15	61.32	60.57	17.97	25.58	16.83
Qwen3.5-397B-A17B-T (Qwen Team, 2026)	64.90	48.71	80.17	50.43	72.51	62.17	61.74	15.69	22.10	15.17
Medical-related Models										
HuatuoGPT-Vision-7B (Chen et al., 2024)	26.13	48.43	60.56	47.19	55.88	53.77	48.57	7.75	10.95	6.77
HuluMed-7B (Jiang et al., 2025)	32.49	33.87	52.19	51.32	49.42	52.34	44.84	4.61	5.39	4.12
HuluMed-32B (Jiang et al., 2025)	34.03	37.74	60.68	50.56	55.35	53.49	48.56	5.66	7.06	5.06
Lingshu-7B (LASA Team et al., 2025)	23.70	55.19	66.51	50.22	59.81	57.03	51.86	5.28	7.07	4.96
Lingshu-32B (LASA Team et al., 2025)	24.72	53.22	61.44	49.52	54.34	56.01	49.61	5.67	7.36	5.30
MedGemma1.5-4B (Sellergren et al., 2026)	22.83	49.83	56.45	50.06	52.50	51.64	46.65	7.55	9.76	6.84
MedGemma-27B (Sellergren et al., 2025)	37.79	45.96	57.57	47.02	51.26	54.40	48.98	14.30	20.30	13.18

4 Experiment

4.1 Experiment Settings

Evaluation Models. We benchmark 23 MLLMs from three groups: proprietary models, open-source generalist models, and medical-domain models. The proprietary group includes GPT (OpenAI, 2026, 2025b), Gemini (Google DeepMind, 2025; Google, 2026), Claude (Anthropic, 2026), and GLM (Z.AI, 2026) families. The open-source group covers representative models from the GLM4.1V(GLM-V Team et al., 2026), LLaVA-OneVision(Li et al., 2024a), InternVL3.5(Wang et al., 2025b), Qwen2.5-VL(Bai et al., 2025b), Qwen3-VL(Bai et al., 2025a), and Qwen3.5(Qwen Team, 2026) families, with model sizes ranging from 7B to 397B parameters. The medical-related group includes HuatuoGPT-V(Chen et al., 2024), HuluMed(Jiang et al., 2025), Lingshu(LASA Team et al., 2025), and MedGemma(Sellergren et al., 2025).

Evaluation Setting. We use a unified visual input budget by uniformly sampling 64 frames from each video for all models. We use GPT-5-mini (OpenAI, 2025a) as the judge model for evaluation, as it shows close agreement with human expert judgments on the evaluation subset (See Appendix D.1). All open-source models are evaluated on 8 RTX 4090 GPUs. Additional implementation and evaluation details are provided in Appendix Sec F.

4.2 Main Results

General-purpose proprietary models show the strongest surgical reasoning performance. As shown in Table 3, proprietary models achieve the strongest overall performance across most semantic CoT metrics. GPT-5.4-Thinking obtains the best overall F1 score of 67.35%, while Gemini-3-Flash-Preview and Gemini-2.5-Pro also maintain competitive performance among proprietary models. Among open-source models, the Qwen3-VL and Qwen3.5 families show the most competitive performance, with Qwen3-VL-32B-Thinking reaching 63.60% overall F1 and Qwen3.5-397B-A17B-Thinking reaching 61.74%. In contrast, medical-domain models do not show a consistent advantage over general-purpose MLLMs, suggesting that medical knowledge specialization alone is insufficient for robust surgical video reasoning.

Knowledge-driven reasoning is stronger than visually grounded perception. Across model families, performance on Medical Knowledge and

Causal Reasoning is generally higher than performance on Visual Perception. For example, GPT-5.2 achieves 66.68% F1 on Medical Knowledge and 74.64% F1 on Causal Reasoning, but only 60.21% F1 on Visual Perception. A similar pattern appears in most open-source and medical-domain models. This gap indicates that current MLLMs are more effective at leveraging clinical knowledge and reasoning patterns than at extracting fine-grained visual evidence from surgical videos.

Precision and recall reveal different reasoning behaviors. The semantic reasoning metrics show that models differ not only in overall F1 but also in how they generate reasoning traces. Several proprietary models achieve high recall, indicating broad coverage of key reasoning steps, but their precision is often lower, suggesting that richer reasoning traces may include less targeted content. In contrast, Qwen3-VL-32B-Thinking shows a more balanced profile, achieving 65.43% overall recall and 61.87% overall precision. This suggests that finer-grained visual perception can significantly improve the performance of surgical intelligence in clinical reasoning. **Fine-grained interaction modeling and temporal synthesis tasks remain challenging.** The fine-grained results in Fig 5 show that model performance is highly uneven across task types. Many models perform well on tasks with explicit procedural or semantic cues, such as instrument detection (ID), phase recognition (PR), and surgical action analysis (SA). For example, GPT-5.2-Thinking achieves 87.7 on ID, 87.0 on PR, and 90.2 on SA. However, performance drops substantially on tasks requiring fine-grained interaction modeling and temporal synthesis. Even the best-performing model reaches only 37.4 on triplet extraction (TE) and 38.4 on surgical summarization (SS). These results suggest that current MLLMs can recognize salient surgical states and procedural cues, but still struggle to integrate instrument-tissue interactions, temporal context, and clinical knowledge into coherent surgical video understanding.

Temporal grounding remains a major bottleneck. As shown in Table 3, all model families struggle with temporal localization. Even the best-performing model, GPT-5.4-Thinking, reaches only 23.70% mIoU and 21.77% IoU@0.5. Open-source models perform competitively in some cases, with Qwen3-VL-32B-Thinking achieving 20.08% mIoU, but most medical-domain models remain substantially lower. These results suggest that identifying when relevant evidence occurs in surgical videos remains difficult for current MLLMs, even when they can produce plausible semantic reasoning.

Table 4: **Effect of input settings.** Delta rows indicate the performance change compared with the previous setting. Green and red denote gains and drops, respectively.

Setting	mIoU	Short	Mid	Long
Gemini-2.5-Flash				
T Only	-	22.10	23.93	25.88
T + 1F	3.26	42.22	43.51	47.14
+ 64F	15.93	50.16	**49.77**	52.01
Δ	+12.67	+7.9	+6.2	+4.9
+ Timestamp	18.30	49.94	48.48	55.03
Δ	+2.37	-0.2	-1.3	+3.0
+ CoT	**18.49**	**52.90**	47.93	**55.30**
Δ	+0.19	+3.0	-0.5	+0.3
Qwen3-VL-8B-Ins				
T Only	-	20.45	19.87	26.74
T + 1F	3.30	31.82	36.32	31.83
+ 64F	8.56	41.93	42.74	42.45
Δ	+5.26	+10.1	+6.4	+10.6
+ Timestamp	18.59	43.47	43.33	41.72
Δ	+10.03	+1.5	+0.6	-0.7
+ CoT	**19.40**	**43.74**	**43.42**	**43.36**
Δ	+0.81	+0.3	+0.1	+1.6

Impact of temporal context and prompting strategy. We study how visual context, timestamp information, and CoT prompting affect model performance using Gemini-2.5-Flash and Qwen3-VL-8B-Ins. As shown in Table 4, replacing single-frame input with 64-frame video input yields the largest improvement for both models. Gemini-2.5-Flash improves from 3.26 to 15.93 mIoU, while Qwen3-VL-8B-Ins improves from 3.30 to 8.56 mIoU, with consistent gains across short, mid-length, and long videos. Adding explicit timestamps further improves temporal grounding, especially for Qwen3-VL-8B-Ins, whose mIoU increases from 8.56 to 18.59. CoT prompting brings additional but smaller gains, with the clearest benefit on long videos, where Qwen3-VL-8B-Ins improves from 41.7 to 43.3. These results show that UniSurgBench depends strongly on visual-temporal evidence, while timestamp cues and step-wise reasoning further help models localize evidence and answer long-form surgical questions.

5 Conclusion

This paper introduces UniSurgBench, a unified surgical video benchmark for evaluating open-ended reasoning in clinical workflows. Through expert-curated video question-answer samples, structured rationales, temporal evidence annotations, and a two-stage evaluation pipeline, UniSurgBench distinguishes itself from existing surgical

benchmarks by evaluating MLLMs from both process and outcome perspectives. By assessing 23 proprietary, open-source, and medical-domain MLLMs, we characterize current model performance across visual perception, medical knowledge, causal reasoning, answer correctness, and temporal grounding. Our findings reveal a clear gap between knowledge-driven reasoning and visually grounded surgical video understanding, providing valuable insights for developing more reliable MLLMs for surgical intelligence.

6 Limitations

Data coverage. UniSurgBench provides broad coverage across surgical procedures and anatomical structures, but its data distribution is still constrained by publicly available educational videos. The benchmark is constructed from professionally narrated videos collected from WebSurg and medical YouTube channels, which often have clearer views, cleaner narration, and more structured workflows than raw operating-room recordings. As a result, UniSurgBench may not fully capture variations in camera systems, surgeon styles, institutional protocols, rare complications, or noisy intraoperative environments. Future extensions should incorporate more multi-center, multi-device, and less curated surgical videos to better approximate real clinical deployment settings.

Annotation subjectivity. Although UniSurgBench uses expert review and human-in-the-loop quality control, some annotations inevitably involve clinical judgment. This is especially true for high-level tasks such as intent inference, safety assessment, pathology inference, and surgical summarization, where multiple clinically plausible interpretations may exist. The provided question-answer-clue triplets and structured rationales are therefore best viewed as expert-aligned reference annotations rather than the only possible reasoning path. Expanding expert panels and reporting finer-grained inter-rater agreement would further strengthen annotation reliability.

Automatic evaluation. SurgExpEval enables scalable evaluation of open-ended responses by extracting structured reasoning components and applying category-aware semantic matching. However, the framework relies on an LLM judge for response parsing and semantic scoring. Although we validate the judge model against clinical experts on a held-out subset, automatic evaluation may still be affected by model-family bias, response-style sensitivity, or imperfect calibration of partial matches. Future work could improve robustness through judge ensembles, larger expert-validated subsets, and more detailed analyses of disagreement cases.

Clinical deployment. UniSurgBench is designed as a diagnostic benchmark for surgical video reasoning, not as a validation study for clinical deployment. The current evaluation measures answer correctness, reasoning quality, and temporal grounding under a fixed input setting, but it does not assess real-time interaction, surgeon-AI collaboration, downstream patient outcomes, or safety-critical decision-making. Therefore, strong performance on UniSurgBench should be interpreted as evidence of improved surgical video reasoning, rather than direct evidence of clinical readiness.

7 Ethical Considerations

Data privacy and intended use. UniSurgBench is constructed from publicly available educational surgical videos collected from WebSurg (WebSurg, 2026) and medical YouTube (YouTube, 2026) channels. We use videos intended for professional education or public medical communication, and all collected videos are reviewed to remove irrelevant, unsafe, or privacy-sensitive content. We further de-identify videos when necessary and do not intentionally retain patient-identifiable information. UniSurgBench is intended for research evaluation of surgical video reasoning, and any future extension involving hospital-owned or raw operating-room recordings should require appropriate institutional review, data-use agreements, and patient consent procedures.

Clinical safety. UniSurgBench should not be interpreted as a clinical deployment benchmark. The dataset and evaluated MLLMs are designed for controlled research evaluation, not for guiding diagnosis, treatment, intraoperative actions, or patient management. Model outputs may be incomplete, temporally misgrounded, or clinically misleading in safety-critical settings. Therefore, benchmark performance should be viewed as evidence of surgical video reasoning capability under the defined evaluation protocol, rather than evidence of clinical readiness.

Human annotation and automatic evaluation. The construction of UniSurgBench involves expert review and human-in-the-loop quality con-

trol, but clinical reasoning annotations may still reflect expert judgment and task-specific assumptions. SurgExpEval further relies on LLM-based structured extraction and semantic matching to evaluate open-ended responses at scale. Although we validate the judge model against clinical experts on a held-out subset, automatic evaluation may still introduce model-family bias, response-style sensitivity, or inconsistent calibration of partial matches. High-stakes use cases should therefore include independent expert review and additional validation.

Potential misuse. UniSurgBench is intended to support reliable and interpretable research on MLLMs for surgical intelligence. However, benchmark scores could be misused to overstate model capability or imply clinical approval. We encourage users to report results together with the benchmark scope, data limitations, and evaluation protocol, and to avoid using UniSurgBench as evidence that a model is safe for real-world surgical assistance without prospective clinical validation.

Reproducibility Statement. We plan to release UniSurgBench and its evaluation resources to the research community after the acceptance of the manuscript through a controlled application-and-review process. Since the benchmark is derived from publicly available medical educational surgical videos and may involve potential privacy, copyright, and safety-related risks, all applicants will be required to complete an officially authorized application form specifying their research purpose, intended scope of use, data protection measures, and expected usage. Upon approval, the dataset may only be used for non-commercial research, model evaluation, and validation purposes. The data must not be used for clinical diagnosis, treatment decision-making, commercial product development, or any unauthorized redistribution. We also plan to release the evaluation protocol, task definitions, evaluation scripts, prompt templates, and necessary metadata to facilitate reproducibility and fair comparison.

8 Acknowledgments

This research was supported in part by the National Natural Science Foundation of China (grant 62576358, and 62176249). The AI-driven experiments, simulations and model training were performed on the robotic AI-Scientist platform of Chinese Academy of Sciences.

References

Anthropic. 2025. Introducing Claude Sonnet 4.5. Accessed: 2026-05-12.

Anthropic. 2026. Introducing claude sonnet 4.6. Accessed: 2026-05-17.

Mohammed Baharoon, Jun Ma, Congyu Fang, Augustin Toma, and Bo Wang. 2025. Exploring the design space of 3d mllms for ct report generation. In *International MICCAI*, pages 237–246. Springer.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

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, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and Benyou Wang. 2024. Towards injecting medical visual knowledge into multimodal llms at scale. In *EMNLP*, pages 7346–7370.

Ying Chen, Guoan Wang, Yuanfeng Ji, Yanjun Li, Jin Ye, Tianbin Li, Ming Hu, Rongshan Yu, Yu Qiao, and Junjun He. 2025a. Slidechat: A large vision-language assistant for whole-slide pathology image understanding. In *CVPR*, pages 5134–5143.

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

Zhixuan Chen, Luyang Luo, Yequan Bie, and Hao Chen. 2025c. Dia-llama: Towards large language model-driven ct report generation. In *MICCAI*, pages 141–151. Springer.

Jiajun Cheng, Xianwu Zhao, Sainan Liu, Xiaofan Yu, Ravi Prakash, Patrick J Codd, Jonathan Elliott Katz, and Shan Lin. 2026. Surgxbench: Explainable vision-language model benchmark for surgery. In *WACV*, pages 8188–8198.

Tri Dao. 2024. Flashattention-2: Faster attention with better parallelism and work partitioning. In *ICLR*, volume 2024, pages 35549–35562.

Rohan Doshi. 2025. Gemini 3 pro: The frontier of vision ai. Google DeepMind blog post. Accessed: 2026-05-13.

Mauro Orazio Drago, Luca Carlini, Pelinsu Celebi Balyemez, Dennis Pierantozzi, Chiara Lena, Cesare Hassan, Danail Stoyanov, Elena De Momi, Sophia Bano, and Mobarak I Hoque. 2025. Surgvivqa: Temporally-grounded video question answering for surgical scene understanding. *arXiv preprint arXiv:2511.03325*.

GLM-V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, and 74 others. 2026. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *Preprint*, arXiv:2507.01006.

Haozhen Gong, Xiaozhong Ji, Yuansen Liu, Wenbin Wu, Xiaoxiao Yan, Jingjing Liu, Kai Wu, Jiazhen Pan, Bailiang Jian, Jiangning Zhang, Hu Xiaobin, and Bran Li Hongwei. 2026. Med-cmr: A fine-grained benchmark integrating visual evidence and clinical logic for medical complex multimodal reasoning. In *CVPR*, pages 41224–41234.

Google. 2026. Introducing Gemini 3 Flash: Benchmarks, global availability. Accessed: 2026-05-12.

Google DeepMind. 2025. Gemini 2.5: Our most intelligent AI model. Accessed: 2026-05-12.

Guangfu Guo, Xiaoqian Lu, and Yue Feng. 2025. Medvragent: A framework for medical visual reasoning-enhanced agents. In *EMNLP*, pages 18613–18627.

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

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

Runlong He, Mengya Xu, Adrito Das, Danyal Z Khan, Sophia Bano, Hani J Marcus, Danail Stoyanov, Matthew J Clarkson, and Mobarakol Islam. 2024. Pitvqa: Image-grounded text embedding llm for visual question answering in pituitary surgery. In *MICCAI*, pages 488–498. Springer.

Zexue He, Yu Wang, An Yan, Yao Liu, Eric Chang, Amilcare Gentili, Julian McAuley, and Chun-Nan Hsu. 2023. Medeval: A multi-level, multi-task, and multi-domain medical benchmark for language model evaluation. In *EMNLP*, pages 8725–8744.

Ming Hu, Peng Xia, Lin Wang, Siyuan Yan, Feilong Tang, Zhongxing Xu, Yimin Luo, Kaimin Song, Jurgen Leitner, Xuelian Cheng, Jun Cheng, Chi Liu, Kaijing Zhou, and Zongyuan Ge. 2024a. Ophnet: A large-scale video benchmark for ophthalmic surgical workflow understanding. In *ECCV*, pages 481–500. Springer.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024b. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *CVPR*, pages 22170–22183.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, Jin Hao, Zi-Jian Chen, Ruijia Wu, Tao Tang, Jun bang Lv, Hongxia Xu, Hongwei Wang, Jun Xiao, Bin Feng, and 6 others. 2025. Hulumed: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*.

Juseong Jin and Chang Wook Jeong. 2024. Surgical-llava: Toward surgical scenario understanding via large language and vision models. *arXiv preprint arXiv:2410.09750*.

Hao Jing, Liang Yuci, Lin Lizhuo, Fan Yuxuan, Zhou Wenkai, Guo Kaixin, Ye Zanting, Sun Yanpeng, Zhang Xinyu, Yang Yanqi, Li Qiankun, Tang Hao, K. H. Tsoi James, Shen Linling, and Feng Hung Kuo. 2025. Oralgpt-omni: A versatile dental multimodal large language model. *CVPR 2026*.

