TOWARDS HIGHER-LEVEL SURGICAL INTELLIGENCE: A COMPREHENSIVE BENCHMARK AND MULTI-AGENT FRAMEWORK FOR SURGICAL INTENT UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years, surgical video understanding has progressed from instrument, action, and phase recognition to multimodal video question answering. However, to achieve higher-level surgical intelligence that can assist surgical maneuvers and support clinical decision-making, models need to understand the clinical intent underlying surgical actions: why an ongoing maneuver is performed and which subsequent surgical objective it prepares for. Therefore, we introduce the **Surgical Intent Understanding** task for endoscopic surgical videos, which jointly evaluates four clinical capabilities: anatomical localization, interpretation of instrument roles, interpretation of instrument coordination, and inference of subsequent surgical objectives. In consultation with experienced surgeons, we define four corresponding evaluation dimensions and construct the **SurgIntent-VidQA** dataset, comprising 603 video clips and 2,520 surgeon-annotated and reviewed five-option, single-answer questions across four procedure types in liver and lung surgery. Based on this dataset, we establish **SurgIntent-Bench** for systematic evaluation. We further propose **SurgIntent-Agent**, a knowledge-guided multi-agent framework that generates candidate intent hypotheses using surgical knowledge, acquires video evidence through targeted questions, and iteratively verifies and refines its predictions. We train the agents using supervised fine-tuning and direct preference optimization. Systematic evaluation on SurgIntent-Bench covers mainstream general-purpose, open-source video, medical, and task-specific fine-tuned multimodal large language models. SurgIntent-Agent achieves an overall accuracy of 67.62%, outperforming the strongest general-purpose model (61.32%) and task-specific SFT baseline (63.72%) by 6.30 and 3.90 percentage points, respectively. Ablation studies further support the contributions of surgical knowledge, visual evidence extraction, and preference optimization.

1 INTRODUCTION

Surgical intelligence aims to enable computational systems to perceive surgical scenes, understand operative processes, and support clinical decision-making. Surgical video analysis has progressed from instrument and anatomy recognition to action and phase recognition (Twinanda et al., 2017; Nwoye et al., 2022), while Multimodal Large Language Models (MLLMs) (Wu et al., 2023) have further extended these capabilities to open-ended surgical video question answering (Li et al., 2024; Wang et al., 2025a). As illustrated in Figure 1, LLaVA-Surg (Li et al., 2024) organizes surgical video understanding into four levels: object (Level 1), action (Level 2), reason (Level 3), and plan (Level 4). Most existing studies focus on the foundational understanding addressed by Levels 1 and 2 (Hao et al., 2025; Perez et al., 2026). However, to further assist surgeons in performing surgical maneuvers, support clinical decision-making, and plan subsequent actions, models need to understand the clinical intent underlying surgical maneuvers. This entails explaining why an ongoing maneuver is performed and inferring the subsequent surgical objective it prepares for, corresponding to Levels 3 and 4. For example, Level 2 in Figure 1 identifies an ultrasonic scalpel cutting a ligamentous attachment. Level 3 explains that the maneuver aims to limit bleeding and mobilize the left liver. Level 4 infers the subsequent objective of further mobilizing the left liver and exposing

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

its lateral surface. Such understanding is important for autonomous surgery (Kim et al., 2025) and surgical education (DaRosa et al., 2008), yet research on these higher levels remains at an early stage (Li et al., 2024; Perez et al., 2026).

To enable models to understand the intent behind surgical maneuvers, we first need to identify which capabilities of surgeons these models need to learn. We therefore examine how surgeons interpret and perform surgical maneuvers from a clinical perspective. In surgical practice, surgeons typically need to identify specific anatomical landmarks and clarify the spatial relationships among relevant structures before performing many maneuvers (Strasberg & Brunt, 2010). These anatomical cues help surgeons determine the operative site and target, and relate the maneuver to its objective within the ongoing procedure (Cristancho et al., 2013). Surgeons must also understand instrument functions and the specific role each instrument serves in the current scene. When multiple instruments participate in a maneuver, surgeons need to understand how they work together and what shared objective their coordination serves. They must further determine which subsequent surgical objective the ongoing maneuver prepares for. In consultation with practicing surgeons, we identify four interrelated capabilities underlying these clinical judgments: anatomical localization, interpretation of instrument roles, interpretation of instrument coordination, and inference of subsequent surgical objectives.

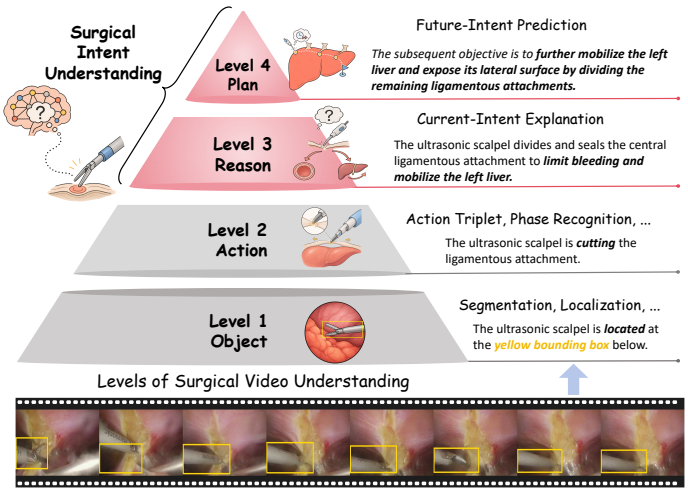


Figure 1: **Levels of surgical video understanding.** Levels 1 and 2 localize visible entities and recognize instrument–action–target relations, whereas Levels 3 and 4 explain the current clinical purpose and predict the subsequent surgical objective. We define Levels 3 and 4 as *Surgical Intent Understanding*. The hierarchy follows LLaVA-Surg (Li et al., 2024).

To systematically assess these capabilities in models, we introduce the **Surgical Intent Understanding** task. We define this task as using surgical video, clinical context, and relevant surgical knowledge to explain the clinical purposes of ongoing maneuvers and infer the subsequent surgical objectives they prepare for. We define four corresponding evaluation dimensions (Figure 2): **(1) Anatomical Intent Understanding** identifies the anatomical structures, landmarks, and spatial relationships relevant to the current maneuver, providing the anatomical grounding for interpreting its purpose. **(2) Functional Intent Understanding** explains the clinical purpose that an individual instrument is intended to accomplish within the current surgical stage and specific scene. **(3) Coordinative Intent Understanding** explains the shared surgical objective served by the coordinated roles of multiple instruments. **(4) Prospective Intent Understanding** infers the subsequent surgical objective that the ongoing maneuver prepares for. Together, these complementary dimensions assess whether models can connect anatomical evidence, instrument roles, and surgical objectives to understand the clinical intent underlying visible maneuvers.

Studying Surgical Intent Understanding presents two main challenges. First, reliable intent annotations require specialized clinical knowledge and surgical experience. Surgeons must interpret and verify current purposes and subsequent objectives by considering video content, anatomical relationships, and procedural context. The need for expert interpretation and review makes constructing reliable datasets demanding. Second, models must combine surgical knowledge with video observations to understand why surgeons perform a maneuver at a particular stage. Surgical knowledge provides the clinical rationale: why a step is performed, how it serves the current objective, and what conditions it creates for subsequent steps. Video observations reveal the current anatomy, instrument actions, and operative progress, allowing models to determine how this knowledge applies to the specific surgical scene.

To support systematic investigation of this task, we construct the **SurgIntent-VidQA** dataset, comprising 603 video clips and 2,520 question–answer pairs across the four evaluation dimensions. The dataset covers four procedure types in liver and lung surgery, using hospital recordings and online surgical teaching videos. Five surgeons authored and cross-checked the questions, reference answers, and clinical contexts, with two senior surgeons conducting the final review. Each question provides five answer options with a single correct answer supported by the video and its clinical context. The remaining four are clinically plausible distractors that do not fit the specific video context. Building on this dataset, we establish **SurgIntent-Bench**, which uses a standardized evaluation protocol to assess overall performance and performance across the four dimensions. We also curate a surgical knowledge base from surgical atlases, medical textbooks, and related medical resources to support interpretation of the clinical rationale behind maneuvers.

To help models connect surgical knowledge with the current scene, we propose **SurgIntent-Agent**, a knowledge-guided multi-agent framework (Figure 3). The framework first retrieves relevant surgical knowledge using a video description. The Hypothesis Agent (HA) then combines the video with retrieved knowledge to form candidate intent hypotheses. The Surgical Intent Reasoning Agent (SIRA) formulates visual questions that distinguish these hypotheses, while the Visual Evidence Agent (VEA) inspects the video and returns relevant observations. SIRA uses this evidence to decide whether to retain or revise the current answer and whether to continue probing or stop. We apply supervised fine-tuning (SFT) (Zhang et al., 2026) to VEA and HA to develop their capabilities for visual evidence extraction and hypothesis initialization, respectively. Using a separate training subset, we construct preference data and train SIRA through direct preference optimization (DPO) (Rafailov et al., 2023), targeting discriminative questioning and decisions supported by evidence.

