SurgAgents: A Multi-agent Framework for Embodied Cataract Surgical Intent Prediction

Shengyuan Wang[1*], **Yanzhe Ji**[1*], **Yue Yao**[1†], **Zhenyue Qin**[1,2], **Yang Liu**[3]
[1]The Australian National University [2]Yale University [3]Carnegie Mellon University
{u8164358, u8164360}@anu.edu.au
{yaoyorke, kf.zy.qin}@gmail.com yangliu7@andrew.cmu.edu

Abstract

Predicting next surgical phase, viewed as a form of surgical procedural intent, with high accuracy is essential for robotic real-time surgical systems. Previous approaches struggle with short-term dynamics, occlusions, and phase persistence; end-to-end large vision-language models (LVLMs) can reason from visual inputs but often fail to exploit structured temporal observations. Thus, we introduce **SurgAgents**, a multi-agent framework for embodied surgical intent prediction. SurgAgents decomposes prediction into specialized agents: (1) a ResNet-based single-frame classifier agent extracts per-frame phase information; (2) a YOLO-based instrument specialist agent extracts per-frame instrument sets; (3) a history descriptor agent composes these facts into frame-level descriptions; (4) a history summarizer agent compresses long histories into a compact temporal narrative; (5) a phase predictor agent assembles the narrative and the current frame into a formatted LVLM prompt; (6) a retrieval agent with a confidence-gated Retrieval-augmented generation (RAG) module selectively retrieves similar case patterns when prediction confidence is low. On Cataract-1K, SurgAgents attains **46.54% macro-F1** with GPT-4o, outperforming the baseline, with systematic ablation studies showing that each agent materially contributes to performance. Latency measurements demonstrate feasibility for robotic real-time surgical support during surgery procedures.

1 Introduction

Robotic real-time surgical assistance systems are rapidly moving from research to practice. In the near future, these systems are likely to become standard in operating rooms, providing surgical support, improving safety, and enabling partial automation of routine intraoperative tasks (Qiu et al., 2025). Accurate intent prediction of surgical phase, referring to what will occur in the next few seconds, has immediate clinical value for robotic real-time surgical support and safe automation (Li et al., 2024; Garrow et al., 2021). In ophthalmic procedures such as cataract surgery, reliable intent predictions of surgical operations or phases (for example, viscoelastic injection, capsulorhexis, or phacoemulsification) enable targeted alerts, adaptive user interfaces, and context-aware automation that can reduce cognitive load and improve patient safety (Ghamsarian et al., 2024). Delivering such capability in practice, however, requires resolving two closely coupled technical challenges.

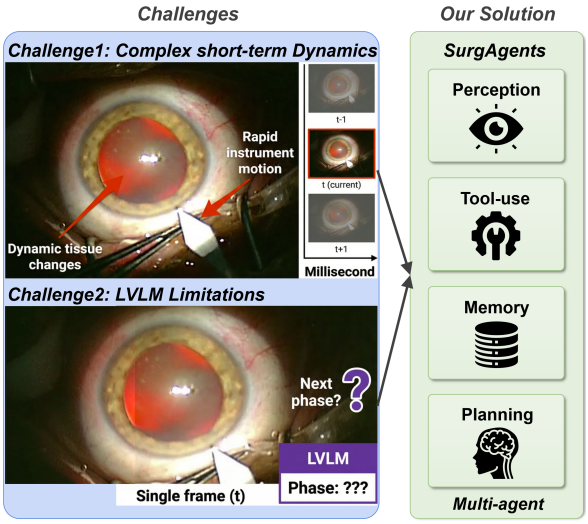


Figure 1: Motivation of our multi-agent framework. Complex short-term dynamics in cataract surgery and the limitations of single-frame LVLMs motivate SurgAgents for surgical intent prediction.

First, surgical workflows present complex short-term dynamics that defeat direct visual inference. Phases often persist over many frames; instruments regularly occlude anatomy and each other (Jalal et al., 2023); surgeons may pause, repeat, or briefly regress to earlier steps; and subtle interleavings of instrument motion and tissue response carry predictive signal. Moreover, visual appearance and work-

[*]Equal contribution.
[†]Corresponding author.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2958–2973
November 6-10, 2026 ©2026 Association for Computational Linguistics

flow state do not have a one-to-one correspondence: the same phase may exhibit substantial visual variation, whereas adjacent phases can appear similar around a transition. Reliable anticipation therefore depends on how these observations evolve over time, rather than on a single static appearance. These characteristics make it difficult for a single end-to-end model to learn robust, generalizable short-horizon dynamics from raw pixels alone: models can latch on to incidental visual correlations, degrade under occlusion, or produce forecasts lacking verifiable intermediate evidence (King et al., 2025; Morita et al., 2019).

Second, large vision–language models (LVLMs) offer powerful and flexible visual reasoning (Hannan et al., 2024), but they are not optimized to extract compact temporal facts directly from dense video streams (Chiu et al., 2025; Rasekh et al., 2025). End-to-end pipelines typically omit explicit intermediate artifacts that clinicians or downstream systems can inspect (Hannan et al., 2024; Cores et al., 2025), and they provide limited mechanisms to incorporate case-level memory selectively when uncertainty is high (Qi et al., 2025). As a result, purely end-to-end strategies face a tradeoff between raw pattern recognition and interpretability, and they may be brittle when confronted with real surgical variability (Rasekh et al., 2025).

To address these issues, we adopt a multi-agent framework that explicitly decomposes surgical intent prediction into modular, interacting agents. "Agent" generally refers to a language-enabled architecture composed of four core components: tool-use (specialized perceptual modules that extract verifiable facts), perception (low-level visual processing), memory (mechanisms to retrieve and integrate relevant past cases or statistics), and planning (an orchestrator that selects and schedules tool invocations while managing latency and resource tradeoffs) (Bandi et al., 2025; Schofield, 2025). This decomposition yields multiple practical advantages. Specialized tools can be designed and validated independently to emit structured, verifiable signals (for example, per-frame phase priors or instrument lists) that materially constrain plausible futures (Zisimopoulos et al., 2018). An explicit memory allows selective retrieval of similar historical patterns when local evidence is ambiguous rather than blind consultation of large corpora (Schofield, 2025). A planner enforces runtime policies, including scheduling, caching, and latency management, so the system can meet real-time surgical timing requirements. Finally, modularity produces an auditable evidence chain: intermediate artifacts offer interpretable explanations for predictions, improving clinical trust, and enabling targeted error analysis (Neha et al., 2025; Baur et al., 2025).

Building on this agentic rationale, we introduce SurgAgents, an instantiation of the multi-agent framework tailored to cataract surgical intent prediction. SurgAgents operationalizes the four core components above by combining validated, task-specific agents used as our tools with a compact, language-friendly intermediate representation and a managed retrieval policy. At a high level, the system extracts structured frame-level facts from the history up to the current time, condenses those facts into a concise temporal narrative and conditions a large vision-language model on the narrative together with the current visual frame to produce a surgical intent prediction. An orchestrator governs tool selection and execution, enforces latency and resource constraints, and records intermediate artifacts.

The design behind SurgAgents is guided by three principles. First, prefer verifiable facts to opaque end-to-end inference: agent-extracted signals restrict hypothesis space and facilitate diagnostic checks. Second, represent multi-frame evidence in a compact, prompt-friendly linguistic form (the temporal narrative) so the LVLM can apply its language-conditioned reasoning without being overloaded by redundant pixel data. Third, use retrieval judiciously: augment local evidence with historical patterns only when model uncertainty warrants the additional cost. Together, these principles balance robustness and computational practicality. In summary, our key contributions are as follows:

- We formalize a multi-agent framework for surgical intent prediction that unifies modular agents, perception, selective retrieval, and planner-driven orchestration.
- We introduce the temporal narrative as a compact intermediate representation that translates multi-frame perceptual facts into language-conditioned evidence suitable for LVLM reasoning.
- We develop a confidence-gated RAG module that selectively augments low-confidence predictions with retrieved historical patterns,

thereby concentrating retrieval cost on ambiguous cases while preserving overall latency constraints.

