# OPHMEM: LEVERAGING CROSS-CASE EXPERIENCE TO MITIGATE VISUAL AMBIGUITY IN OPHTHALMIC SURGICAL VIDEO UNDERSTANDING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Ophthalmic surgical video understanding can assist clinicians in analyzing surgical workflows, identifying abnormalities, and supporting postoperative assessment. However, a key challenge lies in visual ambiguity, where visually similar observations may correspond to different surgical states. Existing methods mainly rely on intra-case contextual information, which does not always provide reliable discriminative cues. To address this limitation, we propose `OphMem`, a framework that continually accumulates and reuses cross-case procedural experience. `OphMem` consists of two core modules: Experience-Conditioned Agentic Reasoning, which retrieves applicable experience to guide video inspection and tool use, and Persistent Experience Memory, which transforms completed reasoning trajectories into medically grounded reusable experience and consolidates them into an evolving experience bank. To further assess visual ambiguity, we construct a dedicated benchmark with three tasks: History-Conditioned Step Disambiguation, Ambiguous Complication Verification, and Ophthalmic Temporal VQA, together with a sequential experience protocol. Across the three tasks, `OphMem` achieves a mean relative gain of $\sim 8\%$ in primary accuracy over various benchmarks. *Code is available at OphMem.*

## 1 INTRODUCTION

Ophthalmic surgery involves precise manipulation of the eye and its surrounding tissues to treat ocular diseases and restore or improve visual function. During surgery, imaging systems connected to surgical microscopes typically record the entire procedure as video, capturing rich visual information about surgical operations. In recent years, deep learning techniques have been increasingly applied to ophthalmic surgical video understanding to automatically analyze and identify key surgical content (Hu et al., 2025; Paranjape et al., 2026; Shah et al., 2025), thereby assisting clinicians in analyzing surgical workflows (Mueller et al., 2025), detecting abnormalities (Sachdeva et al., 2025), and supporting postoperative review and quality assessment (Ahmadi et al., 2026).

However, accurately identifying and analyzing key content in ophthalmic surgical videos remains challenging. Since ophthalmic procedures are typically performed within the same small anatomical region, similar tissue appearances, instruments, and surgical actions may recur across different stages, resulting in high visual similarity between surgical phases (Fang et al., 2022; Ghamsarian et al., 2024). We refer to this phenomenon as ***visual ambiguity in ophthalmic surgical videos***. Consequently, relying solely on a single frame or short video segment is often insufficient to effectively handle such visual ambiguity, thereby hindering the accurate identification and analysis of key surgical content (Shah et al., 2023). To alleviate this issue, existing methods typically retain long-range historical information or retrieve video segments relevant to a specific query, thereby incorporating broader contextual information to facilitate more accurate interpretation of the local visual content. For example, MS-TCN++ combines information from nearby and more distant time points to recognize phases in cataract surgery videos (Mueller et al., 2025). CatVLM associates video features with timestamps to locate the intervals in which a queried surgical step occurs (Paranjape et al., 2026).

Although the above methods have made progress in alleviating the visual ambiguity in ophthalmic surgical videos, they mainly rely on contextual information within the current case, which does not



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.

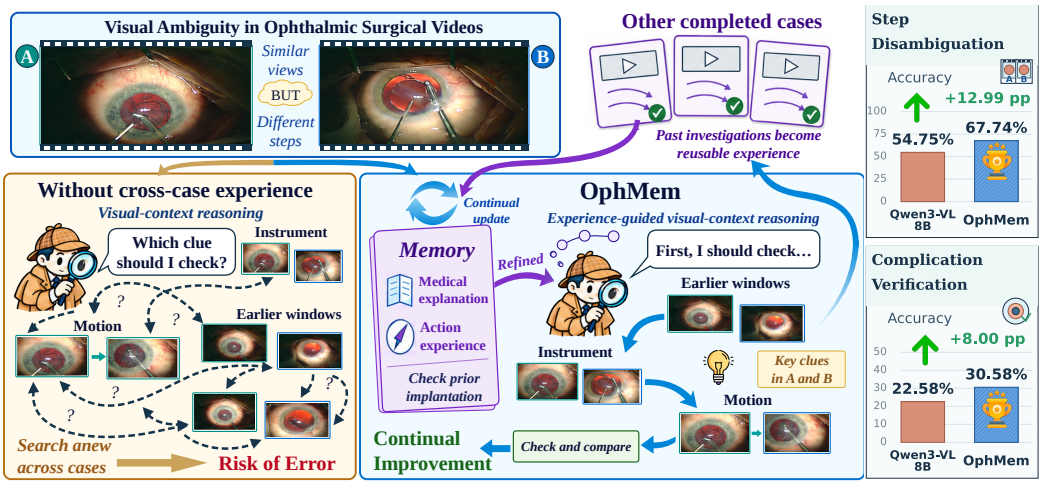



Figure 1: Visually similar observations may correspond to different surgical steps, resulting in visual ambiguity in ophthalmic surgical videos. Without cross-case experience, reasoning relies primarily on contextual cues from the current case, making it difficult to determine which cues are consistently discriminative and which are case-specific. OphMem continually accumulates reusable experience from completed cases and uses it to guide evidence inspection and comparison in subsequent cases. This experience is further updated after each interaction, enabling OphMem to progressively identify more stable and discriminative contextual cues for resolving visual ambiguity.

always provide reliable discriminative cues. This is because the instruments and actions surrounding a particular surgical state may either reflect stable characteristics of that state or simply correspond to case-specific procedural details (Mueller et al., 2025; Holm et al., 2025). As a result, relying on a single case makes it difficult to determine which contextual cues are consistently discriminative and which merely reflect case-specific variations. In contrast, comparing cases can help identify more stable and discriminative contextual cues by examining the commonalities and differences of similar surgical states across cases (Kumar et al., 2022; Dwibedi et al., 2019; Khosla et al., 2020).

In this paper, we propose OphMem, an ophthalmic surgical video understanding framework that continually accumulates and reuses procedural experience from prior analyses (Figure 1). OphMem consists of two core modules: **Experience-Conditioned Agentic Reasoning** and **Persistent Experience Memory**. For the current question, the reasoning module retrieves applicable experience from a persistent bank to guide what video evidence should be inspected and which tools should be used. Importantly, the retrieved experience provides an inspection strategy rather than a stored answer; the agent iteratively updates its judgment from the actual observations returned by the selected tools. After the question is completed, the memory module abstracts successful tool-action steps from the reasoning trajectory into reusable experience, grounds medically relevant claims when necessary, and consolidates the resulting candidates into the experience bank. The bank therefore remains fixed while solving the current question and evolves only between questions, allowing experience accumulated from completed interactions to guide subsequent analyses. By carrying forward reusable inspection strategies across cases, OphMem helps the agent identify more reliable contextual evidence for resolving visual ambiguity in ophthalmic surgical videos.

To further assess the ability of OphMem to handle visual ambiguity in ophthalmic surgical videos, a dedicated benchmark is needed. However, existing benchmarks for ophthalmic surgical video understanding primarily evaluate overall task performance and are not specifically designed to assess visual ambiguity. To fill this gap, we construct a benchmark centered on visual ambiguity and introduce a sequential experience protocol. Based on existing ophthalmic surgical videos and expert annotations, we formulate three tasks: *History-Conditioned Step Disambiguation*, *Ambiguous Complication Verification*, and *Ophthalmic Temporal VQA*, covering context-aware step identification, complication verification, and temporal reasoning, respectively. The protocol specifies the question order, available information, and access to a persistent experience state, allowing experience accumulated from previous interactions to inform subsequent decisions. Compared with Reason-

ingBank on Step Pair Acc, the tied LLaVA-Video-7B and MedGemma-1.5-4B-IT on Complication Pair Acc, and Qwen3-VL-8B-Thinking on Temporal VQA accuracy, `OphMem` improves *History-Conditioned Step Disambiguation* by $+10.07\%$, *Ambiguous Complication Verification* by $+5.34\%$, and *Ophthalmic Temporal VQA* by $+8.33\%$, for a mean relative gain of $+7.92\%$.

The contributions of this paper are summarized as follows:

① We introduce an agentic reasoning loop that retrieves applicable cross-case strategies at each step to guide video inspection and tool use, while deriving answers from the resulting observations.

② We turn successful tool-action steps from completed trajectories into reusable strategies, ground their medical claims, and consolidate candidates into a bank that evolves between questions.

③ We construct three complementary tasks from existing ophthalmic surgical videos and expert annotations, together with a sequential experience protocol that evaluates visual ambiguity reasoning and the benefit of experience accumulated across interactions.