Overall, our contributions are as follows:

- We introduce, to the best of our knowledge, the first Surgical Intent Understanding task for endoscopic surgical videos, jointly evaluating four clinical capabilities: intent-relevant anatomical localization, interpretation of instrument roles, interpretation of instrument coordination, and inference of subsequent surgical objectives. Together, these dimensions assess models' understanding of current operative purposes and subsequent objectives.
- We construct SurgIntent-VidQA, an expert-annotated dataset comprising 603 video clips and 2,520 five-option, single-answer questions across four procedure types in liver and lung surgery. We also establish SurgIntent-Bench and a supporting surgical knowledge base, providing dedicated data, knowledge, and evaluation resources.
- We design SurgIntent-Agent, a knowledge-guided multi-agent framework that connects candidate intent generation, targeted visual questioning, and evidence assessment to iteratively verify and refine intent predictions. We use SFT to train visual evidence extraction and hypothesis initialization, and DPO to train the reasoning agent's questioning and decision-making capabilities.
- We systematically evaluate representative multimodal models and analyze their performance across the four intent dimensions. SurgIntent-Agent achieves an overall accuracy of 67.62%, outperforming the strongest task-specific SFT baseline by 3.90 percentage points. Ablation studies further support the contributions of surgical knowledge, visual evidence extraction, and preference optimization within the framework.

2 Related Work

Surgical video understanding. Surgical video understanding has expanded from instrument and phase recognition (Twinanda et al., 2017; Gao et al., 2021) and instrument–action–target triplet recognition (Nwoye et al., 2022) to multimodal question answering over surgical images and videos (Li et al., 2024; Jin & Jeong, 2024; Li et al., 2025). More recent work explores fine-grained temporal video understanding (Wang et al., 2025a) and extends evaluation to surgical reasoning, forecasting, and clinical intent understanding (Li et al., 2024; Perez et al., 2026).

Agent-based video reasoning. Video agents use task decomposition, specialized roles, and visual tools to gather evidence (Wang et al., 2025c; Shang et al., 2024; Shi et al., 2025), while feedback,

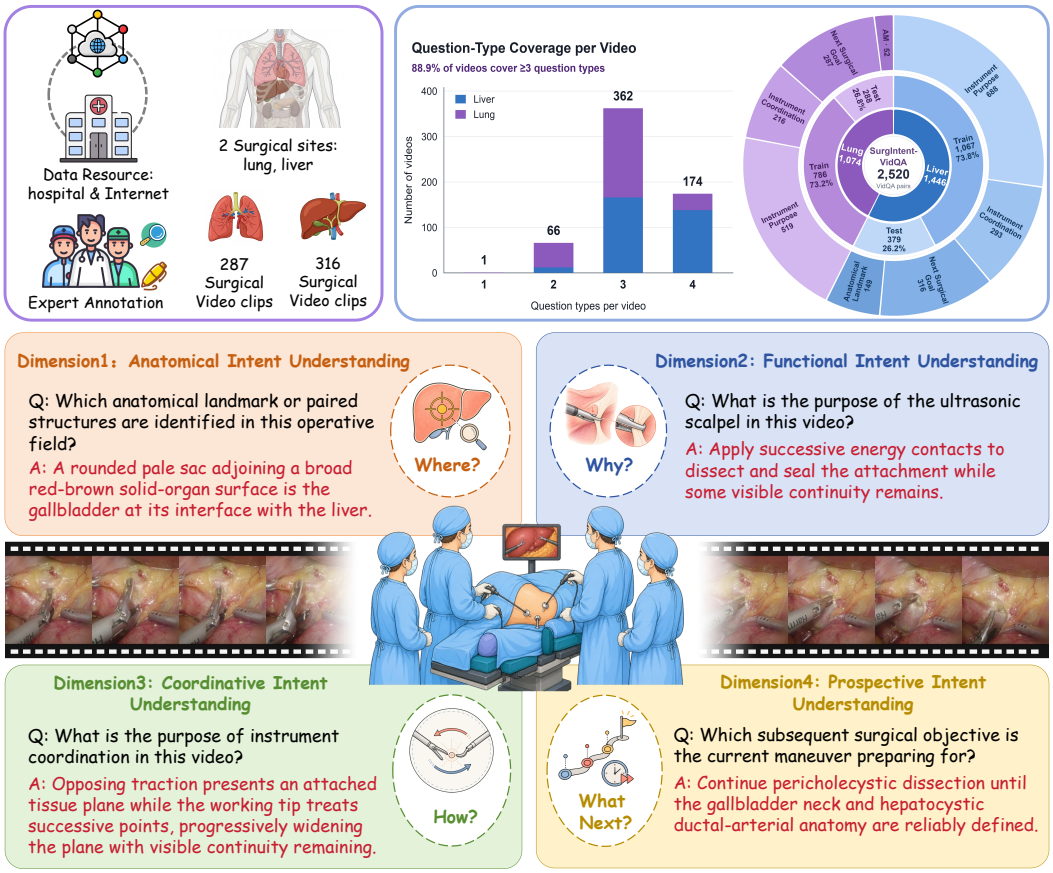


Figure 2: **Overview of SurgIntent-VidQA.** Top left: data sources and expert annotation, covering 603 clips from liver (316) and lung (287) procedures. Top middle: question-type coverage; 88.9% of clips cover at least three intent dimensions. Top right: distribution of 2,520 question–answer pairs across surgical sites, data splits, and question types. Bottom: examples from the same clip illustrate Anatomical, Functional, Coordinative, and Prospective Intent Understanding through anatomical context, instrument purpose, coordination goals, and subsequent surgical objectives, respectively. All questions are five-option, single-answer; only questions and reference answers are shown.

reflection, and verification support iterative reasoning (Zhou et al., 2025; Wang et al., 2026). Surgical intent inference requires interpreting this evidence through knowledge of anatomy, instrument functions, and surgical workflows. SurgIntent-Agent uses retrieved surgical knowledge to guide candidate intent generation and targeted visual questioning to determine which interpretation is supported by the current scene.

Preference optimization for multimodal reasoning. Direct preference optimization (DPO) learns from preferred and rejected responses without a separate reward model (Rafailov et al., 2023), and has been applied to video question answering using caption-supported feedback (Zhang et al., 2025b). We use SFT for visual evidence extraction in VEA and hypothesis initialization in HA, and DPO to train SIRA for discriminative visual probing and evidence-based decisions on answer revision, continued probing, and stopping.

3 Surgical Intent Understanding

3.1 Task Definition

We define Surgical Intent Understanding as explaining the clinical purposes of ongoing maneuvers and predicting the subsequent surgical objectives they prepare for, based on surgical video, clinical

context, and surgical knowledge. The task comprises four complementary dimensions: Anatomical, Functional, Coordinative, and Prospective Intent Understanding. Figure 2 illustrates questions and reference answers for these dimensions using the same surgical clip.

We formulate the task as five-option, single-answer multiple-choice question answering. Each instance comprises a video clip V, surgeon-written clinical context C, a question q, and candidate answers $\mathcal{A} = \{a_1, \ldots, a_5\}$. The model selects the unique correct answer supported by the video evidence and clinical context. Because short clips may not capture the full operative context, we provide question-specific clinical context to supply background information necessary for understanding the current maneuver but not available within the clip itself.

3.2 DATASET

We construct SurgIntent-VidQA, comprising 603 video clips and 2,520 question–answer pairs covering all four intent dimensions across four procedure types involving the liver and lung. We partition the dataset into training and test sets containing 1,853 and 667 questions, respectively, with no overlap between their video clips. All questions associated with the same clip are assigned to the same split. Figure 2 summarizes the dataset composition, question-type coverage, and examples.

Data Collection. We collect surgical recordings from hospitals and publicly available online teaching resources. Surgeons select clinically meaningful operative steps and manually extract the corresponding clips for annotation. We also curate a textual surgical knowledge base from surgical atlases, medical textbooks, and online medical resources, covering procedure-specific operative knowledge, general surgical principles, and anatomical and medical concepts.

Data Annotation and Quality Control. Five surgeons from multiple hospitals annotate questions, reference answers, and question-specific clinical contexts, followed by cross-checking within the team. Two chief physicians conduct the final review. For each question, four distractors are drawn from surgeon-annotated answers to questions of the same type in other videos, making them clinically plausible but inappropriate for the current video. GPT-5.6-sol assists only with balancing correct-answer positions across options A–E, followed by surgeon review. We control differences in option length, phrasing, clinical terminology, and anatomical specificity to reduce the possibility of answering through textual cues or answer-position biases.

Reviewers inspect multiple video frames to exclude semantically redundant or simultaneously valid options, ensuring that each question has a unique correct answer supported by the video. Disputed annotations are revised through expert discussion, and samples with unresolved disagreement or insufficient visual evidence are excluded. We also check whether clinical contexts directly reveal the answer or allow the question to be answered without viewing the video. Detailed annotation rules and review procedures are provided in Appendix A.2.

