EVIAGENT: EVIDENCE-GAIN-GUIDED ACTIVE EVIDENCE ACQUISITION FOR LONG-HORIZON SURGICAL VIDEO REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon surgical video reasoning remains challenging because query-relevant evidence is often sparse, heterogeneous, and dynamically dependent on what has already been established. Existing long-video methods improve context coverage through compression, retrieval, or adaptive observation, but the relationship between established evidence, unresolved evidence, and the next acquisition decision often remains implicit. We therefore formulate long-horizon surgical video reasoning as **active evidence acquisition** and propose **EviAgent**, an evidence-seeking agent that explicitly targets unresolved, query-dependent evidence in long surgical videos. EviAgent maintains a structured evidence state and uses an evidence-gain-guided policy to adaptively select temporal retrieval, local verification, or stopping as reasoning progresses. We evaluate EviAgent on four surgical subsets of MedHorizon under procedure-level cross-fitted evaluation, showing improved performance in both fine-grained surgical understanding and semantic reasoning. Anonymous code is available at `https://anonymous.4open.science/r/eviagent-CC46`.

1 INTRODUCTION

Surgical videos provide a continuous record of procedural execution, capturing both the progression of surgical workflow and fine-grained interactions among instruments, actions, and anatomical structures. Recent advances have extended surgical video understanding from predefined recognition tasks toward more holistic scene understanding, vision–language representation learning, and visual question answering Seenivasan et al. (2022); Yuan et al. (2025); Ayobi et al. (2025); Drago et al. (2026). Despite this progress, reasoning over a complete surgical procedure remains challenging. A surgery may span tens of minutes or longer, while the observations relevant to a particular query may occupy only a small fraction of the video and may occur at substantially different temporal and semantic scales. Increasing the amount of accessible visual context improves coverage, but does not directly address the problem of identifying the evidence that is actually required for a specific reasoning task.

In long surgical procedures, query-relevant evidence is often **sparse**: most frames contribute little to a particular question, while similar instruments, actions, and interactions may recur at multiple points in the same procedure. Long-context modeling and visual compression improve the efficiency of processing such videos Zhang et al. (2025a), but additional context can still introduce substantial redundancy and does not necessarily disambiguate the relevant occurrence. Evidence requirements are also **heterogeneous**. Surgical reasoning may involve coarse procedural states, fine-grained instruments and actions, spatial relations, temporal locations, or combinations of these signals Ayobi et al. (2025). Consequently, a single observation granularity or generic retrieval strategy is unlikely to provide equally informative evidence for different queries. More importantly, the evidence required during reasoning is **dynamic**: once one component of a query has been resolved, the remaining uncertainty may shift to another semantic dimension or temporal occurrence. Recent agentic video methods already support adaptive navigation, selective observation, and iterative evidence seeking Qiu et al. (2026); Li et al. (2026b); Wang et al. (2026d). However, they typically do not explicitly represent which task-relevant evidence has already been established and which remains unresolved. As illustrated in Figure 1(a), such methods reason over progressively acquired

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.

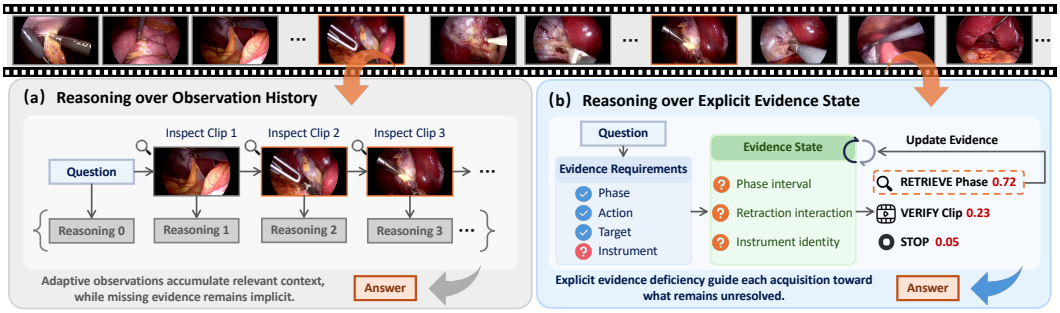


Figure 1: Conceptual comparison between reasoning over observation history and reasoning over an explicit evidence state. (a) Existing agents accumulate observations while leaving missing evidence implicit, whereas (b) EviAgent explicitly tracks established and unresolved evidence to guide subsequent acquisition.

observations, while the evidence still required by the question remains implicit in the reasoning history. As a result, the connection between the current evidence deficiency and the next acquisition decision is not explicitly modeled.

This motivates an **active evidence acquisition** view of long-horizon surgical video reasoning: rather than accumulating observations alone, reasoning should explicitly track what evidence has already been established and what remains unresolved, and use this evolving evidence state to determine what should be acquired next. Under this view, the value of a new observation depends on how much unresolved evidence it resolves, as illustrated in Figure 1(b).

Based on this principle, we propose **EviAgent**, an evidence-seeking agent for long-horizon surgical video reasoning. EviAgent maintains a *structured evidence state* that explicitly represents established and unresolved evidence together with temporal proposals and answer hypotheses. Given this state, an *evidence-gain-guided acquisition policy* adaptively selects temporal retrieval, local verification, or stopping according to the evidence that remains unresolved. The resulting observations are incorporated back into the evidence state, enabling the acquisition strategy to evolve as reasoning progresses. These components respectively address the sparse temporal distribution, heterogeneous evidence requirements, and dynamically evolving information needs of long surgical procedures.

Our contributions are threefold: (1) We formulate **long-horizon surgical video reasoning as active evidence acquisition**, shifting the focus from increasing accessible video context toward identifying and progressively resolving query-dependent evidence requirements. (2) We introduce **EviAgent**, which reasons over an explicit structured evidence state and uses **evidence-gain-guided acquisition** to connect each acquisition decision to the unresolved evidence it is intended to resolve. (3) We evaluate EviAgent on four surgical subsets of MedHorizon Du et al. (2026) under procedure-level cross-fitting, demonstrating improved overall accuracy over existing baselines.