## 2 Related Work

### 2.1 Ophthalmic Surgical Video Understanding

Research on ophthalmic surgical videos spans workflow recognition, complication detection, and question answering. Early resources provide complementary annotations: Cataract-101 (Schoeffmann et al., 2018) labels surgical phases, CATARACTS (Al Hajj et al., 2019) labels instrument presence, and CaDIS (Grammatikopoulou et al., 2021) adds pixel-level instrument and anatomy labels. Cataract-1K (Ghamsarian et al., 2024) provides annotations for phase recognition, scene segmentation, and irregularity detection, while OphNet (Hu et al., 2024) extends workflow analysis to diverse ophthalmic procedures with hierarchical, time-localized annotations. To distinguish visually similar phases, GL-MSTCN (Fang et al., 2022) combines local anatomical and instrument features with temporal convolutions over longer sequences. More recent approaches learn broader surgical representations through pretraining. OphCLIP (Hu et al., 2025) aligns clips and full videos with hierarchical language descriptions and retrieves related silent videos during pretraining. OVFM (Tu et al., 2026) learns spatiotemporal representations from large-scale ophthalmic videos for multiple downstream tasks. Beyond workflow recognition, CataractCompDetect (Sachdeva et al., 2025) combines phase-aware localization, instrument tracking, risk scoring, and vision-language reasoning to detect intraoperative complications. EyePCR (Wang et al., 2025a) evaluates ophthalmic perception, knowledge comprehension, and clinical reasoning through structured visual questions.

### 2.2 Self-Evolving Agent Memory

Agent memory enables past interactions to influence future decisions without requiring updates to model parameters. MemGPT (Packer et al., 2023) manages information across context and external memory, while the Generative Agents architecture (Park et al., 2023) condenses observations into reflections for planning. Early experience-based approaches retain feedback or extract reusable lessons: Reflexion (Shinn et al., 2023) stores verbal reflections for subsequent trials, while ExpeL (Zhao et al., 2024) derives insights from training-task experiences and retrieves them for new tasks. Agent Workflow Memory (Wang et al., 2024) abstracts recurring action sequences into reusable workflows, and ReasoningBank (Ouyang et al., 2026) distills reasoning strategies from self-judged successful and failed trajectories. Voyager (Wang et al., 2023) stores executable skills refined through environment feedback. Learn-by-interact (Su et al., 2025) synthesizes interaction trajectories and retrieves them as demonstrations in its in-context learning setting. MemoryBank (Zhong et al., 2024) updates stored interactions through time- and importance-dependent forgetting and reinforcement. Another line of work studies how memories evolve: A-Mem (Xu et al., 2025b) links new entries to existing notes and updates their contextual representations, whereas MemRL (Zhang et al., 2026) learns retrieval utilities from feedback. Evo-Memory (Wei et al., 2025) evaluates memory adaptation and experience reuse over sequential task streams. In medical imaging, Evo-MedAgent (Shen et al., 2026) applies cross-case learning to chest X-ray interpretation through episodic, procedural, and tool-governance memories.

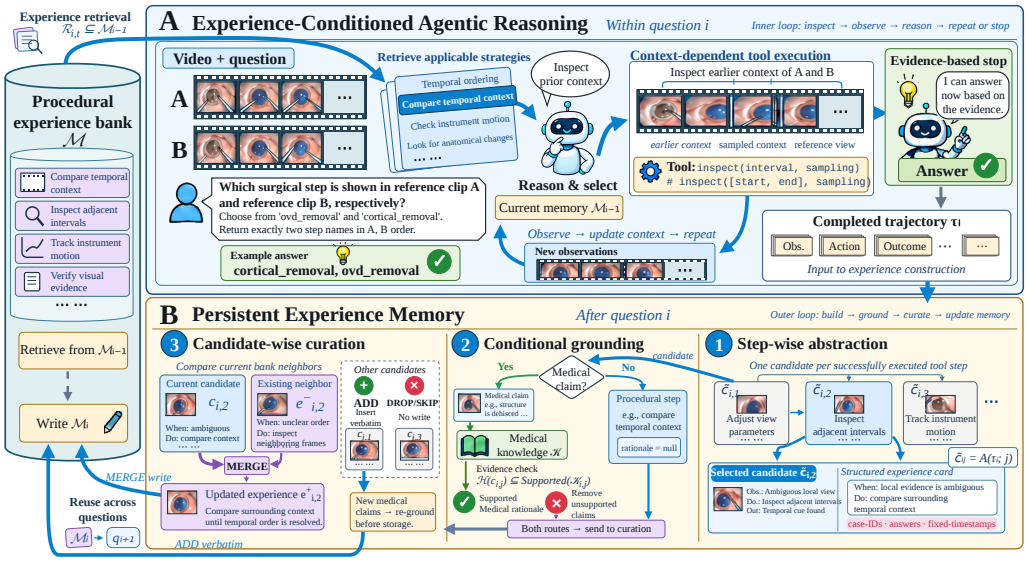



Figure 2: Overview of the proposed OphMem framework. (A) **Experience-Conditioned Agentic Reasoning** retrieves procedural strategies to guide video inspection and answer from current evidence. (B) **Persistent Experience Memory** abstracts completed trajectories, grounds medical claims when needed, and consolidates strategies for later questions.

# 3 METHOD

## 3.1 OVERVIEW

OphMem uses cross-case procedural experience to guide evidence acquisition under visual ambiguity (Figure 2). Its two coupled modules are **Experience-Conditioned Agentic Reasoning** (Section 3.2) and **Persistent Experience Memory** (Section 3.3). The bank guides inspection within a question and is updated from the completed trajectory afterward.

For question $x_i = (V_i, q_i)$, $V_i$ is permitted video evidence and $q_i$ the query. Using bank $\mathcal{M}_{i-1}$, the agent produces trajectory $\tau_i$ and answer $\hat{y}_i$; the post-question update uses $\tau_i$ and medical knowledge $\mathcal{K}$ to produce $\mathcal{M}_i$:

$$p(\tau_{1:N}, \mathcal{M}_{1:N} \mid x_{1:N}, \mathcal{M}_0, \mathcal{K}) = \prod_{i=1}^{N} p_{\mathrm{R}}(\tau_i \mid x_i, \mathcal{M}_{i-1})\, p_{\mathrm{U}}(\mathcal{M}_i \mid \tau_i, \mathcal{M}_{i-1}, \mathcal{K}), \tag{1}$$

where $p_{\mathrm{R}}$ denotes within-question reasoning and $p_{\mathrm{U}}$ the post-question memory update. In the evaluation stream, $\mathcal{M}_0 = \emptyset$; the updated bank $\mathcal{M}_i$ is available from question $i+1$ onward.

## 3.2 EXPERIENCE-CONDITIONED AGENTIC REASONING

### 3.2.1 STANDARD AGENTIC REASONING

At step $t$ of question $i$, the solver uses history $h_{i,t}$ to form inspection intent $z_{i,t}$, select tool $u_{i,t}$ and arguments $\eta_{i,t}$, and append the resulting observation $o_{i,t}$:

$$\begin{aligned} \pi^{(0)}_{\theta,i,t} &:= p_\theta(z_{i,t}, \text{continue} \mid h_{i,t})\, p_{\text{run}}(u_{i,t}, \eta_{i,t} \mid h_{i,t}, z_{i,t}), \\ o_{i,t} &= \mathcal{T}_{u_{i,t}}(V_i; \eta_{i,t}), \qquad h_{i,t+1} = h_{i,t} \oplus (z_{i,t}, u_{i,t}, \eta_{i,t}, o_{i,t}), \end{aligned} \tag{2}$$

where $\mathcal{T}_{u_{i,t}}$ is the video tool, $\eta_{i,t}$ sets its temporal range and sampling, and $\oplus$ extends the history. The cycle continues until sufficient evidence is available. For example, similar irrigation-and-aspiration (I/A) views can indicate cortical or ophthalmic viscosurgical device (OVD) removal; the distinguishing intraocular lens (IOL) implantation may occur much earlier.

#### 3.2.2 EXPERIENCE-CONDITIONED ACTION SELECTION

OphMem retrieves procedural lessons whose applicability conditions match the current history, such as an ambiguous I/A view that requires checking prior IOL implantation. At step $t$, it selects from the fixed bank $\mathcal{M}_{i-1}$:

$$\mathcal{R}_{i,t} \in \underset{\mathcal{R}\subseteq\mathcal{M}_{i-1},\,|\mathcal{R}|=k_{i,t}}{\arg\max} \sum_{e\in\mathcal{R}} s(e, h_{i,t}), \qquad k_{i,t} = \min(k, |\mathcal{M}_{i-1}|), \tag{3}$$

where $s(e, h_{i,t})$ scores relevance to the question and observed history; $k_{i,t}$ limits retrieval to available entries. Retrieval repeats as new observations arrive.