James R Korndorffer Jr, Erika Fellinger, and William Reed. 2010. Sages guideline for laparoscopic appendectomy. *Surgical Endoscopy*, 24(4):757–761.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *SOSP*, pages 611–626.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. 2026. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *IEEE Transactions on Medical Imaging*, 45:2727 – 2737.

LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. 2025. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*.

Seowoo Lee, Jiwon Youn, Hyungjin Kim, Mansu Kim, and Soon Ho Yoon. 2025. Cxr-llava: a multimodal large language model for interpreting chest x-ray images. *European Radiology*, 35(7):4374–4386.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei

Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *NeurIPS*, 36:28541–28564.

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

Qingqiu Li, Zihang Cui, Seong-Woo Bae, Jilan Xu, Runtian Yuan, Yuejie Zhang, Rui Feng, Quanli Shen, Xiaobo Zhang, Junjun He, and Shujun Wang. 2025. Aor: Anatomical ontology-guided reasoning for medical large multimodal model in chest x-ray interpretation. *arXiv preprint arXiv:2505.02830*.

Shi Li, Vinkle Srivastav, Nicolas Chanel, Saurav Sharma, Nabani Banik, Lorenzo Arboit, Kun Yuan, Pietro Mascagni, and Nicolas Padoy. 2026a. Surgtemp: Temporal-aware surgical video question answering with text-guided visual memory for laparoscopic cholecystectomy. *arXiv preprint arXiv:2603.29962*.

Yaoqian Li, Xikai Yang, Dunyuan Xu, Yang Yu, Litao Zhao, Xiaowei Hu, Jinpeng Li, and Pheng-Ann Heng. 2026b. Surgpub-video: A comprehensive surgical video framework for enhanced surgical intelligence in vision-language model. In *AAAI*, volume 40, pages 6628–6635.

Yuci Liang, Xinheng Lyu, Meidan Ding, Wenting Chen, Jipeng Zhang, Yuexiang Ren, Xiangjian He, Song Wu, Sen Yang, Xiyue Wang, Xiaohan Xing, and Linlin Shen. 2025. Wsi-llava: A multimodal large language model for whole slide image. In *ICCV*, pages 22718–22727.

Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Tianyi Zhou, and Zuozhu Liu. 2024. Medcot: Medical chain of thought via hierarchical expert. In *EMNLP*, pages 17371–17389.

Shengyuan Liu, Boyun Zheng, Wenting Chen, Zhihao Peng, Zhenfei Yin, Jing Shao, Jiancong Hu, and Yixuan Yuan. 2026a. Endobench: A comprehensive evaluation of multi-modal large language models for endoscopy analysis. *NeurIPS*, 38.

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

Xiwei Liu, Yulong Li, Xinlin Zhuang, Xuhui Li, Jianxu Chen, Haolin Yang, Imran Razzak, and Yutong Xie. 2026b. Clincot: Clinical-aware visual chain-of-thought for medical vision language models. *arXiv preprint arXiv:2603.01124*.

Chang Han Low, Ziyue Wang, Tianyi Zhang, Zhu Zhuo, Zhitao Zeng, Evangelos B. Mazomenos, and Yueming Jin. 2026. Surgraw: Multi-agent workflow with chain of thought reasoning for robotic surgical video analysis. *IEEE Robotics and Automation Letters*, pages 1–8.

Shentong Mo and Paul Pu Liang. 2024. Multimed: Massively multimodal and multitask medical understanding. *arXiv preprint arXiv:2408.12682*.

Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Sara Pieri, Saeed Yahya Alseiari, Shanavas Cholakkal, Khaled Aldahmani, Fahad Khan, Rao Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. 2025. Bimedix2: Bio-medical expert lmm for diverse medical modalities. *Preprint*, arXiv:2412.07769.

Shuai Niu, Jing Ma, Hongzhan Lin, Liang Bai, Zhihua Wang, Richard Yi Da Xu, Yunya Song, and Xian Yang. 2025. Knowledge-augmented multimodal clinical rationale generation for disease diagnosis with small language models. In *ACL (Volume 1: Long Papers)*, pages 11011–11024.

OpenAI. 2025a. Introducing GPT-5. Accessed: 2026-05-07.

OpenAI. 2025b. Introducing GPT-5.2. Accessed: 2026-05-12.

OpenAI. 2026. Introducing GPT-5.4. Accessed: 2026-05-12.

OpenRouter. 2026. Openrouter api reference. https://openrouter.ai/docs/api/reference/overview. Accessed: 2026-05-17.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In *MICCAI*, pages 337–347. Springer.

Chen Pengcheng, Ye Jin, Wang Guoan, Li Yanjun, Deng Zhongying, Li Wei, Li Tian-Xin, Duan Haodong, Huang Ziyan, Su Yan-Cheng, Wang Benyou, Zhang Shaoting, Fu Bin, Cai Jianfei, Zhuang Bohan, J. Seibel Eric, He Junjun, and Qiao Yu. 2024. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *NeurIPS*, pages 94327–94427.

Alejandra Perez, Chinedu Nwoye, Ramtin Raji Kermani, Omid Mohareri, and Muhammad Abdullah Jamal. 2026a. Surglavi: Large-scale hierarchical dataset for surgical vision–language representation learning. *Medical Image Analysis*, 110:103982.

Alejandra Perez, Anita Rau, Lee White, Busisiwe Mlambo, Chinedu Nwoye, Muhammad Abdullah Jamal, and Omid Mohareri. 2026b. Sureon: A benchmark and vision-language-model for surgical reasoning. *arXiv preprint arXiv:2603.06570*.

Yukun Qi, Yiming Zhao, Yu Zeng, Xikun Bao, Wenxuan Huang, Lin Chen, Zehui Chen, Jie Zhao, Zhongang Qi, and Feng Zhao. 2025. Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning. *arXiv preprint arXiv:2504.07956*.

Guanyi Qin, Xiaozhen Wang, Zhu Zhuo, Chang Han Low, Yuancan Xiao, Yibing Fu, Haofeng Liu, Kai Wang, Chunjiang Li, and Yueming Jin. 2026. Surgo-r1: Benchmarking and modeling contextual reasoning for operative zone in surgical video. *arXiv preprint arXiv:2602.21706*.

Qwen Team. 2026. Qwen3.5: Towards Native Multimodal Agents. Accessed: 2026-05-12.

Mohammad Reza Rezaei, Reza Saadati Fard, Jayson Lee Parker, Rahul G Krishnan, and Milad Lankarany. 2025. Agentic medical knowledge graphs enhance medical question answering: Bridging the gap between LLMs and evolving medical knowledge. In *EMNLP*, pages 12682–12701.

Corentin Royer, Bjoern Menze, and Anjany Sekuboyina. 2024. Multimedeval: A benchmark and a toolkit for evaluating medical vision-language models. *arXiv preprint arXiv:2402.09262*.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya B Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, and 47 others. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.

Lalithkumar Seenivasan, Mobarakol Islam, Adithya K Krishna, and Hongliang Ren. 2022. Surgical-vqa: Visual question answering in surgical scenes using transformer. In *MICCAI*, pages 33–43. Springer.

Andrew B. Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, Liron Yatziv, Tiffany Chen, Bram Sterling, Kenneth Philbrick, Richa Tiwari, Yun Liu, Madhuram Jajoo, Chandrashekar Sankarapu, Swapnil Vispute, and 23 others. 2026. Medgemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*.

Andrew B. Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla P. Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chih-Yao Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, J. T. Lai, Samuel Schmidgall, and 62 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Guohao Sun, Can Qin, Huazhu Fu, Linwei Wang, and Zhiqiang Tao. 2024. Self-training large language and vision assistant for medical question answering. In *EMNLP*, pages 20052–20060.

Yuxuan Sun, Yixuan Si, Chenglu Zhu, Xuan Gong, Kai Zhang, Pingyi Chen, Ye Zhang, Zhongyi Shui, Tao Lin, and Lin Yang. 2025. Cpath-omni: A unified multimodal foundation model for patch and whole slide image analysis in computational pathology. In *CVPR*, pages 10360–10371.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David J. Fleet, P. A. Mansfield, Sushant Prakash, Renee C Wong, Sunny Virmani, and 13 others. 2024. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138.

Guan-Feng Wang, Long Bai, Junyi Wang, Kun Yuan, Zhen Li, Tianxu Jiang, Xiting He, Jinlin Wu, Zhen Chen, Zhen Lei, Hongbin Liu, Jiazheng Wang, Fan Zhang, Nicolas Padoy, Nassir Navab, and Hongliang Ren. 2025a. Endochat: Grounded multimodal large language model for endoscopic surgery. *Medical Image Analysis*, 107CA:103789.

Gui Wang, YongSong Zhou, Kaijun Deng, Wooi Ping Cheah, Rong Qu, Jianfeng Ren, and Linlin Shen. 2026. Surgcot: Advancing spatiotemporal reasoning in surgical videos through a chain-of-thought benchmark. *arXiv preprint arXiv:2604.20319*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 44 others. 2025b. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Yinuo Wang, Baiyang Wang, Robert Mercer, Frank Rudzicz, Sudipta Singha Roy, Pengjie Ren, Zhumin Chen, and Xindi Wang. 2025c. Trustworthy medical question answering: An evaluation-centric survey. In *EMNLP*, pages 27477–27490.

Ziqing Wang, Chengsheng Mao, Xiaole Wen, Luo Yuan, and Kaize Ding. 2025d. Amanda: Agentic medical knowledge augmentation for data-efficient medical visual question answering. In *EMNLP*.

WebSurg. 2026. WebSurg: The Online University of IRCAD. https://www.websurg.com/. Accessed: 2026-05-07.

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

Mengya Xu, Zhongzhen Huang, Jie Zhang, Xiaofan Zhang, and Qi Dou. 2025. Surgical action planning with large language models. In *MICCAI*, pages 563–572. Springer.

Yuxuan Yang, Zhonghao Yan, Yi Zhang, Bo Yun, Muxi Diao, Guowei Zhao, Kongming Liang, Wenbin Li,

and Zhanyu Ma. 2026. Hepato-llava: An expert mllm with sparse topo-pack attention for hepatocellular pathology analysis on whole slide images. *arXiv preprint arXiv:2602.19424*.

YouTube. 2026. YouTube. Accessed: 2026-05-07.

Suhao Yu, Haojin Wang, Juncheng Wu, Luyang Luo, Jingshen Wang, Cihang Xie, Pranav Rajpurkar, Carl Yang, Yang Yang, Kang Wang, Yannan Yu, and Yuyin Zhou. 2025. Medframeqa: A multi-image medical vqa benchmark for clinical reasoning. *arXiv preprint arXiv:2505.16964*.

Kun Yuan, Manasi Kattel, Joël L Lavanchy, Nassir Navab, Vinkle Srivastav, and Nicolas Padoy. 2024a. Advancing surgical vqa with scene graph knowledge. *IJCARS*, 19(7):1409–1417.

Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. 2024b. Hecvl: Hierarchical video-language pretraining for zero-shot surgical phase recognition. In *MICCAI*, pages 306–316. Springer.

Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. 2024c. Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. *NeurIPS*, 37:122952–122983.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, pages 9556–9567.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2025. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *ACL (Volume 1: Long Papers)*, pages 15134–15186.

Z.AI. 2026. GLM-5. Accessed: 2026-05-12.

Zhitao Zeng, Zhu Zhuo, Xiaojun Jia, Erli Zhang, Junde Wu, Jiaan Zhang, Yuxuan Wang, Chang Han Low, Jian Jiang, Zilong Zheng, Xiaochun Cao, Yutong Ban, Qi Dou, Yang Liu, and Yueming Jin. 2025. Surgvlm: A large vision-language model and systematic evaluation benchmark for surgical intelligence. *ArXiv*, abs/2506.02555.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2024. A simple llm framework for long-range video question-answering. In *EMNLP*, pages 21715–21737.

Ling Zhang, Boxiang Yun, Qingli Li, and Yan Wang. 2025a. Historical report guided bi-modal concurrent learning for pathology report generation. In *MICCAI*, pages 343–352. Springer.

Ming Zhang, Yujiong Shen, Zelin Li, Huayu Sha, Binze Hu, Yuhui Wang, Chenhao Huang, Shichun Liu, Jingqi Tong, Changhao Jiang, Mingxu Chai, Zhiheng Xi, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2025b. Llmeval-med: A real-world clinical benchmark for medical llms with physician validation. In *EMNLP*, pages 4888–4914.

Yan Zhu, Te Luo, Pei-Yao Fu, Zhen Zhang, Zi-Long Wang, Yi-Fan Qu, Zi-Han Geng, Jia-Qi Xu, Lu Yao, Li-Yun Ma, Wei Su, Wei-Feng Chen, Quan-Lin Li, Shuo Wang, and Ping-Hong Zhou. 2026. Gi-bench: A panoramic benchmark revealing the knowledge-experience dissociation of multimodal large language models in gastrointestinal endoscopy against clinical standards. *arXiv preprint arXiv:2601.08183*.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*.

Appendix Contents

We organize the appendix content as follow:

Appendix A Related Work
Appendix B Detailed Task Definition
Appendix C Detailed Annotation Pipeline
Appendix D Human Alignment and Ablation on Judge Model
Appendix E Detailed Dataset Statistic
Appendix F Experiment Setting and Frame Sampling Ablation
Appendix G Detailed Experiment Result and More Ablation on Main Result
Appendix H Prompt Templates
Appendix I Case Study

A Related Work

A.1 Surgical Video Understanding

Recent studies have advanced surgical video understanding by constructing surgical QA datasets, pretraining multimodal models on instructional videos, and adapting MLLMs to surgical scenarios (Yuan et al., 2024b,c; Li et al., 2024b; Wang et al., 2025a; Jin and Jeong, 2024; Chen et al., 2025b; Zeng et al., 2025). Early efforts focused on procedure-specific image or video QA datasets built from open-source surgical data, enabling models to recognize basic surgical elements and answer scene-level questions (Seenivasan et al., 2022; Yuan et al., 2024a; He et al., 2024; Wang et al., 2025a; Zeng et al., 2025). Another line of work leverages instructional videos, expert narration, and hierarchical textual information to learn fine-grained surgical representations and procedural context (Yuan et al., 2024b,c; Perez et al., 2026a; Li et al., 2026b). More recent methods explore stronger multimodal capabilities, including long-range temporal modeling, video QA, multi-task learning, and surgical instruction tuning (Li et al., 2024b; Jin and Jeong, 2024; Zeng et al., 2025; Chen et al., 2025b; Zhang et al., 2024). Despite these advances, existing resources remain limited for evaluating clinically grounded surgical video reasoning: many focus on static images, single procedures, or task-specific supervision (Seenivasan et al., 2022; He et al., 2024; Liu et al., 2026a; Zhu et al., 2026; Low et al., 2026), and most evaluations emphasize answer correctness rather than whether responses are supported by temporal evidence and structured reasoning. UniSurgBench addresses this gap by providing an open-ended, workflow-guided benchmark with temporal clues and structured rationales for process-aware evaluation across diverse surgical workflows.

A.2 Reasoning Multimodal Large Language Models

Reasoning has become a core capability of multimodal large language models (MLLMs), particularly in medical scenarios that require diagnostic inference, clinical question answering, and evidence-grounded decision-making (Li et al., 2023; Tu et al., 2024; Saab et al., 2024; He et al., 2023; Royer et al., 2024; Mo and Liang, 2024; Wang et al., 2025d; Mullappilly et al., 2025). Recent medical MLLMs have further advanced this capability through expert-style chain-of-thought prompting, reinforcement learning, and explicit alignment between visual evidence and intermediate clinical rationales (Liu et al., 2024; Lai et al., 2026; Liu et al., 2026b; Sun et al., 2024; Guo et al., 2025; Niu et al., 2025; Rezaei et al., 2025). In surgical video understanding, emerging models place increasing emphasis on procedure-aware reasoning and temporal understanding, adopting paradigms such as plan-execute-reflect reasoning, reinforcement learning, operative zone identification, phase recognition, risk warning, and rationale generation (Hao et al., 2025; Qin et al., 2026; Perez et al., 2026b). This trend reflects an important shift from answer-only prediction toward process-aware clinical intelligence, where models are expected not only to produce correct outputs but also to justify them with temporally and visually grounded evidence. Such grounding is especially critical in surgery, where clinically meaningful judgments often depend on subtle procedural transitions, tool–tissue interactions, and context accumulated over time (Hao et al., 2025; Qin et al., 2026; Perez et al., 2026b). However, existing evaluations still provide limited support for assessing open-ended reasoning quality, temporal grounding, and rationale faithfulness in surgical videos (He et al., 2023; Royer et al., 2024; Mo and Liang, 2024). UniSurgBench addresses this gap by pairing structured rationales with temporal clues and by introducing **SurgExpEval** to evaluate both the reasoning process and the final answer outcome.

A.3 Multimodal Surgery QA Benchmarks

Surgical QA benchmarks provide essential testbeds for evaluating multimodal models in clinically relevant surgical scenarios. Early benchmarks primarily focused on image-based QA, scene re-

Table 5: **Performance comparison on semantic reasoning evaluation**, including visual perception (VP), medical knowledge (MK), and causal reasoning (CR). We report Recall (Rec), Precision (Pre), and F1. Top-1 and Top-2 values in each metric column are highlighted.