2 Related Work

2.1 Long-Horizon Video Understanding

Long-horizon video understanding is constrained by substantial temporal redundancy and limited visual-token budgets. Existing methods improve scalability through adaptive frame selection, visual-token compression, and memory-based representations. Sophia combines adaptive frame pruning with hierarchical attention for hour-scale videos Zhang et al. (2025a), while Q-Frame and XComp further exploit query-aware frame selection and aggressive spatiotemporal compression Zhang et al. (2025b; 2026). Adaptive memory mechanisms provide another route to preserving relevant long-term information Man et al. (2025). These approaches substantially increase the amount of video context accessible to a reasoner. EviAgent instead focuses on how task-relevant evidence should be progressively acquired as reasoning evolves, rather than solely on how long-video content should be selected or compressed.

2.2 Surgical Video Understanding

Surgical video understanding spans multiple temporal and semantic levels, from workflow recognition to fine-grained interaction modeling. CholecT50 represents surgical activities through instrument–verb–target triplets Nwoye et al. (2022), while GraSP jointly characterizes phases, steps, instruments, and atomic actions Ayobi et al. (2025). More recent work extends surgical analysis toward language-conditioned understanding and reasoning, including Surgical-VQA Seenivasan et al. (2022), SurgVLP Yuan et al. (2025), temporally grounded SurgViVQA Drago et al. (2026), and the spatiotemporal reasoning benchmark SurgCoT Wang et al. (2026a). These studies establish rich procedural, interaction, spatial, and temporal semantics for surgical videos. EviAgent differs in organizing such heterogeneous semantics into an explicit evidence state, where established and unresolved evidence jointly inform subsequent acquisition decisions.

2.3 Agentic Evidence Acquisition for Video Reasoning

Recent video agents increasingly replace fixed video preprocessing with adaptive observation. LongVideo-R1 performs hierarchical temporal navigation Qiu et al. (2026), LensWalk allows an agent to dynamically plan how and where to observe Li et al. (2026b), and VideoThinker learns multi-step temporal retrieval and zooming behaviors Li et al. (2026a). More closely related to evidence acquisition, Active Video Perception formulates long-video understanding as iterative evidence seeking Wang et al. (2026d), while VideoARM and VideoHV-Agent organize reasoning around hierarchical memory and hypothesis verification, respectively Yin et al. (2026); Wang et al. (2026c). MedScope further introduces coarse-to-fine tool-based exploration for long clinical videos Li et al.. These methods establish adaptive observation as an effective paradigm for long-video reasoning. EviAgent builds on this direction but explicitly represents *established* and *unresolved* evidence in a structured state, allowing each acquisition decision to be guided by its expected evidence gain.

3 Methodology

3.1 Problem Formulation

Given a full surgical video $V = \{v_t\}_{t=1}^{T}$ and a question q, the goal is to predict an answer $\hat{y}$. Since answer-relevant evidence is often sparse and distributed across a long procedure, identifying such evidence is central to the task. We formulate long-horizon surgical video reasoning as a sequential evidence acquisition process. At reasoning step t, the agent maintains an evidence state S_t and selects an action a_t from a state-dependent candidate set $\mathcal{A}_t$. Executing the action produces an observation o_t, which updates the state as

$$S_{t+1} = \mathcal{U}(S_t, a_t, o_t), \tag{1}$$

where $\mathcal{U}$ denotes the evidence-state update operator. The reasoning process continues until sufficient answer-supporting evidence has been established:

$$S_0 \xrightarrow{a_0,o_0} S_1 \xrightarrow{a_1,o_1} \cdots \xrightarrow{a_{K-1},o_{K-1}} S_K \xrightarrow{\texttt{STOP}} \hat{y}. \tag{2}$$

3.2 Framework Overview

Figure 2 illustrates the overall framework of EviAgent. Given a question, a *Question Parser* extracts its query conditions and answer specification to initialize a structured evidence state. An *Evidence-Gain-Guided Acquisition Policy* reasons over the current state and selects among instantiated `RETRIEVE`, `VERIFY`, and `STOP` actions. `RETRIEVE` searches a broad temporal scope for candidate evidence intervals, whereas `VERIFY` examines selected intervals at a finer granularity to resolve or validate answer-relevant evidence. The resulting structured observations are incorporated back into the evidence state, forming a closed acquisition loop. As the state evolves, subsequent acquisition decisions adapt to what evidence has already been established and what remains unresolved.

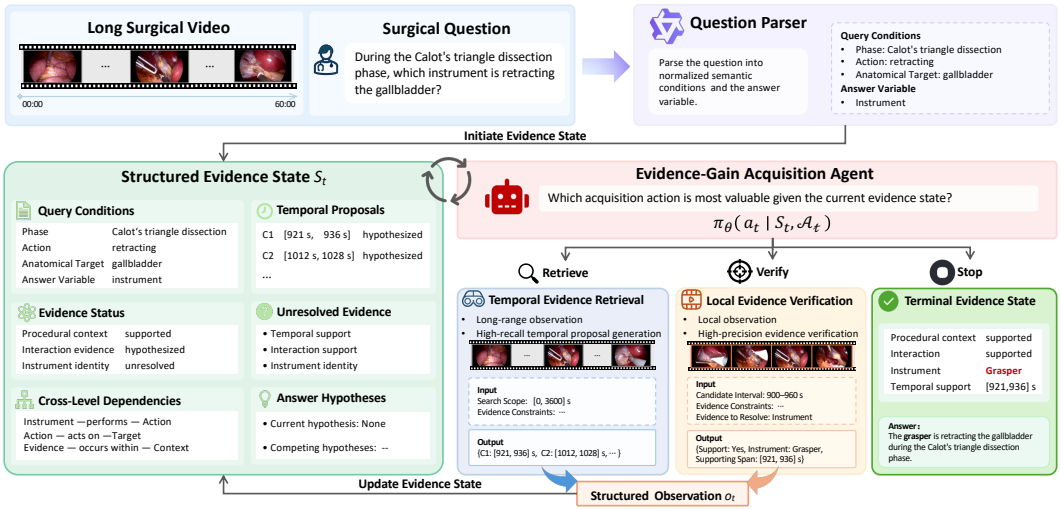


Figure 2: Overview of EviAgent. Given a long surgical video and a question, EviAgent maintains a structured evidence state and uses an evidence-gain-guided policy to adaptively retrieve or verify unresolved evidence. Structured observations update the state in a closed loop to establish sufficient evidence for final answer prediction.