The selected experiences condition the next inspection intent rather than supplying a stored answer. In particular, the standard agentic factor in Eq. (2) becomes:

$$\begin{aligned} \pi_{\theta,i,t} &:= p_\theta(z_{i,t}, \text{continue} \mid h_{i,t}, \mathcal{R}_{i,t})\, p_{\text{run}}(u_{i,t}, \eta_{i,t} \mid h_{i,t}, z_{i,t}), \\ h_{i,t+1} &= h_{i,t} \oplus (\mathcal{R}_{i,t}, z_{i,t}, u_{i,t}, \eta_{i,t}, o_{i,t}), \quad o_{i,t} = \mathcal{T}_{u_{i,t}}(V_i; \eta_{i,t}), \end{aligned} \tag{4}$$

where $\mathcal{R}_{i,t}$ conditions the inspection intent, while $p_{\text{run}}$ selects the tool and arguments. To check prior IOL placement, the agent inspects an earlier interval, then verifies tissue and tool–IOL context in the current clip. Video observations determine the answer.

With $T_i$ executed interactions, the completed trajectory $\tau_i$ includes the retrieved experiences, chosen actions, actual observations, intermediate reasoning, and final answer. Its reasoning factor is

$$p_{\text{R}}(\tau_i \mid x_i, \mathcal{M}_{i-1}) = p_\theta(\hat{y}_i, \text{stop} \mid h_{i,T_i+1}) \prod_{t=1}^{T_i} \pi_{\theta,i,t}\, P_{V_i}(o_{i,t} \mid u_{i,t}, \eta_{i,t}), \tag{5}$$

where $p_\theta$ gives the answer and stopping decision, and $P_{V_i}$ describes tool observations. The completed trajectory enters persistent memory.

### 3.3 PERSISTENT EXPERIENCE MEMORY

After question $i$, OphMem updates the bank through step-level abstraction, medical grounding, and candidate-wise consolidation.

#### 3.3.1 EXPERIENCE ABSTRACTION

The completed trajectory $\tau_i$ records the observations, tool actions, and outcomes of the question. OphMem abstracts each of its $n_i$ successfully executed tool-action steps into a separate candidate:

$$\widetilde{\mathcal{C}}_i = \big[\widetilde{c}_{i,j}\big]_{j=1}^{n_i} = \big[\mathcal{A}(\tau_i; j)\big]_{j=1}^{n_i}, \tag{6}$$

where $\mathcal{A}(\tau_i; j)$ links the $j$-th step's applicability condition, action, and evidence. Abstraction removes case identifiers, answers, timestamps, and fixed tool parameters while preserving reusable inspection guidance. Separate candidates remain independently retrievable.

#### 3.3.2 MEDICAL GROUNDING

A candidate may invoke a medical relation to explain why its inspection is informative. For example, the placement of an IOL between cortical and OVD removal gives a prior-implantation check diagnostic value. OphMem grounds such claims before the candidate enters the bank, composing grounding with step abstraction:

$$\begin{aligned} \mathcal{C}_i = \big[c_{i,j}\big]_{j=1}^{n_i} &= \Big[\big(\mathcal{G}_{\mathcal{K}} \circ \mathcal{A}(\cdot\,; j)\big)(\tau_i)\Big]_{j=1}^{n_i}, \\ \mathcal{H}(c_{i,j}) \subseteq \text{Supported}(\mathcal{K}_{i,j}), &\qquad j = 1, \ldots, n_i, \end{aligned} \tag{7}$$

where $\mathcal{G}_{\mathcal{K}}$ checks the abstracted candidate against medical knowledge, $\mathcal{K}_{i,j}$ is the evidence retrieved for candidate $j$, and $\mathcal{H}(c_{i,j})$ is its set of medical claims. Only supported medical claims are retained; the candidate's inspection strategy remains available even when it has no medical claim. Grounding also applies to claims newly introduced when experiences are merged.

#### 3.3.3 EXPERIENCE CONSOLIDATION

The curator processes the grounded candidates in order, starting from $\mathcal{B}_{i,0} = \mathcal{M}_{i-1}$. For candidate $c_{i,j}$, it consults nearby entries $\mathcal{N}_{i,j} \subseteq \mathcal{B}_{i,j-1}$ and selects $d_{i,j} \in \{\text{ADD}, \text{MERGE}, \text{SKIP}, \text{DROP}\}$. A merge selects one existing entry $e^-_{i,j} \in \mathcal{N}_{i,j}$ and produces a grounded successor $e^+_{i,j}$. All four decisions are expressed by a single bank-update operator:

$$\mathcal{U}_{i,j}(\mathcal{B}) := \big(\mathcal{B} \setminus \{e^-_{i,j} \mid d_{i,j} = \text{MERGE}\}\big) \cup \{c_{i,j} \mid d_{i,j} = \text{ADD}\} \cup \{e^+_{i,j} \mid d_{i,j} = \text{MERGE}\}, \quad (8)$$

where each set $\{a \mid P\}$ contains $a$ when $P$ holds and is empty otherwise. ADD inserts the candidate; MERGE replaces exactly one related entry; SKIP and DROP leave the bank unchanged. Applying $\mathcal{B}_{i,j} = \mathcal{U}_{i,j}(\mathcal{B}_{i,j-1})$ after each decision yields the bank for the next question and the number of accepted writes:

$$\begin{aligned} \mathcal{M}_i &= \big(\mathcal{U}_{i,n_i} \circ \cdots \circ \mathcal{U}_{i,1}\big)(\mathcal{M}_{i-1}), \\ W_i &= \sum_{j=1}^{n_i} \mathbf{1}_{\{\text{ADD},\text{MERGE}\}}(d_{i,j}) \in \{0, \ldots, n_i\}, \end{aligned} \quad (9)$$

where $W_i$ counts additions and merges across the $n_i$ candidates. A merge refines an existing entry without increasing bank size. Subsequent questions retrieve from $\mathcal{M}_i$, completing the experience loop.

## 4 EXPERIMENTS

### 4.1 BENCHMARK & PROTOCOL

**Benchmark design.** We design an ophthalmic video benchmark to test whether agents resolve visual ambiguity through contextual reasoning and experience reuse. Questions pair similar local appearances with different surgical meanings and ask about events that lie earlier in the procedure. Answers come from existing expert annotations, while annotation metadata and fixed visual descriptors guide question selection independently of the evaluated models. Together, these choices focus evaluation on evidence that local frames alone cannot resolve.

**Task construction.** The benchmark probes three complementary forms of contextual evidence. ❶ **Step Disambiguation** pairs visually similar clips from different Cataract-1K (Ghamsarian et al., 2024) cases with different JHU step labels (Ali et al., 2026), asking for both labels from the clips and their respective causal histories. ❷ **Complication Verification** pairs similar observations from the same procedural band in CataComp-104 (Sachdeva et al., 2025) cases with opposite complication labels, asking the agent to judge each case using full-case evidence for posterior capsule rupture, vitreous loss, or iris prolapse. ❸ **Temporal VQA** uses OphNet2024 (OphNet Challenge Organizers, 2025) operation sequences to query event relations, temporal updates, and conditional recall from evidence before each cutoff; consecutive segments of one operation count as a single occurrence, and paired cutoffs include changed-answer questions and unchanged-answer controls. The tasks thus span local step identity, complication presence, and surgical event history.

**Data & evaluation.** Table 1 summarizes the case-separated pools, including 1,179 evaluation questions from 238 cases. Pair Accuracy requires both ordered answers to be correct; macro F1 and Jaccard score the component decisions. Temporal VQA uses normalized answer accuracy, ROUGE-L (Lin, 2004), and METEOR (Banerjee & Lavie, 2005). Each task follows a fixed question stream with an initially empty experience bank that updates after every completed question; memory baselines share the backbone, tools, question order, and inference budget. We report full-stream scores and temporal windows, with construction, input-access, and scoring details in Appendix A.

### 4.2 MAIN RESULTS

We first compare overall performance across the three ambiguity-focused tasks. Table 2 places `OphMem` alongside general vision-language models, medical models, and memory-augmented agents, allowing us to examine both paired visual decisions and temporal question answering.