Model	Visual Perception			Medical Knowledge			Causal Reasoning			Overall		
	Rec	Pre	F1	Rec	Pre	F1	Rec	Pre	F1	Rec	Pre	F1
Proprietary Models												
GPT-5.4-Thinking (OpenAI, 2026)	77.42	43.58	55.77	86.24	57.23	68.80	82.72	70.59	76.17	82.13	57.08	67.35
GPT-5.2-Thinking (OpenAI, 2025b)	72.83	51.32	60.21	79.21	57.57	66.68	79.27	70.53	74.64	77.10	59.75	67.33
Gemini-2.5-pro (Google DeepMind, 2025)	66.26	40.94	50.61	77.09	46.67	58.14	70.69	58.07	63.76	71.35	48.52	57.76
Gemini-3-Flash-Preview (Google, 2026)	65.29	48.48	55.64	71.09	55.72	62.47	70.01	64.56	67.18	68.80	56.20	61.86
Claude-Sonnet-4.6 (Anthropic, 2026)	71.83	36.33	48.25	79.32	40.37	53.51	71.78	55.17	62.39	74.31	43.93	55.22
GLM-5 (Z.AI, 2026)	58.54	41.77	48.75	71.52	40.27	51.52	63.56	51.32	56.79	64.54	44.46	52.65
Open-source Models												
GLM4.1V-Thinking-9B (GLM-V Team et al., 2026)	11.28	52.61	18.57	41.74	59.37	49.02	51.37	63.08	56.63	34.79	60.28	44.12
LLaVA-OneVision-7B (Li et al., 2024a)	22.92	37.46	28.44	51.09	45.36	48.05	45.37	48.67	46.96	39.79	44.27	41.91
InternVL3.5-8B-Thinking (Wang et al., 2025b)	23.98	50.42	32.51	57.26	52.97	55.03	55.49	56.44	55.96	45.58	53.71	49.31
Qwen2.5-VL-7B-Instruct (Bai et al., 2025b)	28.65	44.99	35.01	56.94	48.81	52.56	51.19	50.76	50.97	45.59	48.00	46.77
Qwen3-VL-8B-Thinking (Bai et al., 2025a)	47.38	59.60	52.79	70.92	59.94	64.97	65.25	62.62	63.91	61.18	60.68	60.93
Qwen2.5-VL-32B-Instruct (Bai et al., 2025b)	27.25	37.95	31.72	63.58	45.66	53.15	54.36	52.76	53.55	48.40	46.39	47.37
Qwen3-VL-32B-Thinking (Bai et al., 2025a)	52.61	60.35	56.21	75.13	60.71	67.16	68.54	64.88	66.66	65.43	61.87	63.60
Qwen3.5-35B-A3B (Qwen Team, 2026)	60.70	49.24	54.37	77.30	50.82	61.32	70.43	60.54	65.11	69.48	53.49	60.44
Qwen3-VL-235B-A22B-Thinking (Bai et al., 2025a)	59.01	50.41	54.37	76.38	51.14	61.26	70.15	61.32	65.44	68.52	54.28	60.57
Qwen3.5-397B-A17B-Thinking (Qwen Team, 2026)	64.90	48.71	55.65	80.17	50.43	61.91	72.51	62.17	66.94	72.53	53.75	61.74
Medical-related Models												
HuatuoGPT-Vision-7B (Chen et al., 2024)	26.13	48.43	33.95	60.56	47.19	53.05	55.88	53.77	54.80	47.53	49.66	48.57
HuluMed-7B (Jiang et al., 2025)	32.49	33.87	33.16	52.19	51.32	51.75	49.42	52.34	50.84	44.70	44.98	44.84
HuluMed-32B (Jiang et al., 2025)	34.03	37.74	35.79	60.68	50.56	55.16	55.35	53.49	54.40	50.02	47.18	48.56
Lingshu-7B (LASA Team et al., 2025)	23.70	55.19	33.16	66.51	50.22	57.23	59.81	57.03	58.38	50.01	53.86	51.86
Lingshu-32B (LASA Team et al., 2025)	24.72	53.22	33.76	61.44	49.52	54.84	54.34	56.01	55.16	46.83	52.75	49.61
MedGemma-4B (Sellergren et al., 2025)	22.83	49.83	31.32	56.45	50.06	53.07	52.50	51.64	52.07	43.93	49.73	46.65
MedGemma-27B (Sellergren et al., 2025)	37.79	45.96	41.48	57.57	47.02	51.76	51.26	54.40	52.78	48.87	49.08	48.98

lationship modeling, and local instrument-action interactions (Seenivasan et al., 2022; Yuan et al., 2024a; He et al., 2024). More recent efforts extend evaluation toward spatiotemporal understanding and workflow-aware clinical assessment (Zeng et al., 2025; Drago et al., 2025; Liu et al., 2026a; Zhu et al., 2026). For example, EndoBench (Liu et al., 2026a) and GI-Bench (Zhu et al., 2026) introduce multidimensional evaluations aligned with clinical workflows, while SurgViVQA (Drago et al., 2025) and PitVQA-Anticipation (He et al., 2024) focus on temporally grounded understanding and event anticipation. Other benchmarks and models, such as SUREON (Perez et al., 2026b), SurgCoTBench (Low et al., 2026), and SurGo-R1, further explore rationale generation, operative zone identification, and risk alerting. Despite this progress, existing benchmarks still provide limited support for fine-grained, quantitative evaluation of intermediate reasoning steps and their temporal evidence, which are critical for assessing whether model predictions are grounded in surgical videos (Zhang et al., 2025b; Wang et al., 2025c; Gong et al., 2026). UniSurgBench addresses this gap through open-ended video question-answer-clue triplets, structured rationales, temporal clue annotations, and SurgExpEval for process-aware evaluation.

B Task Definition

UniSurgBench defines 14 workflow-guided tasks across four levels of surgical video reasoning. The taxonomy follows the progression from visual perception to action understanding, clinical intent reasoning, and scenario-specific knowledge analysis. All tasks are formulated as open-ended question answering, enabling evaluation of answer correctness, reasoning quality, and temporal grounding.

B.1 Level 1: Surgical Element Identification

Instrument Detection evaluates whether models can identify visible surgical instruments and describe their screen locations, entry or exit, and state changes. It focuses on evidence-bound recognition of tools and their operative status, such as cutting, coagulation, suction, irrigation, or clip firing.

Anatomy Identification evaluates whether models can identify visible anatomical structures and describe their spatial relations with instruments or surrounding tissue planes. It emphasizes evidence-

Table 6: **Strict answer** accuracy comparison across fine-grained capability dimensions. Top-1 and Top-2 values in each metric column are highlighted.

Model	Element Identification			Operation Analysis			Phase Understanding				Procedure Understanding				Overall
	ID	AI	AR	AcR	TE	IDA	PR	II	SA	NSP	PI	AVI	SS	DSA	
Proprietary Models															
GPT-5.4-Thinking (OpenAI, 2026)	48.7	77.1	76.9	78.5	37.4	90.3	87.0	83.3	86.6	81.7	86.7	100.0	39.6	78.3	71.4
GPT-5.2-Thinking (OpenAI, 2025b)	43.5	72.9	63.0	79.0	37.4	87.7	87.0	81.2	90.2	80.5	70.0	62.5	38.4	80.6	68.5
Gemini-2.5-pro (Google DeepMind, 2025)	25.3	56.6	45.8	63.0	26.1	79.2	77.7	64.6	72.0	50.0	56.7	100.0	16.7	56.2	52.1
Gemini-3-Flash-Preview (Google, 2026)	37.7	59.6	39.5	65.8	20.4	80.1	82.4	66.7	69.5	54.9	58.3	37.5	15.9	65.9	53.8
Claude-Sonnet-4.6 (Anthropic, 2026)	34.4	65.1	45.0	68.0	36.0	84.3	78.6	69.4	80.5	61.0	43.3	100.0	15.9	59.4	55.9
GLM-5 (Z.AI, 2026)	24.0	49.4	27.7	61.2	19.0	78.0	63.4	66.7	79.3	52.4	55.0	62.5	6.5	50.2	45.7
Open-source Models															
GLM4.1V-Thinking-9B (GLM-V Team et al., 2026)	23.4	34.3	32.4	46.6	6.2	55.1	67.2	52.8	51.2	45.1	45.0	62.5	3.7	32.7	36.5
LLaVA-OneVision-7B (Li et al., 2024a)	12.3	31.9	18.1	38.8	10.4	42.4	50.4	48.6	41.5	43.9	28.3	37.5	2.4	28.6	29.30
InternVL3.5-8B-Thinking (Wang et al., 2025b)	20.8	31.9	26.9	37.4	6.6	43.2	58.0	50.7	51.2	47.6	38.3	25.0	4.9	27.2	32.0
Qwen2.5-VL-7B-Instruct (Bai et al., 2025b)	9.7	22.9	24.4	32.4	4.7	40.3	52.9	50.0	41.5	39.0	40.0	12.5	0.8	22.6	27.3
Qwen3-VL-8B-Thinking (Bai et al., 2025a)	29.9	47.6	34.9	48.4	10.9	62.7	63.9	63.9	58.5	45.1	45.0	37.5	8.2	39.6	41.2
Qwen2.5-VL-32B-Instruct (Bai et al., 2025b)	16.2	27.1	14.3	39.7	7.6	44.5	55.0	55.6	61.0	41.5	41.7	25.0	2.4	22.6	29.8
Qwen3-VL-32B-Thinking (Bai et al., 2025a)	34.4	42.2	37.8	57.5	18.5	67.4	74.4	66.0	67.1	54.9	48.3	62.5	11.4	45.6	46.3
Qwen3.5-35B-A3B (Qwen Team, 2026)	33.8	54.2	47.1	61.2	17.5	74.2	81.1	68.8	68.3	58.5	71.7	50.0	15.1	54.8	52.1
Qwen3-VL-235B-A22B-Thinking (Bai et al., 2025a)	34.4	53.0	37.0	60.7	16.1	75.8	76.1	70.1	70.7	61.0	51.7	75.0	11.4	54.8	49.8
Qwen3.5-397B-A17B-Thinking (Qwen Team, 2026)	37.7	59.6	45.4	66.7	21.3	77.5	84.5	66.7	74.4	61.0	65.0	87.5	18.0	57.6	54.8
Medical-related Models															
HuatuoGPT-Vision-7B (Chen et al., 2024)	19.5	34.3	26.5	39.7	7.1	46.2	58.0	52.1	47.6	48.8	45.0	50.0	3.7	24.9	32.5
HuluMed-7B (Jiang et al., 2025)	17.5	27.7	23.5	34.2	2.8	37.3	55.9	46.5	45.1	36.6	28.3	50.0	4.5	23.0	28.2
HuluMed-32B (Jiang et al., 2025)	16.2	42.2	21.4	44.3	5.2	50.4	54.6	51.4	61.0	43.9	35.0	25.0	4.1	23.5	32.3
Lingshu-7B (LASA Team et al., 2025)	18.8	33.7	35.7	46.1	5.7	45.3	61.8	49.3	43.9	47.6	46.7	25.0	4.1	24.9	33.7
Lingshu-32B (LASA Team et al., 2025)	18.8	31.9	40.3	43.8	7.1	50.8	61.8	42.4	48.8	39.0	43.3	50.0	2.4	26.3	33.8
MedGemma1.5-4B (Sellergren et al., 2026)	16.2	30.7	31.9	37.9	7.1	41.1	54.2	34.0	51.2	32.9	50.0	37.5	1.6	28.1	30.1
MedGemma-27B (Sellergren et al., 2025)	22.1	41.0	15.1	49.8	11.8	43.2	58.0	63.9	70.7	45.1	51.7	37.5	2.4	37.8	35.7

supported anatomical recognition rather than unsupported naming of ambiguous structures.

Anomaly Detection evaluates whether models can detect abnormal findings or visual impediments in the surgical field. It covers bleeding, oozing, smoke, fog, lens contamination, tissue charring, fluid seepage, critical occlusion, and other cues that may affect operative progress or risk.

B.2 Level 2: Surgical Operation Analysis

Action Recognition evaluates whether models can recognize dominant surgical maneuvers and their temporal transitions. It covers actions such as retraction, dissection, hemostasis, irrigation, clipping, cutting, and tissue handling, together with their visible purpose.

Triplet Extraction evaluates whether models can represent key operations as structured {instrument, action, target} triplets. Each triplet should be grounded in the video segment and, when applicable, associated with a time window.

Interaction Dynamics evaluates whether models can explain instrument-tissue and multi-instrument coordination over time. It covers traction, countertraction, exposure, tension modulation, occlusion management, and role specialization between instruments.

B.3 Level 3: Surgical Phase Understanding

Phase Recognition evaluates whether models can infer the current surgical phase or step from the observed action sequence, instrument usage, and exposure state. The predicted phase should be consistent with the visible workflow and supported by segment-level evidence.

Intent Inference evaluates whether models can infer the surgeon's near-term operative intention. It covers goals such as improving exposure, clearing tissue, preparing for clipping, controlling bleeding, verifying anatomy, or reducing operative risk.

Safety Assessment evaluates whether models can assess whether the current view supports safe continuation or transition to a definitive maneuver. It focuses on risks such as anatomical misidentification, poor exposure, thermal spread, excessive traction, and proximity to critical structures.

Next-Step Prediction evaluates whether models can predict the most likely immediate next maneuver from exposure, instrument readiness, and action trajectory. It emphasizes short-horizon predictions grounded in the current segment rather than generic procedure-level expectations.

Table 7: **Lenient accuracy** comparison across fine-grained capability dimensions. Overall is computed as the sample-weighted average across all subtasks. Top-1 and Top-2 values in each metric column are highlighted.

Model	Element Identification			Operation Analysis			Phase Understanding				Procedure Understanding				Overall
	ID	AI	AR	AcR	TE	IDA	PR	II	SA	NSP	PI	AVI	SS	DSA	
Proprietary Models															
GPT-5.4-Thinking (OpenAI, 2026)	90.3	96.4	95.8	90.9	89.6	97.5	93.7	98.6	98.8	90.2	93.3	100.0	81.2	95.4	92.8
GPT-5.2-Thinking (OpenAI, 2025b)	83.8	92.8	95.0	93.6	89.1	97.5	94.1	97.9	98.8	93.9	85.0	87.5	82.0	96.3	92.3
Gemini-2.5-Pro (Google DeepMind, 2025)	79.9	88.0	89.1	89.5	83.9	99.2	89.1	95.1	95.1	74.4	71.7	100.0	72.7	88.5	86.8
Gemini-3-Flash-Preview (Google, 2026)	78.6	93.4	86.6	91.8	82.5	97.5	92.9	94.4	96.3	84.1	85.0	87.5	69.8	91.7	87.8
Claude-Sonnet-4.6 (Anthropic, 2026)	78.6	90.4	82.8	90.0	88.2	97.5	91.2	94.4	96.3	78.0	66.7	100.0	68.6	90.3	86.5
GLM-5 (Z.AI, 2026)	68.8	80.1	74.4	85.4	73.0	98.3	79.0	93.8	98.8	78.0	78.3	75.0	32.7	82.5	76.9
Open-source Models															
GLM4.1V-Thinking-9B (GLM-V Team et al., 2026)	83.8	83.1	81.5	87.2	68.2	97.0	88.2	96.5	93.9	80.5	81.7	100.0	80.4	86.2	85.1
LLaVA-OneVision-7B (Li et al., 2024a)	57.8	65.1	53.4	75.3	46.9	89.0	64.3	84.0	80.5	67.1	56.7	62.5	38.0	74.7	64.7
InternVL3.5-8B-Thinking (Wang et al., 2025b)	75.3	78.3	75.2	73.5	70.6	94.9	82.4	90.3	82.9	81.7	68.3	62.5	69.4	89.9	79.6
Qwen2.5-VL-7B-Instruct (Bai et al., 2025b)	62.3	68.7	66.0	74.9	57.3	89.4	75.2	93.1	84.1	79.3	70.0	87.5	39.2	79.7	70.8
Qwen3-VL-8B-Thinking (Bai et al., 2025a)	81.8	79.5	78.6	85.8	75.4	97.5	85.7	95.1	95.1	78.0	70.0	100.0	73.5	89.9	83.9
Qwen2.5-VL-32B-Instruct (Bai et al., 2025b)	61.0	76.5	58.4	78.5	49.8	95.3	80.3	96.5	92.7	81.7	75.0	75.0	31.0	84.8	71.6
Qwen3-VL-32B-Thinking (Bai et al., 2025a)	74.7	83.1	79.8	90.0	83.4	94.9	87.8	96.5	91.5	80.5	70.0	100.0	78.0	90.8	85.5
Qwen3.5-35B-A3B (Qwen Team, 2026)	83.1	86.1	88.7	90.0	82.9	95.3	92.0	91.7	90.2	81.7	88.3	75.0	65.7	91.2	86.5
Qwen3-VL-235B-A22B-Thinking (Bai et al., 2025a)	81.8	86.1	87.0	88.6	80.1	96.6	87.4	97.2	97.6	81.7	75.0	100.0	64.9	90.3	85.6
Qwen3.5-397B-A17B-Thinking (Qwen Team, 2026)	83.8	88.0	82.4	87.7	83.4	95.8	94.5	95.1	93.9	86.6	88.3	100.0	72.7	90.3	87.4
Medical-related Models															
HuatuoGPT-Vision-7B (Chen et al., 2024)	76.0	80.7	76.1	84.5	60.7	93.2	80.7	91.7	86.6	74.4	80.0	100.0	55.5	81.6	77.8
HuluMed-7B (Jiang et al., 2025)	63.0	79.5	69.3	82.2	39.3	91.1	82.8	93.1	84.1	72.0	60.0	100.0	46.9	83.4	72.6
HuluMed-32B (Jiang et al., 2025)	68.8	85.5	69.3	86.3	44.5	96.6	77.3	91.7	92.7	76.8	76.7	87.5	53.9	80.6	75.6
Lingshu-7B (LASA Team et al., 2025)	66.9	79.5	76.5	84.9	71.1	94.5	84.9	93.8	79.3	80.5	76.7	87.5	60.4	86.2	79.7
Lingshu-32B (LASA Team et al., 2025)	55.8	68.7	78.6	78.1	58.8	91.5	79.4	81.2	86.6	72.0	71.7	100.0	38.4	77.4	71.6
MedGemma1.5-4B (Sellergren et al., 2026)	65.6	77.1	76.1	80.8	54.0	90.7	73.9	86.8	86.6	68.3	78.3	75.0	45.3	84.3	73.5
MedGemma-27B (Sellergren et al., 2025)	59.1	78.9	33.6	75.3	60.7	55.1	79.8	93.8	93.9	81.7	80.0	75.0	40.0	90.8	67.1

B.4 Level 4: Surgical Procedure Understanding

Pathology Inference evaluates whether models can infer pathology-related findings from visible tissue appearance or transcript-supported evidence. It covers inflammation, edema, fibrosis, adhesions, friability, exudate, and abnormal tissue planes when these cues are supported by the segment.

Anatomical Variation Inference evaluates whether models can identify evidence suggestive of anatomical variation and reason about its surgical implications. It focuses on unusual structure course, branching, exposure, or isolation difficulty and their impact on strategy or risk control.