- We provide a comprehensive and reproducible system design that operationalizes the multi-agent framework in a deployable pipeline, records intermediate artifacts, and yields an auditable evidence chain for each prediction.

2 Related Work

2.1 Surgical workflow recognition and online phase analysis

Surgical workflow understanding has been widely studied as a foundation for real-time surgical support, skill assessment, and post-operative quality control (Liu et al., 2023b, 2025). A central task is surgical phase recognition, where the goal is to infer the current workflow state from video streams (Liu et al., 2023b; Gao et al., 2021). Prior work has explored a broad spectrum of temporal modeling strategies, including recurrent models, temporal convolutional networks, and Transformer-based approaches (Liu et al., 2023b; Mondal et al., 2019; Pan et al., 2023; Liu et al., 2023a; Shi et al., 2020; Gao et al., 2021). These methods often target online recognition settings where only past frames are available (Liu et al., 2023b; Kadkhodamohammadi et al., 2022). A recurring challenge is the long duration of surgical videos and the need to retain informative temporal context without incurring prohibitive latency (Liu et al., 2023a,b).

Recent efforts have emphasized efficiency and stability for real-time deployment. For example, SKiT proposes a fast key-information aggregation mechanism for online phase recognition, aiming to maintain inference time independent of input length while improving recognition accuracy on laparoscopic benchmarks (Liu et al., 2023b). Such designs highlight the importance of capturing "global appeared evidence" under strict runtime constraints, but they still focus on recognizing the current phase rather than anticipating future workflow steps (Liu et al., 2023b). Near phase boundaries, a correct current-phase label does not indicate whether a transition is imminent. Forecasting therefore requires evolving temporal cues. In contrast, SurgAgents explicitly extracts per-frame facts and composes them into a temporal narrative for short-horizon forecasting, while preserving real-time feasibility.

2.2 Surgical workflow anticipation and short-horizon prediction

Beyond recognizing the current phase, workflow anticipation aims to predict what will happen next, which is especially relevant for proactive real-time surgical support (e.g., reminders, resource preparation, safety checks) (Boels et al., 2024; Rivoir et al., 2020; Zhang et al., 2022; Yuan et al., 2022). Prior anticipation work has explored both short-horizon and long-horizon settings (Zhang et al., 2024). For example, instrument-interaction-aware models anticipate future workflow evolution by explicitly modeling tool usage dynamics (Yuan et al., 2022), while generative formulations such as SWAG predict future phase trajectories over longer horizons (Boels et al., 2024).

Our setting differs from most prior work in two key aspects. First, we target *short-horizon phase prediction* (5 seconds ahead), where subtle near-term dynamics and transient occlusions can dominate the signal. Second, rather than relying solely on end-to-end temporal encoders over raw pixels, we build an explicit intermediate evidence chain that is designed to be inspectable: tool-derived facts are summarized into a narrative representation and then consumed by a multimodal reasoner for next-step prediction.

2.3 Instrument-aware modeling as workflow evidence

A long-standing line of work leverages surgical instruments as structured cues for workflow understanding, either via multi-task learning (joint tool/phase modeling) or via two-stage pipelines where tool predictions inform temporal phase inference (Twinanda et al., 2016a,b; Jin et al., 2019; Zisimopoulos et al., 2018). In cataract surgery in particular, instrument presence and transitions can strongly disambiguate visually similar phases, motivating models that treat tool predictions as first-class signals rather than latent features (Zisimopoulos et al., 2018; Al Hajj et al., 2019; Paracchini et al., 2025).

However, most instrument-aware approaches implicitly encode tool evidence within temporal networks and offer limited interpretability about which concrete observations drive a predicted next step (Paracchini et al., 2025; Zhang et al., 2022, 2024). In contrast, our approach separates *evidence extraction* (single-frame phase prior + instrument detections) from *decision making* (multimodal pre-

diction), yielding a modular design where each component can be ablated, audited, and potentially replaced with stronger specialists as they become available.

2.4 Surgical vision–language models and tool-augmented reasoning

Vision–language modeling for surgery has advanced rapidly with the emergence of surgical video–language pretraining. Works such as SurgVLP learn aligned representations from large-scale surgical lecture videos and ASR-derived text, enabling stronger transfer to downstream surgical tasks (Yuan et al., 2025; Honarmand et al., 2024). More recent procedure-aware VLP frameworks further incorporate hierarchical knowledge augmentation to better capture surgical semantics (Yuan et al., 2024).

Despite these advances, directly applying VLMs to raw surgical videos or isolated frames often makes it difficult to (i) incorporate verifiable temporal evidence under latency constraints and (ii) provide an interpretable trail explaining why a particular next step is predicted. Retrieval-augmented generation (RAG) has been explored in healthcare settings to improve factual grounding and traceability by attaching retrieved evidence to generation, and surveys highlight its potential for transparent decision support (Neha et al., 2025; Baur et al., 2025; Amugongo et al., 2025; Brown et al., 2025). Building on these ideas, we use a *confidence-gated* retrieval strategy based on similar past cases that invokes RAG only when the predictor's confidence falls below a threshold, trading occasional retrieval cost for improved grounding and traceability.

3 Method

In this section, we provide a detailed workflow of **SurgAgents**, a multi-agent framework for cataract surgical intent prediction. We describe each agent, the data flow, and the design choices that enforce causal evaluation, produce verifiable intermediate artifacts, and control latency for robotic real-time surgical use. Notation: video $V = \{f_1, \ldots, f_T\}$, query time t (the last observed time point), sampling stride δ (experiments use $\delta = 3$ s), prediction horizon Δ (experiments use $\Delta = 5$ s).

3.1 Sampling

SurgAgents constructs a causal history by uniformly sampling frames at fixed stride δ from the start of the procedure up to and including time t. Formally, the sampled index set is

$$S_t = \{k\delta \mid k \in N,\ k\delta \le t\}, \tag{1}$$

and the resulting history frames are

$$H_F(t) = \{f(s) \mid s \in S_t\}. \tag{2}$$

Fixed-stride sampling enforces a strict no-lookahead protocol, regularizes temporal alignment across cases, and bounds the length of downstream textual histories. These properties facilitate fair model comparison and allow system designers to allocate predictable compute and token budgets. In practice, δ and Δ are chosen to balance temporal resolution, signal strength for short-horizon prediction, and allowable latency.

3.2 History Evidence Extraction

For every history frame $f(s) \in H_F(t)$, SurgAgents extracts structured, verifiable facts using specialized perception agents. The principal components are a single-frame phase classifier agent and an instrument detector agent(instrument specialist agent). The classifier maps an image to a probability vector over the canonical phase vocabulary,

$$\mathbf{p}_s = \mathrm{Classifier}(f(s)) \in \Delta^{C-1}, \tag{3}$$

where C is the number of target phases and Δ^{C-1} denotes the $(C-1)$-simplex. We preserve compact numeric summaries such as top-k phases and the distribution entropy

$$H(\mathbf{p}_s) = -\sum_c \mathbf{p}_s[c] \log \mathbf{p}_s[c], \tag{4}$$

for later use. The instrument specialist yields a (possibly multisets) collection of detected instrument classes with confidences,

$$\mathcal{I}_s = \{(i_j, \rho_j)\}, \tag{5}$$

where i_j denotes an instrument class and ρ_j its detection confidence. Low-confidence detections ($\rho_j < \varepsilon$) are removed and short-window smoothing reduces transient noise.

Both agents apply light anomaly filtering and minimal temporal smoothing to suppress spurious tokens that would degrade downstream summaries. Each agent records raw and filtered outputs separately so that calibration, error analysis, and targeted improvements remain possible without changing the production pipeline.