Table 1: Benchmark data sources and existing pools (cases / QAs). Pool names denote historical splits, not online learning stages.

| Task | Video / annotation source | Acquisition | Development | Evaluation |
|---|---|---|---|---|
| ❶ Step Disambiguation | Cataract-1K / JHU step labels | 17/85 | 19/92 | 89/442 |
| ❷ Complication Verification | CataComp-104 | 11/38 | 18/61 | 49/237 |
| ❸ Temporal VQA | OphNet2024 | 20/100 | 20/100 | 100/500 |
| **Total** | – | 48/223 | 57/253 | 238/1,179 |

Table 2: Main results (%). Baseline scores retain their original benchmark coverage. Pair Acc requires both ordered answers to be correct.

| Method | Step Disambiguation | | | Complication Verification | | | Temporal VQA | | |
|---|---|---|---|---|---|---|---|---|---|
| | Pair Acc ↑ | F1 ↑ | Jaccard ↑ | Pair Acc ↑ | F1 ↑ | Jaccard ↑ | Acc ↑ | ROUGE-L ↑ | METEOR ↑ |
| *General Vision-Language Models* | | | | | | | | | |
| `Qwen3-VL-8B-Thinking` | 54.75 | 55.72 | 38.77 | 22.58 | 45.30 | 29.31 | 50.40 | 55.40 | 40.65 |
| `InternVL3.5-8B` | 50.23 | 51.73 | 35.14 | 19.35 | 37.07 | 23.20 | 44.60 | 51.10 | 35.52 |
| `LLaVA-Video-7B` | 51.13 | 51.87 | 35.21 | 29.03 | 49.25 | 33.27 | 49.00 | 56.60 | 38.98 |
| *Medical Models* | | | | | | | | | |
| `Lingshu-7B` | 45.29 | 50.73 | 33.24 | 9.68 | 27.27 | 18.75 | 42.40 | 51.90 | 35.49 |
| `HuatuoGPT-Vision-7B` | 50.00 | 51.21 | 34.52 | 25.81 | 43.58 | 27.91 | 41.00 | 51.27 | 33.55 |
| `MedGemma-1.5-4B-IT` | 50.23 | 52.14 | 35.43 | 29.03 | 55.25 | 39.19 | 43.40 | 52.10 | 35.76 |
| *Memory-Augmented Agents* | | | | | | | | | |
| `A-Mem` | 53.17 | 53.53 | 36.69 | 27.42 | 47.22 | 30.95 | 42.21 | 45.90 | 26.19 |
| `ReasoningBank` | 61.54 | 61.39 | 44.54 | 24.20 | 43.57 | 27.86 | 40.80 | 39.86 | 24.19 |
| **`OphMem`** | 67.74 | 71.00 | 61.62 | 30.58 | 53.77 | 39.66 | 54.60 | 57.80 | 42.62 |

**Primary accuracy across tasks.** Table 2 shows that `OphMem` attains the highest reported primary accuracy on all three ambiguity-focused tasks. On Step Disambiguation, Pair Accuracy rises from 61.54% (`ReasoningBank`) to 67.74%, a relative 10.07% ↑ gain; on Complication Verification, it rises from the strongest baseline score of 29.03% to 30.58% (5.34% ↑). For Temporal VQA, Accuracy increases 50.40% → 54.60% (8.33% ↑), with `OphMem` also reaching 57.80% ROUGE-L and 42.62% METEOR. The mean relative improvement across the three primary accuracy metrics is +7.92%, spanning paired visual decisions and temporal recall.

**Transferable procedural experience.** `OphMem` also improves over the memory-augmented agents on the primary score of each task (Table 2). Against `ReasoningBank`, Step Jaccard climbs 44.54% → 61.62% (+38.35%), while Pair Accuracy reaches 67.74% from 61.54%. Against `A-Mem`, Complication F1 improves 47.22% → 53.77% (+13.87%) and Temporal-VQA Accuracy improves 42.21% → 54.60% (+29.35%). The gains across distinct metrics show that reusable inspection strategies transfer beyond the cases from which they were learned.

### 4.3 Online Learning and Experience Evolution

The experience bank changes after each completed question, so its trajectory is as informative as its final score. We follow accuracy along the fixed streams and then examine how new experiences are added, merged, and retrieved as later questions arrive.

**Accuracy along the stream.** Figure 3 tracks the fixed question streams as `OphMem` accumulates experience from completed questions. Step Disambiguation rises early and again near the endpoint, reaching 67.74%; Complication Verification records its largest increase later and ends at 30.58%. Temporal VQA gains earlier, fluctuates around ∼55%, and finishes at 54.60%. These task-specific trajectories show that accumulated experience becomes useful at different stages as new visual and temporal evidence appears.

**From construction to reuse.** Figure 4 connects these accuracy trajectories to changes in the experience bank. At the start of the stream, `ADD` accounts for the largest share of post-question curation decisions as the initially empty bank acquires procedures. Later, `MERGE` becomes more prominent,

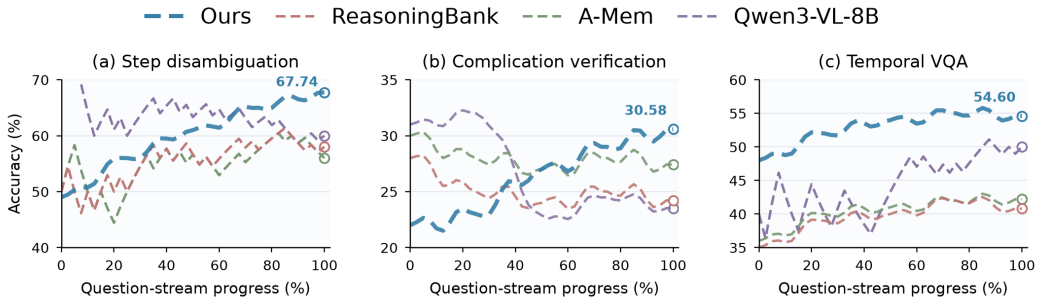



Figure 3: Accuracy trajectories across the question streams for (a) step disambiguation, (b) complication verification, and (c) temporal VQA. Each panel has its own vertical scale, and annotated values mark the displayed `OphMem` endpoints.

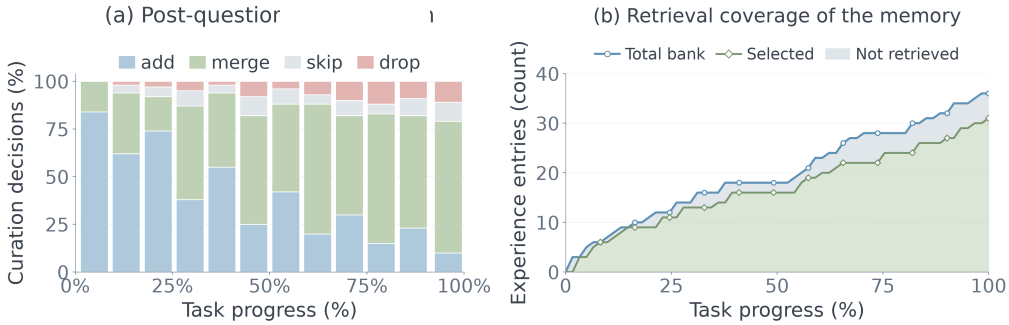



Figure 4: Experience-bank evolution across the question stream. (a) Proportions of post-question candidate decisions assigned to `ADD`, `MERGE`, `SKIP`, and `DROP`. (b) Growth of the current bank, distinguishing selected entries from entries not retrieved.

Table 3: Component ablation (%). Metrics match Table 2.

| Method | Step Disambiguation | | | Complication Verification | | | Temporal VQA | | |
|---|---|---|---|---|---|---|---|---|---|
| | Pair Acc ↑ | F1 ↑ | Jaccard ↑ | Pair Acc ↑ | F1 ↑ | Jaccard ↑ | Acc ↑ | ROUGE-L ↑ | METEOR ↑ |
| **OphMem** | 67.74 | 71.00 | 61.62 | 30.58 | 53.77 | 39.66 | 54.60 | 57.80 | 42.62 |
| `No Memory` | 54.75 | 55.72 | 38.77 | 22.58 | 45.30 | 29.31 | 50.40 | 55.40 | 40.65 |
| `Raw Case Memory` | 52.10 | 58.82 | 45.93 | 15.79 | 35.71 | 24.35 | 42.60 | 46.47 | 32.65 |
| w/o `Medical Grounding` | 52.63 | 60.80 | 46.39 | 25.79 | 51.97 | 38.36 | 39.20 | 43.31 | 30.14 |
| w/o `Experience Evolution` | 42.84 | 58.48 | 45.28 | 26.32 | 44.28 | 29.19 | 48.80 | 51.36 | 37.48 |

integrating new observations with existing experiences; `SKIP` and `DROP` occupy smaller shares throughout. The retrieval panel shows that the bank grows alongside the number of selected experiences, while a portion of its entries remains unselected at each stage. Thus, bank expansion is accompanied by use of stored procedures in subsequent questions, and later questions increasingly refine existing entries rather than only adding new ones.

### 4.4 Ablation Studies

We isolate the contributions of structured experience, medical grounding, and between-question bank evolution. Table 3 compares the full system with variants that remove or replace each component, tracing which parts of the memory loop drive the observed gains.