Surgical Summarization evaluates whether models can generate a concise segment-level intraoperative summary. The summary should cover key actions, findings, difficulties or complications, mitigation steps, and segment-level outcome.

Difficulty-Specific Assessment evaluates whether models can identify what makes a segment more difficult or different from a routine case. It covers factors such as dense adhesions, persistent oozing, smoke-impaired view, unclear anatomy, frequent instrument changes, poor exposure, or high-risk proximity to critical structures.

Across all tasks, answers should be grounded in visible video evidence, timestamped clues, or transcript-supported observations. When key cues are missing, occluded, or ambiguous, models may answer *Unverifiable* or *Not supported* while specifying the missing evidence.

C Annotation Pipeline

C.1 Annotation Model Selection

Before the formal annotation process, we systematically compared the video captioning capabilities of mainstream proprietary models, strong open-source models, and medical domain-specific models in surgical video description scenarios, including GPT-5 (OpenAI, 2025a), Gemini-3-pro (Doshi, 2025), Claude-4.5-Opus (Anthropic, 2025), Qwen3-VL-235B-A22B-Thinking (Bai et al., 2025a), and MedGemma-27B (Sellergren et al., 2025). In this setting, we use carefully designed video caption prompts to instruct candidate models to generate very detailed video captions for 45 representative surgical videos with various surgical procedures. After that, we invited two senior surgeons with more than 10 years of experience to conduct a comprehensive evaluation of the video descriptions generated by the five candidate models, considering hallucination frequency, knowledge accuracy, visual consistency, spatial-temporal localization ability, and clinical adoption confidence. As shown in Table 10, the evaluation results show that GPT-5-Thinking

Table 8: **Performance comparison on temporal grounding and temporal answer accuracy** across proprietary, open-source, and medical-related models on curation accuracy and temporal grounding. Top-1 and Top-2 values in each metric column are highlighted.

Model	Acc@Short	Acc@Mid	Acc@Long	mIoU	IoU@0.3	IoU@0.5
Proprietary Models						
GPT-5.4-Thinking (OpenAI, 2026)	**70.51**	**73.18**	**71.34**	**23.70**	**34.73**	**21.77**
GPT-5.2-Thinking (OpenAI, 2025b)	67.06	71.73	67.58	23.03	33.05	21.66
Gemini-2.5-pro (Google DeepMind, 2025)	50.47	55.94	54.12	18.64	26.23	17.21
Gemini-3-Flash-Preview (Google, 2026)	53.21	55.56	52.98	21.97	31.38	20.76
Claude-Sonnet-4.6 (Anthropic, 2026)	54.63	60.09	54.85	22.95	33.17	19.97
GLM-5 (Z.AI, 2026)	44.48	48.93	49.07	20.88	30.77	18.72
Open-source Models						
GLM4.1V-Thinking-9B (GLM-V Team et al., 2026)	37.32	35.47	35.79	3.90	4.98	3.52
LLaVA-OneVision-7B (Li et al., 2024a)	30.14	27.68	27.43	4.01	5.00	3.77
InternVL3.5-8B-Thinking (Wang et al., 2025b)	32.89	31.15	28.76	8.49	11.23	7.76
Qwen2.5-VL-7B-Instruct (Bai et al., 2025b)	27.71	26.49	27.46	9.92	13.42	9.26
Qwen3-VL-8B-Thinking (Bai et al., 2025a)	42.96	39.42	35.08	18.97	26.21	18.30
Qwen2.5-VL-32B-Instruct (Bai et al., 2025b)	29.83	30.37	29.31	5.31	6.96	4.94
Qwen3-VL-32B-Thinking (Bai et al., 2025a)	48.52	44.09	40.84	20.08	27.50	19.46
Qwen3.5-35B-A3B (Qwen Team, 2026)	51.87	52.41	53.56	15.46	22.12	14.18
Qwen3-VL-235B-A22B-Thinking (Bai et al., 2025a)	48.84	52.58	47.83	17.97	25.58	16.83
Qwen3.5-397B-A17B-Thinking (Qwen Team, 2026)	56.05	54.02	49.09	15.69	22.10	15.17
Medical-related Models						
HuatuoGPT-Vision-7B (Chen et al., 2024)	33.08	32.74	26.82	7.75	10.95	6.77
HuluMed-7B (Jiang et al., 2025)	28.61	27.36	27.49	4.61	5.39	4.12
HuluMed-32B (Jiang et al., 2025)	31.97	33.42	33.86	5.66	7.06	5.06
Lingshu-7B (LASA Team et al., 2025)	34.35	32.29	36.91	5.28	7.07	4.96
Lingshu-32B (LASA Team et al., 2025)	35.14	34.68	21.77	5.67	7.36	5.30
MedGemma1.5-4B (Sellergren et al., 2026)	29.26	30.83	35.04	7.55	9.76	6.84
MedGemma-27B (Sellergren et al., 2025)	35.09	37.65	33.81	14.30	20.30	13.18

achieved superior performance in hallucination frequency, knowledge accuracy, and clinical adoption confidence, while Gemini-3-pro performed better in spatial-temporal localization. Therefore, in the subsequent annotation process, we used GPT-5-Thinking as the primary annotation expert and introduced Gemini-3-pro to ensure the accuracy of spatial-temporal localization and multimodal perception. In addition, to avoid potential biases introduced by model priors, we excluded GPT-5-Thinking and Gemini-3-pro from the evaluation and instead assessed other model variants released around the same period.

C.2 Hierarchical Video Captioning

Hierarchical video captioning generates visual descriptions at three temporal granularities, namely Task, Step, and Phase, following the same hierarchy as the surgical segmentation.

Task-Level Captioning. For each Task-level clip, we uniformly sample 32 frames and combine them with the ASR transcript as input to GPT-5 (OpenAI, 2025a). The model generates a dense, evidence-first visual caption covering instruments (morphology, position, operational state, and contact intervals), tissues (appearance and pathological cues), and visual phenomena (bleeding, smoke, and occlusion). All timestamps are expressed relative to the start of the segment.

Step-Level Captioning. For each Step-level segment, all Task-level captions within the Step are provided together with 64 uniformly sampled frames. The model aggregates and deduplicates the child captions' *instrument–action–target* triplets while preserving monotonic temporal ordering and their containment relationship with Task-level intervals. The output emphasizes action-chain continuity and causal transitions between tasks within the Step.

Phase-Level Captioning. For each Phase-level segment, all Step-level captions and 64 densely sampled frames are used as input. The model consolidates multiple Steps into a coherent description that highlights stage-wise intent shifts (e.g., exposure → key dissection → control/removal → inspection/closure) and maintains consistency with lower-level temporal anchors.

Table 9: **Judge-model comparison on a human-alignment subset** of 200 video-QAC samples balanced across tasks and surgical specialties. The subset includes responses from three representative models: a proprietary model, an open-source model, and a medical-domain model. GPT-5-mini achieves the strongest agreement with human experts, with a per-sample Pearson correlation coefficient of $R = 0.935$ computed against the averaged human ratings.

Judge	GPT-5.2-Thinking		Qwen3-VL-8B-thinking		MedGemma-27B	
	CoT F1	Ans. Acc.	CoT F1	Ans. Acc.	CoT F1	Ans. Acc.
Human Expert						
Senior surgeon	68.92	70.88	55.21	38.57	42.09	29.96
Surgical AI expert	65.37	59.24	60.46	39.84	45.31	26.22
LLM Judges						
GPT-5-mini (OpenAI, 2025a)	67.33	68.64	60.93	41.32	48.98	35.67
Gemini-3-Flash-Preview (Google, 2026)	56.63	75.02	50.22	54.18	45.90	48.24
Claude-Haiku-4.5-20251001 (Anthropic, 2025)	46.30	25.36	37.52	6.68	34.06	6.68

Table 10: Human preference evaluation for selecting the annotation model. DH, MK, VC, TL, and CAC denote description hallucination control, medical knowledge, visual consistency, temporal localization, and clinical adoption confidence, respectively. Higher scores indicate better human preference.

Model	DH	MK	VC	TL	CAC	Overall
GPT-5-Thinking (OpenAI, 2025a)	**5**	**5**	**5**	4	**5**	**24**
Gemini-3-Pro (Doshi, 2025)	4	4	**5**	**5**	4	22
Claude-4.5-Opus (Anthropic, 2025)	4	4	4	3	4	19
Qwen3-VL-235B-A22B-Thinking (Bai et al., 2025a)	3	4	3	4	3	17
MedGemma-27B (Sellergren et al., 2025)	3	3	3	2	3	14

C.3 Key-Frame Verification

Each QAC triplet contains Observation items with explicit timestamp intervals. To verify whether the claimed visual evidence is present in the corresponding video segment, we implement a key-frame verification step using Gemini-3-Pro (Doshi, 2025) as the judge.

Adaptive Frame Sampling. Frames are sampled using a two-density strategy: (1) *dense sampling* within the timestamp windows cited by the CoT Observation items; and (2) *sparse sampling* over the remaining portions of the segment. This strategy ensures that critical visual moments are well represented while keeping the total number of frames manageable.

Verification Judgment. Given the sampled frames and the visual caption, the model evaluates each QAC triplet along seven dimensions: factuality, visual grounding, temporal grounding, answer uniqueness, question quality, professionalism, and surgical suitability. Each dimension is scored using a carefully designed 1–3 Likert scale. Samples receiving a REJECT verdict with a total score below 20 are removed from the annotation pool. This verification step filters out 701 QAC pairs.

C.4 Language-only Leakage Filtering

A central design requirement of UniSurgBench is that correct answers should rely on visual and contextual evidence from the video rather than solely on linguistic priors in the question text. To enforce this requirement, we apply a language-only leakage test to all remaining candidates after key-frame verification.

Procedure. Each question is submitted to GPT-5-mini (OpenAI, 2025a) with only the question text as input. The model determines whether the question can be answered using generic surgical knowledge and linguistic cues alone. If the language-only answer semantically matches the ground-truth answer, as judged by the SurgExpEval Conclusion scoring rubric, the sample is flagged as language-leaking and removed.

Prompt for Language-only Leakage Detection

Role. You are a surgical knowledge expert. You will be given a surgical video question *without any visual input*. Your task is to assess whether the question can be answered correctly based on general surgical knowledge and language priors alone.

Instructions.

1. Determine whether a knowledgeable surgeon could answer this question correctly without any video evidence.

2. If yes, provide the most likely answer based on surgical common sense.
3. If no, respond with `"Visual evidence required."`

Question. `{question}`

Output.

```
{
  "answerable_without_video": true|false,
  "language_only_answer": "...",
  "reasoning": "..."
}
```

This filtering step removes 324 additional QAC pairs, resulting in the final benchmark of 2,302 high-quality video-QAC triplets.

C.5 Evaluation Rubric in Quality Control

The human-in-the-loop quality control uses a two-stage iterative verification-and-refinement protocol. Both stages use the same nine-dimensional Likert-scale rubric, with each dimension scored from 1 to 5 and a total score ranging from 9 to 45.

Rubric Dimensions:

1. **Surgical Scene Validity**: Does the segment show a real intraoperative scene, either endoscopic, open, or minimally invasive, with visible instruments and tissue? Score 5: unambiguous real surgical scene; Score 1: animation or non-surgical content.
2. **Question Clarity**: Is the question unambiguous, causally clear, and self-contained within the segment? Score 5: unambiguous; Score 1: not determinable or seriously ambiguous.
3. **Segment Match**: Does the segment contain sufficient visual information to support the answer? Score 5: key evidence fully within the segment boundaries; Score 1: clearly mismatched.
4. **Visual Evidence Accuracy**: Are the Observation items accurately described with correct timestamps and object descriptions? Score 5: all Observation items are accurate; Score 1: fabricated or substantially incorrect evidence.
5. **Clinical Factuality**: Are the medical facts, anatomical references, and instrument characterizations correct? Score 5: consistent with surgical standards; Score 1: clear errors or hallucinations.
6. **Causal Chain**: Is the reasoning chain logically complete, with correct cause–effect relationships? Score 5: complete and accurate; Score 1: reversed causality or missing critical links.
7. **Reasoning Acceptability**: Is the CoT structured coherently and does it appropriately follow the Observation / Knowledge / Assessment / Reasoning format? Score 5: exemplary; Score 1: incoherent or missing key components.
8. **Answer Correctness**: Is the final Answer logically derived from the CoT and factually correct? Score 5: correct and fully supported; Score 1: wrong or unsupported.
9. **Overall Quality**: Is the sample usable and reliable for the benchmark as a whole? Score 5: excellent and ready to use; Score 1: unusable.

Round-1 Protocol (Automatic). GPT-5 (OpenAI, 2025a) acts as the Verifier, and Gemini-3-Pro (Doshi, 2025) acts as the Refiner. Multiple rounds of automatic scoring and revision are conducted for each sample. Samples are assigned to one of three outcomes: *accept* (total score ≥ 40), *revise* (35–40, up to three rounds), or *reject* (< 35). Only samples assigned *revise* or *accept* proceed to Stage 2.

Round-2 Protocol (Expert). Two senior surgeons independently reviewed the Stage-1 outputs under a blinded protocol, following a unified evaluation rubric. Both annotators were from general surgery and had more than 15 years of clinical experience in hepatobiliary and pancreatic surgery, as well as in various laparoscopic and open surgical procedures.

For specialty-specific issues arising in other types of surgery, we asked the surgeons to review the relevant departmental surgical guidelines and expert recommendations in advance to calibrate their judgments and conclusions.

In this phase, reviewers are required to watch the full surgical videos and provide Stage-1 revisions alongside the original transcripts, captions, and QACs. They then assign scores according to the rubric. For samples scoring above 35 points, reviewers manually correct ambiguous anatomical references, logical inconsistencies, inappropriate clinical terminology, and selected inaccurate timestamp markings to improve sample quality; samples deemed low quality or exhibiting clear issues are removed. Inter-rater agreement is quantified using Cohen's κ across nine dimensions, and reviewers discuss any cases with substantial score discrepancies to reach consensus.

In total, 508 samples were labeled as high quality, 1,794 were assessed as requiring revision, and 272

were labeled as low quality. The Cohen's κ between two surgeon is 0.94. Finally, we supply GPT-5 with the expert scores, revision suggestions, first-round revised QACs, and video captions, enabling the annotation model to perform targeted revisions based on expert feedback.

Alignment with SurgExpEval. The five structured components used during annotation, namely Observation, Knowledge, Assessment, Reasoning, and Conclusion, are the same five categories evaluated by SurgExpEval at inference time. This design ensures consistency between the annotation rubric and the evaluation protocol.

D Ablation on the Judge Model

SurgExpEval relies on an LLM judge to score open-ended model responses. We therefore conduct an ablation study on the choice of judge model. This analysis addresses two questions: whether the default judge agrees with human judgments, and whether the final scores are sensitive to the judge model.

D.1 Human Alignment Set

We construct a human alignment set to compare LLM-based scoring with expert review. The set contains 200 video-QA samples selected from UniSurgBench, covering all four task groups, 14 subtasks, and 8 main surgical specialist. For each sample, we include responses from three representative models: GPT-5.2-Thinking (OpenAI, 2025b), Qwen3-VL-8B-Thinking (Bai et al., 2025a), and MedGemma-27B (Sellergren et al., 2025). This design allows the alignment set to cover diverse task types, surgical scenarios, and model capability levels.

D.2 Human Review Protocol

Human reviewers evaluate the alignment set under a blinded protocol. For each sample, reviewers are shown the video clip, question, ground-truth annotation, model response, and matched reasoning items produced by SurgExpEval. Model identities are anonymized as Model A, Model B, and Model C. Reviewers score each matched item using the same 0/1/2 scale as the automatic judge, corresponding to mismatch, partial match, and exact match. This protocol enables direct fine-grained comparison between human judgments and LLM judge scores.

D.3 Judge Model Comparison

We used the same prompt template and scoring rubric to compare GPT-5-mini, Gemini-3-Flash, and Claude-Haiku-4.5. As shown in Table 9, GPT-5-mini produces scores that are closest to those of the human experts, supporting its use as the default judge model in SurgExpEval. The other judge models exhibit notable evaluation biases: Gemini-3-Flash assigns higher answer-accuracy scores but lower CoT F1 scores, whereas Claude-Haiku-4.5 yields substantially lower overall scores, especially in answer correctness. These results indicate that the choice of judge model affects absolute scores, and that GPT-5-mini achieves the highest agreement with human expert judgments in performance and model ranking among the evaluated judge models.

E Dataset Statistics

This section reports additional statistics for UniSurgBench. Table 11 summarizes the distribution of task types, and Table 12 reports the anatomical and procedural coverage of the benchmark. After quality control, UniSurgBench contains 2,302 video-QAC samples from 254 complete surgical videos, covering 45 surgical procedures and a diverse set of anatomical structures. Each sample is represented as an open-ended question-answer-clue triplet, paired with temporally localized evidence and structured intermediate rationales.

Table 11 shows the distribution of the 14 tasks across the four capability levels. At the level of task groups, the dataset contains 559 samples for Surgical Element Identification, 667 for Surgical Operation Analysis, 546 for Clinical Intent and Scene Reasoning, and 530 for Scenario-Specific Analysis. This distribution prevents the benchmark from being dominated by a single capability level and supports evaluation across a progression from short-horizon perception to higher-level clinical reasoning. At the subtask level, however, the distribution remains naturally long-tailed. Some tasks correspond to relatively infrequent or highly context-dependent clinical phenomena. For example, Anatomical Variation Inference contains only a small number of samples. For these rare task types, we applied additional manual verification to ensure that the questions are clinically meaningful, visually grounded, and supported by explicit temporal evidence.

Table 12 further reports the distribution of anatomical scopes and their corresponding pro-

cedures. The dataset includes common abdominal procedures, such as colectomy, gastrectomy, liver resection, cholecystectomy, splenectomy, and pancreatic surgery, as well as procedures from thoracic, urological, endocrine, orthopedic, ophthalmic, and neurosurgical settings. These counts should be interpreted as a coverage summary rather than a target for uniform sampling. The final distribution reflects the availability of public educational surgical videos that satisfy our quality criteria, while retaining procedure-level diversity across different anatomical regions and workflow patterns.