3.3 Structured Evidence Representation

Query Parsing and State Initialization. EviAgent first maps the question to an evidence requirement that specifies the query conditions, the information to be resolved, and the reasoning required to derive the final answer from acquired evidence. For example, the question *"Which instrument is present during the opening moments of the Bladder Neck Identification phase?"* specifies the procedural phase *Bladder Neck Identification* and a temporal condition referring to its onset, while requiring the instrument identity to be resolved from evidence acquired near the phase onset. Importantly, query conditions constrain evidence acquisition but are not themselves treated as visually verified evidence. These parsed requirements are used to initialize the evidence state.

Structured Evidence State. At step t, the current reasoning state is represented as

$$S_t = (\mathcal{C}_q, \mathcal{E}_t, \mathcal{U}_t, \mathcal{P}_t, H_t), \tag{3}$$

where $\mathcal{C}_q$ contains the query-derived conditions and answer requirements, $\mathcal{E}_t$ contains currently supported or hypothesized evidence, $\mathcal{U}_t$ denotes unresolved evidence requirements, $\mathcal{P}_t$ contains candidate temporal proposals, and H_t represents the current answer hypotheses. Evidence may be hypothesized, supported, conflicting, or rejected as reasoning progresses. The state further preserves semantic dependencies among procedural context, instruments, actions, and anatomical targets, allowing heterogeneous observations to be interpreted jointly. Thus, S_t represents not only what has been observed, but also what evidence still needs to be resolved.

3.4 Multi-Granular Evidence Acquisition

Temporal Evidence Retrieval. When the temporal location of required evidence is uncertain, EviAgent performs query-conditioned retrieval over a temporal scope Ω_t. Given a retrieval request r_t^{ret}, the temporal retriever $\mathcal{L}$ produces

$$\mathcal{P}_t^{\text{ret}} = \mathcal{L}\left(V[\Omega_t], r_t^{\text{ret}}\right) = \{\hat{\tau}_i\}_{i=1}^{K}, \tag{4}$$

where $\hat{\tau}_i$ denotes a candidate temporal interval. The request is determined by the evidence currently sought rather than by a fixed search hierarchy. Procedural structure can provide a localization prior when available, but EviAgent does not assume access to a complete or accurate procedural timeline. The retrieved intervals are treated as temporal proposals rather than established evidence, allowing long-range retrieval to prioritize candidate discovery while leaving fine-grained evidence assessment to subsequent reasoning.

Local Evidence Verification. A selected temporal proposal can be examined at a finer granularity using the local verifier $\mathcal{I}$. Given candidate interval $\hat{\tau}_i$ and verification request r_t^{ver},

$$o_t = \mathcal{I}\left(V[\hat{\tau}_i], r_t^{\text{ver}}\right). \tag{5}$$

A `VERIFY` action performs staged local verification: a support stage first determines whether the candidate satisfies the requested evidence conditions, after which a resolution stage, when applicable, extracts its semantic content and temporal support. Verification can therefore confirm a hypothesis, resolve missing evidence, reject an unsupported candidate, or refine its temporal support. Retrieval and verification provide complementary observation granularities without imposing a fixed global acquisition schedule. Subject to their dependency constraints, the policy may interleave retrieval and verification across temporal scopes as the evidence state evolves.

3.5 Evidence-Gain-Guided Acquisition Policy

Candidate Actions and Evidence Gain. For each state, EviAgent constructs a finite set of valid actions

$$\mathcal{A}_t = \mathcal{G}(S_t) = \mathcal{A}_t^{\text{ret}} \cup \mathcal{A}_t^{\text{ver}} \cup \{\texttt{STOP}\}, \tag{6}$$

where retrieval actions are instantiated from unresolved evidence requirements and verification actions from existing temporal proposals. The policy therefore selects among state-dependent acquisition candidates rather than freely generating arbitrary actions. Let $\Phi^*(S_t; q)$ denote an annotation-derived, training-time measure of evidence sufficiency. The immediate evidence gain of an acquisition action is

$$\Delta\Phi(S_t, a) = \mathbb{E}\left[\Phi^*(S_{t+1}; q) - \Phi^*(S_t; q) \mid S_t, a\right]. \tag{7}$$

This quantity measures the immediate progress of an action toward sufficient answer-supporting evidence. However, acquisition decisions are not made greedily from one-step evidence gain alone, since an action may enable useful subsequent acquisitions without immediately increasing verified evidence. We therefore assign each legal action a finite-horizon oracle value $Q^*(S_t, a)$ that accounts for both immediate evidence progress and the downstream utility of subsequent acquisitions, favoring correct and evidence-sufficient termination while penalizing unnecessary acquisition. Both Φ^* and Q^* are available only during training. Immediate evidence gain captures local progress, whereas $Q^*(S_t, a)$ additionally credits actions whose utility is realized through subsequent acquisitions.

State-Dependent Acquisition Policy. The acquisition policy

$$\pi_\theta(a_t \mid S_t, \mathcal{A}_t) \tag{8}$$

reasons over the current evidence state and selects among the valid acquisition actions according to their learned long-horizon utility. Each acquisition induces the closed-loop transition

$$S_t \to a_t \to o_t \to S_{t+1}. \tag{9}$$

Because each observation changes what has been established and what remains unresolved, the relative value of subsequent actions also changes. The policy is trained to prefer `STOP` when the current evidence is sufficient to support an answer; otherwise, acquisition continues until stopping or budget exhaustion.

3.6 Learning and Inference

Evidence Module Learning. During training, source surgical annotations are converted into evidence requests and their corresponding supporting observations. The temporal retriever is trained to localize intervals satisfying a given evidence request, while the local verifier is trained to determine whether a candidate interval supports the requested evidence and to resolve its semantic content when applicable. MedHorizon question–answer pairs and answer labels are not used to train the evidence modules, and annotation-derived supervision is unavailable during inference. Details of evidence-supervision construction are provided in Appendix B.3.