**Distinct roles of experience components.** Table 3 separates the effects of structured storage, medical grounding, and bank evolution, with relative changes measured against the full `OphMem` model. Replacing structured entries with `Raw Case Memory` reduces Step Pair Accuracy from 67.74% to 52.10% (23.09% ↓) and Temporal-VQA Accuracy from 54.60% to 42.60% (21.98% ↓). Re-

moving `Experience Evolution` gives the largest Step Pair Accuracy decrease, $67.74\% \rightarrow 42.84\%$ ($36.76\%\downarrow$); removing `Medical Grounding` lowers Step Pair Accuracy by $22.31\%$ and Temporal-VQA Accuracy by $28.21\%$. These results identify structured, evolving experience as the main source of gains, while medical grounding has task-dependent effects.

### 4.5 Qualitative Case Study

A paired surgical-step question illustrates how stored experience guides evidence acquisition in a new case. Figure 5 follows the inspections that distinguish cortical from OVD removal and the resulting refinement of the strategy saved for later questions.

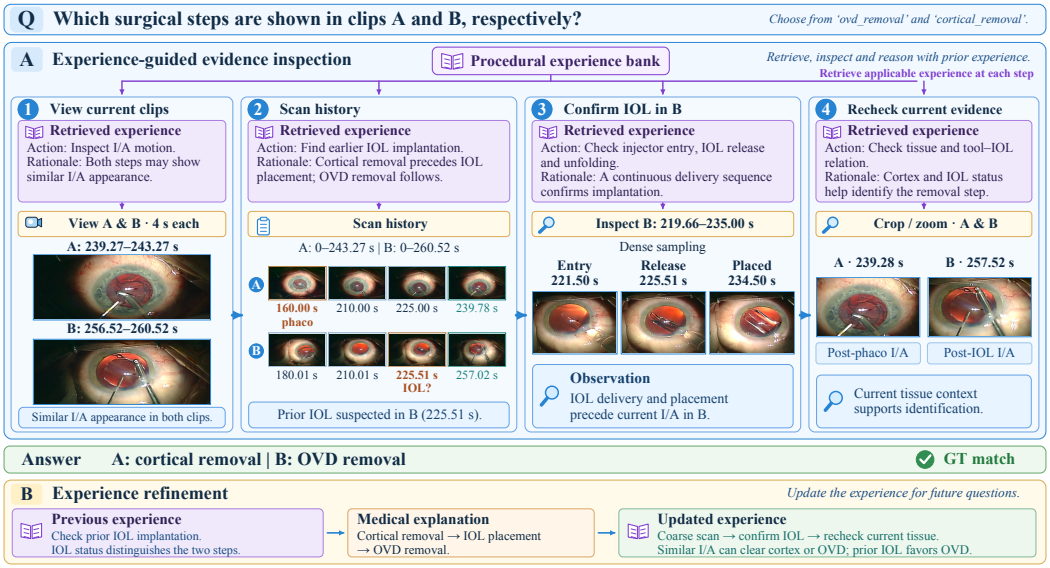



Figure 5: Experience-guided step disambiguation and memory update in a real paired question. The panels show retrieved guidance, video inspections, the final labels (A: cortical removal; B: OVD removal), and the refined experience.

**Experience-guided evidence acquisition.** Figure 5 pairs two I/A clips whose similar local views obscure whether the step is cortical or OVD removal. `OphMem` retrieves a strategy to inspect tip motion and earlier procedural history. In case B, a broad scan locates injector entry, and focused frames confirm IOL delivery, release, and unfolding before the current I/A segment; case A's segment precedes IOL placement. The agent then rechecks tissue and tool–IOL context in the current clips, identifying cortical removal in A and OVD removal in B, as annotated.

**Experience refinement.** The lower panels of Figure 5 show how this completed analysis refines the retrieved experience. The earlier entry asked whether IOL implantation had occurred; the updated entry specifies a broad history scan, focused confirmation of release and unfolding, and a final check of tissue in the current segment. It preserves the medical ordering of cortical removal, IOL placement, and OVD removal while making the visual checks for future cases explicit.

## 5 Conclusion

We present `OphMem`, an experience-based framework for contextual reasoning in ophthalmic surgical videos. It turns completed interaction trajectories into reusable strategies through medical grounding and continual consolidation. Across three ambiguity-focused tasks, `OphMem` improves the primary accuracy metrics over matched memory baselines; stream and ablation analyses show how experience accumulates and is reused. These findings highlight the value of reusable cross-case experience in guiding contextual evidence inspection to resolve visual ambiguity in ophthalmic surgical videos.

### ACKNOWLEDGMENTS

We thank ophthalmologists for their helpful discussions and clinical insights on ophthalmic surgical video understanding.

## AI USE STATEMENT

Generative AI tools were used to assist with language editing and LaTeX formatting during manuscript preparation. All scientific content, including the methodology, experimental results, analyses, and conclusions, was reviewed and verified by the authors.

### REPRODUCIBILITY STATEMENT

Section 3 specifies the reasoning and memory-update procedures. Appendix A documents the data sources, task construction, question order, input access, and scoring rules. Section 4 reports the evaluation settings and results.

### ETHICS STATEMENT

This work uses previously released ophthalmic surgical videos and their associated annotations. We collected no new patient data or clinical labels. `OphMem` is evaluated as a research system for video understanding; its benchmark results do not establish suitability for clinical use.

## REFERENCES

Mohammad Javad Ahmadi, Iman Gandomi, Parisa Abdi, Seyed-Farzad Mohammadi, Amirhossein Taslimi, Mehdi Khodaparast, Hassan Hashemi, Mahdi Tavakoli, and Hamid D. Taghirad. Cataract-LMM large-scale multi-source multi-task benchmark for deep learning in surgical video analysis. *Scientific Data*, 13:1189, 2026. doi: 10.1038/s41597-026-07464-0. URL https://www.nature.com/articles/s41597-026-07464-0.

Hassan Al Hajj, Mathieu Lamard, Pierre-Henri Conze, Soumali Roychowdhury, Xiaowei Hu, et al. CATARACTS: Challenge on automatic tool annotation for cataRACT surgery. *Medical Image Analysis*, 52:24–41, 2019. doi: 10.1016/j.media.2018.11.008. URL https://doi.org/10.1016/j.media.2018.11.008.

Ibrahim Dehshini Ali, Nisarg A. Shah, Mingze Xia, Mike Peven, Nanthini Narayanan, Rishahb Nanawati, Nuren Zhaksylyk, Ibrahim Abboud, Ayesha Javaid, Vishal M. Patel, S. Swaroop Vedula, and Shameema Sikder. Data and code associated with the publication: Labels for automated step recognition in cataract surgery for videos in the Cataract-1K dataset. Johns Hopkins Research Data Repository, 2026. URL https://doi.org/10.7281/T1ZZX1NJ.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72. Association for Computational Linguistics, 2005. URL https://aclanthology.org/W05-0909/.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. Towards injecting medical visual knowledge into multimodal LLMs at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7346–7370, 2024. doi: 10.18653/v1/2024.emnlp-main.418. URL https://aclanthology.org/2024.emnlp-main.418/.

Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1801–1810, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Dwibedi_Temporal_Cycle-Consistency_Learning_CVPR_2019_paper.html.

Lixin Fang, Lei Mou, Yuanyuan Gu, Yan Hu, Bang Chen, Xu Chen, Yang Wang, Jiang Liu, and Yitian Zhao. Global–local multi-stage temporal convolutional network for cataract surgery phase recognition. *BioMedical Engineering OnLine*, 21:82, 2022. doi: 10.1186/s12938-022-01048-w. URL https://doi.org/10.1186/s12938-022-01048-w.

Negin Ghamsarian, Yosuf El-Shabrawi, Sahar Nasirihaghighi, Doris Putzgruber-Adamitsch, Martin Zinkernagel, Sebastian Wolf, Klaus Schoeffmann, and Raphael Sznitman. Cataract-1K dataset for deep-learning-assisted analysis of cataract surgery videos. *Scientific Data*, 11:373, 2024. doi: 10.1038/s41597-024-03193-4. URL https://www.nature.com/articles/s41597-024-03193-4.

Maria Grammatikopoulou, Evangello Flouty, Abdolrahim Kadkhodamohammadi, Gwenolé Quellec, Andre Chow, Jean Nehme, Imanol Luengo, and Danail Stoyanov. CaDIS: Cataract dataset for surgical RGB-image segmentation. *Medical Image Analysis*, 71:102053, 2021. doi: 10.1016/j.media.2021.102053. URL https://doi.org/10.1016/j.media.2021.102053.

Felix Holm, Gözde Ünver, Ghazal Ghazaei, and Nassir Navab. CAT-SG: A large dynamic scene graph dataset for fine-grained understanding of cataract surgery. In *Medical Image Computing and Computer Assisted Intervention*, pp. 96–106, 2025. doi: 10.1007/978-3-032-05114-1_10. URL https://doi.org/10.1007/978-3-032-05114-1_10.