4 METHOD

Surgical Intent Understanding requires connecting visual observations of instrument interactions, anatomical relationships, and tissue changes with surgical knowledge about operative purposes. Similar actions may serve different clinical objectives, while an explanation consistent with general surgical knowledge may not apply to the current video. Models therefore need to use surgical knowledge to form plausible intent hypotheses, actively seek visual evidence that distinguishes these hypotheses, and examine and revise their judgments accordingly.

Following this principle, we propose **SurgIntent-Agent**, a multi-agent framework assigning visual observation, knowledge-guided hypothesis formation, and evidence assessment to specialized agents connected through active probing. As illustrated in Figure 3, the Hypothesis Agent (HA) combines the video with retrieved knowledge to establish an initial intent judgment. The Surgical Intent Reasoning Agent (SIRA) formulates visual questions that distinguish competing hypotheses, and the Visual Evidence Agent (VEA) returns the corresponding observations. SIRA then updates its answer and decides whether to continue probing or stop. We perform SFT to develop visual evidence extraction and hypothesis initialization, followed by DPO for probing and decision-making.

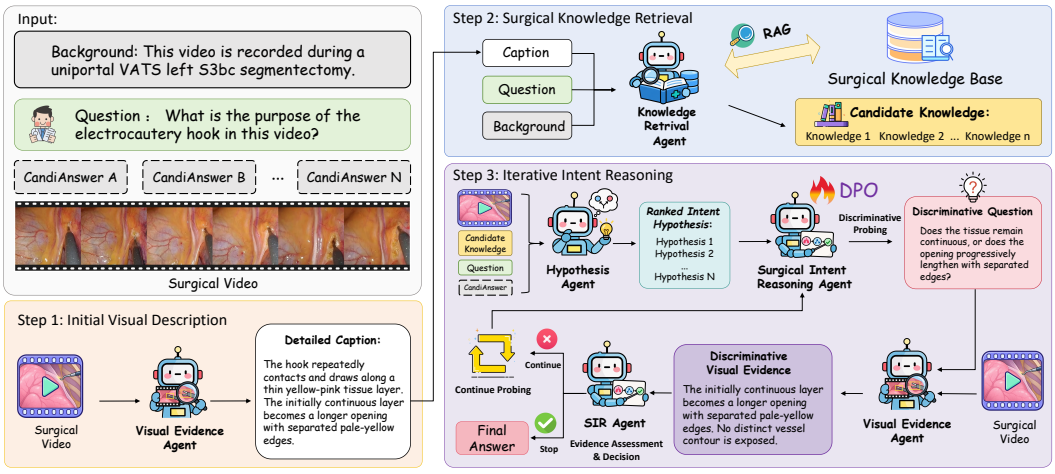


Figure 3: **Overview of SurgIntent-Agent.** (1) The Visual Evidence Agent (VEA) generates an initial video description. (2) The Knowledge Retrieval Agent uses this description, the question, and clinical context to retrieve candidate knowledge from the surgical knowledge base. (3) The Hypothesis Agent combines the video, question, answer options, and retrieved knowledge to rank candidate intents. The Surgical Intent Reasoning Agent (SIRA) poses discriminative visual questions and uses observations returned by VEA to revise its answer and decide whether to continue probing or stop with a final answer. DPO trains SIRA's probing and evidence-assessment actions.

4.1 Knowledge-Guided Intent Hypothesis Formation

Given video V, question q, clinical context C, and candidate answers $\mathcal{A} = \{a_i\}_{i=1}^{M}$ with $M = 5$, VEA generates a visual description d for retrieval from the surgical knowledge base $\mathcal{B}$:

$$K = \mathcal{R}(d, q, C; \mathcal{B}), \tag{1}$$

where $\mathcal{R}$ retrieves instrument functions, anatomical relationships, and operative principles to support clinical interpretation of the observations.

Each candidate answer represents an intent hypothesis. HA directly receives V, q, C, $\mathcal{A}$, and K to score these hypotheses. Let z_i be the logit of option a_i's label token at the answer position. The highest-scoring option is

$$i_0 = \arg\max_{1 \le i \le M} z_i. \tag{2}$$

The initial answer is $y_0 = a_{i_0}$, and descending scores define the initial ranking r_0. All candidates and their ranking remain fixed, while new visual evidence may update the current answer.

4.2 Active Evidence Acquisition and Intent Refinement

SIRA iteratively seeks evidence that distinguishes competing hypotheses. Its state at the start of round t is

$$s_t = (q, C, \mathcal{A}, K, r_0, y_t, H_t, T_V, b_t)\,. \tag{3}$$

where y_t is the current answer, $\mathcal{H}_t$ the preceding interaction history, T_V the video duration, and b_t the remaining interaction rounds. SIRA accesses video evidence only through VEA, receiving neither the video nor the initial description d.

Visual probing. SIRA generates a visual question u_t and frame count n_t:

$$(u_t, n_t) = f_\theta^{\text{probe}}(s_t). \tag{4}$$

VEA returns observations from the sampled frames:

$$e_t = f_{\text{VEA}}(\mathcal{S}(V, n_t), u_t)\,, \tag{5}$$

where $\mathcal{S}(V, n_t)$ uniformly samples n_t video frames and e_t denotes the observations. Questions target observable instrument interactions, anatomical relationships, or tissue changes, rather than

intent predictions. VEA receives only the sampled frames and question, without candidate-answer mappings or reasoning history.

Evidence assessment. SIRA uses the observations to update its answer and continuation decision:

$$(y_{t+1}, \delta_t, \rho_t) = f_\theta^{\text{assess}}(s_t, u_t, n_t, e_t), \tag{6}$$

where $\delta_t \in \{\text{continue}, \text{stop}\}$ and ρ_t contains the assessment and supporting VEA references. Both f_θ^{probe} and f_θ^{assess} denote generation and output parsing by the same SIRA model under different prompts, sharing parameters θ.

Revisions should rely on evidence distinguishing the current answer from an alternative, rather than ambiguous observations compatible with multiple hypotheses. Programmatic checks validate action format and consistency, enforce budgets, and verify reference traceability; SIRA judges semantic support. Each round updates the history and remaining budget. The system returns the current answer upon stopping or budget exhaustion. Full interaction rules appear in the Appendix.

4.3 Two-Stage Agent Training

Supervised fine-tuning. We separately fine-tune VEA and HA: VEA learns to extract question-relevant evidence through video-based visual question answering, while HA maximizes correct-answer sequence likelihood given video and knowledge. HA's option-label scores initialize the candidate ranking and answer at inference.

Local-action preference optimization. We construct preference pairs (x, α^+, α^-) from a separate training subset, comparing preferred and rejected actions under the same complete input x. Inputs contain the task prompt and current state; assessment inputs additionally include the current visual question and returned evidence. Candidate questions are generated through greedy and sampling-based decoding and executed by frozen VEA; candidate assessments share the same question and evidence. A teacher model reviews actions under a fixed protocol and supplies replacements when needed. Probing preferences favor visually answerable, discriminative questions. Assessment preferences favor retaining answers under insufficient evidence and revising them when discriminative observations support an alternative. Ground truth is used only for offline review, never as SIRA input; final-answer correctness alone does not determine preference. We optimize SIRA using DPO (Rafailov et al., 2023):

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x,\alpha^+,\alpha^-)\sim\mathcal{D}_{\text{pref}}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(\alpha^+ \mid x)}{\pi_{\text{ref}}(\alpha^+ \mid x)} - \beta\log\frac{\pi_\theta(\alpha^- \mid x)}{\pi_{\text{ref}}(\alpha^- \mid x)}\right)\right]. \tag{7}$$

Here, $\mathcal{D}_{\text{pref}}$ is the preference dataset, π_θ is SIRA's conditional generation distribution, π_{ref} is the frozen adapter-disabled base model, σ is the sigmoid function, and $\beta > 0$ controls regularization relative to the reference model. Log-probabilities sum over generated action tokens, excluding input-prompt tokens. Training uses offline text preference pairs and updates only SIRA's LoRA (Hu et al., 2021) parameters, keeping VEA and HA frozen without video-model calls. Annotation protocols and training configurations are provided in the appendix.

5 Experiment

5.1 Experimental Setup

Dataset and metrics. We evaluate surgical intent understanding on SurgIntent-Bench. The training set contains 1,853 questions, while the test set comprises 667 questions from 156 video clips, including 379 liver and 288 lung questions. We report accuracy across the four intent dimensions, the two surgical sites, and the entire test set, with overall accuracy computed over all test questions. Dataset construction and partitioning are described in Section 3.2.

Baselines and evaluation protocol. We evaluate four groups of models: mainstream general-purpose MLLMs, open-source video MLLMs, medical MLLMs, and task-specific SFT models. In the main evaluation, each model receives the video, question, five answer options, and clinical context. To control access to surgical knowledge, we also provide baseline models with the same candidate surgical knowledge retrieved offline by SurgIntent-Agent.

Table 1: Main results on SurgIntent-Bench. Accuracy is reported in %. IP, IC, NG, and AM denote Functional, Coordinative, Prospective, and Anatomical Intent Understanding, respectively. The best and second-best results in each column are marked in red and blue, including ties.