Table 1: Statistics of the MedHorizon surgical subsets used in our experiments.

Subset	Videos	Hours	QA Pairs
GraSP	13	32.4	112
MultiBypass	125	193.3	403
PitVQA	10	10.5	13
TMVP	59	122.7	340
Total	**207**	**358.9**	**868**

Acquisition Policy Learning. After freezing the evidence modules, policy supervision is constructed over training-time acquisition trajectories and their valid state transitions. These trajectories are constructed only from annotation-derived evidence tasks associated with the training procedures. Annotation-derived oracle evidence is used only to evaluate candidate actions and generate supervision, and is excluded from the policy input. For each training state S_t, the candidate action set $\mathcal{A}_t$ is instantiated, and annotation-derived oracle information is used to evaluate the finite-horizon action value $Q^*(S_t, a)$ of each valid action. The preferred action is defined as

$$a_t^* = \arg\max_{a \in \mathcal{A}_t} Q^*(S_t, a). \tag{10}$$

This construction uses oracle information only to define training targets, while the policy itself conditions on the observable evidence state and candidate actions. The primary action-selection objective is

$$\mathcal{L}_{\text{act}} = -\sum_t \log \pi_\theta \left(a_t^* \mid S_t, \mathcal{A}_t\right). \tag{11}$$

This objective is complemented by auxiliary value-regression, pairwise-ranking, and stopping-calibration losses, with additional implementation details provided in the appendix.

Inference. At inference time, EviAgent receives the surgical video and question; in the multiple-choice setting used in our experiments, the candidate answers are additionally provided. The question is parsed to initialize S_0, after which the acquisition policy iteratively selects retrieval, verification, or stopping actions, with the structured evidence state updated after each observation. The acquisition loop terminates when `STOP` is selected or when the acquisition budget is exhausted. A deterministic finalization procedure then maps the terminal evidence state to the final prediction according to the answer requirement identified from the question; in the multiple-choice setting, it returns exactly one candidate answer. No ground-truth annotation or oracle signal is available during inference.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate EviAgent on four surgical subsets of MedHorizon Du et al. (2026): GraSP Ayobi et al. (2025), MultiBypass Lavanchy et al. (2024), PitVQA He et al. (2024), and TMVP Lan et al. (2025). These subsets span robot-assisted prostatectomy, bariatric surgery, pituitary surgery, and cardiac surgery, comprising 207 full-procedure videos, 358.9 hours of video, and 868 multiple-choice questions, as summarized in Table 1. We adopt two-fold cross-fitting with procedure-disjoint training and evaluation splits across all four subsets. For each split, all trainable EviAgent components are trained using annotation-derived supervision from one procedure fold and evaluated on MedHorizon questions from the held-out fold. Results are aggregated over both folds, ensuring that each question is evaluated without training on annotations from its corresponding surgical procedure. MedHorizon questions and answer labels are excluded from training.

Task & Evaluation Metrics. Each model receives the surgical video together with a question and candidate answers and predicts the correct option. The questions cover both fine-grained surgical understanding and higher-level semantic reasoning. We retain the original MedHorizon questions, answer choices, and ground-truth labels, and report multiple-choice accuracy as the evaluation metric.

Comparable Methods. We compare EviAgent with representative baselines spanning four categories. **(1) General-domain MLLMs** include GPT-5.4 OpenAI (2026), Gemini-2.5-Flash Comanici et al. (2025), Kimi-K3 Team et al. (2026), Qwen3-VL-8B-Instruct Bai et al. (2025), InternVL3.5-8B Wang et al. (2025), and LLaVA-Video-7B Zhang et al. (2024b), covering general-purpose multimodal models without explicit specialization for long-horizon surgical video reasoning. We additionally include **Qwen3-VL-8B-Instruct + Annotation SFT** as a supervision-matched control. It is adapted using clip-level semantic supervision constructed from the same training-fold surgical annotations available to EviAgent, but performs direct video question answering without active evidence acquisition. **(2) Specialized long-video models** include VideoChat-Flash-7B Li et al. (2026c) and LongVA-7B-DPO Zhang et al. (2024a), which are specifically designed to improve long-context video understanding. **(3) Video reasoning and agentic models** include Video-R1-7B Feng et al. (2026), VideoRFT Wang et al. (2026b), VideoChat-R1-7B Li et al. (2025), LongVT-7B-RFT Yang et al. (2026), and the Qwen2.5- and Qwen3-based variants of LongVideo-R1 Qiu et al. (2026), covering reasoning-oriented post-training and adaptive interaction with long videos. Finally, **(4) Medical-domain MLLMs** include MedGRPO Su et al. (2026), Lingshu-7B Xu et al. (2026), Hulu-Med-7B Jiang et al. (2025), and HuatuoGPT-Vision-7B Chen et al. (2024), representing multimodal models specialized for medical understanding and reasoning.

Implementation and Inference Protocol. EviAgent uses **Qwen3-VL-8B-Instruct** Bai et al. (2025) as the shared visual-language backbone for temporal retrieval and local evidence verification, and **Qwen3-4B-Instruct-2507** for question parsing and the evidence-gain acquisition policy. The Retriever and Verifier are trained with annotation-derived supervision from the training procedures, while the policy is trained on acquisition trajectories constructed from the same procedure folds. All trainable EviAgent components follow the same procedure-level data separation. At inference, the policy selects among dynamically constructed `RETRIEVE`, `VERIFY`, and `STOP` actions until the policy selects `STOP` or the acquisition budget is exhausted. For all baselines, we use the officially released checkpoints and inference procedures whenever available. All methods are evaluated on the same MedHorizon questions and answer choices, with multiple-choice accuracy as the evaluation metric.

4.2 Main Results

Table 2 presents the comparison with existing methods on the four surgical subsets of MedHorizon. EviAgent achieves the highest overall accuracy of 50.46%, surpassing the strongest existing baseline, Kimi-K3, by 5.76 percentage points. Improvements are observed in both evaluation categories, with EviAgent achieving 46.31% on fine-grained understanding and 57.83% on semantic reasoning, exceeding Kimi-K3 by 7.03 and 3.52 percentage points, respectively.