Ming Hu, Peng Xia, Lin Wang, Siyuan Yan, Feilong Tang, Zhongxing Xu, Yimin Luo, Kaimin Song, Jurgen Leitner, Xuelian Cheng, Jun Cheng, Chi Liu, Kaijing Zhou, and Zongyuan Ge. OphNet: A large-scale video benchmark for ophthalmic surgical workflow understanding. In *European Conference on Computer Vision*, 2024. doi: 10.1007/978-3-031-73235-5_27. URL https://arxiv.org/abs/2406.07471.

Ming Hu, Kun Yuan, Yaling Shen, Feilong Tang, Xiaohao Xu, Lin Zhou, Wei Li, Ying Chen, Zhongxing Xu, Zelin Peng, Siyuan Yan, Vinkle Srivastav, Diping Song, Tianbin Li, Danli Shi, Jin Ye, Nicolas Padoy, Nassir Navab, Junjun He, and Zongyuan Ge. OphCLIP: Hierarchical retrieval-augmented learning for ophthalmic surgical video-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2411.15421.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.neurips.cc/paper_files/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

Vidit Kumar, Vikas Tripathi, Bhaskar Pant, Sultan S. Alshamrani, Ankur Dumka, Anita Gehlot, Rajesh Singh, Mamoon Rashid, Abdullah Alshehri, and Ahmed Saeed AlGhamdi. Hybrid spatiotemporal contrastive representation learning for content-based surgical video retrieval. *Electronics*, 11(9):1353, 2022. doi: 10.3390/electronics11091353. URL https://doi.org/10.3390/electronics11091353.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, 2004. URL https://aclanthology.org/W04-1013/.

Simon Mueller, Bhuvan Sachdeva, Singri Niharika Prasad, Raphael Lechtenboehmer, Frank G. Holz, Robert P. Finger, Kaushik Murali, Mohit Jain, Maximilian W. M. Wintergerst, and Thomas Schultz. Phase recognition in manual small-incision cataract surgery with MS-TCN++ on the novel SICS-105 dataset. *Scientific Reports*, 15(1):16886, 2025. doi: 10.1038/s41598-025-00303-z. URL https://www.nature.com/articles/s41598-025-00303-z.

OphNet Challenge Organizers. OphNet challenge: Towards fine-grained ophthalmic surgical workflow understanding. Official challenge dataset and documentation, 2025. URL https://ophnet-challenge.github.io/. Accessed September 22, 2026.

Siru Ouyang, Jun Yan, I-Hung Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long T. Le, Samira Daruki, Xiangru Tang, Vishy Tirumalashetty, George Lee, Mahsan Rofouei, Hangfei Lin, Jiawei Han, Chen-Yu Lee, and Tomas Pfister. ReasoningBank: Scaling agent self-evolving with reasoning memory. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.25140.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023. URL https://arxiv.org/abs/2310.08560.

Jay N. Paranjape, Nisarg Shah, Nanthini Narayanan, Shameema Sikder, S. Swaroop Vedula, and Vishal M. Patel. CatVLM: Enhancing temporal understanding in cataract surgery videos with boundary-aware VLM. In *Proceedings of The 9th International Conference on Medical Imaging with Deep Learning*, volume 315 of *Proceedings of Machine Learning Research*, pp. 947–959. PMLR, 2026. URL https://proceedings.mlr.press/v315/paranjape26a.html.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Bhuvan Sachdeva, Sneha Kumari, Rudransh Agarwal, Shalaka Kumaraswamy, Niharika Singri Prasad, Simon Mueller, Raphael Lechtenboehmer, Maximilian W. M. Wintergerst, Thomas Schultz, Kaushik Murali, and Mohit Jain. CataractCompDetect: Intraoperative complication detection in cataract surgery. *arXiv preprint arXiv:2511.18968*, 2025. URL https://arxiv.org/abs/2511.18968.

Klaus Schoeffmann, Mario Taschwer, Stephanie Sarny, Bernd Münzer, Manfred Jürgen Primus, and Doris Putzgruber. Cataract-101: Video dataset of 101 cataract surgeries. In *Proceedings of the 9th ACM Multimedia Systems Conference*, pp. 421–425, 2018. doi: 10.1145/3204949.3208137. URL https://doi.org/10.1145/3204949.3208137.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, et al. MedGemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026. URL https://arxiv.org/abs/2604.05081.

Nisarg A. Shah, Shameema Sikder, S. Swaroop Vedula, and Vishal M. Patel. GLSFormer: Gated–long, short sequence transformer for step recognition in surgical videos. In *Medical Image Computing and Computer Assisted Intervention*, pp. 386–396, 2023. doi: 10.1007/978-3-031-43996-4_37. URL https://doi.org/10.1007/978-3-031-43996-4_37.

Nisarg A. Shah, Chaminda Bandara, Shameema Sikder, S. Swaroop Vedula, and Vishal M. Patel. CSMAE: Cataract surgical masked autoencoder (MAE) based pre-training. In *IEEE International Symposium on Biomedical Imaging*, pp. 1–5, 2025. doi: 10.1109/ISBI60581.2025.10981288. URL https://doi.org/10.1109/ISBI60581.2025.10981288.

Weixiang Shen, Bailiang Jian, Jun Li, Che Liu, Johannes Moll, Xiaobin Hu, Daniel Rueckert, Hongwei Bran Li, and Jiazhen Pan. Evo-MedAgent: Beyond one-shot diagnosis with agents that remember, reflect, and improve. *arXiv preprint arXiv:2604.14475*, 2026. URL https://arxiv.org/abs/2604.14475.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Hongjin Su, Ruoxi Sun, Jinsung Yoon, Pengcheng Yin, Tao Yu, and Sercan Arik. Learn-by-interact: A data-centric framework for self-adaptive agents in realistic environments. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/dd745a0c4f91fe91866fe6788be9cc28-Abstract-Conference.html.

Puxun Tu, Ce Zheng, Xiaoling Xie, Jiao Lv, Meng Xie, Jinming Guo, Shengjie Yin, Kunliang Qiu, Yue Wei, Chongyang Wang, Jingfeng Cai, Wei Mi, Yafu Wang, Xiao Zhang, Danba Jiachu, Kun Peng, Wanqi Zhang, Fengfeng Tong, Huiying Chu, Peiquan Zhao, Mingzhi Zhang, and Xiaojun Chen. An ophthalmic video foundation model for surgical recognition and navigation with wet-lab porcine eye validation. *Nature Biomedical Engineering*, 2026. doi: 10.1038/s41551-026-01622-w. URL https://www.nature.com/articles/s41551-026-01622-w.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023. URL https://arxiv.org/abs/2305.16291.

Gui Wang, Yang Wennuo, Xusen Ma, Zehao Zhong, Zhuoru Wu, Ende Wu, Rong Qu, Wooi Ping Cheah, Jianfeng Ren, and Linlin Shen. EyePCR: A comprehensive benchmark for fine-grained perception, knowledge comprehension and clinical reasoning in ophthalmic surgery. *arXiv preprint arXiv:2509.15596*, 2025a. URL https://arxiv.org/abs/2509.15596.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b. URL https://arxiv.org/abs/2508.18265.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. *arXiv preprint arXiv:2409.07429*, 2024. URL https://arxiv.org/abs/2409.07429.

Tianxin Wei, Noveen Sachdeva, Benjamin Coleman, Zhankui He, Yuanchen Bei, Xuying Ning, Mengting Ai, Yunzhe Li, Jingrui He, Ed H. Chi, Chi Wang, Shuo Chen, Fernando Pereira, Wang-Cheng Kang, and Derek Zhiyuan Cheng. Evo-Memory: Benchmarking LLM agent test-time learning with self-evolving memory. *arXiv preprint arXiv:2511.20857*, 2025. URL https://arxiv.org/abs/2511.20857.

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

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-Mem: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/19909c36f51abc4856b4560aff3d36d6-Abstract-Conference.html.

Shengtao Zhang, Jiaqian Wang, Ruiwen Zhou, Junwei Liao, Yuchen Feng, Weinan Zhang, Ying Wen, Zhiyu Li, Feiyu Xiong, Yutao Qi, Bo Tang, and Muning Wen. MemRL: Self-evolving agents via runtime reinforcement learning on episodic memory. *arXiv preprint arXiv:2601.03192*, 2026. URL https://arxiv.org/abs/2601.03192.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=EElFGvt39K.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19632–19642, 2024. doi: 10.1609/aaai.v38i17.29936. URL https://ojs.aaai.org/index.php/AAAI/article/view/29936.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19724–19731, 2024. doi: 10.1609/aaai.v38i17.29946. URL https://ojs.aaai.org/index.php/AAAI/article/view/29946.