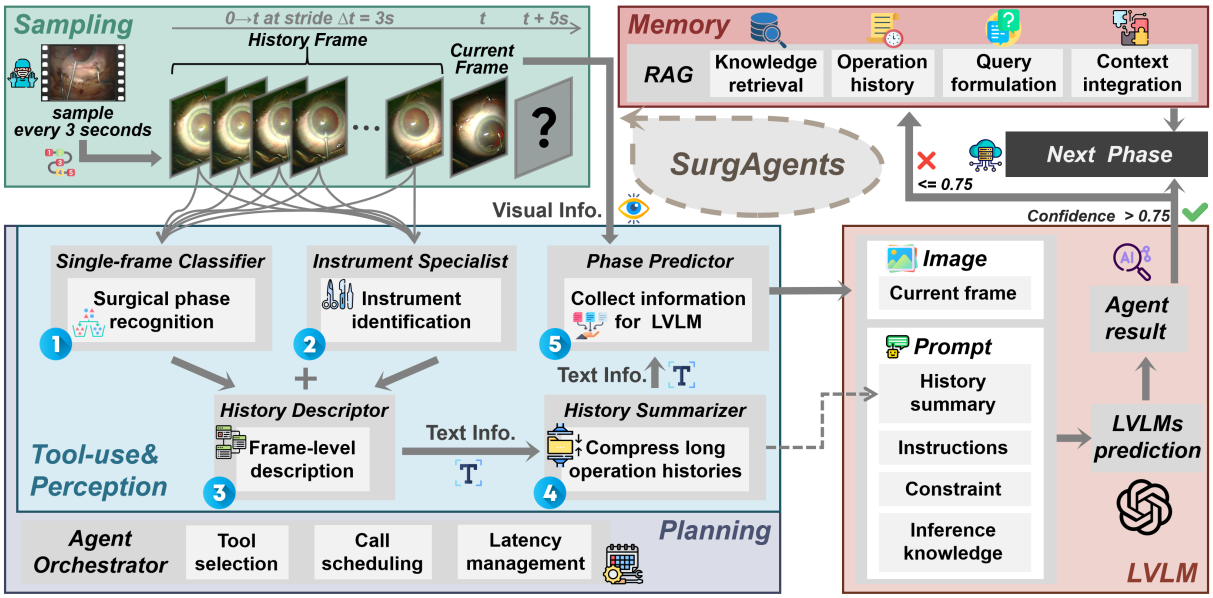


Figure 2: Illustration of SurgAgents framework. History frames are sampled and analyzed by specialized agents to extract phase and instrument information. A history descriptor agent and summarizer agent convert this into a compressed temporal narrative. The phase predictor agent combines the narrative with the current frame and feeds it to an LVLM for surgical intent prediction. Low-confidence predictions trigger a RAG module, while the agent orchestrator coordinates tool selection and agent scheduling under latency constraints.

3.3 History Evidence Structuring

The history evidence extraction agents' outputs are transformed into deterministic frame descriptors that are concise, canonical, and parseable. A descriptor agent

$$d_s = D(\mathbf{p}_s, \mathcal{I}_s) \tag{6}$$

produces one descriptor per sampled time s. Descriptors follow a fixed template and vocabulary to ensure consistency and reproducibility: identical agent outputs always yield identical descriptors, and token length per descriptor is bounded to control narrative size. The ordered descriptor sequence

$$\mathcal{D}_t = (d_{s_1}, \ldots, d_{s_m}) \tag{7}$$

is both the input to the summarizer agent and a persisted audit artifact.

Deterministic templates and strict tokenization reduce the chance of downstream language model hallucination by constraining the input space. Descriptor metadata (agent confidences, smoothing decisions, filter activations) are logged alongside descriptors to support systematic failure analysis.

3.4 Temporal Narrative Reasoning

The summarizer agent compresses $\mathcal{D}_t$ into a compact, informative temporal narrative $\mathcal{N}_t$ that answers "what has happened so far" and "what trend is emerging," while respecting a token budget $L_{\max}$. We adopt a deterministic, rule-based compression strategy: consecutive idle descriptors are collapsed, repeated phase priors are merged into time spans, instrument transitions are marked as events, and short high-confidence phase stretches are highlighted. This rule-based approach guarantees reproducibility; where sufficient paired descriptor–summary data exists, a learned compressor may be evaluated, but the default remains deterministic to preserve auditability.

The summarizer agent prioritizes recent non-idle events and instrument transitions because these elements consistently carry the strongest short-horizon predictive signal. The resulting temporal narrative is intentionally compact and uses canonical operation tokens and explicit time references, so it is both prompt-friendly for LVLM consumption and directly inspectable by clinicians.

3.5 Multimodal Prediction

The phase predictor agent constructs a strict prompt from the textual temporal narrative $\mathcal{N}_t$ and the current frame (visual evidence) $f(t)$. The prompt contains a short canonical label glossary, the narrative, and a concise instruction such as:

> "Predict the operation at $t + \Delta$. Output only the exact canonical label."

Constraining the prompt and label vocabulary reduces ambiguity and encourages outputs that align with the canonical answer set.

The LVLM processes this prompt and returns a score vector

$$\boldsymbol{\ell} = g(\mathcal{N}_t, f(t)), \tag{8}$$

which we convert to a probability distribution via softmax,

$$\hat{\mathbf{q}} = \mathrm{softmax}(\boldsymbol{\ell}). \tag{9}$$

The predicted label is

$$\hat{y} = \arg\max_c \hat{\mathbf{q}}[c], \tag{10}$$

and the associated confidence is

$$\mathrm{conf} = \max_c \hat{\mathbf{q}}[c]. \tag{11}$$

We calibrate these scores on held-out data so that the confidence value is meaningful for downstream gating decisions. Prompt construction and LVLM decoding are kept compact by enforcing token and latency budgets; when time permits, the orchestrator may append a bounded number of recent descriptors, otherwise the narrative with the current frame constitutes the core evidence.

To limit spurious generations, decoding can be constrained to the canonical label set; when available, instruction-tuned LVLMs trained on in-domain prompt–label pairs improve token alignment and reduce hallucination.

3.6 Confidence-Gated Retrieval

When the LVLM returns a low confidence, i.e. $\mathrm{conf} \leq \tau$ (threshold τ set on validation), SurgAgents triggers a targeted retrieval step. A compact retrieval key $\phi(\mathcal{N}_t)$, designed to emphasize recent instrument transitions and dominant phase priors, is used to fetch the top-K historical snippets $\{r_1, \ldots, r_K\}$ from a protected memory store that enforces strict test-case exclusion.

The retrieved snippets are concatenated with the temporal narrative to form an augmented input under an additional token budget L_{RAG}. The LVLM is then re-invoked on this augmented prompt to produce an updated distribution and confidence. The orchestrator caps K and enforces a retrieval time budget T_{ret}; if confidence remains low after retrieval, predefined fallback rules apply (for example, returning the highest prior, using an ensemble vote, or flagging the case for human review). This selective retrieval policy concentrates extra computation on ambiguous cases where historical context yields the largest marginal benefit, while preserving average latency.

Design safeguards include strict test exclusion in the memory index, canonicalized snippet formats to reduce model drift, and retrieval timeouts to avoid violating response deadlines.

4 Experiment

4.1 Experiment Setup

Datasets. We evaluate SurgAgents and baselines on three publicly available cataract surgery corpora: (1) Cataract-1K (Ghamsarian et al., 2023) contains 56 full-length cataract surgery procedures with frame-level labels for 12 canonical phase classes plus an idle class; it additionally provides bounding-box annotations for surgical instruments, enabling explicit training of instrument detectors. (2) Cataract-101 (Schoeffmann et al., 2018) comprises 101 cataract surgeries captured from multiple surgeons and devices, annotated with time-synchronized phase labels. (3) ITEC Cat21 (Primus et al., 2018) is a smaller multi-institutional corpus with 21 annotated cataract surgery videos and predefined training/validation splits. Importantly, only Cataract-1K contains instrument bounding boxes, so the YOLO-based instrument specialist agent is trained on Cataract-1K and applied at test time to all three corpora. This transfer strategy follows common cross-dataset instrument-detection and domain-adaptation practices in surgical vision research (Paranjape et al., 2023). For the ResNet-based single-frame classifier agent, each dataset has its own trained model, which is not transferred across datasets. The detailed train/test split statistics for all datasets are reported in Appendix A.1.

Upstream components. The single-frame phase classifier uses ResNet-50 and is trained separately on each dataset. It achieves held-out accuracies of 85.45% on Cataract-1K, 89.58% on Cataract-101, and 86.29% on ITEC Cat21. The instrument specialist uses YOLOv11, trained on the Cataract-1K bounding-box annotations, and achieves an mAP of 54.10%. We use a detection confidence cutoff of $\epsilon = 0.50$. Since Cataract-101 and ITEC Cat21 do not provide compatible instrument bounding-box annotations, detector mAP cannot be directly evaluated on these datasets.