Compared with its backbone, Qwen3-VL-8B-Instruct, EviAgent improves overall accuracy by 20.05 percentage points. EviAgent also outperforms its annotation-tuned counterpart, demonstrating the advantage of the proposed framework over direct annotation-based adaptation in this evaluation setting.

At the task level, EviAgent achieves the highest accuracy on PR, IR, and RSR, while matching the highest result on CPR. Its performance is particularly notable on PR and IR, which account for the majority of evaluation questions. Although improvements are not uniform across all tasks, the overall results demonstrate the effectiveness of EviAgent across fine-grained understanding and semantic reasoning. These findings motivate a closer examination of structured evidence representation and adaptive acquisition in the following ablation studies.

4.3 Qualitative Results

Figure 3 illustrates how EviAgent acquires and integrates evidence from temporally distributed regions of long surgical videos. In Fig. 3(a), the initial verification establishes evidence from Pedicle Control and produces a provisional hypothesis of *Monopolar Curved Scissors*. Rather than terminating immediately, the Policy continues acquisition across additional temporal regions, accumulating instrument observations from distinct operative phases. The resulting evidence supports the final answer through cross-phase comparison.

Table 2: Comparison with existing methods on four surgical subsets of MedHorizon. We report accuracy (%) on fine-grained understanding and semantic reasoning, together with the overall performance. The best and second-best results across all methods in each column are highlighted in **bold** and underline, respectively.

Model	Fine-grained Understanding			Semantic Reasoning					Overall
	Overall	PR	SD	Overall	IR	WR	RSR	CPR	
GPT-5.4	37.66	37.70	37.50	44.09	44.68	26.92	69.23	30.77	39.98
Gemini-2.5-Flash	30.81	28.22	**41.07**	38.34	37.45	38.46	65.38	19.23	33.53
Kimi-K3	39.28	40.18	35.71	54.31	56.60	26.92	76.92	38.46	44.70
Qwen3-VL-8B-Instruct	26.13	27.54	20.54	38.02	39.57	34.62	57.69	7.69	30.41
InternVL3.5-8B	25.23	25.28	25.00	31.95	31.06	30.77	65.38	7.69	27.65
LLaVA-Video-7B	23.06	23.93	19.64	34.19	29.79	38.46	61.54	42.31	27.07
VideoChat-Flash-7B	23.60	23.93	22.32	23.32	21.70	26.92	15.38	42.31	23.50
LongVA-7B-DPO	25.95	26.41	24.11	31.31	27.23	34.62	53.85	42.31	27.88
Video-R1-7B	21.98	20.32	28.57	35.46	29.36	38.46	69.23	**53.85**	26.84
VideoRFT	21.62	20.32	26.79	37.06	34.04	**42.31**	61.54	34.62	27.19
VideoChat-R1-7B	24.50	22.80	31.25	35.46	33.62	34.62	61.54	26.92	28.46
LongVT-7B-RFT	23.06	20.32	33.93	38.66	35.74	26.92	65.38	50.00	28.69
LongVideo-R1-Qwen2.5	31.71	31.15	33.93	35.78	36.17	26.92	65.38	11.54	33.18
LongVideo-R1-Qwen3	31.89	31.38	33.93	33.55	34.04	30.77	53.85	11.54	32.49
HuatuoGPT-Vision-7B	24.32	23.70	26.79	30.35	28.94	23.08	50.00	30.77	26.50
Lingshu-7B	24.50	22.35	33.04	35.78	34.89	26.92	50.00	38.46	28.57
Hulu-Med-7B	25.95	22.80	38.39	35.78	32.34	**42.31**	61.54	34.62	29.49
MedGRPO	27.57	26.41	32.14	38.66	38.30	23.08	61.54	34.62	31.57
Qwen3-VL-8B + Annotation SFT	34.41	37.47	22.32	41.85	42.55	34.62	61.54	23.08	37.10
EviAgent (Ours)	**46.31**	**47.86**	40.18	**57.83**	**57.87**	34.62	**84.62**	**53.85**	**50.46**

Table 3: Ablation on evidence representation, comparing structured evidence states with accumulated action–observation history under the same acquisition modules and evaluation protocol.

Representation	Fine-grained	Semantic	Overall
Observation History	40.90	48.56	43.66
Structured Evidence State	**46.31**	**57.83**	**50.46**

Figure 3(b) demonstrates how additional evidence can revise an existing answer hypothesis. The Policy first acquires evidence from an intermediate video region, then navigates to a later region before returning to an earlier one. As verified instrument–phase observations accumulate, the hypothesis changes from *Bipolar Forceps* to *Suction Instrument* and subsequently returns to *Bipolar Forceps*. Together, these trajectories illustrate the roles of adaptive temporal acquisition, structured evidence aggregation, and evolving answer hypotheses in long-horizon surgical video reasoning.

4.4 Ablation on Evidence Representation

Table 3 investigates the effect of structured evidence representation. Replacing the structured evidence state with accumulated action–observation history reduces overall accuracy from 50.46% to 43.66%. Performance decreases on both fine-grained understanding (46.31% to 40.90%) and semantic reasoning (57.83% to 48.56%), with a larger reduction observed in the latter. This difference suggests that structured evidence representation is particularly important for reasoning tasks that require integrating observations from multiple temporal regions. By organizing temporal proposals, verified semantic evidence, and evolving answer hypotheses, the structured state provides a more explicit basis for subsequent acquisition and reasoning than accumulated observation history alone.