We also analyze temporal and linguistic properties of the benchmark. Figure 6 shows the duration distribution across the four capability levels. Lower-level perception and operation-analysis tasks are typically associated with shorter evidence windows, whereas reasoning-oriented tasks often require longer temporal context. Figure 8 reports the number of clues per QAC sample. Most samples contain multiple supporting clues, with an average of approximately six clues per QAC. The average clue duration is 19.19 seconds, suggesting that the evidence is temporally localized while preserving sufficient context for surgical interpretation.

Finally, Figure 7 presents the word distribution of questions, answers, and reasoning steps. Questions contain 39.4 words on average, answers contain 51.4 words, and reasoning steps contain 224.7 words. These statistics reflect the open-ended formulation of UniSurgBench: we do not reduce surgical video understanding to short classification labels, but instead requires models to produce explanatory answers supported by visual evidence and clinical reasoning. The high-frequency terms are mainly associated with surgical workflows, anatomical structures, instruments, operative actions, and clinical assessment, indicating that the collected QAC samples are centered on clinically relevant surgical concepts.

Table 11: **Task and subtask distribution across the four levels.** We sought to maintain a relatively balanced distribution of task counts. For scarce sample categories, we adopted more conservative selection and filtering procedures to preserve the validity of long-tail instances.

Level	Task	Count
Level 1: Surgical Element Identification		
	Instrument Detection (ID)	154
	Anatomy Identification (AI)	166
	Anomaly Recognition (AR)	239
	Subtotal	**559**
Level 2: Surgical Operation Analysis		
	Action Recognition (AcR)	220
	Triplet Extraction (TE)	211
	Interaction Dynamics Analysis (IDA)	236
	Subtotal	**667**
Level 3: Surgical Phase Understanding		
	Phase Recognition (PR)	238
	Intent Inference (II)	144
	Safety Assessment (SA)	82
	Next-Step Prediction (NSP)	82
	Subtotal	**546**
Level 4: Surgical Procedure Understanding		
	Pathology Inference (PI)	60
	Anatomical Variation Inference (AVI)	8
	Surgical Summarization (SS)	245
	Difficulty-Specific Assessment (DSA)	217
	Subtotal	**530**
Total		**2302**

F Experimental Setting

During the experimental phase, we use uniformly sampled 64 frames as the default setting. For Thinking models, we allocate the maximum available reasoning budget and close web search. For non-reasoning models, we explicitly instruct the model to produce a complete reasoning process in the prompt, so as to mitigate potential bias in reasoning-related scores caused by overly sparse responses. For closed-source models, all evaluations are conducted through their officially released APIs. For models with fewer than 70B parameters, inference is performed using vLLM (Kwon et al., 2023) with a Transformer implementation accelerated by FlashAttention2 (Dao, 2024). For models with more than 70B parameters (e.g., Qwen3-297B-A22B-Thinking), inference is conducted through the OpenRouter API (OpenRouter, 2026).

F.1 Frame Sampling Budget

Fig 9 reports the effect of the number of sampled input frames on reasoning quality, answer accuracy, and temporal grounding. The results show that model performance generally improves as the number of frames increases, especially when moving from single-frame input to multi-frame input. For Gemini-2.5-Flash, CoT F1 increases from 41.72 with 1 frame to 61.74 with 128 frames, answer accuracy increases from 36.5 to 53.0, and mIoU increases from 0.83 to 18.93. Qwen3-VL-8B-Ins. shows a similar trend, with CoT F1 increasing from 34.71 to 55.26, answer accuracy from 23.7 to 43.1,

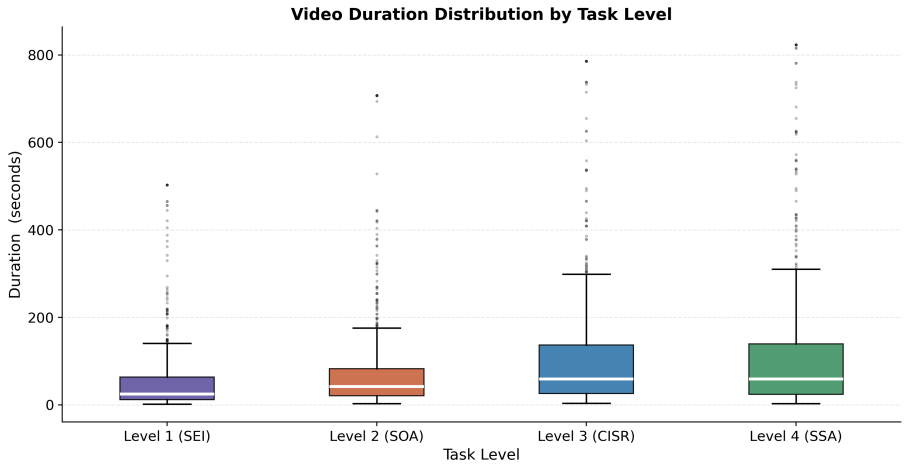


Figure 6: Duration distribution across four different levels. We assigned tasks of varying complexity and visual difficulty to reduce over-inference and avoid low-quality data.

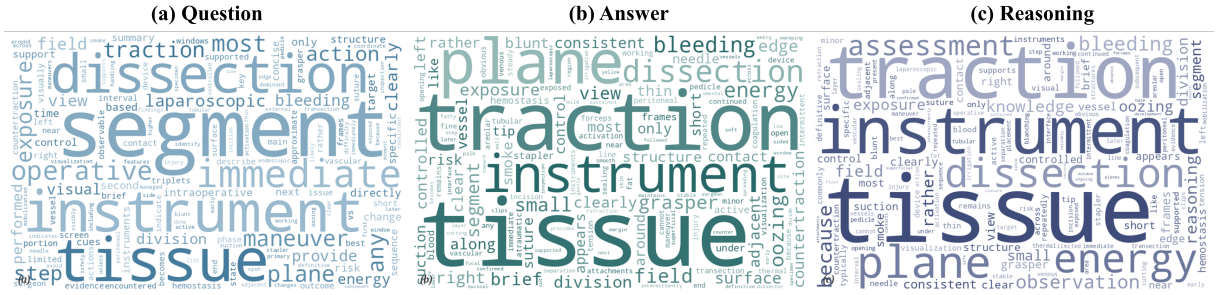


Figure 7: Word cloud of question, answer, and reasoning steps. The high frequency word in UniSurgBench are highly consistent with clinical workflow.

and mIoU from 0.85 to 18.04. These results indicate that sparse visual input is insufficient for open-ended surgical video reasoning, since models need temporally distributed evidence to recognize surgical actions, infer clinical intent, and localize supporting evidence.

In addition, the performance gains do not increase linearly with the number of sampled frames. Most gains come from increasing the input from 1 frame to 16 or 32 frames, while performance tends to saturate after 64 frames. For Gemini-2.5-Flash, 64 frames achieve the best answer accuracy, mIoU, and duration-specific accuracies, with scores of 53.2, 19.44, and 51.8, 54.6, and 53.4 on short, mid-length, and long videos, respectively. When the input is further increased to 128 frames, CoT F1 improves only slightly, while the other metrics decrease slightly. This may be because overly dense visual input interferes with the model's knowledge prior, creating a mismatch between its language prior and visual signals, which may lead to over-inference or more conservative descriptions under uncertainty. Overall, uniform sampling with 64 frames provides a good balance between performance and input efficiency. Therefore, we use this setting in our evaluation for fair comparison.

G Experiment Result

G.1 Results on Semantic CoT Evaluation

Table 5 reports model performance on semantic CoT evaluation across Visual Perception (VP), Medical Knowledge (MK), Causal Reasoning (CR), and overall reasoning quality. Proprietary models achieve the strongest F1 scores overall, with GPT-5.4-Thinking obtaining the best overall F1 of 67.35%. Among open-source models, Qwen3-VL-32B-Thinking achieves the highest overall F1 of 63.60%, approaching the leading proprietary models and outperforming all medical-related models. Medical-related models show moderate performance but do not exhibit a consistent advantage over strong general-purpose MLLMs, suggesting that domain-specific tuning alone is insufficient for robust surgical video reasoning.

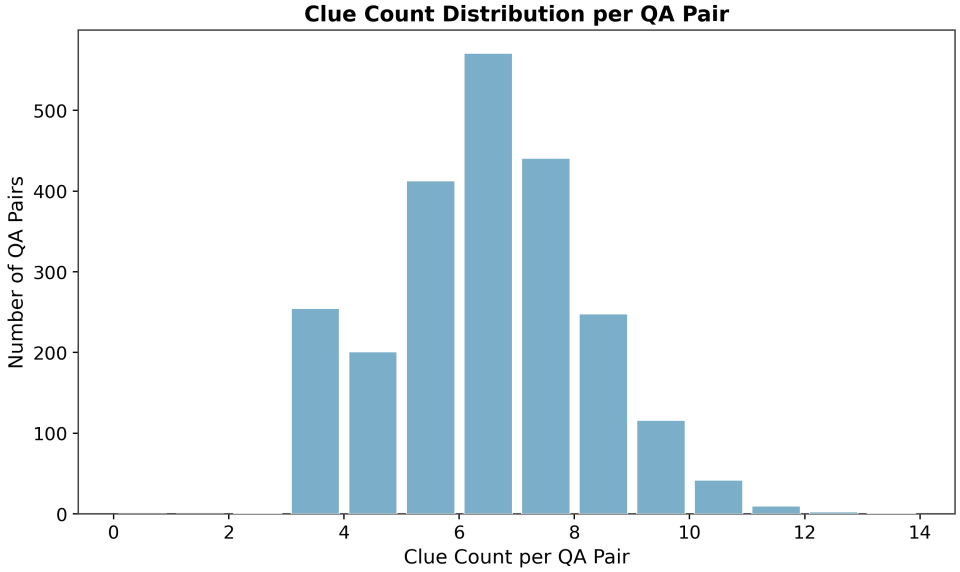


Figure 8: Clue number distribution. Most QAC triplet contains an average of 6 clues in total.

The component-level results further reveal a clear mismatch between knowledge-driven reasoning and visually grounded perception. Most models obtain higher F1 scores on MK and CR than on VP, indicating that current MLLMs are better at leveraging clinical knowledge and reasoning patterns than extracting fine-grained visual evidence from surgical videos. For example, GPT-5.2-Thinking achieves 74.64% F1 on CR and 66.68% on MK, but only 60.21% on VP. This gap is even more pronounced in many open-source and medical-related models, where VP is often the lowest-performing component. These results support the need to evaluate surgical MLLMs beyond final answers, since a model may produce plausible clinical reasoning while still missing key visual evidence.

Precision and recall further reveal different reasoning behaviors across model families. Several proprietary models achieve high recall, indicating broad coverage of key reasoning steps, but their precision is often lower, suggesting that richer reasoning traces may also include less targeted content. In contrast, Qwen3-VL-32B-Thinking achieves the highest overall precision of 61.87% while maintaining competitive recall and F1, reflecting a more compact and better-aligned reasoning profile. These results show that semantic CoT evaluation provides a finer-grained view of model reasoning than answer-level metrics alone.

G.2 Results on Answer Accuracy

Table 7 reports answer accuracy across 14 fine-grained task types grouped into four capability dimensions. The results show that model performance varies substantially across task types, rather than decreasing monotonically with task level. Models generally perform well on tasks with explicit procedural or semantic cues, such as instrument detection (ID), phase recognition (PR), and safety assessment (SA). For example, GPT-5.2-Thinking reaches 87.7 on ID, 87.0 on PR, and 90.2 on SA, indicating that strong models can recognize salient surgical states and infer clinically meaningful scene information when the supporting evidence is explicit.

In contrast, tasks requiring fine-grained interaction modeling or temporal synthesis remain difficult. Triplet extraction (TE) and surgical summarization (SS) show consistently lower performance across model families, even among top-performing proprietary models. GPT-5.2-Thinking reaches only 37.4 on TE and 38.4 on SS, while most open-source and medical-related models perform substantially worse. These results suggest that current MLLMs still struggle to organize instrument–action–target relations and summarize temporally extended surgical evidence into coherent clinical descriptions.

Table 8 further reports answer accuracy across short, mid-length, and long videos. Proprietary models maintain the strongest performance across all duration groups, with GPT-5.2-Thinking achieving 67.0, 71.7, and 67.5 on short, mid-length, and long videos, respectively. Strong open-source models remain competitive in some settings, especially Qwen3.5-397B-A17B-Thinking and Qwen3.5-35B-

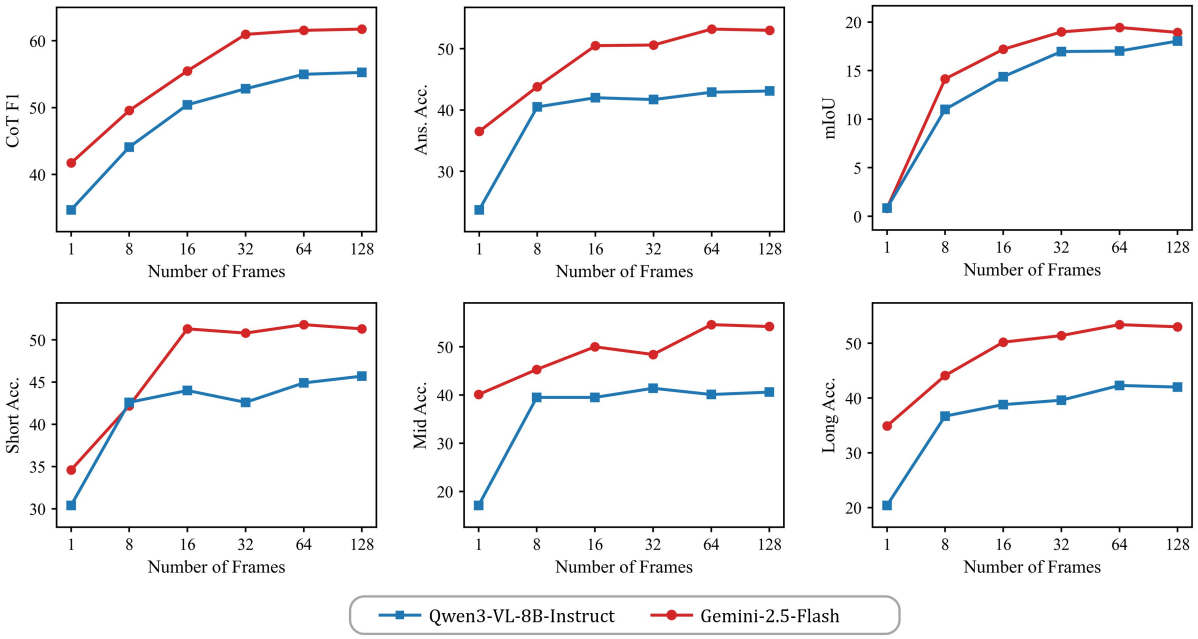


Figure 9: Ablation Study on the Frame Sampling Strategy.

A3B, but they still trail the best proprietary models. Medical-related models generally perform worse across duration groups, indicating that surgical-domain adaptation alone does not yet guarantee robust open-ended surgical QA performance.

As shown in Fig. 11, the category-level heatmap reveals clear performance heterogeneity across organ and tissue types. Some categories, such as lung, brain, and abdominal wall, are handled relatively consistently by stronger models, with proprietary systems achieving both high scores and comparatively stable behavior across these columns. This suggests that when visual patterns are more salient or anatomical cues are more readily identifiable, current MLLMs can form more reliable judgments. In contrast, categories such as tonsils, joint, ureter, and skin & soft tissue remain more challenging, especially for open-source and medical-related models, which show both lower scores and larger cross-category fluctuations. Notably, the gap between proprietary and non-proprietary models is not uniform across categories: it becomes more pronounced on structurally complex or visually ambiguous regions such as brain, retroperitoneum, bone, and soft tissue, while narrower differences are observed on a few comparatively easier categories. These results suggest that answer accuracy is shaped not only by overall model capability, but also by category-specific demands, including visual distinctiveness, anatomical complexity, and the amount of contextual clinical reasoning required for correct prediction.

G.3 Results on Temporal Grounding

Table 8 shows that temporal grounding remains challenging across all model families. Even the strongest model, GPT-5.4-Thinking, achieves only 23.70 mIoU and 21.77 IoU@0.5, indicating that accurate evidence localization remains difficult even when semantic reasoning and answer accuracy are relatively strong. Other proprietary models show similar limitations: Claude-Sonnet-4.5 achieves the highest IoU@0.3 of 33.17 but only 19.97 IoU@0.5. This gap suggests that models can often identify roughly relevant temporal regions, but still struggle with precise evidence localization.

Open-source models show more varied temporal grounding performance. Qwen3-VL-32B-Thinking achieves 20.08 mIoU and 19.46 IoU@0.5, making it the strongest open-source model on temporal grounding and competitive with several proprietary systems. However, many smaller open-source models obtain very low localization scores. For example, GLM4.1V-Thinking-9B and LLaVA-OneVision-7B both have mIoU values close to 4. Medical-related models also remain weak overall, although MedGemma-27B performs better than other medical-related models, with 14.30 mIoU and 13.18 IoU@0.5.

These results indicate that temporal grounding

is a distinct bottleneck beyond answer generation. Models with reasonable answer accuracy may still fail to locate the supporting evidence in time, which is critical for auditing surgical reasoning. This finding supports the design of UniSurgBench and SurgExpEval, where temporal clues are evaluated explicitly rather than treated as auxiliary annotations.

G.4 Step-wise Reasoning Quality Correlates with Answer Accuracy

As shown in Figure 10, the 23 evaluated models exhibit a strong positive correlation between step-wise reasoning F1 and answer accuracy ($r = 0.902$, $p < 0.0001$). This result indicates that stronger final-answer performance is closely associated with higher-quality intermediate reasoning, supporting our decomposition-based evaluation design. Proprietary models show a tighter alignment between reasoning quality and answer correctness, suggesting that stronger models are better able to use intermediate reasoning steps to produce clinically informative responses. In contrast, most medical-related models and some open-source models show weaker consistency between reasoning quality and final answers, which may reflect less reliable visual grounding or greater reliance on language priors rather than video evidence.

H Prompt Templates

H.1 Prompt Templates for Data Annotation

Video Segmentation.

The following prompt is applied to timestamp-aligned ASR transcripts to perform hierarchical surgical video parsing, yielding a three-level temporal structure together with a procedure-level summary.