Baseline. Our work focuses on a new task formulation rather than proposing a new method over

Method	Cataract-1K		Cataract-101		ITEC Cat21	
	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
LVLM-only (Baseline)						
- MEDGEMMA-4B-IT	4.87	0.76	3.77	1.45	2.72	0.55
- INTERNVL3-8B-INSTRUCT	6.01	0.94	4.97	1.93	3.14	0.62
- GPT-4O	10.53	7.62	8.33	4.81	4.35	1.18
SurgAgents (Ours)						
- MEDGEMMA-4B-IT	52.41	23.89	11.53	7.14	10.90	3.37
- INTERNVL3-8B-INSTRUCT	25.57	17.14	22.94	11.12	16.67	8.58
- GPT-4O	**53.92**	**46.54**	**40.83**	**36.72**	**32.48**	**19.01**

Table 1: Performance of baseline (LVLM-only) and SurgAgents across different LVLMs (medgemma-4b-it, InternVL3-8B-Instruct, GPT-4o) on three cataract datasets in terms of Accuracy(%) and Macro-F1(%).

existing baselines. Based on this new task setting, we introduce a basic method (LVLM-only) as our baseline for comparison with SurgAgents. The basic method is straightforward and commonly used in existing systems, e.g., SV2P (Babaeizadeh et al., 2018). **LVLM-only.** Performs no history evidence extraction: there is neither classifier nor instrument specialist, nor history descriptor, history summarizer, or RAG module. The phase predictor concatenates the history frames and current frame into a horizontal strip image, which is fed directly to the LVLM for surgical intent prediction. We use Accuracy/macro-F1 scores as the evaluation metrics. Implementation details of the baseline are provided in Appendix A.2.

4.2 Experimental Results

We evaluate SurgAgents and the baseline (LVLM-only) on the test splits of Cataract-1K, Cataract-101, and ITEC Cat21. Sampling stride and prediction horizon are $\delta = 3$s and $\Delta = 5$s. Intent prediction is performed with three LVLMs: GPT-4o (OpenAI, 2024), InternVL3-8B-Instruct (Zhu et al., 2025), and medgemma-4b-it (Sellergren et al., 2025); GPT-4o is queried via hosted API while the others are invoked through Hugging Face endpoints.

Table 1 reports the quantitative comparison between the LVLM-only baseline and SurgAgents across three LVLMs and three cataract datasets. SurgAgents yields large and consistent gains over the baseline for all evaluated models and datasets. Most notably, when paired with GPT-4o, SurgAgents improves Cataract-1K accuracy from 10.53% to 53.92% (+43.39%) and macro-F1 from 7.62% to 46.54% (+38.92%). Additionally, comparable absolute gains are observed on Cataract-101 (accuracy +32.50%, macro-F1 +31.91%) and ITEC Cat21 (accuracy +28.13%, macro-F1 +17.83%). With the open-source InternVL3-8B-Instruct model, SurgAgents also provides improvements (e.g., Cataract-1K accuracy 25.57% vs. 6.01% baseline; macro-F1 17.14% vs. 0.94%). The medgemma-4b-it model shows that SurgAgents achieves a large improvement on Cataract-1K (accuracy 52.41% vs. 4.87%), but smaller gains on Cataract-101 and ITEC Cat21. Overall, these results demonstrate that the multi-agent framework consistently improves surgical intent prediction, with the magnitude of improvement depending on the underlying LVLM.

We observe that macro-F1 is consistently lower than accuracy across datasets and evaluated LVLMs (Table 1). This gap is expected in surgical intent prediction for three reasons.

First, phase labels are typically imbalanced: frequent, long-duration phases contribute many more samples than rare or short phases, so accuracy is dominated by majority classes, whereas macro-F1 weights each class equally and is therefore more sensitive to minority-phase errors.

Second, the task exhibits strong phase persistence over a 5-second horizon: for many timestamps, the correct label at $t{+}5$s remains the same as the current phase. A model can thus obtain moderate accuracy by over-predicting persistent phases, while still missing brief transitions and short phases that matter for per-class F1.

Third, we observe a characteristic *phase-lag* behavior: when the model's understanding of the surgical intent forecasting objective is limited, it tends to produce temporally shifted predictions. Such systematic temporal misalignment disproportionately harms per-class precision/recall for transition-associated phases, leading to large class-wise F1

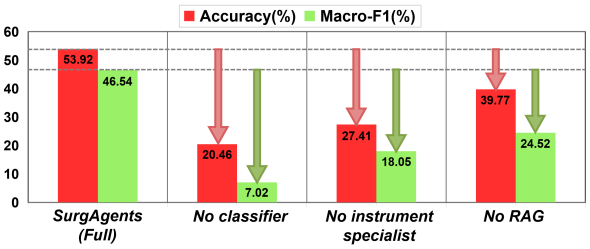


Figure 3: Results of leave-One-Out agent ablation study. The bar chart reports accuracy (%) in red and macro-F1 (%) in green for the full SurgAgents system and for variants with one component agent removed (no classifier, no instrument specialist, and no RAG).

variance and, consequently, a lower macro-F1 even when overall accuracy appears relatively higher.

4.3 Ablation Study

Leave-One-Out agent ablation study. We conduct a leave-one-out agent ablation study to quantify the contribution of each agent. Figure 3 reports accuracy (%) and macro-F1 (%) on Cataract-1K for four configurations: *SurgAgents (Full)*, *SurgAgents (No classifier)*, *SurgAgents (No instrument specialist)*, and *SurgAgents (No RAG)* with GPT-4o.

Removing any major agent leads to substantial degradation, confirming that SurgAgents's intent prediction performance relies on complementary evidence sources and agent orchestration within a multi-agent framework.

SurgAgents (Full) achieves 53.92% accuracy and 46.54% macro-F1. When the classifier agent is removed, performance drops to 20.46% accuracy and 7.02% macro-F1, corresponding to a decrease of 33.46% in accuracy and 39.52% in macro-F1. This indicates that the classifier agent provides a critical phase prior that stabilizes downstream agent reasoning. Disabling the instrument specialist agent also causes a marked decline to 27.41% accuracy and 18.05% macro-F1 (-26.51% accuracy, -28.49% macro-F1), demonstrating that explicit instrument cues are highly informative for anticipating near-future surgical actions. In cataract surgery, many phases exhibit subtle appearance changes, while instrument presence and transitions act as strong discriminative signals; removing these cues makes the predictor more prone to confusing adjacent or visually similar phases. Finally, disabling RAG reduces performance to 39.77% accuracy and 24.52% macro-F1 (-14.15% accuracy, -22.02% macro-F1). Although RAG is only invoked under low confidence, the drop suggests that uncertainty-driven retrieval meaningfully improves hard cases by providing relevant historical patterns to refine the LVLM's prediction.

Method	Accuracy	Macro-F1
TeCNO	9.85	1.50
SV-RCNet	5.28	1.68
SurgAgents (GPT-4o)	**53.92**	**46.54**

Table 2: Comparison of non-LVLM models and SurgAgents (GPT-4o) for surgical intent prediction on Cataract-1K. Accuracy and macro-F1 are reported as percentages.

Overall, the agent ablation results obtained on Cataract-1K with GPT-4o show that SurgAgents benefits from (i) robust phase priors from the classifier, (ii) interpretable instrument evidence, and (iii) selective retrieval for ambiguous contexts, and that these components are not redundant but jointly necessary for strong surgical intent prediction.

End-to-end latency. The measured end-to-end latency of SurgAgents on Cataract-1K is 5.3 s per sample, averaged over the test set, when using GPT-4o. Because our prediction horizon Δ is 5 s, the observed latency closely matches the target interval, implying that the system's inference completes at approximately the moment the predicted horizon is reached. This alignment reduces temporal misalignment risk but leaves little slack for additional processing: therefore, confidence-gated RAG is constrained by orchestrator-enforced retrieval budgets T_{ret}. In deployment we enforce retrieval timeouts and deterministic fallback rules to ensure deadlines are not violated.