# A Benchmark Construction and Details

## A.1 Data Sources and Annotation Basis

The benchmark organizes existing surgical videos and annotations into question-answering instances. It adds no clinician annotation. The three sources supply different forms of supervision: temporal step intervals for Step Disambiguation, case-level complication labels for Complication Verification, and ordered operation occurrences for Temporal VQA. Construction uses annotation metadata and fixed visual descriptors, without selecting questions according to any evaluated model's correctness.

**Cataract-1K and JHU step annotations.** Task 1 uses Cataract-1K (Ghamsarian et al., 2024) videos and the JHU manual step annotations (Ali et al., 2026). Each annotation identifies a case, a step, an interval, and an official split. The frozen label map contains ten retained steps: side incision, main incision, viscoelastic injection, capsulorhexis, hydrodissection, phacoemulsification, cortical removal, lens insertion, ophthalmic viscosurgical device removal, and wound closure/hydration. We retain the source names as answer strings, including `visco_injection` and `ovd_removal`; background and non-cataract labels are excluded. Cases with missing or unreadable media, inconsistent case/split metadata, invalid intervals, or annotations extending beyond the video are excluded. Case selection uses step-label coverage with deterministic tie ordering before visual pair mining.

**CataComp-104.** Task 2 uses CataComp-104 (Sachdeva et al., 2025), with posterior capsule rupture (`pcr`), vitreous loss, and iris prolapse as three separate binary targets. Ground truth comes from the official CataCompDetect case-level CSV files. Five validation cases have conflicting iris-prolapse values between those files and the hosted metadata; cases 007, 008, 009, 011, and 027 of the official validation split were excluded before sampling. The CSV also supplies cortical-wash and lens-insertion anchors used to organize visual mining. These coarse anchors and the selected observations do not provide complication-onset labels.

**OphNet2024.** Task 3 uses the OphNet2024 challenge release (OphNet Challenge Organizers, 2025), whose cases and splits differ from the original OphNet paper release (Hu et al., 2024). In the frozen annotation CSV, the benchmark identifies event types by `surgery_id` and merges consecutive segments with the same value into one operation occurrence. The `phase_id` field is retained only as provenance. Thus, consecutive phase segments within an operation do not count as repeated events. References identify an operation through a representative source segment: the longest constituent segment, with the earlier annotation index breaking ties. The operation interval and its constituent segment intervals are retained separately.

## A.2 Curation of the Paired Tasks

**Step Disambiguation.** For a valid step interval $[s, e]$, the query time is its midpoint $t = (s+e)/2$. Queries before 10 seconds are discarded. The local observation is $[\max(s, t-4), t]$, and accessible history is $[0, t]$. We compute a fixed descriptor from the start, midpoint, and end of this observation. Each frame is resized to $160\times120$ pixels; features concatenate a normalized $8\times4\times4$ HSV histogram, an $8\times8$ grayscale thumbnail, and eight vertical-strip edge statistics (Canny thresholds 80/160). The three frame descriptors are averaged and normalized.

For each query, we find the most similar descriptor from another case in the same benchmark split with a different step label. A pair is retained if cosine similarity is at least 0.88 and exceeds that of a deterministically chosen different-label comparison by at least 0.01. Eligible questions are ranked by similarity within the primary case, with question ID breaking ties, and at most five are retained. The two candidate step names are placed in a deterministic hash-based order; their display order does not define the answer. Reference A is the query and reference B its selected counterpart. The GT is the ordered pair of their source step labels. Both references provide their own local observation and causal history.

**Complication Verification.** Mining partitions each case into four bands using the available anchors: before cortical wash, cortical wash, after cortical wash and before lens insertion, and after lens insertion. Up to three six-second windows are placed evenly in each band, with five evenly

spaced requested frames per window. Windows that cannot supply all five frames are rejected. Each frame uses a normalized $8 \times 4 \times 2$ HSV histogram, an $8 \times 8$ grayscale thumbnail, and a $4 \times 4$ edge grid (Canny thresholds 50/150), computed at $160 \times 120$ resolution. Window features concatenate the frame-wise mean, standard deviation, mean absolute adjacent-frame change, and first/last descriptors, followed by normalization.

For each complication, candidate windows are compared across cases in the same split and procedural band. Pairs have high descriptor similarity, exceed the split-specific same-band null threshold, and occur in local neighborhoods containing both complication labels. The selected set contains 237 evaluation questions from 49 cases: 83 no/yes, 80 yes/no, 40 no/no, and 34 yes/yes pairs. The agent judges the two cases independently in A/B order using their full-case evidence. The selected local windows identify visually similar observations; complication ground truth remains case-level.

**Task 2 time coordinates.** Observation frames use verified source presentation timestamps. The full-case evidence scope follows each source video's timeline, and frozen source-frame hashes identify the local observations used in the questions.

### A.3 Temporal VQA Construction and Complete Question Formats

A case must support at least eight operation occurrences, three distinct operation types, one recurring type, seven transitions, and five valid questions at its assigned case order. Annotation order must be consistent; adjacent operations may overlap by at most 0.75 seconds. Cases are selected by template coverage, repeated types, distinct types, and occurrence count, with seeded ties.

Every selected case yields five questions, including a two-question temporal-update pair. A deterministic schedule based on case order modulo four alternates changed-answer pairs and unchanged-answer controls, then fills the remaining positions with eligible relation or recall templates. Invalid templates are skipped; the frozen selection requires five valid questions. Across the full split, all three reasoning families are represented. A temporal-update pair contributes two independently answered QAs, each with its own cutoff, rather than one ordered-pair answer as in Tasks 1 and 2.

For cutoff $c$, accessible history contains the source segments of operation occurrences ending no later than $c$. References are resolved to their actual constituent media segments within that history. Source operation/phase IDs, source filenames, annotation rows, GT, and answer-changing pair metadata are withheld from the public input. The loader exposes opaque media tokens, chronological order, the question, and the cutoff. An event type in the following templates is specified by its reference clip. Counts refer to merged operation occurrences; "between" excludes the bounding occurrences.

**Event Relation: presence.** *"Between the two most recent occurrences matching reference clip A, did an event matching reference clip B occur? Answer yes or no."* The cutoff is the end of the later A occurrence. GT is yes exactly when a B-type occurrence lies strictly between these two A occurrences. Negative instances require a visible B reference whose type is absent from that interval. This is template `ER_LAST_TWO_A_CONTAINS_B`.

**Event Relation: latest intervening event.** *"Between the two occurrences matching reference clip A, which reference event occurred later: B or C?"* The two most recent A occurrences at the cutoff bound an interval with at least two distinct intervening event types. One choice represents the final intervening type; the other is a different intervening type. GT is `reference_b` or `reference_c`, with deterministic choice-order permutation. This is `ER_LAST_BETWEEN_A_PAIR`.

**Event Relation: count.** *"How many events matching reference clip B occurred between the two most recent occurrences matching reference clip A? Answer with an integer."* GT counts B-type occurrences strictly between consecutive A occurrences, with the cutoff at the latter A's end. The construction schedule prefers counts of one or more than one when available. This is `ER_COUNT_B_BETWEEN_A`.

**Temporal Update: recurrence, changed answer.** *"After the shown reference occurrence, did an event matching it occur again in this case by the causal cutoff? Answer yes or no."* The reference is the first occurrence of a recurring type. The earlier cutoff ends immediately before its second

occurrence; the later cutoff includes that second occurrence. The two GTs are no and yes. This is `TU_RECURRED_CHANGED`.

**Temporal Update: recurrence, unchanged answer.** The question is exactly the same recurrence question above. A no/no control uses two cutoffs after the reference and before recurrence; a yes/yes control uses two cutoffs at or after the second occurrence. These are instances of `TU_RECURRED_UNCHANGED`. The pair relationship is for analysis and is not supplied as an answer hint.

**Temporal Update: most recent reference type.** *"By the causal cutoff, which referenced event type occurred more recently: reference clip A or reference clip B?"* Two recurring types provide the references. The paired cutoffs include a later A occurrence and then a later B occurrence, changing GT from `reference_a` to `reference_b`. This is `TU_LAST_EVENT_CHANGED`.

**Conditional Recall: preceding different event.** *"Immediately before the occurrence matching reference clip A, did the most recent different event match reference clip B or C?"* The cutoff is the end of the target A occurrence. One candidate represents its immediately preceding different operation type, and the distractor is an earlier type distinct from both the target and correct predecessor. GT is `reference_b` or `reference_c`. This is `CR_PREVIOUS_DIFFERENT`.