4.5 Ablation on Acquisition Strategy

Table 4 evaluates the effect of different evidence-acquisition strategies. The one-step-gain policy improves overall accuracy from 42.74% to 46.31% compared with the fixed schedule, indicating the

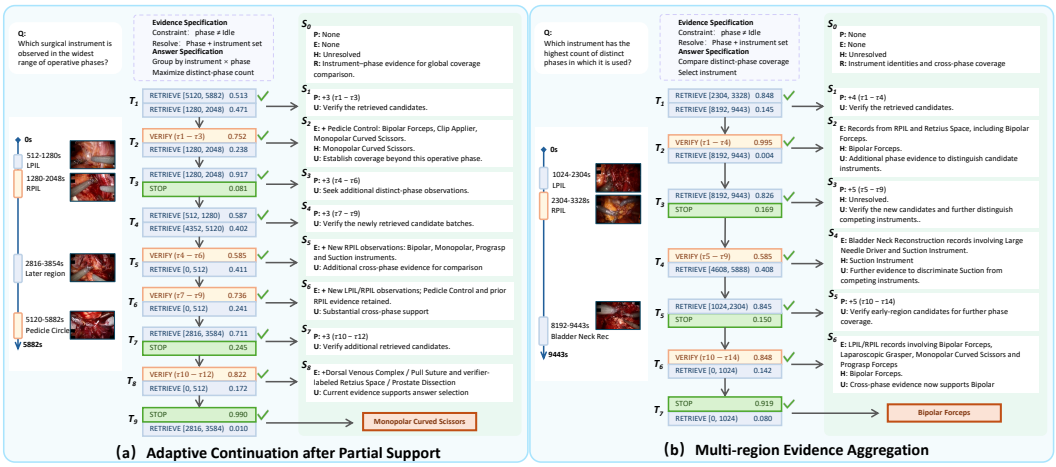


Figure 3: Qualitative acquisition trajectories of EviAgent. The left column shows the retrieved regions along the chronological video timeline, while the middle column presents the Policy execution sequence and action scores. T_i denotes the i-th decision step, and S_i denotes the resulting evidence state, summarized in the right column. P, E, H, and U represent temporal proposals, established evidence, the current answer hypothesis, and unresolved evidence requirements, respectively. (a) illustrates adaptive continuation beyond an initial hypothesis, while (b) demonstrates non-chronological evidence aggregation and hypothesis revision.

Table 4: Ablation on acquisition strategies, comparing fixed scheduling, one-step evidence-gain supervision, and finite-horizon action-value supervision under the same evaluation protocol.

Acquisition Strategy	Fine-grained	Semantic	Overall
Fixed Schedule	40.54	46.65	42.74
One-Step-Gain Policy	44.32	49.84	46.31
Finite-Horizon Policy	**46.31**	**57.83**	**50.46**

benefit of adapting acquisition decisions to the current evidence state. The proposed finite-horizon policy further increases overall accuracy to 50.46%, with improvements in both fine-grained understanding and semantic reasoning. The gain is particularly pronounced in semantic reasoning, where the finite-horizon policy achieves 57.83%, compared with 49.84% for the one-step-gain policy. This result is consistent with the motivation for considering the downstream value of acquisition actions. In particular, temporal retrieval may produce candidate proposals without immediately establishing answer-supporting evidence, while subsequent verification can resolve their semantic content. Finite-horizon action-value supervision accounts for such delayed contributions, providing a broader basis for acquisition decisions than immediate evidence gain alone.

5 Conclusion

We presented EviAgent, an active evidence acquisition framework for long-horizon surgical video reasoning. By explicitly tracking supported and rejected observations, unresolved evidence requirements, and evolving answer hypotheses, EviAgent supports adaptive retrieval and verification through a structured evidence state. Experiments on four surgical subsets of MedHorizon demonstrate improvements in fine-grained understanding and semantic reasoning, while ablation studies examine the contributions of evidence representation and acquisition strategy. These findings demonstrate the potential of active evidence acquisition for reasoning over long surgical videos with temporally distributed evidence.

Reproducibility Statement

An anonymized repository accompanies the submission with implementation code, configurations, and evaluation scripts. Detailed training and inference settings are provided in the Appendix.

Ethics Statement

This work uses publicly available or authorized surgical video datasets. EviAgent is intended solely for research and evaluation and should not be used for autonomous surgery or direct clinical decision-making. Model outputs may be inaccurate and require expert interpretation.

AI Use Statement

Generative AI tools were used in this work solely for language polishing and improving the clarity of writing. No LLMs contributed to research ideation, experimental design, analysis, or the development of core scientific content. All conceptual and technical contributions are original to the authors.

References

Nicolás Ayobi, Santiago Rodríguez, Alejandra Pérez, Isabela Hernández, Nicolás Aparicio, Eugénie Dessevres, Sebastián Peña, Jessica Santander, Juan Ignacio Caicedo, Nicolás Fernández, et al. Pixel-wise recognition for holistic surgical scene understanding. *Medical Image Analysis*, pp. 103726, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, et al. Towards injecting medical visual knowledge into multimodal llms at scale. pp. 7346–7370, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

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. Surgvivqa: temporally grounded video question answering for surgical scene understanding. *International Journal of Computer Assisted Radiology and Surgery*, pp. 1–9, 2026.

Bodong Du, Bowen Liu, Yang Yu, Xinpeng Ding, Zhiheng Wu, Shuning Wang, Shuo Nie, Naiming Liu, Qifeng Chen, Yangqiu Song, et al. Medhorizon: Towards long-context medical video understanding in the wild. *arXiv preprint arXiv:2605.06537*, 2026.

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. volume 38, pp. 99114–99137, 2026.

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

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

Meng Lan, Weixin Si, Xinjian Yan, and Xiaomeng Li. A new dataset and versatile multi-task surgical workflow analysis framework for thoracoscopic mitral valvuloplasty. *Medical Image Analysis*, pp. 103724, 2025.

Joël L Lavanchy, Sanat Ramesh, Diego Dall'Alba, Cristians Gonzalez, Paolo Fiorini, Beat P Müller-Stich, Philipp C Nett, Jacques Marescaux, Didier Mutter, and Nicolas Padoy. Challenges in multi-centric generalization: phase and step recognition in roux-en-y gastric bypass surgery. *International journal of computer assisted radiology and surgery*, 19(11):2249–2257, 2024.

Chenglin Li, Qianglong Chen, Feng Han, Yikun Wang, Xingxi Yin, Yan Gong, Ruilin Li, Yin Zhang, and Jiaqi Wang. Videothinker: Building agentic videollms with llm-guided tool reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 8226–8236, June 2026a.