Model	IP	IC	NG	AM	Liver	Lung	Overall
Mainstream MLLMs							
GPT-5.6-sol (OpenAI, 2026)	59.88	66.67	57.69	67.31	64.91	56.60	61.32
Gemini-3-Pro (Google DeepMind, 2026)	55.25	69.63	58.33	78.85	63.85	56.60	60.72
Claude-Opus-5 (Anthropic, 2026)	60.49	57.04	70.51	28.85	63.85	54.17	59.67
Qwen3-VL-235B (Bai et al., 2025a)	48.77	54.81	35.26	44.23	53.83	36.81	46.48
Kimi-K2.5 (Kimi Team et al., 2026)	54.32	57.78	46.15	44.23	58.05	44.79	52.32
Open-source Video MLLMs							
Apollo-7B-t32 (Zohar et al., 2024)	28.09	18.52	35.26	23.08	31.93	21.53	27.44
InternVideo2.5-Chat-8B (Wang et al., 2025d)	19.75	33.33	35.26	15.38	26.39	25.00	25.79
LLaVA-Video-7B-Qwen2 (Zhang et al., 2024b)	27.16	24.44	20.51	15.38	30.08	16.32	24.14
VideoLLaMA3-7B (Zhang et al., 2025a)	30.25	22.96	33.33	36.54	36.41	21.53	29.99
Video-R1-7B (Feng et al., 2025)	41.05	23.70	23.72	30.77	38.52	25.00	32.68
LongVA-7B-DPO (Zhang et al., 2024a)	29.63	34.07	33.33	23.08	35.88	24.31	30.88
Qwen3-VL-8B-Instruct (Bai et al., 2025a)	38.58	31.11	33.33	40.38	34.56	37.85	35.98
Qwen2.5-VL-7B-Instruct (Bai et al., 2025b)	39.81	24.44	26.28	40.38	39.31	26.04	33.58
MiniCPM-V-4.5 (Yu et al., 2025)	43.83	37.04	42.31	28.85	47.76	31.94	40.93
GLM-4.1V-9B-Thinking (GLM-V Team et al., 2025)	27.47	19.26	21.79	15.38	25.86	20.49	23.54
InternVL3.5-8B (Wang et al., 2025b)	28.40	24.44	26.28	30.77	30.87	22.57	27.29
Ovis2.5-9B (Lu et al., 2025)	27.78	20.74	46.15	28.85	37.20	22.22	30.73
Medical MLLMs							
NEXUS-MedVLM-7B (Su et al., 2026)	39.20	39.26	45.51	30.77	43.01	36.11	40.03
Lingshu-7B (Xu et al., 2025)	40.12	22.22	48.08	50.00	44.85	31.60	39.13
Hulu-Med-7B (Jiang et al., 2025)	36.73	28.15	53.85	25.00	42.74	31.94	38.08
Fleming-VL-8B (Shu et al., 2025)	28.70	14.81	25.64	21.15	28.50	19.44	24.59
SurgVidLM (Wang et al., 2025a)	29.01	12.59	42.31	50.00	39.58	18.40	30.43
Task-specific MLLMs							
Qwen3-VL-8B (SFT) (Bai et al., 2025a)	51.23	72.59	72.44	51.92	63.85	56.25	60.57
InternVL3.5-8B (SFT) (Wang et al., 2025b)	60.19	55.56	74.36	44.23	69.39	50.69	61.32
Qwen3.5-4B (SFT) (Qwen Team, 2026)	62.04	62.22	78.21	34.62	70.18	55.21	63.72
SurgIntent-Agent (Ours)	64.81	68.89	80.13	44.23	72.30	61.46	67.62

Implementation and training details. VEA, HA, and SIRA are based on Qwen3.5-4B (Qwen Team, 2026), while the Knowledge Retrieval Agent uses GPT-5.6-sol. Initial visual description and knowledge retrieval are performed offline; online inference comprises hypothesis initialization, visual probing, and evidence assessment. We divide the 1,853 training questions into two disjoint subsets: 1,394 questions for SFT of HA and 459 questions for constructing action preference data for SIRA. For comparison, the Qwen3.5-4B SFT baseline uses the full training set. We additionally construct 2,176 surgical scene understanding question–answer pairs to train VEA. We then train SIRA using LoRA and DPO, keeping VEA and HA frozen. All three agents use a learning rate of 2×10^{-6}, and each test question allows at most three rounds of visual probing. The remaining training hyperparameters are provided in the appendix.

5.2 Main Results

Insights from SurgIntent-Bench. Table 1 reveals three findings about surgical intent understanding in existing models. **(1) Medical MLLMs do not yet show an advantage.** The leading general-purpose model, GPT-5.6-sol, achieves an overall accuracy of 61.32%, while the best open-source video and medical models achieve 40.93% and 40.03%, respectively. These results indicate that current medical MLLMs remain insufficient for accurately understanding the clinical intent behind surgical maneuvers. **(2) Performance varies across the four clinical capabilities.** For example, Gemini-3-Pro achieves 78.85% on Anatomical Intent Understanding, but only 55.25% and 58.33% on Functional and Prospective Intent Understanding, respectively. Strong anatomical understanding is therefore not accompanied by comparable performance in explaining operative purposes and inferring subsequent objectives, indicating that performance on a single dimension does not adequately represent overall surgical intent understanding. **(3) Task-specific supervision effectively**

Table 2: Ablation results for SurgIntent-Agent. Accuracy (%); red values show percentage-point changes from the full framework.

Setting	Liver	Lung	Overall
SurgIntent-Agent (Full)	**72.30**	**61.46**	**67.62**
w/o surgical knowledge	60.16 −12.14	52.08 −9.38	56.67 −10.95
w/o SIRA (HA only)	65.44 −6.86	54.17 −7.29	60.57 −7.05
w/o VEA-SFT	64.12 −8.18	51.74 −9.72	58.77 −8.85
w/o SIRA-DPO	67.28 −5.02	56.25 −5.21	62.52 −5.10

Table 3: Effects of video and candidate knowledge. Accuracy (%); the best result for each model is in bold.

Video	Knowledge	Qwen3.5-4B (SFT)	GPT-5.6-sol
–	–	37.78	41.68
–	✓	45.28	52.32
✓	–	52.32	60.57
✓	✓	**63.72**	**61.32**

improves performance. SFT increases the overall accuracy of Qwen3-VL-8B and InternVL3.5-8B from 35.98% and 27.29% to 60.57% and 61.32%, respectively. These gains demonstrate the value of dedicated surgical intent supervision, although fine-tuned models still exhibit differences across capabilities and leave room for improvement.

Performance of SurgIntent-Agent. SurgIntent-Agent achieves the highest overall accuracy of 67.62%, outperforming the strongest task-specific SFT baseline, Qwen3.5-4B, by 3.90 percentage points and GPT-5.6-sol by 6.30 points. It also achieves the highest accuracy on both liver and lung questions, reaching 72.30% and 61.46%, respectively. Across the four intent dimensions, SurgIntent-Agent performs best on Functional and Prospective Intent Understanding, with accuracies of 64.81% and 80.13%, respectively. Qwen3-VL-8B (SFT) leads on Coordinative Intent Understanding, while Gemini-3-Pro leads on Anatomical Intent Understanding. The framework's strengths therefore lie in overall performance and in interpreting operative purposes and subsequent objectives, rather than consistently leading across all dimensions.

5.3 Ablation Studies and Analysis

Surgical knowledge and iterative verification. Table 2 reports the framework ablations. *w/o surgical knowledge* removes retrieved surgical knowledge, reducing overall accuracy from 67.62% to 56.67%, a decrease of 10.95 percentage points. *w/o SIRA (HA only)* directly returns HA's initial answer without subsequent visual probing or evidence assessment, reducing accuracy to 60.57%. These results support the roles of surgical knowledge and iterative verification in intent inference.

Agent training. *w/o VEA-SFT* uses VEA without supervised fine-tuning on visual question–answer data, reducing accuracy to 58.77%, a decrease of 8.85 percentage points. *w/o SIRA-DPO* retains the full reasoning workflow but uses SIRA without DPO training, reducing accuracy to 62.52%, a decrease of 5.10 points. These results show that both training for visual evidence extraction and preference optimization for probing and decision-making improve performance.

Effects of video and surgical knowledge. Table 3 compares video and candidate knowledge inputs while retaining the question, options, and clinical context. Because knowledge is retrieved from video descriptions, "without video" removes direct video access but may retain indirect visual cues. Even without video or retrieved knowledge, above-chance performance is expected, as models can use surgical priors to identify plausible answers; video determines which interpretation fits the current scene. Adding video without knowledge improves Qwen3.5-4B (SFT) and GPT-5.6-sol by 14.54 and 18.89 points, respectively, while adding knowledge to video yields further gains of 11.40 and 0.75 points. Thus, direct observation provides the primary gain, with retrieved knowledge offering additional, model-dependent benefits.