Prompt for Hierarchical Video Segmentation

You are a professional surgical video analysis expert and a medical documentation specialist. Your input consists of timestamped surgical narration subtitles (ASR output, may include noise or non-surgical content).

Global Rules: Each Phase, Step, and Task must include a Confidence field $\in [0, 1]$. Preparation or introduction audio should be separated from actual surgical procedures. All timestamps in HH:MM:SS; Start $<$ End.

Task: Analyze the provided timestamped narration and complete: (1) *Validity Determination*—identify whether the video contains clinically significant surgical procedures. (2) *Multi-Granularity Segmentation*—segment hierarchically into Phase $>$ Step $>$ Task, timestamps strictly based on subtitles. (3) *Knowledge-Augmented Description*—generate professional surgical subtitles. (4) *Case Report Generation*—generate a standard postoperative summary.

Hierarchy Definitions:

- **Phase**: Macroscopic, goal-oriented process blocks. Subtitles describe intent (exposure, division, resection, reconstruction, hemostasis) and may reference anatomy/pathology.
- **Step**: Clear subgoals to achieve the Phase objective. Subtitles explain operative logic and relation to adjacent steps.
- **Task**: Specific instrument-tissue actions. Subtitles use the *instrument + action + anatomical object* triplet.

Output (strict JSON, all timestamps in HH:MM:SS):

```
{
  "Surgical_Video_Judgement": {
    "Overall": 0-100,
    "Start_Time": "HH:MM:SS",
    "End_Time":   "HH:MM:SS"
  },
  "Surgical_Segment": {
    "Phases": [
      {"Start":"..", "End":"..",
       "Label":"..", "Subtitle":".."}
    ],
    "Steps": [ ... ],
    "Tasks": [ ... ]
  },
  "Report": "<500-word surgical summary>"
}
```

User template: `Video ID: {video_id}\nSubtitles:\n{srt_content}`

Visual Caption.

Visual captions are generated at two granularities. For **Task-level** clips, uniformly sampled frames and transcript subtitles are used as input. For **Step-** and **Phase-level** clips, child-level captions are aggregated with broader-segment frames to produce hierarchically consistent descriptions.

Prompt for Task-Level Visual Caption

You are a medical expert specialized in surgical video analysis, extracting key visual information and conducting evidence-based interpretations.

Task: Generate detailed, professional captions covering visual descriptions and scene analysis with clear timestamps. All timestamps are RELATIVE to segment start (00:00:00 = segment start).

Visual Information Extraction:

1. **Visual Targets:**
   - **Instruments**: type, morphology, position, operational state (open/-close/energized/clamping/cutting/suturing), contact site and time interval.
   - **Tissues & Structures**: morphology, color, texture; pathological cues (charring, edema, fibrosis) with uncertainty qualifiers.
2. **Temporal Recognition:**
   - **Action Recognition**: *instrument–action–target* triplet with time interval `[hh:mm:ss-hh:mm:ss]`, tissue response, and layer changes.
   - **Operation Recognition**: identify current surgical Phase and Step and its relation to the overall workflow.
3. **Phenomena & Quality Issues**: bleeding, smoke, occlusion, blur—state severity, source, and risk.
4. **Uncertain Elements**: for unconfirmable instruments/structures/actions, use "possible/suspected" with visible evidence only.

Prompt for Step/Phase-Level Hierarchical Caption

You are a medical expert specialized in integrating multi-granularity evidence from surgical action-level captions and sampled frames to produce rigorous, time-aligned long video captions.

Use Cases: *Step-level*—inputs include all Task-level captions under this Step plus sampled frames. Goal: merge action chains and causal relations into a comprehensive description spanning the entire Step. *Phase-level*—inputs include all Step-level captions plus sampled frames. Goal: consolidate multiple Steps, highlighting scene transitions and Phase objectives.

Alignment Rules:

- **Step-level**: all action intervals must lie within Step boundaries and cover essential content of all child Tasks.

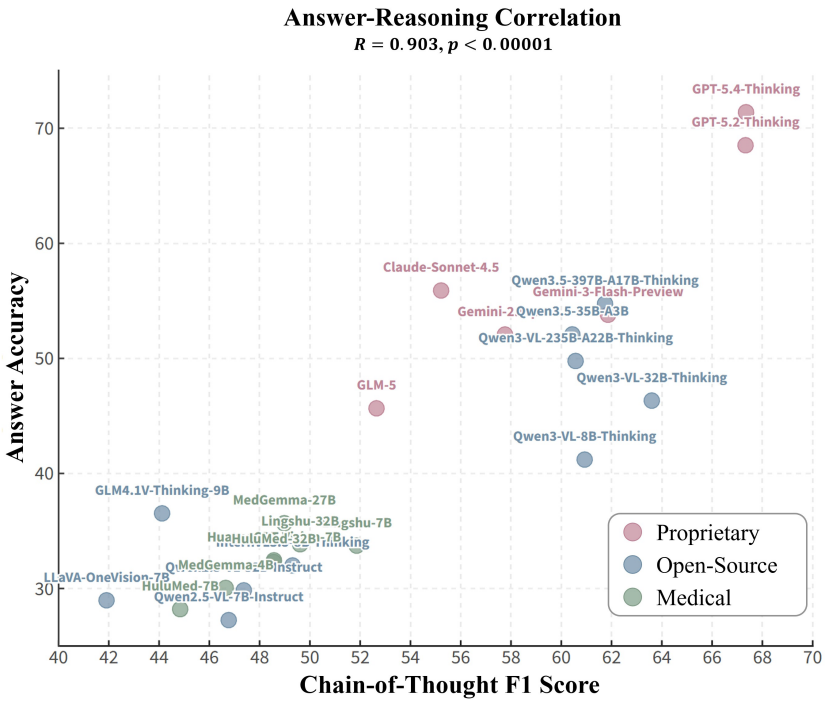


Figure 10: Correlation between semantic reasoning F1 score and answer correctness.

- **Phase-level**: ordering and transitions across Steps must be explicit; emphasize Stage-wise intent shifts (exposure → key dissection → control/removal → closure).
- All time anchors RELATIVE to segment start (00:00:00); monotonic ordering throughout.

QA Generation.

Given a video segment's captions, transcript, and procedure-level summary, the following prompt generates open-ended QAC triplets (Question, Answer, Chain-of-Thought) aligned with the 14 workflow-guided tasks of UniSurgBench.

System Prompt for QAC Generation

You are a highly professional surgical expert specialized in multiple surgical operations.

Task: Generate high-quality, difficult, exam-level Open-Ended questions and detailed step-by-step solutions for a given surgical video clip, to be used in **UniSurgBench** for evaluating multimodal large language models and senior clinicians.

1. Question Generation

- Write 2–5 Open-Ended questions per surgical segment (count reflects spatiotemporal complexity).
- Each question **MUST**: (a) be based on one allowed task from the taxonomy; (b) be strictly constrained to the reference inputs; (c) require multi-step, visual evidence-based reasoning; (d) reference specific time steps from input without including explicit timestamps in the stem.

2. Answer Formulation

- Provide a **Direct Answer**: a concise phrase, specific timestamp, anatomical name, or clear action description.
- The answer must be factually supported by the input evidence.

3. Chain-of-Thought (CoT) Format — use the following EXACT headings in order:

- **Observation:** Cite relevant time anchors and describe visual/temporal evidence.
- **Knowledge:** State ONLY the surgical principles relevant to this case.
- **Assessment:** Assess the surgical procedure, operative target, and risk.
- **Reasoning:** Multi-step reasoning directly referencing Observation and Knowledge.
- **Conclusion:** State the correct answer with evidence-based rationale.

Output Rules: Every factual claim MUST be supported by inputs. Return ONLY valid JSON arrays. Each object MUST contain: `"id"`, `"dimension"`, `"question"`, `"multiple-choice": false`, `"answer"`, `"cot"`.

User Template for QAC Generation

Please analyze the following information and generate the corresponding Question, Answer, and CoT:

1. Video captions: `{Caption}`

2. Transcript ASR:

NOTE: DO NOT follow the timestamps in this part.

`{Subtitle}`

3. Global operative report: `{Background}`

4. Question Taxonomy: `{QuestionType}`

5. Allowed Tasks for THIS Segment: `{AllowedTasksStr}`

Quality Control.

Prior to human-in-the-loop quality control, an automatic key-frame verification step uses Gemini-3-Pro to evaluate each QAC sample against sampled video frames and visual captions, filtering samples with incorrect visual evidence, temporal localization, or unsupported reasoning.

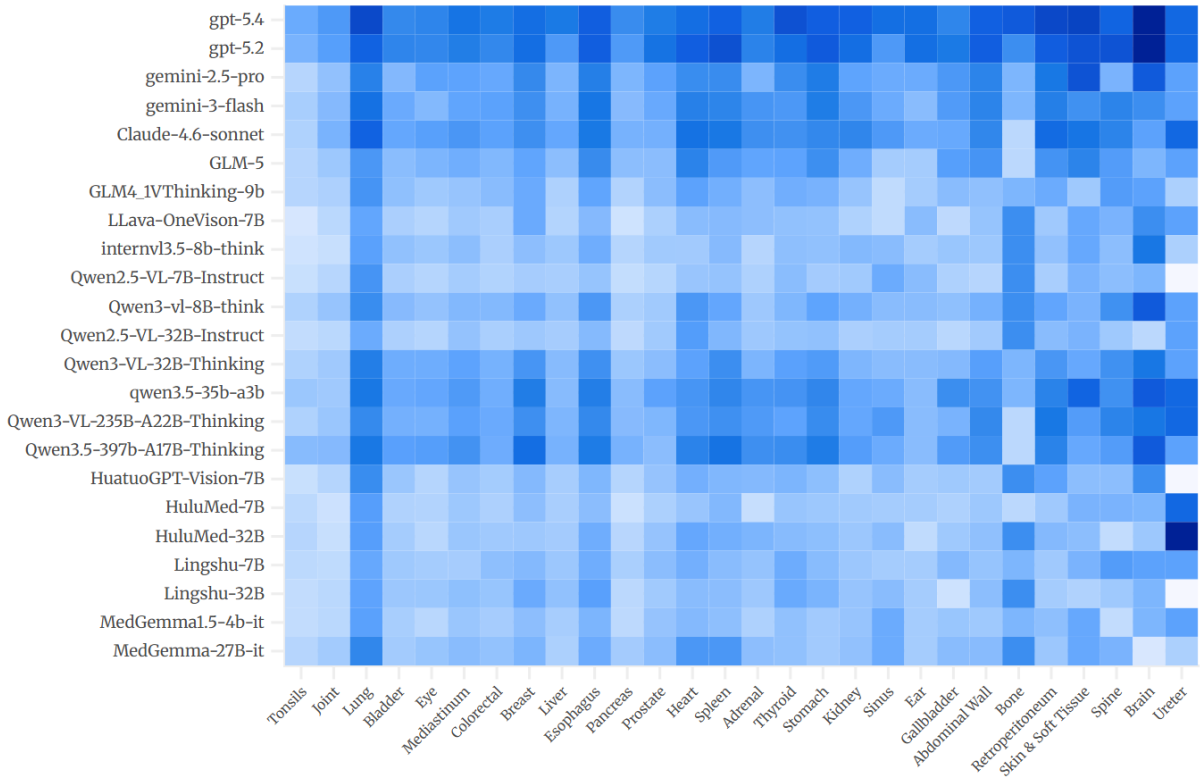


Figure 11: Heatmap of the answer accuracy performance of different models across 29 organ and tissue categories. Each row corresponds to a model and each column corresponds to an organ/tissue category. Darker colors indicate higher scores.

System Prompt for Key-Frame Verification

You are a rigorous, evidence-bound QA evaluator for surgical video Question-Answer pairs.

Task: Evaluate whether the Question, Answer, and Reasoning CoT are supported by the provided evidence and will be **good examples** for UniSurgBench.

Evidence Sources (ONLY):

- **EVIDENCE A (Caption)**: time-aligned text describing visual events and audio background.
- **EVIDENCE B (Key Frames)**: images sampled from the segment clip with RELATIVE timestamps.

Rules: (1) All timestamps RELATIVE to segment start (00:00:00). (2) Do NOT invent missing details; if insufficient, mark as "Unverifiable." (3) Use medical knowledge only to detect clear impossibilities, NOT to fill missing evidence.

Scoring Rubric (1–3):

1. **Factuality**: all core claims Supported (3); minor issues (2); any core claim Contradicted (1).
2. **Visual Grounding**: key claims clearly in frames (3); partial (2); not supported (1).
3. **Temporal Grounding**: actions align with evidence (3); correct gist but weak anchoring (2); wrong order (1).
4. **Answer Uniqueness**: evidence supports Answer as best conclusion (3); competing answers (2); generic (1).
5. **Question Quality**: clinically meaningful, not guessable from common sense (3); partially (2); generic (1).
6. **Professionalism**: correct terminology, objective (3); minor issues (2); incorrect (1).
7. **Surgical Suitability**: clear real surgical scene (3); mostly real with limitations (2); no scene (1).

Rejection Policy: REJECT if any core claim is Contradicted or Unverifiable, or if no real surgical scene is present.

User Template for Key-Frame Verification

[Item to Check]
- Question: `{question}`
- Answer: `{answer}`
- Reasoning CoT: `{cot}`

[EVIDENCE A: Caption] `{caption}`

[EVIDENCE B: Key Frames]
Frames sampled at CoT-cited timestamps (dense) and elsewhere (sparse).
Timestamps RELATIVE to segment start.
Frame timestamps: `{frame_ts_block}`

[Output] Return ONLY a JSON object:

```
{
  "scores": {
    "factuality": 1-3,
    "visual_grounding": 1-3,
    "temporal_grounding": 1-3,
    "answer_uniqueness": 1-3,
    "question_quality": 1-3,
    "professionalism": 1-3,
    "surgical_suitability": 1-3
  },
  "total_score": <sum>,
  "fatal_issues": ["..."],
  "recommendation": "ACCEPT" or "REJECT"
}
```

Verifier.

In the human-in-the-loop quality control stage, GPT-5 acts as the Verifier, scoring each QAC sample on a 9-dimension Likert scale (1–5 per dimension; maximum 45 points). Samples with total scores $\geq$40 are accepted; 35–40 are queued for revision; below 35 are rejected.

System Prompt for the Verifier (Stage 1 Quality Control)

You are the evaluation committee for UniSurgBench quality review.

CRITICAL FIRST CHECK — Surgical Scene Validation: Before any detailed evaluation, determine if frames show a REAL surgical procedure (endoscopic, open, or microscopic field with visible instruments/tissue). If not (animation, slides, non-surgical content), set all scores to 1, `recommendation="REJECT"`, and stop.

Evidence Sources: Question, Answer, Reasoning (Observation / Knowledge / Assessment / Reasoning) and video frames (ordered, with relative timestamps).

9-Dimension Scoring Rubric (1–5 each):

1. **Surgical Scene Validity**: Is a real surgical scene visible?
2. **Question Clarity**: Is the question unambiguous and self-contained within the segment?
3. **Segment Match**: Does the segment provide sufficient visual information for the answer?
4. **Visual Evidence Accuracy**: Are Observations accurately described with correct timestamps?
5. **Clinical Factuality**: Are medical facts and clinical assessments correct?
6. **Causal Chain**: Is the reasoning chain logically complete and causally valid?
7. **Reasoning Acceptability**: Is the overall CoT structured and coherent?
8. **Answer Correctness**: Is the final Answer correctly derived from the reasoning?
9. **Overall Quality**: Overall usability and reliability for UniSurgBench.

Recommendation Rules:

- **ACCEPT**: real surgical scene, all key claims Supported, overall_quality ≥ 4, visual_evidence_accuracy ≥ 4.
- **REVISE**: real scene, fixable issues, can become a quality QAC triplet.
- **REJECT**: no real scene OR key claim Contradicted OR evidence seriously mismatched.

Output JSON:

```
{
  "surgical_scene_valid": true|false,
  "scores": {
    "surgical_scene_validity": 1-5,
    "question_clarity": 1-5,
    "segment_match": 1-5,
    "visual_evidence_accuracy": 1-5,
    "clinical_factuality": 1-5,
    "causal_chain": 1-5,
    "reasoning_acceptability": 1-5,
    "answer_correctness": 1-5,
    "overall_quality": 1-5
  },
  "total_score": 0-45,
  "recommendation": "ACCEPT"|"REVISE"|"REJECT",
  "critical_issues": ["..."]
}
```

Refiner.

Samples flagged for revision are passed to the Refiner (Gemini-3-Pro in Stage 1; GPT-5 in Stage 2 for expert-directed revision), which rewrites the QAC triplet based on the Verifier's structured feedback.

System Prompt for the Refiner

You are the QA Rewrite Editor for UniSurgBench. Improve the overall quality of a QAC triplet (Question, Answer, Chain-of-Thought) using the Verifier's feedback and the original caption as reference.

Sources (ONLY): Original QAC triplet; Verifier feedback (problems, suggestions); Video frames (ordered, with relative timestamps); Original visual caption.

Rewrite Strategy:

1. **Determinability**: if the question is undeterminable, rewrite it as a reliably-judgeable question.
2. **Answer Uniqueness**: write the conclusion best-supported by evidence; add uncertainty qualifiers if needed.
3. **Professional Expression**: mention surgical risk only if warranted by visual evidence.

Reasoning Structure: The `cot` field must preserve the 5-component format:

- **Observation**: cite relevant time anchors and describe visual/temporal evidence.
- **Knowledge**: state ONLY the surgical principles relevant to this case.
- **Assessment**: assess the operative target and risk based on evidence.
- **Reasoning**: multi-step reasoning referencing Observation and Knowledge.
- **Conclusion**: the revised direct answer with evidence-based rationale.

H.2 Prompt Templates for QA Evaluation

The following prompts are used during inference. Models receive uniformly sampled frames together with the question and are instructed to reason step by step. A **Chain-of-Thought (CoT) variant** includes explicit instructions to structure the response following the Observation / Knowledge / Assessment / Reasoning / Conclusion format used in annotation.

Inference Prompt (Standard)

You are a professional surgical video analysis expert.
Answer ONLY based on the provided video.
Question:
`{question}`
Use timestamps `[HH:MM:SS]` to support your analysis.
Please think it step by step.