Comparison between non-LVLM models and SurgAgents. To evaluate the advantage of our LVLM-based multi-agent framework, we compared SurgAgents with two representative non-LVLM surgical phase predictors, TeCNO (Czempiel et al., 2020) and SV-RCNet (Jin et al., 2018), under the same evaluation protocol on Cataract-1K. Both models rely on end-to-end temporal modeling over visual frames without leveraging structured intermediate evidence or language-conditioned reasoning. TeCNO achieves a macro-F1 of 1.50%, while SV-RCNet reaches 1.68%, demonstrating that conventional visual-only architectures struggle to capture short-horizon dynamics and instrument-informed cues. In contrast, SurgAgents attains 46.54% macro-F1 using GPT-4o, highlighting the critical contribution of modular evidence extraction, temporal narrative reasoning, and confidence-

Dataset	Acc. before	Acc. after	Gain	Degraded	Raw ECE	Cal. ECE
Cataract-1K	39.77%	53.92%	+14.15 pp	7.62%	11.42%	3.18%
Cataract-101	30.25%	40.83%	+10.58 pp	10.14%	13.87%	3.94%
ITEC Cat21	23.67%	32.48%	+8.81 pp	11.58%	16.26%	4.87%

Table 3: Confidence-gated RAG analysis with GPT-4o across the three datasets. "Degraded" denotes the proportion of retrieval-augmented queries for which the retrieved evidence degrades the prediction.

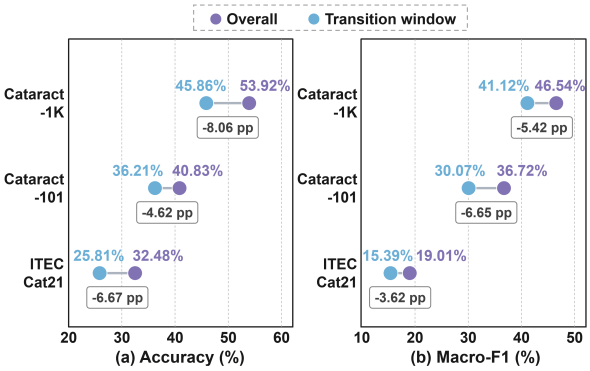


Figure 4: Overall and transition-window performance with GPT-4o across the three datasets. Annotations show the decrease from overall to transition-window performance in percentage points.

gated retrieval. This performance gap underscores the importance of the multi-agent framework for accurate surgical intent prediction.

Transition-window evaluation. To control for phase persistence and evaluate boundary anticipation, we perform a transition-window analysis. A query is included in the transition-window subset if its ground-truth phase changes within the following 5 s. This setting removes most stable-phase queries and focuses the evaluation on cases where simply retaining the current phase is less likely to produce a correct prediction. As shown in Figure 4, SurgAgents achieves 45.86% accuracy and 41.12% macro-F1 on the Cataract-1K transition windows, compared with 53.92% accuracy and 46.54% macro-F1 on the complete test set. This corresponds to decreases of 8.06 and 5.42 percentage points, respectively. The corresponding transition-window results are 36.21% accuracy and 30.07% macro-F1 on Cataract-101, representing decreases of 4.62 and 6.65 percentage points. On ITEC Cat21, SurgAgents obtains 25.81% accuracy and 15.39% macro-F1, compared with the overall results of 32.48% and 19.01%, respectively. Although performance decreases near phase boundaries, the transition-window results remain relatively close to the overall test-set performance across all three datasets. This indicates that the aggregate gains of SurgAgents are not explained solely by phase persistence. The model retains substantial predictive capability near workflow boundaries, where the current phase provides a weaker prior and instrument changes, recent phase progression, and other temporal cues become more important for anticipating the upcoming phase.

Retrieval reliability. The RAG store is built only from training cases, with strict split-level exclusion such that no test videos are indexed. Each memory item is a compact case snippet derived from the same intermediate artifacts used at inference. Retrieved snippets are injected only when the LVLM confidence is below the gating threshold. Table 3 reports the retrieval behavior across datasets. Accuracy improves from 39.77% to 53.92% on Cataract-1K, from 30.25% to 40.83% on Cataract-101, and from 23.67% to 32.48% on ITEC Cat21. Retrieval degrades the prediction in 7.62%, 10.14%, and 11.58% of retrieval-augmented cases, respectively, showing that retrieved evidence is beneficial overall but not uniformly reliable. Confidence calibration reduces ECE from 11.42% to 3.18%, from 13.87% to 3.94%, and from 16.26% to 4.87%, respectively, providing a more reliable control signal for deciding when to trigger retrieval at inference time.

5 Conclusion and Future Work

In this paper, we investigated cataract surgical intent prediction and proposed SurgAgents, a multi-agent framework, which integrates single-frame classifier agent, instrument specialist agent, history descriptor agent, history summarizer agent, phase predictor agent, and a retrieval agent (a RAG module) to construct a temporal narrative for LVLM reasoning. Extensive experiments on multiple cataract datasets with three LVLMs validate the effectiveness of our approach and the necessity of each agent. Future work will focus on enhancing agent specialization via embodied perception–action integration, optimizing narrative compression, and extending SurgAgents to longer-horizon intent predictions and diverse surgical procedures.

Limitations

Despite the strong empirical performance of SurgAgents, the current framework has several inherent limitations:

- **Upstream-agent dependency and error propagation.** SurgAgents's modular pipeline depends on upstream agents (single-frame classifier, instrument specialist). Errors like misclassifications, missed detections, or temporally inconsistent outputs are materialized as concrete tokens in the temporal narrative and can systematically bias LVLM reasoning. While modularity facilitates targeted diagnostics and correction, it does not eliminate the need for robust error detection, uncertainty-aware fusion, or recovery mechanisms for upstream failures.

- **Limited cross-domain generalization.** Evaluation is concentrated on cataract datasets (Cataract-1K, Cataract-101, ITEC Cat21). Differences in procedure semantics, instrument inventories, imaging hardware, and institutional workflows mean that measured gains may not transfer to other surgical domains. Addressing this limitation requires larger, heterogeneous multi-center corpora, domain-adaptation strategies, and calibration procedures for new instrument sets and workflow idiosyncrasies.

Ethical Considerations

SurgAgents is presented as a clinical decision-support multi-agent rather than an autonomous controller. Two main ethical concerns deserve explicit attention.

Clinical safety and human oversight. Outputs are advisory and must be presented with calibrated confidence and transparent provenance (frame evidence, classifier outputs, narrative summary). A human-in-the-loop workflow is required: clinicians retain final authority, receive training on typical failure modes, and can inspect intermediate artifacts to assess trustworthiness. Alerting policies should be conservative to avoid distracting or misleading clinicians and to reduce alert fatigue in the operating room.

Bias, generalization, and validation. Current evaluations are concentrated on ophthalmic datasets with limited device and population diversity; model performance may degrade across institutions, instrument sets, or patient subgroups. Responsible deployment therefore requires (i) stratified evaluation and reporting of subgroup performance, (ii) domain adaptation or re-calibration for new settings, and (iii) prospective clinical validation and human factors studies before operational use. Documentation of limitations and staged rollouts with monitoring and rollback procedures is essential to mitigate harm.

References

H Al Hajj, M Lamard, P-H Conze, S Roychowdhury, X Hu, G Marsalkaite, O Zisimopoulos, MA Dedmari, F Zhao, J Prellberg, M Sahu, A Galdran, T Araujo, DM Vo, C Panda, N Dahiya, S Kondo, Z Bian, A Vahdat, and 19 others. 2019. CATARACTS: Challenge on automatic tool annotation for cataract surgery. *Medical Image Analysis*, 52:24–41.

L. M. Amugongo, P. Mascheroni, S. Brooks, S. Doering, and J. Seidel. 2025. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*.

Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H. Campbell, and Sergey Levine. 2018. Stochastic variational video prediction. In *International Conference on Learning Representations (ICLR)*. Poster.

Ajay Bandi, Bhavani Kongari, Roshini Naguru, Sahitya Pasnoor, and Sri Vidya Vilipala. 2025. The rise of agentic AI: A review of definitions, frameworks, architectures, applications, evaluation metrics, and challenges. *Future Internet*, 17(9):404.