6 Conclusion

In this paper, we introduce Surgical Intent Understanding based on four clinical capabilities of surgeons and construct SurgIntent-VidQA and SurgIntent-Bench for systematic evaluation. Our results reveal limitations in existing models' ability to explain operative purposes and infer subsequent objectives, as well as the value of task-specific supervision. We further propose SurgIntent-Agent, which improves surgical intent understanding through surgical knowledge guidance and iterative visual verification. This work provides a foundation for evaluating and developing models that understand the purposes of surgical maneuvers, with potential applications in surgical education and future intelligent surgical assistance.

AI USE STATEMENT

Generative AI tools were used only to improve the language, clarity, and readability of the manuscript. They were not used to formulate the research methodology, construct the datasets (All datasets were manually annotated by surgeons. AI was used only to assist with balancing correct-answer positions across options A–E, followed by surgeon review), generate or analyze experimental results, or draw scientific conclusions. All content was reviewed and verified by the authors.

ETHICS STATEMENT

The clinical surgical videos used in this study were collected under institutional ethics approval and de-identified before research use. Public surgical teaching videos were processed to remove audio, subtitles, and teaching annotations. The dataset is intended solely for research on surgical video understanding and should not be interpreted as providing clinical recommendations or evidence of clinical benefit.

REPRODUCIBILITY STATEMENT

Upon publication, we will release the SurgIntent-VidQA annotations, fixed splits, candidate surgical knowledge, agent prompts, model configurations, and code for training, inference, and evaluation. Video assets will be shared where permitted by institutional approvals and source licences. Appendix A describes dataset construction and statistics, Appendix B details the framework and training procedure, Appendix C specifies the experimental protocols, and Appendix D provides representative reasoning and failure cases.

REFERENCES

Anthropic. Introducing Claude Opus 5, 2026. URL https://www.anthropic.com/news/claude-opus-5.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL Technical Report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL Technical Report, 2025b. URL https://arxiv.org/abs/2502.13923.

Sayra M. Cristancho, Meredith Vanstone, Lorelei Lingard, Marie-Eve LeBel, and Michael Ott. When surgeons face intraoperative challenges: a naturalistic model of surgical decision making. *The American Journal of Surgery*, 205(2):156–162, 2013. doi: 10.1016/j.amjsurg.2012.10.005.

Debra DaRosa, David A. Rogers, Reed G. Williams, Linnea S. Hauge, Heather Sherman, Kenric Murayama, Alexander Nagle, and Gary L. Dunnington. Impact of a structured skills laboratory curriculum on surgery residents' intraoperative decision-making and technical skills. *Academic Medicine*, 83(10 Suppl):S68–S71, 2008. doi: 10.1097/ACM.0b013e318183cdb1.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing Video Reasoning in MLLMs, 2025. URL https://arxiv.org/abs/2503.21776.

Xiaojie Gao, Yueming Jin, Yonghao Long, Qi Dou, and Pheng-Ann Heng. Trans-SVNet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer, 2021. URL https://arxiv.org/abs/2103.09712.

GLM-V Team et al. GLM-4.5V and GLM-4.1V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning, 2025. URL https://arxiv.org/abs/2507.01006.

Google DeepMind. Gemini 3 Pro model card, 2026. URL https://deepmind.google/models/model-cards/gemini-3-pro/. Updated May 2026.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Songtao Jiang, Yuan Wang, Sibo Song, et al. Hulu-Med: A Transparent Generalist Model towards Holistic Medical Vision-Language Understanding, 2025. URL https://arxiv.org/abs/2510.08668.

Juseong Jin and Chang Wook Jeong. Surgical-LLaVA: Toward surgical scenario understanding via large language and vision models, 2024. URL https://arxiv.org/abs/2410.09750.

Ji Woong (Brian) Kim, Juo-Tung Chen, Pascal Hansen, Lucy Xiaoyang Shi, Antony Goldenberg, Samuel Schmidgall, Paul Maria Scheikl, Anton Deguet, Brandon M. White, De Ru Tsai, Richard Jaepyeong Cha, Jeffrey Jopling, Chelsea Finn, and Axel Krieger. SRT-H: A hierarchical framework for autonomous surgery via language-conditioned imitation learning. *Science Robotics*, 10(104):eadt5254, 2025. doi: 10.1126/scirobotics.adt5254.

Kimi Team et al. Kimi K2.5: Visual Agentic Intelligence, 2026. URL https://arxiv.org/abs/2602.02276.

Jiajie Li, Garrett Skinner, Gene Yang, Brian R Quaranto, Steven D Schwaitzberg, Peter C W Kim, and Jinjun Xiong. LLaVA-Surg: Towards multimodal surgical assistant via structured surgical video learning. *arXiv preprint arXiv:2408.07981*, 2024. URL https://arxiv.org/abs/2408.07981.

Yaoqian Li, Xikai Yang, Dunyuan Xu, Yang Yu, Litao Zhao, Xiaowei Hu, Jinpeng Li, and Pheng-Ann Heng. SurgPub-Video: A comprehensive surgical video dataset for enhanced surgical intelligence in vision-language model, 2025. URL https://arxiv.org/abs/2508.10054.

Shiyin Lu, Yang Li, Yu Xia, et al. Ovis2.5 Technical Report, 2025. URL https://arxiv.org/abs/2508.11737.

Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022. doi: 10.1016/j.media.2022.102433.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition, 2026. URL https://openai.com/index/gpt-5-6/.

Alejandra Perez, Anita Rau, Lee White, Busisiwe Mlambo, Chinedu Nwoye, Muhammad Abdullah Jamal, and Omid Mohareri. SUREON: A benchmark and vision-language-model for surgical reasoning. *arXiv preprint arXiv:2603.06570*, 2026. URL https://arxiv.org/abs/2603.06570.

Qwen Team. Qwen3.5: Towards native multimodal agents, 2026. URL https://qwen.ai/blog?id=qwen3.5.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

Chuyi Shang, Amos You, Sanjay Subramanian, Trevor Darrell, and Roei Herzig. TraveLER: A modular multi-LMM agent framework for video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9740–9766. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.544. URL https://aclanthology.org/2024.emnlp-main.544/.

Yudi Shi, Shangzhe Di, Qirui Chen, and Weidi Xie. Enhancing video-LLM reasoning via agent-of-thoughts distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8523–8533, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Shi_Enhancing_Video-LLM_Reasoning_via_Agent-of-Thoughts_Distillation_CVPR_2025_paper.html.

Yan Shu, Chi Liu, Robin Chen, Derek Li, and Bryan Dai. Fleming-VL: Towards Universal Medical Visual Reasoning with Multimodal LLMs, 2025. URL https://arxiv.org/abs/2511.00916.

Steven M. Strasberg and L. Michael Brunt. Rationale and use of the critical view of safety in laparoscopic cholecystectomy. *Journal of the American College of Surgeons*, 211(1):132–138, 2010. doi: 10.1016/j.jamcollsurg.2010.02.053.

Yuhao Su, Anwesa Choudhuri, Zhongpai Gao, et al. MedGRPO: Multi-Task Reinforcement Learning for Heterogeneous Medical Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2512.06581.

Andru P. Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. EndoNet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 36(1):86–97, 2017. doi: 10.1109/TMI.2016.2593957.

Guankun Wang, Junyi Wang, Wenjin Mo, et al. SurgVidLM: Towards Multi-grained Surgical Video Understanding with Large Language Model, 2025a. URL https://arxiv.org/abs/2506.17873.

Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency, 2025b. URL https://arxiv.org/abs/2508.18265.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *Computer Vision – ECCV 2024*, volume 15138 of *Lecture Notes in Computer Science*, pp. 58–76. Springer Nature Switzerland, 2025c. doi: 10.1007/978-3-031-72989-8_4. URL https://link.springer.com/chapter/10.1007/978-3-031-72989-8_4.

Yi Wang, Xinhao Li, Ziang Yan, et al. InternVideo2.5: Empowering Video MLLMs with Long and Rich Context Modeling, 2025d. URL https://arxiv.org/abs/2501.12386.

Zheng Wang, Haoran Chen, Haoxuan Qin, Zhipeng Wei, Tianwen Qian, and Cong Bai. Think, then verify: A hypothesis-verification multi-agent framework for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33784–33793, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_Think_Then_Verify_A_Hypothesis-Verification_Multi-Agent_Framework_for_Long_Video_CVPR_2026_paper.html.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. Multimodal large language models: A survey. In *2023 IEEE international conference on big data (bigdata)*, pp. 2247–2256. IEEE, 2023.

Weiwen Xu, Hou Pong Chan, Long Li, et al. Lingshu: A Generalist Foundation Model for Unified Multimodal Medical Understanding and Reasoning, 2025. URL https://arxiv.org/abs/2506.07044.

Tianyu Yu, Zefan Wang, Chongyi Wang, et al. MiniCPM-V 4.5: Cooking Efficient MLLMs via Architecture, Data, and Training Recipe, 2025. URL https://arxiv.org/abs/2509.18154.

Boqiang Zhang, Kehan Li, Zesen Cheng, et al. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding, 2025a. URL https://arxiv.org/abs/2501.13106.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long Context Transfer from Language to Vision, 2024a. URL https://arxiv.org/abs/2406.16852.