**Conditional Recall: occurrence index.** *"Counting matching occurrences from the start of the case, which occurrence of the event shown in the reference clip was most recent by the causal cutoff? Answer with an integer."* The reference shows the first occurrence of the selected recurring type. GT is the one-based index of its latest occurrence visible at the cutoff, equivalently the number of completed matching occurrences. This is `CR_MOST_RECENT_OCCURRENCE`.

### A.4 Splits, Input Access, and Sequential Protocol

**Case separation and counting.** Benchmark roles are assigned at the source-case level before pairing. Task 1 uses the source train, validation, and test splits for acquisition, development, and evaluation. Task 2 uses disjoint cases from the source validation split for acquisition and development, and source train cases for evaluation. Task 3 uses source validation, test, and train cases for the respective roles. Case IDs and media identities are checked for separation across roles. Table 1 counts the cases and questions represented in each pool; a case may contribute several distinct questions.

**Sequential experience protocol.** Each task uses a fixed question order and an initially empty experience bank. Within a question, the bank is fixed while the agent inspects permitted evidence and produces its answer. The completed trajectory is then abstracted and curated, and accepted experience is available to the next question. A paired question, containing both A and B references, is one update unit. The run records question order, input access, answers, curator decisions, and bank state after each question.

**Compared models.** The general vision-language baselines are Qwen3-VL-8B-Thinking (Bai et al., 2025), InternVL3.5-8B (Wang et al., 2025b), and LLaVA-Video-7B (Zhang et al., 2025). The medical baselines are Lingshu-7B (Xu et al., 2025a), HuatuoGPT-Vision-7B (Chen et al., 2024), and MedGemma-1.5-4B-IT (Sellergren et al., 2026). The memory baselines are A-Mem (Xu et al., 2025b) and ReasoningBank (Ouyang et al., 2026).

**Video input.** Task 1 samples 16 history frames and four local frames per reference. Task 2 uses 32 uniformly spaced full-case frames plus the five exact frozen observation frames per reference; the release permits a ceiling of 256 uniform full-case frames. Task 3 samples two frames from each accessible source history segment and two from each representative reference segment. The reference resolver uses the representative segment's own interval, not the longer merged-operation duration. Media adapters preserve the permitted interval or verified source timestamps and record input provenance. These are sampled video inputs; full-case access in Task 2 specifies the evidence scope, not exhaustive processing of every frame.

**Reasoning and resource accounting.** The agent can inspect video iteratively, retrieve experience, use tools, and make multiple model calls within each question. Calls, inspected frames, retrieved entries, and memory updates are recorded separately.

### A.5 Scoring and Aggregation

For a paired question $i$, write GT as $(y_{iA}, y_{iB})$ and prediction as $(\hat{y}_{iA}, \hat{y}_{iB})$. Pair accuracy is

$$\text{PairAcc} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\hat{y}_{iA} = y_{iA} \ \wedge \ \hat{y}_{iB} = y_{iB}\}.$$

The order matters. Macro F1 and Jaccard flatten the $N$ pairs into $2N$ component decisions. For each class $k$, the scores are $2TP_k/(2TP_k + FP_k + FN_k)$ and $TP_k/(TP_k + FP_k + FN_k)$, then averaged over classes; a zero denominator contributes zero. Task 1 uses the step classes present in GT. Task 2 pools all complication questions and averages over the two labels yes/no. Its main-table macro scores are not an average of the three per-complication scores. Per-complication scores and component accuracy remain supplementary scorer outputs.

Temporal VQA reports exact accuracy after answer normalization, together with mean per-question ROUGE-L (Lin, 2004) and METEOR (Banerjee & Lavie, 2005). Normalization applies Unicode NFKC, lowercasing, underscore-to-space conversion, punctuation removal, and whitespace collapse. A leading article is removed; A/B/C and clip-reference variants map to the corresponding reference label. ROUGE-L is the token longest-common-subsequence F1. METEOR uses the NLTK single-reference implementation on normalized tokens. No clinical synonym judge or LLM grading is used. Each cutoff question carries equal weight, including both members of a temporal-update pair. Main-table values are multiplied by 100.

The answer adapter accepts supported lossless encodings of an ordered pair. Unresolved or invalid paired responses are retained as two empty answer components and scored as incorrect; they are not dropped from the denominator. Well-formed pairs receive component-level credit for any correct component even when pair accuracy is zero. A complete task must have exactly one prediction for each expected question ID, with no extra or duplicate IDs. The run audit also checks input coverage, causal restrictions, question order, and the continuity of persistent updates before exporting scores.

### A.6 Representative Acquisition Examples

The following examples reproduce released acquisition questions and GT. Annotation-derived explanations establish label provenance; annotation tables and GT are not model inputs. Times are seconds. No evaluation-instance answer is reproduced.

**Step Disambiguation: two histories, two labels.** Question `t1-case_2220-002`: *"Which surgical step is shown in reference clip A and reference clip B, respectively? Choose from 'wound_closure_hydration' and 'visco_injection'. Return exactly two step names in A, B order."*

| Reference | Case | Local interval | Accessible history |
|---|---|---|---|
| A | `case_2220` | $[13.640, 16.755]$ | $[0, 16.755]$ |
| B | `case_2726` | $[323.775, 327.775]$ | $[0, 327.775]$ |

GT is [`visco_injection`, `wound_closure_hydration`]. Source CSV row 4 of case 2220 labels $[13.64, 19.87]$ as viscoelastic injection; row 12 of case 2726 labels $[321.20, 334.35]$ as wound closure/hydration. Both query times lie inside those intervals. Descriptor cosine similarity is 0.9668 and the hard-versus-comparison margin is 0.1202. This establishes the mined visual contrast and label trace, without asserting that these histories necessarily improve a model's answer.

**Complication Verification.** Question `t2-catacomp-val-001-pcr-0`: *"Considering the complete evidence in each case, is pcr present in reference case A and reference case B, respectively? Answer yes or no in A, B order."* Reference A is `catacomp-val-001`; B is `catacomp-val-046`. Their full-case scopes end at 870.709 and 4046.532 seconds. The verified observation timestamps are:

| | | | | | |
|---|---|---|---|---|---|
| A | 608.729 | 610.292 | 611.791 | 613.327 | 614.858 |
| B | 1029.617 | 1031.189 | 1032.686 | 1034.223 | 1035.719 |

GT is [no, yes], from the official validation CSV's PCR flags for these cases. The selected observations have cosine 0.9511, neighborhood entropy 0.9544, and similarity above the same-band null's 95th percentile (0.9502). The GT concerns each complete case; it does not claim that PCR is visible in either selected observation.

**Event Relation: counting repeated operations.** Question `t3-case_0344-er-count`: *"How many events matching reference clip B occurred between the two most recent occurrences matching reference clip A? Answer with an integer."* Reference A represents operation type 3 over $[509.08, 516.20]$; B represents type 2 over $[596.68, 625.98]$. At cutoff 1147.02, the bounding A occurrences are $[509.08, 516.20]$ and $[1136.08, 1147.02]$. Three intervening type-2 occurrences occupy $[596.68, 625.98]$, $[853.08, 901.76]$, and $[1082.96, 1105.76]$, so GT is `3`. The first of these combines two consecutive source segments and counts once.

**Temporal Update.** Questions `t3-case_0885-tu_recurred_changed-0` and `-1` both ask: *"After the shown reference occurrence, did an event matching it occur again in this case by the causal cutoff? Answer yes or no."* The reference is the type-0 occurrence $[43.79, 90.51]$. At cutoff 96.23, no later matching occurrence has completed, so GT is `no`; at 268.40, the next matching occurrence has completed, so GT is `yes`. Each question receives only its own accessible history.

For contrast, questions `t3-case_0230-tu_recurred_unchanged-0` and `-1` use the same recurrence wording with reference $[9.90, 98.43]$. Cutoffs 98.43 and 103.84 both precede a later completed occurrence of that type, yielding `no` and `no`. A later cutoff therefore does not itself imply an answer change.

**Conditional Recall.** Question `t3-case_0230-cr-previous-0`: *"Immediately before the occurrence matching reference clip A, did the most recent different event match reference clip B or C?"* At cutoff 103.84, the visible merged sequence is type 5 over $[0, 9.90]$, type 0 over $[9.90, 98.43]$, and type 3 over $[98.43, 103.84]$. A, B, and C refer to the third, second, and first occurrences respectively. GT is `reference_b`. The six consecutive source segments comprising the type-0 occurrence do not become six distinct events.

### A.7 Release Identity and Reproduction

The benchmark manifest records source splits, case identities, annotation references, question IDs, media intervals, construction hashes, and question order. Reproduction uses the source resources, frozen question manifest, verified media mapping, public-input loader, runner configuration, and scorer.

## B Limitations & Future Work

Our evaluation covers the task types defined in the current benchmark; future work will extend the study to additional question formulations. OphMem represents experiences as text, and future work will explore richer structured and multimodal memory representations.