Inference Prompt (CoT Variant with Frame Legend)

You are a professional surgical video analysis expert.
You will be given `{frame_count}` frames.
Answer ONLY based on the provided frames.
Question:
`{question}`
Use timestamps `[HH:MM:SS]` from the frame legend below. **DO NOT** use "Frame N:" as reference.
Frame timestamps: `{frame_legend}`
Please think step by step.

H.3 Prompt Templates for Evaluation

SurgExpEval consists of two sequential stages: (1) **Structured Prediction Extraction**, which parses free-form model responses into the five structured components; and (2) **Category-wise Semantic Matching**, which scores each predicted item against the corresponding ground-truth annotation.

Structured Prediction Extraction.

Prompt for Structured Prediction Extraction

You are a professional information extraction assistant.

Extract structured content from a surgical video QA reasoning/answer text.

Rules: Output STRICT JSON only. Do NOT add facts, infer, or paraphrase aggressively. Preserve original meaning and order. Empty categories → empty list `[]`. Each Observation MUST include timestamps.

Categories:

- **Observation**: visual evidence with timestamps. Format: `{start, end, evidence}`.

- **Knowledge**: medical knowledge/principles. Format: `{claim}`.
- **Assessment**: situation assessment based on visual evidence. Format: `{claim}`.
- **Reasoning**: reasoning steps connecting evidence to answer. Format: `{claim}`.
- **Conclusion**: final answer. Format: `{answer}`.

Output Schema:

```
{
  "Observation": [
    {"start":"HH:MM:SS","end":"HH:MM:SS",
     "evidence":"..."}
  ],
  "Knowledge":   [{"claim":"..."}],
  "Assessment":  [{"claim":"..."}],
  "Reasoning":   [{"claim":"..."}],
  "Conclusion":  [{"answer":"..."}]
}
```

[Problem]: `{question}`
[Text to Extract]: `{prediction_text}`

Category-wise Semantic Matching.

Prompt for Category-wise Semantic Matching

You are a highly rigorous expert in information evaluation and structured understanding.
Task: Evaluate the matching between Ground Truth (GT) and predicted (Pred) items in a strict *one-to-one* manner for each category.
Evaluation Dimensions:

- **Nouns**: are key entities/objects/concepts correctly matched?
- **Actions**: are main actions/events described accurately?
- **Causality**: are cause–effect relationships and logical dependencies preserved?
- **Temporal Localization**: are start/end times and event order aligned?

Scoring (0–2):
For Observation and Conclusion (temporal-aware):

- **[2]**: timestamps very close (IoU > 0.5) AND semantically equivalent on all key dimensions.
- **[1]**: significant temporal difference with correct semantics, OR correct temporal with notable errors.
- **[0]**: timestamps largely mismatched AND/OR major semantic errors.

For Knowledge, Assessment, Reasoning (semantic-only):

- **[2]**: semantically equivalent; all key nouns, actions, causal links correct.
- **[1]**: core concept correct, minor omissions or vague descriptions.
- **[0]**: wrong/missing key entities, contradiction, or hallucination.

Output (each gt_idx and pred_idx appears at most once per category):

```
{
  "pairs": {
    "Observation": [
      {"gt_idx":0,"pred_idx":1,
       "label":2,"reason":"..."}
    ],
    "Knowledge":  [...],
    "Assessment": [...],
    "Reasoning":  [...],
    "Conclusion": [...]
  }
}
```

[Problem]: `{question}`
[PredJSON]: `{pred_json}`
[GTJSON]: `{gt_json}`

I Case Study

I.1 Case Study on Surgical Element Identification

Figure 12 shows a case of surgical element identification. The ground-truth answer is the left lobe of the liver. The key visual evidence is the smooth and uniformly glistening liver capsule, while the nearby omentum appears as yellow lobulated fat. GPT-5.4-Thinking correctly uses the capsular surface as evidence and identifies the visible structure as the left hepatic lobe. In contrast, Qwen3-VL-32B-Thinking is distracted by the more salient yellow fatty tissue and describes the region as omental fat or gastric serosa. This leads to a wrong final answer, although some local visual descriptions are plausible. The case shows that this task is not only about recognizing common laparoscopic textures. The model also needs to locate which visible structure is being asked about and separate the target anatomy from adjacent tissues.

I.2 Case Study on Surgical Operation Analysis

Figure 13 presents a case that requires recognizing the ongoing surgical action. The correct interpretation is external exposure and mobilization around the gastroesophageal junction and hiatus. The instruments are working on tissue planes outside the esophagus, rather than passing an instrument inside the lumen. GPT-5.4-Thinking correctly describes the main action as traction and release around the hiatal region. It also links the action to the observed instrument–tissue interaction. MedGemma-27B makes a different error. It treats a visible tubular structure as a bougie or dilator entering the esophagus, and then explains the scene as intraluminal calibration. This error starts from a wrong visual identification and then affects the predicted action. The case illustrates why operation analysis is harder than element identification: the model must not only name the instrument or tissue, but also infer how the instrument is interacting with the tissue over time.

I.3 Case Study on Clinical Intent and Scene Reasoning

Figure 14 shows a case where the model must infer the clinical intent behind the current maneuver. The important cue is gentle, capsule-parallel sweeping near the pancreas. The purpose of this maneuver is to expand the dissection plane while avoiding injury to the pancreatic capsule or parenchyma.

GPT-5.4-Thinking gives a well-supported answer by connecting the local dissection pattern with the risk of pancreatic leak or fistula. Qwen3-VL-32B-Thinking also reaches the correct intent, but its temporal evidence is less precise and includes broader segments that are not all equally diagnostic. MedGemma-27B gives a clinically reasonable explanation, but its visual support is weaker. It describes the scene more generally as freeing tissue planes and avoiding vessels, rather than focusing on the pancreatic surface and the specific sweeping motion. This case shows that a model may answer the clinical intent correctly while still giving incomplete visual evidence. Therefore, answer-level accuracy alone can overestimate scene understanding.

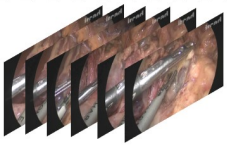
I.4 Case Study on Scenario-Specific Analysis

Figure 15 shows a scenario-specific failure shared by all three models. The visual evidence mainly shows yellow fatty tissue, thin septa, and routine traction. More importantly, the dynamic evidence does not show clear tearing, persistent oozing, or major hemostatic intervention. However, the models interpret the tissue state as more abnormal than it appears. GPT-5.4-Thinking describes the tissue as fibrotic or chronically inflamed. Qwen3-VL-32B-Thinking further links the tissue to adhesions and increased bleeding risk. MedGemma-27B over-infers edematous, neoplastic, or necrotic tissue. These answers are medically plausible, but they are not well supported by the video evidence. The shared error is that the models rely too much on static appearance, such as yellow fat and red areas, and use too little negative temporal evidence, such as smooth separation and the absence of active bleeding. This case highlights a key challenge in scenario-specific surgical reasoning: the model must judge tissue state from both appearance and tissue behavior, rather than turning common laparoscopic textures into high-risk clinical findings.

Table 12: Statistic information of anatomy and surgical procedures in the UniSurgBench.

Organ	Procedure	Count	Organ	Procedure	Count
Colorectal	Colectomy	93	***Pancreas***	Distal Pancreatectomy	49
	Anterior Resection	71		Pancreaticoduodenectomy	43
	Total Mesorectal Excision	67		Pancreatectomy	13
	Subtotal	**273**		**Subtotal**	**105**
Esophagus	Myotomy	117	***Eye***	Ophthalmic Surgery	52
	Esophagectomy	93		Cataract Surgery	40
	Subtotal	**210**		**Subtotal**	**92**
Liver	Liver Resection	129	***Thyroid***	Thyroidectomy	47
	Hepatectomy	52		Lobectomy	42
	Subtotal	**181**		**Subtotal**	**89**
Bladder	Cystectomy	105	***Lung***	Lobectomy	78
	TURBT	70		Pneumonectomy	1
	Subtotal	**175**		**Subtotal**	**79**
Abdominal Wall	Herniorrhaphy	74	***Heart***	Valve Replacement	74
	Hernioplasty	70		**Subtotal**	**74**
	Subtotal	**144**	***Retroperitoneum***	Tumor Resection	66
				Subtotal	**66**
Stomach	Gastrectomy	101	***Kidney***	Nephrectomy	39
	Subtotal Gastrectomy	33		Partial Nephrectomy	24
	Subtotal	**134**		**Subtotal**	**63**
Spleen	Splenectomy	115	***Gallbladder***	Cholecystectomy	58
	Subtotal	**115**		**Subtotal**	**58**
Joint	Arthroscopy	100	***Tonsils***	Tonsillectomy	46
	Arthroplasty	14		**Subtotal**	**46**
	Subtotal	**114**	***Breast***	Mastectomy	20
				Breast Reconstruction	10
				Lumpectomy	7
				Subtotal	**37**
Mediastinum	Thymectomy	96	***Adrenal***	Adrenalectomy	37
	Tumor Resection	11		**Subtotal**	**37**
	Subtotal	**107**	***Ear***	Tympanoplasty	11
				Subtotal	**11**
Prostate	Prostatectomy	28	***Sinus***	Endoscopic Sinus Surgery	11
	Subtotal	**28**		**Subtotal**	**11**
Spine	Spinal Fusion	17	***Brain***	Craniotomy	10
	Subtotal	**17**		**Subtotal**	**10**
Skin & Soft Tissue	Flap Surgery	17	***Bone***	Internal Fixation	5
	Subtotal	**17**		**Subtotal**	**5**
Ureter	Nephroureterectomy	4			
				Total	**2302**





[00:00:00]-[00:00:05]

Task 1: Surgical Element Identification | Anatomy Identification | Gastrectomy

Question: What is the main structure being retracted by the suspension suture with interposed gauze at the beginning of the clip, and what visual cues support this identification ?

CoT:

- Observation: [00:00:00-00:05] whitish monofilament + gauze span upper field, gauze lifts a broad, smooth, shiny tan-brown surface; surrounding inferior tissue is yellow and lobulated (fatty).
- Knowledge: In laparoscopic gastrectomy, **gauze-protected suspension** sutures are commonly used to retract the LEFT HEPATIC LOBE to expose the EGJ / lesser curvature. **Liver capsule = smooth + uniformly glistening; omentum = lobulated yellow fat with septations**; gastric serosa = thinner pale-pink/tan, not a capsular dome being slung.
- Assessment: Configuration (suture + gauze elevating a smooth capsular surface) most strongly supports left-lobe liver retraction.
- Reasoning: The visibly slung tan-brown capsular surface at the top of the field, plus the adjacent but separate yellow lobulated fat, makes "left hepatic lobe" the best-supported identification.

Answer:

The suspension suture with gauze is retracting the LEFT LOBE OF THE LIVER. It elevates a smooth, uniformly glistening tan-brown capsular surface (hepatic capsule), contrasting with adjacent lobulated yellow omental fat; gastric serosa would appear smoother but is not what the gauze-suture is directly lifting.

GPT-5.4-Thinking LLM Judge: Answer: ✓ Visual Perception: ✓ Medical Knowledge: ✓ Reasoning: ✓

The structure being retracted by the suspension suture with interposed gauze at the beginning is the **left lobe of the liver**, not omental fat or gastric serosa.
Visual cues supporting liver identification:
- At the start, the retracted tissue is a **broad, smooth, tan-pink solid organ surface** rather than a loose fatty apron. This is evident in the superior field around **[00:00:01]–[00:00:08].**
- The surface appears **capsular and gently convex**, consistent with hepatic parenchyma under Glisson's capsule, rather than the **lobulated yellow fat** seen in the adjacent omentum on the right side of the field at [00:00:01]–[00:00:10].
- It is being elevated anteriorly/superiorly with a **suspension stitch using gauze as a bolster**, a classic way to retract the **left liver lobe** to expose the hiatus/GE junction. This relationship is clear as the hiatus is progressively exposed beneath it by [00:00:17]–[00:00:31].
Why not omental fat?
- Omental fat is **yellow, lobulated, and floppy**, with frond-like fatty appendages; these features are seen laterally/rightward, not in the retracted structure [00:00:01]–[00:00:10].
- The suspended structure is too **uniform and capsular** to be omentum.
Why not gastric serosa?
- Gastric serosa is **shiny, smooth, and conforms to the rounded stomach wall**, seen inferiorly [00:00:12]–[00:00:21].
- The retracted structure lies **superior/anterior to the hiatus exposure** and behaves like a solid hepatic lobe, not a hollow viscus wall.
So, based on the frames, the suspension suture with gauze is primarily retracting the **left hepatic lobe**.

Qwen3-VL-32B-Thinking LLM Judge: Answer: ✗ Visual Perception: ✓ Medical Knowledge: ✓ Reasoning: ✗

So, let's analyze the video. The question is about the main structure retracted by the suspension suture with interposed gauze at the beginning. First, look at the initial frames. The structure has a distinct texture—**gastric serosa is smooth, omental fat is more lobulated**. The retracted structure here has a **yellowish, lobulated appearance** but wait, no—wait, the key is the suture. At the beginning, the suture is holding a structure. Let's check the timestamps. The first few seconds (0:00-0:05) show the suture. The structure being retracted is the omentum? No, wait. Wait, the video shows a suture with gauze. The structure is the omentum (omental fat) or gastric serosa? Wait, the visual cues: **gastric serosa is smooth, shiny, with visible blood vessels. Omental fat is more yellow, lobulated**. But the question says "as opposed to omental fat or gastric serosa"—so what's it? Wait, the correct answer is the **greater omentum**. But let's check the video. At 0:00, the suture is retracting a structure. The structure's texture: the omentum has a fatty, lobulated appearance. Wait, no—wait, the greater omentum is a fold of peritoneum with fat, so it's yellowish, lobulated. But the question is to distinguish. Wait, the key is the presence of the suture. The main structure retracted is the greater omentum (omental fat). But wait, no—wait, the problem says "as opposed to omental fat or gastric serosa"—so the correct answer is **omental fat**? No, wait, the question is asking what it is, so the answer is the greater omentum. Wait, but let's check the visual cues. The structure is yellowish, lobulated, which is characteristic of omental fat. But the question is to confirm. Wait, the suture is holding the omentum. Let's look at the first frame (0:00). The structure being retracted has a yellowish color, and the surrounding tissue. Also, the gauze is interposed, so the suture is through the omentum. The visual cues: the texture is lobulated, not smooth like gastric serosa (which is smooth, shiny, with visible vessels). Omental fat has a more granular, lobulated look. So the main structure is the greater omentum (omental fat). But wait, the question is to identify what it is. The answer should be **omental fat**. But let's check the timestamps. The beginning of the clip (**0:00-0:05**) shows the suture retraction. The visual cues: the structure is yellow, lobulated, which is omental fat. Gastric serosa would be smooth, with a different texture. So the answer is **omental fat**, but wait, no—wait, the problem says "as opposed to omental fat or gastric serosa", so the correct answer is that it's omental fat. Let's confirm the visual cues. The structure has a fatty, yellowish appearance, which is omental fat. The gastric serosa is a smooth, shiny surface with visible blood vessels, which is not the case here. So the answer is omental fat, with visual cues from the beginning (0:00-0:05) showing the yellow, lobulated tissue.

Error Analysis: The models overcalled fibrosis or edema in a normal loose fatty plane. Routine traction and thin septa were misread as pathological adhesion, leading to inflated estimates of dissection difficulty and bleeding risk despite smooth separation and no visible oozing.

Figure 12: Comparison between GPT-5.4-Thinking and Qwen3-VL-32B-Thinking in Gastrectomy surgery.



Task 2: Surgical Operation Analysis | Surgical Action Recognition | Myotomy

During the stable mid-clip interval ([00:00:12--00:00:18]) while the distal esophagus/GEJ is exposed, what is the dominant operative maneuver being performed: primarily external traction/retraction for exposure, or intraluminal calibration (bougie passage)? State the best-supported choice and cite the instrument–tissue cues visible in the field (and any corroborating context) that indicate direction/purpose.

CoT:

- **Observation:** [00:00:12--00:00:18] Two laparoscopic graspers remain engaged on soft tissue near the hiatus/GEJ and keep the field open with sustained tension and small repositioning movements. No smoke, charring, or flashing consistent with energy activation is seen, and no intraluminal device is directly visible in the operative field during this interval; surrounding tissues show only subtle contour shifts.
- **Knowledge:** In laparoscopic Heller myotomy/hiatal exposure, the common dominant maneuver is atraumatic traction/counter-traction on fundus/omentum or periesophageal tissues to improve visualization. Intraluminal bougie calibration may cause luminal/serosal contour change but is often not directly visible; energy cutting/hemostasis typically shows active tip application with thermal cues (smoke/plume, blanching/charring).
- **Assessment:** The most reliably verifiable action in this clip segment is external retraction for exposure, because the graspers' sustained hold and directional tension are clearly seen, while bougie passage is not directly seen and any luminal effect is subtle/indirect. The absence of thermal cues argues against cutting or coagulation as the dominant maneuver.
- **Reasoning:** Since the clearest, continuous visible action from [00:00:12--00:00:18] is graspers maintaining traction to open the hiatus/GEJ workspace, and there are no definitive visual signs of energy use or a clearly visible bougie in-frame, the best-supported interpretation is that traction/retraction is the dominant maneuver (with any bougie calibration, although mentioned in surrounding context, not the dominant directly evidenced action in this interval).

Answer:

Primarily external traction/retraction to maintain exposure at the distal esophagus/GEJ; the graspers hold and tension soft tissue without visible energy activation, and the bougie itself is not directly visualized (any luminal calibration is only indirectly suggested by subtle contour change).