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander G Hauptmann, Yonatan Bisk, and Yiming Yang. Direct preference optimization of video large multimodal models from language model reward. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 694–717. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.naacl-long.30. URL https://aclanthology.org/2025.naacl-long.30/.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, et al. Instruction tuning for large language models: A survey. *ACM Computing Surveys*, 58(7):1–36, 2026.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video Instruction Tuning With Synthetic Data, 2024b. URL https://arxiv.org/abs/2410.02713.

Yiyang Zhou, Yangfan He, Yaofeng Su, Siwei Han, Joel Jang, Gedas Bertasius, Mohit Bansal, and Huaxiu Yao. ReAgent-V: A reward-driven multi-agent framework for video understanding. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-5066. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/de72a54c3d39523b0dd06b64ffa6d50d-Abstract-Conference.html.

Orr Zohar, Xiaohan Wang, Yann Dubois, et al. Apollo: An Exploration of Video Understanding in Large Multimodal Models, 2024. URL https://arxiv.org/abs/2412.10360.

Towards Higher-Level Surgical Intelligence: A Comprehensive Benchmark and Multi-Agent Framework for Surgical Intent Understanding

Appendices

A SurgIntent-VidQA Construction and Statistics 15

A.1 Video Collection and Clip Selection 15

A.2 Expert Annotation and Quality Control 15

A.3 Distractor Construction 15

A.4 Dataset Statistics and Splits 16

A.5 Surgical Knowledge Resources 16

B Framework and Training Details 16

B.1 Agent Interfaces and Inference Procedure 16

B.2 Supervised Fine-Tuning 17

B.3 Preference Construction and Optimization 17

B.4 Decision Constraints and Stopping Criteria 18

B.5 Agent Prompts 18

C Experimental Protocols 19

C.1 Baseline Configurations 19

C.2 Training and Inference Settings 20

C.3 Evaluation and Ablation Settings 20

D Qualitative Analysis 21

D.1 Trajectory Selection and Presentation 21

D.2 Evidence-Grounded Reasoning Cases 21

D.3 Failure Cases 22

D.4 Cross-Case Observations 23

E Scope, Ethics, and Reproducibility 23

E.1 Scope and Limitations 23

E.2 Ethics and Data Governance 23

E.3 Reproducibility and Release 23

A SurgIntent-VidQA Construction and Statistics

A.1 Video Collection and Clip Selection

SurgIntent-VidQA combines de-identified clinical surgical recordings with publicly accessible surgical teaching videos. It covers four procedures: left hemihepatectomy, uniportal video-assisted thoracoscopic left upper lobe S3bc subsegmentectomy, thoracoscopic right upper lobectomy, and thoracoscopic left lower lobectomy.

Surgeons reviewed the source videos, identified clinically meaningful operative steps, and extracted short clips for intent understanding. Clip selection focused on anatomical landmarks, individual instrument roles, instrument coordination, and the subsequent objectives supported by ongoing maneuvers. Audio, subtitles, and instructional annotations were removed during preprocessing to reduce answer cues from narration or teaching overlays.

The final dataset contains 603 clips with a total duration of approximately 133.18 minutes. The mean and median clip durations are 13.25 and 13.07 seconds, respectively. All videos have a frame rate of 30 fps and retain their original spatial resolutions.

A.2 Expert Annotation and Quality Control

Five practicing surgeons from multiple hospitals manually annotated the questions, reference answers, and clinical contexts, followed by cross-checking within the team. They considered anatomical relationships, instrument actions, tissue changes, and procedural context when determining whether each interpretation was supported by the corresponding video.

Questions were provided with the clinical background required to interpret short clips. This information was reviewed to avoid directly revealing the correct answer. Two senior surgeons conducted the final review and adjudicated unresolved cases. Samples with unresolved disagreement or insufficient visual support were excluded.

Every retained question–answer pair was verified against its corresponding video. This verification formed part of the annotation quality-control process. The additional visual question–answer data used to train the Visual Evidence Agent are described in Appendix B.2.

A.3 Distractor Construction

Each question contains one correct option and four clinically plausible distractors. The distractors represent interpretations that could occur in related surgical contexts but are inconsistent with the current video. Their construction varies semantic elements such as instrument function, target structure, inter-instrument relationships, tissue effects, anatomical relationships, and procedural position. Table 4 summarizes the design principles.

Table 4: **Distractor construction across question types.** Distractors preserve clinical plausibility while varying the interpretation of the observed operative scene.

Question type	Distractor construction principle
Instrument purpose	Alternative plausible uses, targets, or tissue effects of the same instrument in relevant procedures.
Instrument coordination	Alternative functional roles or interaction patterns involving the same instruments.
Next-goal enabling	Earlier, later, or adjacent objectives within the same surgical pathway.
Anatomy marker	Plausible neighboring structures or alternative anatomical relationships within the operative field.

Distractors were drawn from surgeon-annotated answers to questions of the same type in other videos. GPT-5.6-sol assisted only with balancing correct-answer positions, followed by surgeon review. Options were matched in length, sentence structure, technical detail, and anatomical specificity to reduce textual shortcuts. Review removed redundant options, simultaneously valid interpretations, and alternatives that could not be distinguished from the available video. Each retained

question has one answer supported by the video and its clinical context. Correct-answer positions are approximately balanced across options A–E.

A.4 Dataset Statistics and Splits

SurgIntent-VidQA contains 2,520 five-option multiple-choice questions, with 1,853 training questions and 667 test questions. Liver and lung procedures contribute 316 and 287 clips, respectively. Table 5 reports the distribution across splits and question types.

Table 5: **SurgIntent-VidQA statistics.** All questions associated with the same video clip are assigned to the same split.

Item	Train	Test	Total
Video clips	447	156	603
Instrument purpose	883	324	1,207
Instrument coordination	374	135	509
Next-goal enabling	447	156	603
Anatomy marker	149	52	201
Total questions	**1,853**	**667**	**2,520**

We construct clip-level temporal blocks and assign all questions from the same clip to one split. Adjacent clips at split boundaries are excluded to reduce local temporal redundancy. No video clip appears in both the training and test sets. The resulting splits preserve the distributions of organs and question types.

A.5 Surgical Knowledge Resources

The structured surgical knowledge resources organize information on instrument functions, anatomical relationships, applicable operative contexts, and procedural steps. They were compiled with reference to publicly available surgical literature, surgical teaching materials, and device documentation.

Question-specific candidate knowledge is prepared for the knowledge-augmented evaluation setting. It covers instrument functions, anatomical relationships, and procedural workflows. All evaluated models receive the same candidate knowledge for a given question and must assess its applicability against the current video. Anatomy-marker questions receive no candidate knowledge and rely on visual anatomical evidence alone.

B Framework and Training Details

SurgIntent separates knowledge-guided hypothesis initialization, targeted visual observation, and evidence-based decision-making. The Hypothesis Agent (HA), Visual Evidence Agent (VEA), and Surgical Intent Reasoning Agent (SIRA) are based on Qwen3.5-4B. Detailed optimization settings are provided in Appendix C.2.

B.1 Agent Interfaces and Inference Procedure

Offline knowledge preparation. Before iterative reasoning, an initial video description, the question, and the available clinical context are used to prepare question-specific candidate surgical knowledge. This stage is performed offline. The resulting candidate knowledge is supplied to HA and SIRA during formal inference, and the reasoning loop does not perform online web search.

Hypothesis initialization. HA directly processes the surgical video, question, answer options, clinical context, and candidate knowledge. Each answer option is treated as a candidate intent hypothesis. HA ranks these hypotheses using their relative answer-token scores, and the highest-ranked option initializes the current answer. The complete ranking remains available throughout inference, and lower-ranked candidates are not discarded.

SIRA begins from the candidate ranking, current answer, and associated textual inputs. It does not receive an initial video caption. The ranking serves as a prior rather than visual evidence.

Targeted visual inspection. SIRA receives the question, options, clinical context, candidate knowledge, initial ranking, current answer, video duration, and remaining interaction budget. It invokes `inspect_video` to investigate an observable distinction between competing hypotheses. The action specifies a neutral visual question, the compared options, and an appropriate visual focus.

VEA receives only the sampled video frames and the visual question. It does not receive the answer-option mapping, hypothesis ranking, or SIRA's reasoning history. VEA reports visible instruments, actions, tissue relationships, direct visual effects, and a focused response to the requested check. Separate fields record events not observed and unresolved uncertainty.

Evidence assessment. After each inspection, SIRA invokes `submit_assessment` to evaluate the returned observations. It identifies the affected hypotheses, retains or updates the current answer, and decides whether to continue or stop. SIRA retains the textual history of earlier questions, observations, and assessments. Ground-truth answers and annotation-only evidence are excluded from all inference inputs.

B.2 Supervised Fine-Tuning

Visual evidence supervision. VEA is trained on 2,176 visual question–answer examples constructed from training videos. GPT-5.6-sol produced the initial examples, which were subsequently reviewed by surgeons. The supervision targets factual scene understanding, including instrument motion, tissue interactions, and directly visible effects, rather than intent prediction.

Training follows the structured observation format used during inference. VEA first describes the visible scene and then addresses the focused visual question. The targets distinguish observed events, events not observed in the sampled content, and events that cannot be assessed because of limited visibility. This distinction discourages treating missing observations as definitive negative evidence.