David Baur, Jörg Ansorg, Christoph-Eckhard Heyde, and Anna Voelker. 2025. Development and evaluation of a retrieval-augmented generation chatbot for orthopedic and trauma surgery patient education: Mixed-methods study. *JMIR Artificial Intelligence*, 4:e75262.

Maxence Boels, Yang Liu, Prokar Dasgupta, Alejandro Granados, and Sebastien Ourselin. 2024. SWAG: Long-term surgical workflow prediction with generative-based anticipation. *arXiv preprint arXiv:2412.18849*.

Andrew Brown, Muhammad Roman, and Barry Devereux. 2025. A systematic literature review of retrieval-augmented generation: Techniques, metrics, and challenges. *arXiv preprint*. ArXiv:2508.06401.

Bo-Cheng Chiu, Jen-Jee Chen, Yu-Chee Tseng, Feng-Chi Chen, and An-Zi Yen. 2025. DaMO: A data-efficient multimodal orchestrator for temporal reasoning with video LLMs. *arXiv preprint*. ArXiv:2506.11558.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. 2025. Lost

in time: A new temporal benchmark for video LLMs. In *CVPR 2025 Workshops*.

Tobias Czempiel, Magdalini Paschali, Matthias Keicher, Walter Simson, Hubertus Feussner, Seong Tae Kim, and Nassir Navab. 2020. TeCNO: Surgical phase recognition with multi-stage temporal convolutional networks. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2020, Lecture Notes in Computer Science*, volume 12263, pages 343–352. Springer.

Xiaojie Gao, Yueming Jin, Yonghao Long, Qi Dou, and Pheng-Ann Heng. 2021. Trans-SVNet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. *arXiv preprint*. ArXiv:2103.09712.

Carly R. Garrow, Karl-Friedrich Kowalewski, Linhong Li, Martin Wagner, Mona W. Schmidt, Sandy Engelhardt, Daniel A. Hashimoto, Hannes G. Kenngott, Sebastian Bodenstedt, Stefanie Speidel, Beat P. Müller-Stich, and Felix Nickel. 2021. Machine learning for surgical phase recognition: A systematic review. *Annals of Surgery*, 273(4):684–693.

Negin Ghamsarian, Yosuf El-Shabrawi, Sahar Nasirihaghighi, Doris Putzgruber-Adamitsch, Martin Zinkernagel, Sebastian Wolf, Klaus Schoeffmann, and Raphael Sznitman. 2023. Cataract-1K: Cataract surgery dataset for scene segmentation, phase recognition, and irregularity detection. *arXiv preprint*. ArXiv:2312.06295.

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

Tanveer Hannan, Md Mohaiminul Islam, Jindong Gu, Thomas Seidl, and Gedas Bertasius. 2024. ReVisionLLM: Recursive vision-language model for temporal grounding in hour-long videos. *arXiv preprint*. ArXiv:2411.14901.

Mohammadmahdi Honarmand, Muhammad Abdullah Jamal, and Omid Mohareri. 2024. VidLPRO: A video-language pre-training framework for robotic and laparoscopic surgery. In *NeurIPS Workshop on Advancements in Medical Foundation Models*.

Nour Aldeen Jalal, Tamer Abdulbaki Alshirbaji, Paul David Docherty, Herag Arabian, Bernhard Laufer, Sabine Krueger-Ziolek, Thomas Neumuth, and Knut Moeller. 2023. Laparoscopic video analysis using temporal, attention, and multi-feature fusion based approaches. *Sensors*, 23(4):1958.

Yueming Jin, Qi Dou, Hao Chen, Lequan Yu, Jing Qin, Chi Wing Fu, and Pheng Ann Heng. 2018. SV-RCNet: Workflow recognition from surgical videos using recurrent convolutional network. *IEEE Transactions on Medical Imaging*, 37(5):1114–1126.

Yueming Jin, Huaxia Li, Qi Dou, Hao Chen, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. 2019. Multi-task recurrent convolutional network with correlation loss for surgical video analysis. *arXiv preprint*. ArXiv:1907.06099.

Abdolrahim Kadkhodamohammadi, Imanol Luengo, and Danail Stoyanov. 2022. PATG: position-aware temporal graph networks for surgical phase recognition on laparoscopic videos. *International Journal of Computer Assisted Radiology and Surgery*, 17:849–856.

Anni King, George E. Fowler, Rhiannon C. Macefield, Hamish Walker, Charlie Thomas, Sheraz Markar, Ethan Higgins, Jane M. Blazeby, and Natalie S. Blencowe. 2025. Use of artificial intelligence in the analysis of digital videos of invasive surgical procedures: scoping review. *BJS Open*, 9(4):zraf073.

Yunlong Li, Zijian Zhao, Renbo Li, and Feng Li. 2024. Deep learning for surgical workflow analysis: a survey of progresses, limitations, and trends. *Artificial Intelligence Review*, 57:291.

Yang Liu, Maxence Boels, Luis C. Garcia-Peraza-Herrera, Tom Vercauteren, Prokar Dasgupta, Alejandro Granados, and Sebastien Ourselin. 2023a. LoViT: Long video transformer for surgical phase recognition. *arXiv preprint*. ArXiv:2305.08989.

Yang Liu, Jiayu Huo, Jingjing Peng, Rachel Sparks, Prokar Dasgupta, Alejandro Granados, and Sebastien Ourselin. 2023b. SKiT: a fast key information video transformer for online surgical phase recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zhenzhong Liu, Kelong Chen, Shuai Wang, Yijun Xiao, and Guobin Zhang. 2025. Deep learning in surgical process modeling: a systematic review of workflow recognition. *Journal of Biomedical Informatics*, 162:104779.

Shanka Subhra Mondal, Rachana Sathish, and Debdoot Sheet. 2019. Multitask learning of temporal connectionism in convolutional networks using a joint distribution loss function to simultaneously identify tools and phase in surgical videos. *arXiv preprint*. ArXiv:1905.08315.

Shoji Morita, Hitoshi Tabuchi, Hiroki Masumoto, Tomofusa Yamauchi, and Naotake Kamiura. 2019. Real-time extraction of important surgical phases in cataract surgery videos. *Scientific Reports*, 9:16590.

Fnu Neha, Deepshikha Bhati, and Deepak Kumar Shukla. 2025. Retrieval-augmented generation (RAG) in healthcare: A comprehensive review. *AI*, 6(9):226.

OpenAI. 2024. GPT-4o system card. Https://cdn.openai.com/gpt-4o-system-card.pdf.

Xiaoying Pan, Xuanrong Gao, Hongyu Wang, Wuxia Zhang, Yuanzhen Mu, and Xianli He. 2023. Temporal-based Swin Transformer network for workflow recognition of surgical video. *International Journal of Computer Assisted Radiology and Surgery*, 18:139–147.

Sara Paracchini, Cristina Taliento, Giulia Pellecchia, Veronica Tius, Madalena Tavares, Chiara Borghi, Alessandro Antonio Buda, Adrien Bartoli, Nicolas Bourdel, and Giuseppe Vizzielli. 2025. Artificial intelligence in the operating room: A systematic review of AI models for surgical phase, instruments and anatomical structure identification. *Acta Obstetricia et Gynecologica Scandinavica*, 104:2054–2064.

Jay N. Paranjape, Shameema Sikder, Vishal M. Patel, and S. Swaroop Vedula. 2023. Cross-dataset adaptation for instrument classification in cataract surgery videos. *arXiv preprint*. ArXiv:2308.04035.

Manfred Jürgen Primus, Doris Putzgruber-Adamitsch, Mario Taschwer, Bernd Münzer, Yosuf El-Shabrawi, László Böszörmenyi, and Klaus Schoeffmann. 2018. Frame-based classification of operation phases in cataract surgery videos. In *MultiMedia Modeling – 24th International Conference (MMM 2018)*, volume 10704 of *Lecture Notes in Computer Science*, pages 241–253. Springer.