GPT-5.4-Thinking — LLM Judge: Answer: ✓ Visual Perception: ✓ Medical Knowledge: ✓ Reasoning: ✓

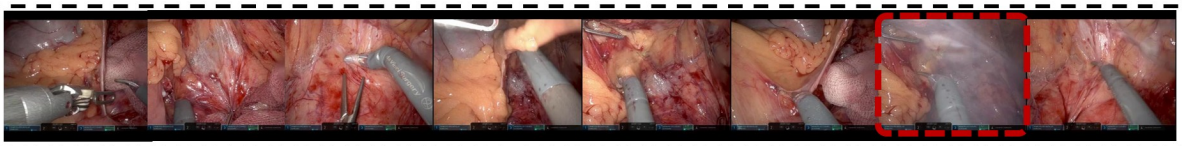
Best-supported choice: primarily external traction/retraction for exposure, not intraluminal calibration (bougie passage).
Why:
- In the key interval **[00:00:12]–[00:00:18]**, the field shows **laparoscopic instruments acting on external tissue planes** around the exposed distal esophagus/GEJ:
  - a **grasper/retractor** is holding and tensioning tissue at the hiatus/GEJ,
  - a**Harmonic device** is applied along the **outer periesophageal/hiatal tissue** and phrenoesophageal attachments for dissection/release,
  - the maneuver direction is **outward/upward lateral traction** to open the window and maintain exposure.
- The visible structure in the center is the **externally seen esophagus/GEJ**, not a smooth cylindrical instrument traversing inside the lumen. There is **no clear bougie tip/shaft entering the esophagus or creating a longitudinal intraluminal bulge used for sizing** during this interval.**[00:00:12]–[00:00:14]**: the hiatus and distal esophagus are exposed while one instrument maintains **traction on the tissue edge** and the Harmonic works on the **white fibrous hiatal/phrenoesophageal tissue**, consistent with external mobilization.**[00:00:15]–[00:00:16]**: the grasper continues to **pull the GEJ region downward/anteriorly**, widening the dissection plane; the energy device stays **outside the lumen**, dividing attachments.**[00:00:17]–[00:00:18]**: exposure remains stable with the same pattern—**retraction plus external dissection**. No dedicated calibration maneuver is seen.
Corroborating context:- After **[00:00:18]**, the view widens and the stomach/fundic surface is manipulated externally, again supporting that the operative focus is **exposure and mobilization of the GEJ/hiatus**, not bougie passage.
So, the dominant maneuver in the stable mid-clip interval is **external traction/retraction for exposure**.

MedGemma-27B — LLM Judge: Answer: ✗ Visual Perception: ✗ Medical Knowledge: ✓ Reasoning: ✗

The user wants me to analyze the provided surgical video frames (specifically the interval [00:00:12--00:00:18]) and determine the dominant operative maneuver. The options are external traction/retraction for exposure or intraluminal calibration (bougie passage).
1. **Identify the relevant time interval:** The interval is from [00:00:12] to [00:00:18]. This corresponds to frames 13 to 19.
2. **Examine the visual information in the frames:**
   * **[00:00:12] (Frame 13):** A tubular instrument (likely a bougie or dilator) is being inserted into the esophagus. The tip is visible entering the distal esophagus/GEJ area.
   * **[00:00:13] (Frame 14):** The tubular instrument continues to be advanced into the esophagus.
   * **[00:00:14] (Frame 15):** The tubular instrument is further advanced.
   * **[00:00:15-00:00:18] (Frame 16):** The tubular instrument is still being advanced.
3. **Analyze the instruments and actions:**
   * **A long, thin, flexible instrument (bougie) is clearly visible being inserted into the lumen of the esophagus.**
   * The instrument is being pushed forward, suggesting calibration or dilation.
   * While there might be some retraction happening elsewhere in the field, the primary focus and action within the distal esophagus/GEJ during this interval is the advancement of the intraluminal instrument.
4. **Conclusion:** **The dominant maneuver during the specified time interval is intraluminal calibration (bougie passage). The instrument-tissue cues (bougie tip entering/advancing within the esophagus) directly indicate this action.**

Error Analysis for MedGemma: The model misidentified a visually prominent tubular structure as a bougie and over-relied on this incorrect visual cue to infer intraluminal calibration. This initial misperception then propagated through the reasoning chain, while insufficient surgical context prevented it from prioritizing the more consistently visible external traction/retraction.

Figure 13: Comparison between GPT-5.4-Thinking and MedGemma-27B in Myotomy surgery.



Task 3: Surgical Phase Understanding | Intent Inference | Anterior Resection

In this mesocolon-to-pancreas/lesser-sac dissection segment, once an initial separation window is created, what is the surgeon trying to do next, and what key complication risk explains the very gentle, capsule-parallel sweeping technique?

CoT:

- **Observation:** In the early portion [00:00:19--00:01:10], an atraumatic grasper maintains steady countertraction while a curved scissors performs repeated low-amplitude, plane-parallel sweeping/peeling motions to extend the same tissue interface rather than targeting a single large structure. A sponge/landmark becomes visible in the opened space by ~[00:00:47], consistent with a created pocket. **Later [00:02:41--00:02:55], there is low-volume oozing that is managed with brief compression/energy application,** without brisk hemorrhage.
- **Knowledge:** After entering the correct mesocolon–pancreas plane to open the lesser sac during splenic flexure mobilization, the standard next step is controlled expansion of that plane using traction–countertraction and short, capsule-parallel strokes, with selective sealing of small bridging vessels. Excess depth or tearing can injure the pancreatic capsule and/or small adherent vessels, leading to troublesome bleeding and potential pancreatic trauma.
- **Assessment:** The repeated, gentle, surface-parallel sweeps under constant traction and the use of spot hemostasis for minor oozing support that the operative goal at this moment is plane enlargement (pocket expansion) rather than division of a named major vessel. The technique appears deliberately constrained to protect an underlying smooth capsular surface consistent with the pancreas.
- **Reasoning:** Because the dissection repeatedly advances along the same opened interface with controlled traction and only addresses small bridging strands when encountered, the most supported intention is to widen the existing lesser-sac/mesocolic plane. Given the capsule-parallel approach and careful management of minor oozing, the dominant risk shaping technique is avoiding pancreatic capsular breach and related bleeding from small adherent vessels that can occur if the surgeon strays deeper or takes thicker bites.

Answer:

The surgeon's immediate intention is to enlarge the already-entered mesocolic–pancreatic interface (lesser sac pocket) by progressively extending the avascular plane with traction–countertraction and short, surface-parallel sweeps while sealing small bridging strands. The guiding risk is inadvertent pancreatic capsular injury (and associated bleeding/ooze from small adherent vessels) from drifting too deep or taking thick bites out of the correct plane.

GPT-5.4-Thinking — LLM Judge: Answer: ✓ Visual Perception: ✓ Medical Knowledge: ✓ Reasoning: ✓

After creating the first small window between mesocolon and pancreas, the surgeon is trying to **extend that plane broadly across the anterior pancreatic surface / into the lesser sac**, essentially "peeling" the mesocolon or fused areolar attachments off the pancreatic capsule to develop a continuous safe retro-mesocolic dissection plane. This is seen as the instrument repeatedly **lifts thin translucent attachments and sweeps parallel to the smooth pale pancreatic surface** rather than diving into tissue **[00:00:21–00:01:17],** then progressively enlarges the exposed area **[00:01:48–00:03:05],** and later continues this mobilization along a wider pancreatic segment [00:03:56–00:05:28].

The reason for the **very gentle, capsule-parallel sweeping is the risk of **pancreatic capsule/parenchymal injury**, which can cause a **pancreatic leak/fistula**** (and associated troublesome bleeding from small adherent vessels). The frames show the surgeon staying tangential to the gland, teasing off flimsy strands instead of using deep bites or aggressive traction, particularly once the smooth pancreatic surface becomes visible [00:00:56–00:01:32], again while working around denser adhesions/fatty bridges [00:02:24–00:03:20], and during continued delicate separation over the gland [00:04:22–00:05:13]. This capsule-preserving motion suggests the goal is to stay on the mesocolic side of the fusion plane and avoid violating the pancreas.

Qwen3-VL-32B-Thinking — LLM Judge: Answer: ✓ Visual Perception: ✓ Medical Knowledge: ✓ Reasoning: ✓

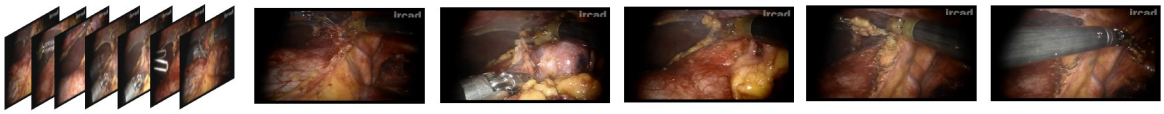
So, let's analyze the video. The question is about what the surgeon does next after creating an initial separation window in mesocolon-to-pancreas/lesser-sac dissection, and the key complication risk for the gentle, capsule-parallel sweeping technique.
First, look at the timestamps. The initial separation window—when does that happen? The video shows the start of dissection. Then, after that, the surgeon is expanding the plane. The technique is gentle, parallel to the capsule. The key risk here is damaging the pancreas, because the pancreas is nearby, and the capsule is delicate. If you go too deep or not parallel, you might injure the pancreas.
Looking at the video: from around **0:23 to 0:34,** the tools are working in the plane. The sweeping motion is parallel to the capsule. The complication risk is pancreatic injury. So the next step is to expand the dissection plane along the pancreas, and the risk is avoiding pancreatic injury.Now, make sure to use timestamps from the video. The question says "what is the surgeon trying to do next" and the technique. So the next step is to extend the dissection plane. The key complication risk is pancreatic injury.
Check the video: in the dissection, the surgeon is creating a wider plane by sweeping parallel to the capsule. The risk is that if you go too deep, you hit the pancreas. So the answer is:
[0:45-1:00] Expand the dissection plane along the pancreas; key risk: Pancreatic injury (to avoid capsule violation).

MedGemma-27B — LLM Judge: Answer: ✓ Visual Perception: ✗ Medical Knowledge: ✓ Reasoning: ✓

Step-by-step analysis using timestamps: * **Initial separation:** The sequence shows the surgeon using an instrument (likely a blunt dissector or grasper) to sweep between tissue layers. **Gentle technique:** The movements are deliberate and controlled, not forceful. This is evident in the way the instrument is moved along the tissue surface. **Goal:** The surgeon is attempting to create a space or plane between the mesocolon and the underlying structures (pancreas or lesser sac). **Risk:** The proximity to the pancreas and major vessels necessitates a cautious approach. Injury to the pancreas can lead to pancreatitis, fistula formation, or pseudocyst development. Injury to major vessels can cause significant hemorrhage. **Reason for technique:** The gentle, sweeping motion aims to separate the tissue layers without tearing or perforating structures, thereby minimizing the risk of injury to the pancreas or vessels.
Conclusion: After creating an initial separation window in the mesocolon-to-pancreas/lesser-sac dissection, **the surgeon is trying to progressively free the mesocolon by carefully delineating the tissue planes. The gentle, capsule-parallel sweeping technique is employed to minimize the risk of injuring the delicate pancreas or major blood vessels located in close proximity.**

Error Analysis: The visual evidence are not enough in the high-risk scenario, the energy activate is important for the prediction of surgical intent.

Figure 14: Comparison between GPT-5.4-Thinking, Qwen3-VL-32B-Thinking, and MedGemma-27B in Anterior Resection surgery.



Task 4: Surgical Procedure Reasoning | Pathology Inference | Thymectomy

What does the tissue appearance and behavior suggest about the local tissue state (edema, fibrosis, or normal loose plane), and how does this influence dissection difficulty and bleeding risk in this step?

CoT:

Observation: The field shows organized yellow adipose lobules with thin fibrous septa and pale pink–tan pericardial serosa, with no charring, eschar, active bleeding, or coagulative changes; tissue separates under gentle traction and sweeping across [00:00:06–00:00:24].
Knowledge: Loose areolar/fatty planes separate readily with minimal traction and minimal bleeding, whereas fibrosis shows dense adherence and tearing; edema/congestion presents with friability and oozing.
Assessment: The mechanical maneuvers achieve separation without tearing or hemostatic intervention; no oozing is seen.
Reasoning: The ease of plane development and clean field argue against fibrosis or edema; the low vascular response supports a normal loose plane.

Answer:

Findings are consistent with a normal loose plane: organized yellow adipose with thin septa, lightly vascular pericardial serosa, and easy mechanical separation without oozing. This favors lower dissection difficulty and minimal bleeding risk during mechanical clearance.

GPT-5.4-Thinking LLM Judge: Answer: ✗ Visual Perception: ✗ Medical Knowledge: ✗ Reasoning: ✗

The tissue in this step appears to be **predominantly **fibrotic with some chronic inflammatory change**,** rather than a normal loose areolar plane or marked edema. - At **[00:00:07]–[00:00:14]**, the tissues look **dense, whitish, and tethered**, with strands that do not separate easily under traction. Instead of opening into a clear avascular plane, the tissue stretches and remains adherent. - At **[00:00:18]–[00:00:23]**, traction-countertraction exposes a **thickened band-like layer** that requires energy-based peeling rather than simple blunt spreading, again suggesting fibrosis. - At **[00:00:31]–[00:00:40]**, the dissection proceeds along a **scarred, irregular surface** with charred spots from repeated cautery, consistent with a difficult adherent plane rather than normal loose tissue. - At **[00:00:43]–[00:00:54]**, the fatty tissue is **stuck and cohesive**, coming off in chunks with energy application instead of falling away in a soft edematous fashion. - By **[00:00:55]–[00:01:00]**, the plane remains **poorly defined and tenacious**, again favoring fibrosis/chronic adhesions. **This does **not strongly suggest edema**, because the tissue is not markedly boggy, translucent, or easily split by blunt dissection. It also does **not look like a normal loose plane**,** since there is no effortless areolar separation.
Implication for dissection difficulty:
This local state implies **increased dissection difficulty**. Fibrotic tissue obscures the natural plane, requires more sharp/energy dissection, and reduces tissue mobility.
Implication for bleeding risk:
The bleeding risk is ****moderately increased****. Although there is no major active hemorrhage seen in these frames, fibrotic inflamed tissue tends to bleed from small raw surfaces and makes vessel identification/control less straightforward. So this step appears **technically demanding with higher risk of oozing than a normal loose plane**, but not massively hypervascular or edematous.

Qwen3-VL-32B-Thinking LLM Judge: Answer: ✗ Visual Perception: ✗ Medical Knowledge ✗ Reasoning: ✗

So, let's analyze the video. First, look at the tissue appearance. **The tissue has a yellowish, fatty look with some red areas (vascular).** But there's also some fibrous texture. Let's check the dissection. **The tools are working on tissue that seems to have some adhesions or fibrosis.** Wait, in the video, the tissue is being manipulated, and there's some resistance. Also, the presence of dense, yellowish tissue with some scarring might indicate fibrosis. Edema would be more swollen, but here it's more firm. Normal loose plane is easier to dissect.
Looking at timestamps: from 0:00 to 0:10, the tissue being dissected has **a tough, fibrous texture**. The instruments have to work a bit to separate, which suggests fibrosis. Fibrosis makes dissection harder because the tissue is dense and less pliable. Bleeding risk: in fibrotic tissue, there might be more vessels due to the scarring, **so higher risk.** Let's confirm. At 0:00-0:15, the tissue is being pulled and cut; it's not easy, so **fibrosis**. The tissue isn't edematous (no excessive fluid), and not normal loose (which would be easy to separate). So local state is fibrosis. Influence: dissection difficulty is high because fibrotic tissue is dense. **Bleeding risk is elevated because fibrosis can involve more vascular structures.** The Final answer is: Fibrosis **(0:00-0:30)** evidenced by dense, yellowish, and adherent tissue with resistance to dissection. This increases dissection difficulty **(0:10-0:45)** and elevates bleeding risk due to dense vascularization within scarred tissue.

MedGemma-27B LLM Judge: Answer: ✗ Visual Perception: ✗ Medical Knowledge: ✗ Reasoning: ✗

The provided video frames show a surgical dissection procedure, likely involving the **removal of a tumor or mass.** The tissue appears to be a **yellowish, somewhat friable, and edematous tissue, possibly representing neoplastic tissue or tissue undergoing necrosis. The dissection plane is not clearly defined**, and the tissue seems adherent in places, suggesting fibrosis or inflammation. Here's a breakdown based on the provided frames: 1. **Tissue Appearance:** The tissue being dissected has a pale yellow or tan color, which can be indicative of edema, ischemia, or necrosis, often associated with tumors or inflammation. **It appears somewhat soft and potentially friable.** [00:00:00] - [00:00:58] 2. **Behavior During Dissection:** The tissue does not seem to separate easily along clear planes. The dissection tool (forceps or similar) often requires significant traction and pressure to separate the tissue from the surrounding structures. [00:00:00] - [00:00:58] 3. **Local Tissue State:** The combination of the tissue's appearance (pale yellow, possibly friable) and its adherence during dissection suggests it is not a normal, loose plane. It is likely eitherz: Edematous:** Swollen tissue due to fluid accumulation, which can make it friable and difficult to dissect cleanly. Fibrotic:** Scar tissue formation can cause tissues to become dense and adherent, making dissection difficult and increasing the risk of tearing. **Neoplastic/Inflamed:** Tumor tissue itself** or tissue involved in an inflammatory process can exhibit abnormal adherence and friability. Based on the appearance and behavior, the tissue state is likely a combination of edema and potential fibrosis or neoplastic involvement, making it difficult to dissect with clear margins. Influence on Dissection Difficulty:** The lack of a clear plane and the tissue's adherence significantly increase the difficulty of dissection. The surgeon must be careful to avoid tearing the tissue while ensuring complete removal. This requires slow, deliberate movements and potentially the use of different instruments (e.g., blunt dissection, sharp dissection with caution). 5. **Influence on Bleeding Risk:** **The tissue appears somewhat vascularized (reddish areas visible in some frames), and its friability suggests that it may bleed easily upon manipulation or tearing. Dissection in this type of tissue requires careful hemostasis to control bleeding.** The dissection process itself can potentially damage small vessels within the tissue or adjacent structures, increasing bleeding risk.
In summary, the tissue appears edematous and potentially fibrotic or neoplastic, resulting in poor definition of dissection planes, increased dissection difficulty, and a higher risk of bleeding due to friability and potential vascularity.

All three models failed at the initial tissue-state recognition step. They overinterpreted yellow fatty tissue, thin septa, and routine traction as pathological adhesion or edema, while underusing dynamic negative evidence such as smooth separation, no tearing, no oozing, and no hemostatic intervention. As a result, the downstream reasoning consistently shifted toward higher dissection difficulty and bleeding risk.

Figure 15: Comparison between GPT-5.4-Thinking, Qwen3-VL-32B-Thinking, and MedGemma-27B in Thymectomy surgery.