Hypothesis initialization supervision. HA is fine-tuned on 1,394 surgical intent questions using the video, question, options, clinical context, and candidate knowledge. The target is the correct option enclosed in `<answer>` tags. This supervision trains HA to identify the supported intent hypothesis but does not provide relative-order labels for the incorrect options.

At inference, logits for the five option tokens are normalized over A–E and sorted to obtain the initial ranking. The relative ordering of incorrect options therefore arises from model predictions rather than explicit ranking supervision. HA and VEA are each trained for one epoch using a learning rate of 2×10^{-6}. Both use LoRA adaptation with a trainable vision tower.

B.3 Preference Construction and Optimization

Preference data are constructed from a separate partition of 459 training questions. Each example compares two local actions under the same agent state rather than treating a complete trajectory as a supervised target. The preferences cover both visual inspection and evidence assessment.

For visual inspection, candidate questions are generated through greedy and sampled decoding from the Qwen3.5-4B policy. Each candidate is executed through the frozen VEA to obtain the corresponding visual observations. Alternative assessment actions are generated from the same returned evidence. GPT-5.6-sol supplies valid alternatives when necessary and assists with semantic review under a fixed protocol. Ground-truth answers are used only as offline audit information. They are not included in SIRA inputs, and final-answer correctness alone does not determine action preference.

The preferences target three complementary behaviors:

- **Discriminative inspection.** Prefer neutral and visually answerable questions that distinguish competing hypotheses over questions that delegate intent prediction to VEA.
- **Conservative retention.** Prefer retaining the current answer when evidence is ambiguous, shared across hypotheses, or insufficiently attributable to the queried instrument.

- **Supported correction.** Prefer updating the answer when directly observed evidence distinguishes a new hypothesis from the current answer.

The resulting examples have the form (s, a^+, a^-), where s is the shared state and a^+ and a^- are the preferred and dispreferred action completions. We optimize SIRA using direct preference optimization (DPO) (Rafailov et al., 2023). Only the SIRA LoRA parameters are updated. HA and VEA remain frozen, and the adapter-disabled Qwen3.5-4B base model serves as the reference policy.

DPO training is performed offline on textual states and action completions. It does not invoke VEA or generate new video interaction trajectories during optimization. The loss is computed over completion tokens using $\beta = 0.1$ and a learning rate of 2×10^{-6}.

B.4 Decision Constraints and Stopping Criteria

Each episode permits at most three inspection–assessment rounds. SIRA selects an adaptive visual inspection budget from a predefined set and samples the video uniformly over its full duration. Each inspection must compare at least two valid options and focus on an observable distinction such as morphology, motion, or tissue effect.

The implementation enforces tool order, interaction budgets, action consistency, and traceable references to VEA observations. An assessment labeled `ambiguous` or `none` cannot change the current answer. An update must declare `direct` evidence and identify the new answer among the favored options. The semantic validity and discriminative value of the evidence are guided by the agent prompts and preference optimization.

SIRA stops when the available evidence is sufficient, no useful new visual check remains, or the round budget is exhausted. Continuing requires a new observable distinction that could alter the current decision. Stopping returns the current answer.

B.5 Agent Prompts

The following excerpts present the task-defining agent instructions. Variable fields and repeated serialization requirements are omitted for readability.

Hypothesis Agent. The prompt requires HA to assess the supplied knowledge against the current video rather than answer from clinical plausibility alone.

```
You are evaluating a short minimally invasive surgical video.
Use the visible video evidence, the supplied surgical context,
and only applicable external knowledge to infer the surgeon's
intent. Candidate knowledge may be irrelevant; do not select
an answer from medical plausibility alone when it is unsupported
by the video.

[...]

Select exactly one option. End with only its letter inside
<answer> tags, for example <answer>A</answer>.
```

Surgical Intent Reasoning Agent. SIRA treats the initial ranking as a prior, seeks observable differences between competing hypotheses, and separates uncertain observations from evidence supporting an answer change.

```
You are SIRA, the surgical intent reasoning and verification
agent. A frozen video model has ranked all answer options.
Start from its first option, but keep every option eligible.
The ranking is a prior, not visual evidence. There is no initial
caption, and you must never use hidden labels.

Call inspect_video and then submit_assessment. Ask a neutral
question about a visible difference between the current answer
and its strongest competitor(s). Request observable contact,
motion, tissue relations, or direct tissue effects. Do not ask
VEA to select an option or state a surgical purpose.

[...]

UPDATE only when direct evidence supports a new favored option.
If evidence is ambiguous, shared, absent, or poorly attributed,
RETAIN the current answer.

Continue only when the next visual check targets a new observable
distinction that could change the answer. Stop when evidence is
sufficient, no useful check remains, or the final round is reached.
```

Visual Evidence Agent. VEA reports factual visual observations without converting the question into an expected answer or inferring an unobserved purpose.

```
You are the Visual Evidence Agent for surgical video
understanding. Report the concrete surgical content that is
visibly present before describing uncertainty.

Do not infer surgical intent, choose an option, or fill unseen
outcomes from medical knowledge. Describe uncertain anatomy
generically. Phrase the focused observation in terms of visible
instruments, motion, tissue relations, or direct tissue effects.

[...]

For a presence/absence check, distinguish an event seen, an event
not seen while the target is visible, and a check that is not
assessable because visibility or sampling is inadequate. Do not
turn "not seen" into proof that an event never occurred.
```

C Experimental Protocols

All reported results use the official SurgIntent-Bench test set of 667 questions from 156 video clips.

C.1 Baseline Configurations

Models and inputs. The benchmark includes general-purpose multimodal models, open-source video models, medical multimodal models, and models fine-tuned for the task. All baselines receive the same question, answer options, and available clinical context. Knowledge-augmented conditions additionally provide the question-specific candidate surgical knowledge, except for anatomy-marker questions.

The answer instruction is kept consistent across models. Video tokenization, chat templates, and preprocessing follow the supported interface of each model. The configuration of each model remains fixed across the complete test set.

Task-specific fine-tuned baselines. Qwen3-VL-8B, InternVL3.5-8B, and Qwen3.5-4B are each fine-tuned on all 1,853 training questions. They use the same question set but retain their model-specific trainable components, as summarized in Table 6.

Table 6: **Training settings for task-specific baselines.** All three models use the same 1,853 training questions.

Model	Trainable components	Epochs
Qwen3-VL-8B	Language LoRA and vision tower	2
InternVL3.5-8B	Language LoRA only	2
Qwen3.5-4B	Language LoRA and vision tower	1

The HA-only ablation uses the HA checkpoint trained on the 1,394-question SFT partition. It therefore differs from the Qwen3.5-4B task-specific baseline, which is trained on all 1,853 training questions.

C.2 Training and Inference Settings

HA, VEA, and SIRA are based on Qwen3.5-4B. HA is fine-tuned on 1,394 surgical intent questions, while VEA is fine-tuned on 2,176 additional surgeon-reviewed visual question–answer examples. SIRA is trained on offline local-action preferences using DPO. HA and VEA remain frozen during SIRA training.

All three agents are trained using eight NVIDIA RTX 4090 GPUs and BF16 precision. Table 7 summarizes the principal optimization settings.

Table 7: **Principal training settings for SurgIntent agents.** SFT denotes supervised fine-tuning; DPO denotes direct preference optimization.

Setting	HA-SFT	VEA-SFT	SIRA-DPO
Learning rate	2×10^{-6}	2×10^{-6}	2×10^{-6}
LoRA rank / alpha	64 / 128	64 / 128	16 / 32
LoRA dropout	0.05	0.05	0
Vision tower	Trainable	Trainable	Frozen
Per-device batch size	1	1	1 pair
Gradient accumulation	2	2	1
Training duration	1 epoch	1 epoch	20 updates
DPO β	–	–	0.1

SIRA uses the frozen adapter-disabled Qwen3.5-4B base model as the DPO reference policy. The maximum sequence length is 4,096 tokens per branch. Preference labels and VEA observations are prepared before optimization, and training does not invoke VEA or an online evaluator.

During inference, HA initializes the candidate ranking from the video and associated textual inputs. SIRA may then request targeted visual reinspection under a fixed three-round interaction budget and may stop earlier when sufficient evidence is available. Baselines follow their model-specific video preprocessing, while all configurations remain fixed across the test set.

C.3 Evaluation and Ablation Settings

Scoring. We report accuracy on five-option multiple-choice questions. Overall accuracy is micro-averaged over all 667 test questions. Liver and Lung accuracy use 379 and 288 questions as their respective denominators. Dimension-specific accuracy is computed over the questions of each type without first averaging over video clips.

For generative baselines, the answer parser checks, in order, for an `<answer>` tag, a `Final Answer` expression, the final output line, and a unique valid option letter. An output that cannot be resolved to one valid option is counted as incorrect. Each question is answered once without voting across sampled responses. Retries address request failures and are not used to select among alternative answers.

Ablation variants. All ablations use the same official test set. Table 8 specifies the change made in each variant.