Keliang Li, Yansong Li, Hongze Shen, Mengdi Liu, Hong Chang, and Shiguang Shan. Lenswalk: Agentic video understanding by planning how you see in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19518–19528, June 2026b.

Wenjie Li, Yujie Zhang, Haoran Sun, Xingqi He, Hongcheng Gao, Chenglong Ma, Ming Hu, Guankun Wang, Shiyi Yao, Renhao Yang, et al. Medscope: Incentivizing" think with videos" for clinical reasoning via coarse-to-fine tool calling.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. In *International Conference on Learning Representations*, volume 2026, pp. 109089–109117, 2026c.

Yuanbin Man, Ying Huang, Chengming Zhang, Bingzhe Li, Wei Niu, and Miao Yin. Adacmˆ 2: On understanding extremely long-term video with adaptive cross-modality memory reduction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8534–8544. IEEE, 2025.

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.

OpenAI. Introducing gpt-5.4, 2026. URL `https://openai.com/index/introducing-gpt-5-4/`. Product release, accessed April 27, 2026.

Jihao Qiu, Lingxi Xie, Xinyue Huo, Qi Tian, and Qixiang Ye. Longvideo-r1: Smart navigation for low-cost long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40505–40515, June 2026.

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

Yuhao Su, Anwesa Choudhuri, Zhongpai Gao, Benjamin Planche, Van Nguyen Nguyen, Meng Zheng, Yuhan Shen, Arun Innanje, Terrence Chen, Ehsan Elhamifar, 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*, pp. 2788–2798, 2026.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Gui Wang, YongSong Zhou, Kaijun Deng, Wooi Ping Cheah, Rong Qu, Jianfeng Ren, and Linlin Shen. Surgcot: Advancing spatiotemporal reasoning in surgical videos through a chain-of-thought benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17011–17021, June 2026a.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. volume 38, pp. 4350–4376, 2026b.

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

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, June 2026c.

Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S Ryoo, and Juan Carlos Niebles. Active video perception: Iterative evidence seeking for agentic long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9088–9099, 2026d.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zuhao Yang, Sudong Wang, Kaichen Zhang, Keming Wu, Sicong Leng, Yifan Zhang, Bo Li, Chengwei Qin, Shijian Lu, Xingxuan Li, et al. Longvt: Incentivizing" thinking with long videos" via native tool calling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33816–33826, 2026.

Yufei Yin, Qianke Meng, Minghao Chen, Jiajun Ding, Zhenwei Shao, and Zhou Yu. Videoarm: Agentic reasoning over hierarchical memory for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24042–24051, 2026.

Kun Yuan, Vinkle Srivastav, Tong Yu, Joel L Lavanchy, Jacques Marescaux, Pietro Mascagni, Nassir Navab, and Nicolas Padoy. Learning multi-modal representations by watching hundreds of surgical video lectures. *Medical Image Analysis*, 105:103644, 2025.

Daoze Zhang, Yuze Zhao, Jintao Huang, and Yingda Chen. Sharper and faster mean better: Towards more efficient vision-language model for hour-scale long video understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4423–4439, 2025a.

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. *arXiv preprint arXiv:2406.16852*, 2024a.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22056–22065, 2025b.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024b.

Zheyu Zhang, Ziqi Pang, Shixing Chen, Xiang Hao, Vimal Bhat, and Yu-Xiong Wang. One token per highly selective frame: Towards extreme compression for long video understanding. volume 38, pp. 57332–57358, 2026.

Appendix

A Additional Method Details

A.1 Structured Evidence Representation

The question parser identifies query conditions, the information to be resolved, and the reasoning requirement for answer aggregation. Candidate answers are available in the multiple-choice setting. Conditions stated in the question constrain acquisition but are not treated as visually verified evidence.

The state $S_t = (\mathcal{C}_q, \mathcal{E}_t, \mathcal{U}_t, \mathcal{P}_t, H_t)$ records query requirements, acquired evidence, unresolved requirements, temporal proposals, and answer hypotheses, respectively. Evidence can be hypothesized, supported, conflicting, or rejected. Dependencies among procedural context, instruments, actions, anatomical targets, and temporal conditions govern how observations are interpreted and combined.

A.2 Multi-Granular Evidence Acquisition

`RETRIEVE` searches a temporal scope for intervals relevant to an evidence request. Retrieved intervals are proposals rather than verified facts. `VERIFY` examines selected proposals through two stages: Support assesses whether the requested conditions are satisfied, and Resolution extracts semantic content and temporal support from supported candidates. Retrieval and verification can be interleaved across temporal scopes, subject to their local dependency constraints.

A.3 Evidence-Gain-Guided Acquisition Policy

The policy ranks a finite set of legal `RETRIEVE`, `VERIFY`, and `STOP` actions. Retrieval candidates are constructed from unresolved requirements, and verification candidates depend on available temporal proposals. The policy receives the observable structured state and candidate actions.

As described in the main text, policy supervision uses annotation-derived finite-horizon action values. These values account for evidence progress, subsequent acquisition utility, and acquisition cost. This permits a retrieval action to receive credit for enabling a later verification even when it does not immediately establish verified evidence. Oracle evidence and action values are used to construct training targets, not as policy inputs.

A.4 State Update and Answer Finalization

Retrieval adds temporal proposals, while verification updates evidence associated with the inspected proposals. Observations are aggregated according to the question's reasoning requirement to update the answer hypothesis. Acquisition terminates when the policy selects `STOP` or exhausts its budget. A deterministic finalization procedure maps the terminal state to one candidate answer.

Termination is a policy or budget decision, not a guarantee that the acquired evidence is complete or the answer is correct. In particular, local supporting observations do not by themselves establish exhaustive coverage for global counting, ranking, or proportion questions. The multiple-choice protocol requires a prediction even when the terminal evidence remains incomplete.