Daiqing Qi, Dongliang Guo, Hanzhang Yuan, Handong Zhao, Mengxuan Hu, Lehan Yang, and Sheng Li. 2025. Improve temporal reasoning in multimodal large language models via video contrastive decoding. In *NeurIPS 2025 Posters*.

Yao Qiu, Xiaolan Chen, Xinyuan Wu, Yunqian Li, Pusheng Xu, Kai Jin, Xianwen Shang, Peranut Chotcomwongse, Mingguang He, and Danli Shi. 2025. Embodied artificial intelligence in ophthalmology. *NPJ Digital Medicine*, 8:351.

Ali Rasekh, Erfan Bagheri Soula, Omid Daliran, Simon Gottschalk, and Mohsen Fayyaz. 2025. Enhancing temporal understanding in Video-LLMs through stacked temporal attention in vision encoders. *arXiv preprint*. ArXiv:2510.26027.

Dominik Rivoir, Sebastian Bodenstedt, Isabel Funke, Felix von Bechtolsheim, Marius Distler, Jürgen Weitz, and Stefanie Speidel. 2020. Rethinking anticipation tasks: Uncertainty-aware anticipation of sparse surgical instrument usage for context-aware assistance. *arXiv preprint*. ArXiv:2007.00548.

Klaus Schoeffmann, Mario Taschwer, Stephanie Sarny, Bernd Münzer, Manfred Jürgen Primus, and Doris Putzgruber. 2018. Cataract-101: Video dataset of 101 cataract surgeries. In *Proceedings of the 9th ACM Multimedia Systems Conference (MMSys '18)*, pages 421–425.

Daniel Schofield. 2025. Defining the autonomous enterprise: Reasoning, memory, and the core capabilities of agentic AI. Unstructured.io blog. Https://unstructured.io/blog/defining-the-autonomous-enterprise-reasoning-memory-and-the-core-capabilities-of-agentic-ai.

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

Xueying Shi, Yueming Jin, Qi Dou, and Pheng-Ann Heng. 2020. LRTD: Long-range temporal dependency based active learning for surgical workflow recognition. *arXiv preprint*. ArXiv:2004.09845.

Andru P. Twinanda, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. 2016a. Single- and multi-task architectures for surgical workflow challenge at M2CAI 2016. *arXiv preprint*. ArXiv:1610.08844.

Andru P. Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. 2016b. EndoNet: A deep architecture for recognition tasks on laparoscopic videos. *arXiv preprint*. ArXiv:1602.03012.

Kun Yuan, Matthew Holden, Shijian Gao, and Wonsook Lee. 2022. Anticipation for surgical workflow through instrument interaction and recognized signals. *Medical Image Analysis*, 82:102611.

Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. 2024. Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. In *Advances in Neural Information Processing Systems*, volume 37.

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

Francis Xiatian Zhang, Jingjing Deng, Robert Lieck, and Hubert P. H. Shum. 2024. Adaptive graph learning from spatial information for surgical workflow anticipation. *arXiv preprint*. ArXiv:2412.06454.

Xiatian Zhang, Noura Al Moubayed, and Hubert P. H. Shum. 2022. Towards graph representation learning based surgical workflow anticipation. *arXiv preprint*. ArXiv:2208.03824.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint*. ArXiv:2504.10479.

Odysseas Zisimopoulos, Evangello Flouty, Imanol Luengo, Petros Giataganas, Jean Nehme, Andre Chow, and Danail Stoyanov. 2018. DeepPhase: Surgical phase recognition in cataracts videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 11073, pages 556–563.

A Appendix

A.1 Datasets

All datasets are split at the case (video) level using an approximate 8:2 train/test ratio, as summarized in Table 4, to avoid any information leakage across videos. For Cataract-1K, the training split is used to train both the dataset-specific single-frame phase classifier and a unified instrument detector shared across all datasets, while the test split is reserved exclusively for evaluating SurgAgents. For Cataract-101 and ITEC Cat21, the training splits are used to train their respective dataset-specific single-frame phase classifiers, and the corresponding test splits are used only for SurgAgents evaluation. Across all datasets, no test cases are used during training of any component involved in evaluation, ensuring a fair and consistent assessment of short-horizon phase prediction performance.

Dataset	Total Cases	Train	Test
Cataract-1K	56	45	11
Cataract-101	101	81	20
ITEC Cat21	21	17	4

Table 4: Case-level train/test splits for all datasets. Splits are performed at the video (case) level using an approximate 8:2 ratio.

A.2 Baseline

For reproducibility and fair comparison, we provide implementation details for the baseline method used in our experiments: LVLM-only. The baseline shares the same evaluation protocol as SurgAgents, including fixed-stride history sampling ($\delta = 3$ s) and short-horizon prediction ($\Delta = 5$ s). The baseline (LVLM-only) bypasses explicit history evidence extraction entirely. The history frames $H_F(t)$ and the current frame $f(t)$ are concatenated into a single horizontal strip image, preserving temporal order, and fed directly to the LVLM for prediction. No textual prompts, descriptors, summarization, or retrieval are applied. The baseline tests the performance of end-to-end multimodal reasoning without modularized evidence extraction or intermediate supervision.

For the baseline, Accuracy/macro-F1 are used as evaluation metrics. Training and fine-tuning use the same optimizer and learning rate as SurgAgents. Inference uses fixed stride sampling to ensure causal consistency and allow direct comparison of latency with SurgAgents. LVLM prompts are constrained to a fixed token budget, and decoding is restricted to canonical phase labels to prevent hallucinations. Historical frame lengths and current frame resolutions match those used in the main pipeline.

The baseline implementation provides controlled comparisons to SurgAgents, highlighting the contributions of modular agents, temporal narrative summarization, and confidence-gated retrieval in surgical intent prediction.

A.3 Prompt Templates

A.3.1 Overview

The figure 5 presents the complete prompt templates used by the phase predictor agent for surgical intent prediction. The template is composed of two parts: a system-level prompt that defines global task rules, reasoning heuristics and output constraints, and a case-specific user prompt that supplies the temporal summary, current observation and optional per-frame details. At runtime these two parts are concatenated and submitted with the current frame image to an LVLM. This design enforces consistent, auditable predictions while allowing per-query injection of summarized surgical history and detailed evidence.

A.3.2 System prompt (core instructions)

The system prompt sets the phase predictor's role and hard constraints. It explicitly instructs the model to forecast the future state at frame $t + 5$ (5 seconds ahead) rather than to describe the current frame. The prompt limits the model's accessible information to the last observed frame, the provided temporal summary and standard workflow knowledge, and it forbids access to the target frame itself. To guide temporal reasoning, the system prompt supplies empirical stage-duration priors and asks the model to consider whether a stage is likely to continue or transition. Critically, the system prompt requires a strict JSON output containing `predicted_stage`, `confidence` and `rationale`, where the rationale must justify the prediction using observable frame evidence

and temporal-progression cues, ensuring results are both machine-parseable and human-auditable.

A.3.3 User prompt (case-specific template)

The user prompt injects instance-specific information for each query. The temporal context field is filled by the history summarizer agent and provides a concise narrative of events up to time t. The current observation identifies the last visible frame and reminds the model that this is the only visual input available. The prediction task reiterates the target frame ($t + 5$) and the prohibition on viewing the future frame. The prompt also specifies the reasoning guidance to use for this instance—e.g., how to weigh stage duration, temporal progression and instrument changes—and it lists the canonical stage labels that `predicted_stage` must exactly match. An optional detailed frame history can provide per-frame descriptions from the history descriptor to support finer-grained reasoning or manual review. Placeholder names used in templates include `temporal_summary`, `observed_frame_num`, `target_frame_num`, and `detailed_frame_analysis`.

A.3.4 Usage and implementation notes

In practice, the system prompt is retained as a session-level instruction while the user prompt is populated for each prediction with the temporal summary, observed frame identifier and any detailed frame analysis; both parts plus the current frame image form the LVLM input. The LVLM must return valid JSON; the system parses `predicted_stage`, `confidence` and `rationale`, and a low confidence triggers the confidence-gated retrieval module. Requiring exact label matching prevents ambiguous outputs and enables automatic scoring. Requiring a structured rationale ensures each prediction is accompanied by explicit temporal and observational evidence, which supports ablation studies, error analysis and system audits. The template therefore provides a reproducible, constrained interface for short-horizon surgical phase forecasting without model fine-tuning.