Table 8: **Definitions of the reported SurgIntent ablations.** Components not named in a row retain their full-framework settings.

Variant	Change relative to the full framework
w/o surgical knowledge	Remove candidate surgical knowledge from both HA and SIRA.
w/o SIRA (HA only)	Return HA's highest-ranked option without subsequent visual inspection or evidence assessment.
w/o VEA-SFT	Replace VEA with the original Qwen3.5-4B model without visual-evidence supervised fine-tuning.
w/o SIRA-DPO	Disable the SIRA adapter while retaining the agent prompts, structural rules, and interaction budget.

Video and knowledge input conditions. The input analysis compares four conditions: text only, text with candidate knowledge, video without candidate knowledge, and video with candidate knowledge. "Text" includes the question, options, and available clinical context in every condition. The same questions, options, and context are retained while video and knowledge availability are varied. Prompt wording is adjusted only where needed to reflect the absence of an input modality. Anatomy-marker questions receive no candidate knowledge under any condition.

D Qualitative Analysis

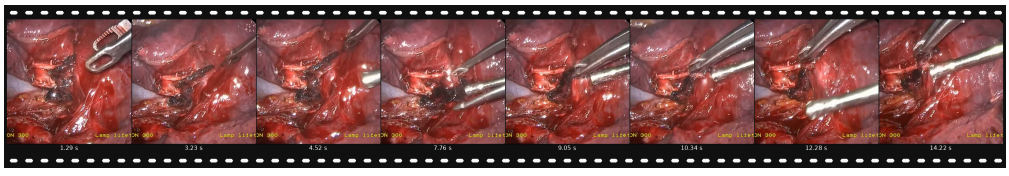
We examine three trajectories from the same 667-question test run: an incorrect initial answer that was corrected, a correct answer that was confirmed, and an initial error that remained unresolved. Questions, observations, and assessment reasons are extracted from the saved agent trajectories and lightly edited for readability without changing their meaning. Expert answers are shown only for retrospective evaluation and were not provided to the agents. The displayed frames are selected for visualization and do not represent the complete video evidence examined by VEA.

D.1 Trajectory Selection and Presentation

The examples in the main paper illustrate the four intent dimensions and their question formats. The trajectories below instead demonstrate how an initial hypothesis motivates a targeted visual check and how the returned observation affects the subsequent decision. They were selected to cover distinct answer-transition patterns and should not be interpreted as representing the frequency of these behaviors across the test set.

D.2 Evidence-Grounded Reasoning Cases

Case 1: Incorrect to correct **Lung, SI-LNG-V02-P025-Q01**



Eight frames selected for display.

Question. What is the purpose of the oval forceps in this video?

HA initialization. Ranking: $B > C > E > D > A$. The initial answer was B, describing lateral retraction. Option C describes upward elevation, while E describes downward traction.

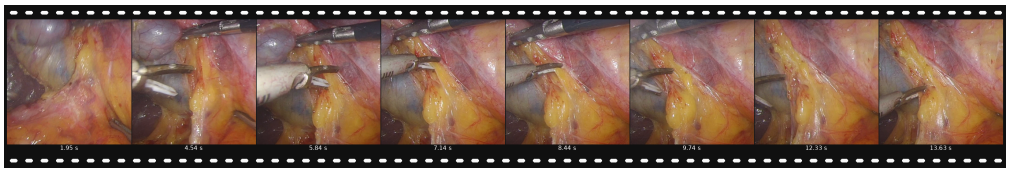
SIRA → VEA (`inspect_video`). "What is the motion and tissue effect of the oval forceps in this video?" The check targeted the direction of traction and the resulting tissue displacement to distinguish B, C, and E.

VEA observation. "The oval forceps grasps and lifts a red tissue edge from the upper side of the recess while the blunt probe presses the adjacent surface." The reported direct effect was that "the lifted edge is displaced upward and away from the central recess". VEA also noted that the exact tissue identity was not visually established.

SIRA assessment. Evidence strength: `direct`; favored option: C; ruled-out options: B and E. SIRA concluded that the upward displacement supported C rather than lateral or downward traction.

Final action: `update` $B \rightarrow C$; `stop`. **Expert answer:** C.

Case 2: Correct answer confirmed **Liver, SI-LIV-V01-P032-Q02**



Eight frames selected for display.

Question. What is the purpose of the left grasping forceps in this video?

HA initialization. Ranking: $A > D > C > B > E$. The initial answer was A, describing upward elevation of the contacted tissue margin.

SIRA → VEA (`inspect_video`). "What is the motion and tissue effect of the left grasping forceps in this video?" The check compared upward elevation with local regripping, downward traction, lateral retraction, and removal of a detached fragment.

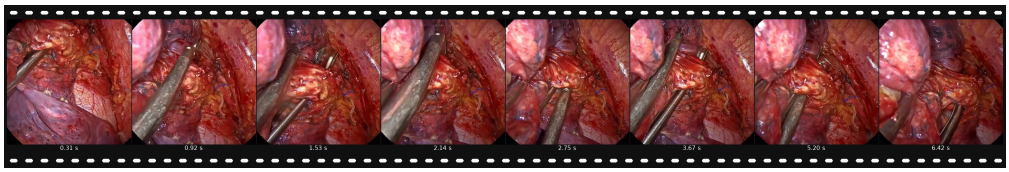
VEA observation. "The left grasping forceps repeatedly contacts and lifts the lower edge of a yellow-red tissue band while the upper and lower margins are held by other instruments." The reported effect was that "the lower edge of the band is displaced upward and the central plane is widened".

SIRA assessment. Evidence strength: `direct`; favored option: A. SIRA concluded that the upward displacement supported the interpretation of upward elevation and retained the initial answer.

Final action: `retain` $A \rightarrow A$; `stop`. **Expert answer:** A.

D.3 Failure Cases

Case 3: Initial error left unresolved **Lung, SI-LNG-V03-P123-Q01**



Eight frames selected for display.

Question. What is the purpose of the serpentine-head instrument in this video?

HA initialization. Ranking: $A > B > D > C > E$. The initial answer was A, describing blunt spreading or sweeping to separate tissue layers. Option B describes lifting, depressing, or retracting a margin to maintain exposure.

Round 1. SIRA asked whether the target instrument performed blunt spreading or sweeping to enlarge the tissue plane. VEA described several visible probes and forceps but could not confidently identify the queried instrument. SIRA assigned `none`, retained A, and continued.

Round 2. SIRA asked whether the target instrument lifted, depressed, or retracted tissue to maintain exposure. VEA reported that a blunt probe pressed and slid along the lower margin

while the tissue remained open. SIRA judged the observation `ambiguous` because the action could not be reliably attributed to the queried instrument or used to separate A from B. It retained A.

Round 3. SIRA checked whether the instrument applied focused compression to a bleeding point. VEA did not observe sustained focal compression. SIRA again assigned `none`, retained A, and stopped at the interaction limit.

Final action: `retain` $A \rightarrow A$; `stop`. **Expert answer:** B.

Failure analysis. The second inspection produced a potentially relevant description of a depressed tissue margin, but the action could not be reliably attributed to the queried instrument. The final inspection shifted to another candidate without resolving the principal competition between A and B.

D.4 Cross-Case Observations

The successful cases relied on concrete visual distinctions that could be attributed to the queried instrument. In both cases, the direction of tissue displacement provided evidence for retaining or updating the current answer. By contrast, the failure case remained unresolved because the visible action could not be reliably assigned to the target instrument, and the subsequent inspection did not return to the main competition.

These trajectories illustrate the importance of both evidence quality and inspection selection. Targeted visual inspection is most useful when an observation is attributable to the relevant instrument and discriminative among competing interpretations. The failure case also shows that conservative evidence assessment can avoid an unsupported update while still leaving an incorrect initial hypothesis unresolved.

E Scope, Ethics, and Reproducibility

E.1 Scope and Limitations

SurgIntent-VidQA focuses on short clips from four minimally invasive liver and lung procedures. The benchmark evaluates surgical intent understanding in a five-option multiple-choice setting and does not cover unrestricted generation of operative explanations. Its current coverage of organs, procedures, and clinical settings remains limited. The evaluation measures generalization to held-out clips; cross-patient and cross-institution generalization remains to be established. SurgIntent is intended as a research framework and not as a system for intraoperative decision-making.

E.2 Ethics and Data Governance

The collection and use of hospital-recorded surgical videos were approved by the relevant institutional ethics review board. The clinical videos were de-identified before research use. Publicly accessible surgical teaching videos were used for research, and audio, subtitles, and teaching annotations were removed during preprocessing. The dataset is intended exclusively for research, and benchmark performance should not be interpreted as evidence of clinical safety or benefit.

E.3 Reproducibility and Release

Upon publication, we will release the SurgIntent-VidQA benchmark package, including its annotations, fixed splits, candidate surgical knowledge, agent prompts, model configurations, and code for training, inference, and evaluation. Video assets will be distributed where permitted by the applicable institutional approvals and source licences. Appendices A–C document the dataset construction, framework implementation, training settings, and evaluation protocol. Appendix D provides representative reasoning trajectories.