B Data and Evidence Supervision

B.1 MedHorizon Surgical Evaluation

MedHorizon Du et al. (2026) is released as an evaluation-only benchmark. We use the four surgical subsets listed in the main text: GraSP, MultiBypass, PitVQA, and TMVP, totaling 207 procedures and 868 questions. Original questions, answer options, and answer labels are retained for evaluation. Our two-fold procedure-level cross-fitting protocol uses source surgical annotations for training and

evaluates questions from held-out procedures; it is distinct from evaluating an unadapted checkpoint on the original benchmark.

Fine-grained Understanding comprises Phase Recognition (PR) and Step Discrimination (SD). Semantic Reasoning comprises Instrument Recognition (IR), Workload Ranking (WR), Ranked Step Retrieval (RSR), and Count-Proportion Reasoning (CPR). Accuracy is computed over the questions in each task, capability group, or overall evaluation set.

B.2 Source Annotations

The source datasets provide complementary supervision: GraSP Ayobi et al. (2025) includes procedural and instrument/action annotations; MultiBypass Lavanchy et al. (2024) provides phase and step annotations; PitVQA He et al. (2024) includes procedural and instrument information; and TMVP-SurVideo Lan et al. (2025) provides phase and instrument annotations. Supervision uses only annotation types available in the corresponding source dataset.

Sparse observations are treated as evidence only at annotated times and are not extrapolated into unobserved intervals. Missing annotations are treated as unknown rather than as negative evidence.

B.3 Evidence Supervision Construction

Following the training protocol in the main text, evidence requests and targets are constructed from source annotations in the training procedure fold. MedHorizon questions and answer labels are excluded from training.

Retriever examples pair an evidence request and temporal search scope with annotation-supported temporal targets. Repeated supported occurrences are retained when available. Retriever frame selection is based on the search scope and frame budget rather than the target evidence location. Sparse targets remain anchored to explicitly annotated observations.

Support examples assess whether candidate observations satisfy the requested conditions. Negative supervision is used only when the available annotations establish a contradiction; an unannotated interval is not automatically a negative example. Resolution examples use supported candidates to supervise semantic content and temporal support.

After freezing the evidence modules, policy supervision is constructed from acquisition trajectories associated with annotation-derived tasks in the training fold. For each intermediate state, the preferred legal action is selected using the finite-horizon values described in the main text. Annotation-derived task answers, oracle evidence, and action values provide supervision but are excluded from the policy-visible input.

B.4 Data Separation

All examples derived from the same surgical procedure remain in the same fold, including clips, retrieval targets, verifier examples, and policy trajectories. All trainable EviAgent components follow this procedure-level separation. Each fold is evaluated using components trained on the opposite fold, and predictions are pooled across the two held-out evaluations. Ground-truth annotations, oracle temporal spans, and oracle action values are unavailable at inference.

C Implementation Details

C.1 Model Configuration

The visual evidence modules use Qwen3-VL-8B-Instruct Bai et al. (2025). Retriever, Support, and Resolution use separate language-side LoRA adapters with rank 32 and scaling factor 64; the vision encoder remains frozen. LoRA dropout is zero for Retriever and 0.02 for Support and Resolution. Adapters are applied to attention and feed-forward projection layers.

The Parser and Policy use Qwen3-4B-Instruct-2507 Yang et al. (2025). The Parser uses LoRA rank 32 and scaling factor 64; the Policy uses rank 16 and scaling factor 32. The Policy is a state–action

Table 5: Multiple-choice accuracy (%) with single-stage and staged local verification.

Verification	Fine-grained	Semantic	Overall
Single-stage	38.20	46.33	41.13
Support → Resolution	**46.31**	**57.83**	**50.46**

ranker: it scores legal candidates conditioned on the current state and selects the highest-scoring action.

C.2 Training Configuration

The visual modules are trained independently with parameter-efficient fine-tuning, bfloat16 arithmetic, AdamW, and gradient checkpointing. Their losses are computed on assistant output tokens. Retriever training uses temporal search scopes, while Support and Resolution use local observations for condition assessment and semantic extraction.

The policy is trained by supervised learning on oracle-derived action targets, rather than reinforcement learning. Its objective combines the action-selection loss in the main text with auxiliary value-alignment, pairwise-ranking, and stopping-calibration terms:

$$\mathcal{L}_{\text{policy}} = \mathcal{L}_{\text{act}} + \lambda_{\text{val}}\mathcal{L}_{\text{val}} + \lambda_{\text{rank}}\mathcal{L}_{\text{rank}} + \lambda_{\text{stop}}\mathcal{L}_{\text{stop}}. \tag{12}$$

The auxiliary terms align scores with oracle values, preserve action preferences, and supervise the preference between stopping and continuing, respectively. Oracle-derived supervision does not imply that training covers every error pattern of the learned visual modules.

C.3 Inference and Evaluation Protocol

EviAgent receives the full procedure video, question, and candidate answers. The Parser initializes the state, and the Policy alternates acquisition and state updates until stopping or budget exhaustion. The implementation uses a chronological 1-FPS frame representation. Retriever requests uniformly sample at most 256 frames from the current temporal scope; local verification requests use at most 64 frames from the selected interval. Parser and visual-module outputs use deterministic decoding.

Direct video baselines receive frames sampled uniformly from the complete procedure, subject to their input budgets, together with the same question and answer options. Their frame selection does not depend on the correct answer or its annotated evidence location. Agentic baselines use their released inference procedures whenever available. The annotation-SFT control is adapted on the same training-fold source annotations and performs direct video question answering, as described in the main text.

Outputs are normalized to one answer option. Missing predictions and outputs that cannot be mapped to a valid option are counted as incorrect. VideoChat-Flash scores candidate option tokens directly. These results compare predictive accuracy under the stated inference procedures; they do not establish equal computational cost across methods.

C.4 Compute and Reproducibility

Training and evaluation use NVIDIA H200 GPUs. The accompanying anonymized repository provides implementation code, configurations, and evaluation scripts.

D Additional Experiments

Effect of Staged Verification. Table 5 compares single-stage verification with separate support assessment and semantic resolution. Staged verification improves overall accuracy from 41.13% to 50.46%, with gains in both capability groups. This comparison supports the use of staged verification in the evaluated system, but does not directly measure evidence faithfulness or total inference cost.