A.4 History Summarizer Logic

The history summarizer converts a time-ordered sequence of per-frame facts (frame id, single-frame stage prior, detected instruments) into a compact, human-readable temporal narrative that summarizes "what happened up to time t". The summarizer is rule-based and deterministic: it preserves

Algorithm 1 History Summarizer — Rule-based Temporal Narrative Generation

Require: sequence of frame facts $F = [(f_i, s_i, I_i)]_{i=1..m}$
Ensure: temporal summary string `temporal_summary`

Sort F by frame id ascending.
$segments \leftarrow []$
$current_seg \leftarrow (start = f_1,\ end = f_1,\ stage = s_1,\ instruments = I_1)$
for $i \leftarrow 2$ to m **do**
  if $s_i = current_seg.stage$ **then**
    $current_seg.end \leftarrow f_i$
    $current_seg.instruments \leftarrow current_seg.instruments \cup I_i$
  else
    append $current_seg$ to $segments$
    $current_seg \leftarrow (start = f_i,\ end = f_i,\ stage = s_i,\ instruments = I_i)$
  end if
end for
append $current_seg$ to $segments$
if $|segments| = 1$ **then**
  `temporal_summary` ← format("Surgery focused on {stage} (using {instruments}).")
else if $|segments| = 2$ **then**
  `temporal_summary` ← format("Surgery progressed from {stage1} ... to {stage2} ...")
else
  Build parts list and join with ", then "
  `temporal_summary` ← "Surgery progressed through " + joined parts + "."
end if
return `temporal_summary`

temporal ordering, collapses stable runs of the same predicted stage into stage segments, records instrument sets observed during each segment, and emits a short natural-language description that is easy for an LVLM to consume as conditioning context.

Inputs and output are simple and explicit. The input is a sequence of tuples $\{(f_i, s_i, I_i)\}_{i=1}^{m}$ where f_i is frame id, s_i is the canonical stage label predicted for frame f_i, and I_i is the set of instruments detected in f_i. The output is a single textual string ‘temporal_summary‘ that narrates stage segments in chronological order, optionally including frame ranges and prominent instruments per segment. The summary aims to be brief (one or a few sentences) while retaining the key temporal structure and instrument evidence useful for short-horizon forecasting.

The following algorithm implements the summarizer. It first sorts frames by id, detects contiguous stage runs (stage segments), aggregates instrument sets per segment, then formats one of several template patterns depending on the number of segments found. The algorithm is intentionally simple to guarantee reproducibility and to make the resulting narrative easy to parse by both humans and LVLMs.

Example output (one-line summary):

> “Surgery progressed through idle (frames 0–15), then Incision (frames 16–28 using scalpel, forceps), then Viscoelastic injection (frames 29–45 using cannula, BSS syringe).”

Key implementation notes and design choices: the summarizer preserves temporal ordering and includes frame ranges to provide stage-duration awareness; instrument sets are unioned per segment and reported to highlight salient tools that often signal upcoming transitions; transitions are explicitly marked so the downstream LVLM can attend to recent boundary activity; template selection (single / two / multiple stages) keeps summaries concise while adapting to the amount of observed variation. The rule-based design favors reproducibility and interpretability; it also makes the narrative amenable to automatic inspection and ablation studies.

A.5 LVLMs

In our experiments on surgical intent prediction across Cataract-1K, Cataract-101, and ITEC Cat21, we employ three representative large vision–language models (LVLMs) for multimodal reasoning. Specifically, we use GPT-4o (OpenAI, 2024), a proprietary multimodal model with strong zero-shot visual–language reasoning capabilities; InternVL3-8B-Instruct (Zhu et al., 2025), an open-source instruction-tuned LVLM optimized for multimodal understanding; and medgemma-4b-it (Sellergren et al., 2025), a lightweight medical-domain LVLM designed for efficient visual–text inference.

The phase predictor in SurgAgents is implemented as a modular agent that supports multiple LVLMs. After extension, it flexibly routes the temporal summary and the current frame to a user-specified LVLM, enabling consistent evaluation and fair comparison across different models.

A.6 GPU Configuration

All the experiments in this paper are conducted on NVIDIA GeForce RTX 3090 GPUs with 24 GB VRAM. CUDA 12.4 and NVIDIA driver version 550.144.03 are used.

Complete Prompts for Surgical Phase Prediction

- **Phase Predictor System Prompt Template** (Core system instructions for t+5 future frame prediction)

You are a surgical phase prediction expert for cataract surgery. Your task is to PREDICT THE FUTURE surgical phase that will occur 5 frames (5 seconds) ahead, based on: 1. The temporal context from previous frames (what has happened so far). 2. The last observed frame image (current state at frame t). 3. Typical surgical workflow patterns.

Critical Understanding:

You are predicting the FUTURE (frame t+5), NOT analyzing the current frame (frame t)!

Key principles:

1. You can ONLY see frame t (the last observed frame). 2. You need to predict what stage will be happening at frame t+5. 3. Surgical phases have typical durations (20-50 frames). 4. Consider stage transition probability based on temporal progression

Prediction Strategy:

1. Analyze Current State (frame t): What stage and instruments do you see? 2. Review Temporal History: What stages have occurred? How long has current stage lasted? 3. Estimate Stage Duration: Is current stage likely to continue for 5+ more frames? 4. Consider Transitions: If near stage boundary, what stage typically comes next?

Stage Duration Guidelines:

- Short stages (5-15 frames): Incision, Lens positioning
- Medium stages (20-40 frames): Capsulorhexis, Hydrodissection, Lens Implantation
- Long stages (40-80 frames): Phacoemulsification, Irrigation/Aspiration
- Very long (80+ frames): idle (before/after main procedure)

Output Requirements:

Your response must be valid JSON with:

- predicted_stage: The stage you predict will occur at frame t+5
- confidence: 0.0-1.0 (certainty of your prediction)
- rationale: Detailed explanation including: 1. What you observe in current frame t. 2. Temporal progression analysis. 3. Why you predict this stage for frame t+5. 4. Whether you expect stage continuation or transition.

- **Phase Predictor User Prompt Template** (Structured prediction task with temporal summary)

Temporal Context from Surgery History: {temporal_summary}

Current Observation (Frame t={observed_frame_num}): You are currently viewing frame {observed_frame_num}, which is the LAST OBSERVED frame.

Prediction Task:

Predict the surgical stage at frame {target_frame_num} (which is {prediction_horizon} frames in the FUTURE).

Important Instructions: 1. You can ONLY see frame {observed_frame_num} (current state). 2. You CANNOT see frame {target_frame_num} (that's what you're predicting). 3. Use temporal progression and workflow knowledge to forecast. 4. Consider stage duration: typical stages last 20-50 frames.

Reasoning Framework:

- If current stage started recently (< 20 frames ago) → Likely to CONTINUE
- If current stage has lasted long (> 30 frames) → May TRANSITION soon
- Use instrument changes as transition indicators
- Consider typical surgical sequence (but don't blindly follow it)

Available Stages: idle, Incision, Viscoelastic, Capsulorhexis, Hydrodissection, Phacoemulsification, Irrigation/Aspiration, Capsule Polishing, Lens Implantation, Lens positioning, Viscoelastic_Suction, Anterior_Chamber Flushing, Tonifying/Antibiotics

Detailed Frame History: {detailed_frame_analysis}

Output Format (strict JSON):

```
{
 "predicted_stage": "<exact_stage_name_from_list>",
 "confidence": <0.0-1.0>,
 "rationale": "Explain: (1) Current state at frame t={observed_frame_num}, (2) Temporal progression pattern, (3) Why you predict this specific stage for frame t+5={target_frame_num}, (4) Stage continuation vs transition reasoning"
}
```

Remember: You are forecasting the FUTURE state at frame t+5, not describing the current frame t!

Figure 5: Complete system and user prompt templates for the phase predictor agent. The system prompt defines the prediction task and output requirements, while the user prompt supplies the temporal history, current observation, and available phase labels. The agent returns a predicted phase, confidence score, and rationale in JSON format.