EOR: EVIDENCE-OBLIGATION REASONING FOR VERIFICATION-DRIVEN LONG-VIDEO QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-video question answering requires reasoning over sparse and temporally dispersed evidence. A central challenge is that retrieving video segments relevant to a question does not necessarily provide sufficient evidence to justify an answer. Existing retrieval-based methods mainly determine what to inspect from question relevance or intermediate reasoning cues, while recent hypothesis-verification approaches make evidence acquisition more goal-directed but typically do not maintain explicit, persistent verification requirements that govern subsequent reasoning.We propose EOR, an Evidence-Obligation Reasoning framework that formulates long-video QA as iterative hypothesis verification. Each candidate hypothesis is associated with persistent evidence obligations, i.e., observable conditions that should be established before the hypothesis is sufficiently grounded. Their states are maintained across reasoning rounds, allowing unresolved, contradicted, and discriminative requirements to determine what evidence should be inspected next. EOR uses a lightweight temporal event graph to localize candidate regions, verifies the corresponding source-video evidence, and allocates verification effort across a compact set of complementary hypotheses rather than repeatedly inspecting redundant candidates. For open-ended QA, accumulated verified evidence can further revise an inadequate hypothesis space instead of merely re-ranking a fixed set of initial candidates. We evaluate EOR in both multiple-choice and open-ended long-video question answering, examining answer quality, evidence-seeking efficiency, and robustness to incomplete initial hypotheses.

1 INTRODUCTION

Long-video question answering (LVQA) requires models to identify decision-critical evidence from videos containing thousands of frames. The challenge is not merely input length: relevant evidence may be brief, temporally distant, and distributed across multiple events whose temporal or causal relations jointly determine the answer. Recent benchmarks, including EgoSchema, LongVideoBench, Video-MME, MLVU, and LVBench, increasingly expose these difficulties over extended temporal contexts (Mangalam et al., 2023; Wu et al., 2024; Fu et al., 2025; Zhou et al., 2025; Wang et al., 2025a). In such settings, retrieving a semantically relevant clip is not equivalent to obtaining sufficient evidence. A segment may contain the correct person, object, or action while still omitting the event needed to establish why, when, or under what condition an answer is correct. This gap is especially important for vision-language models, which may produce plausible answers even when the underlying visual evidence remains incomplete.

The distinction between relevance and sufficiency changes what a long-video reasoning system should optimize for. If retrieval is driven only by similarity to the question, the system may repeatedly observe visually related events without resolving the specific uncertainty that separates competing answers. For example, seeing a person take medication may be highly relevant to a question about why the medication was needed, yet it does not by itself establish the cause. A reliable answer may additionally require evidence about an earlier injury, the affected body part, or the temporal relation between the injury and the observed action. Long-video reasoning therefore requires not only locating useful content, but also determining which verification requirements remain unresolved after each observation.

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.

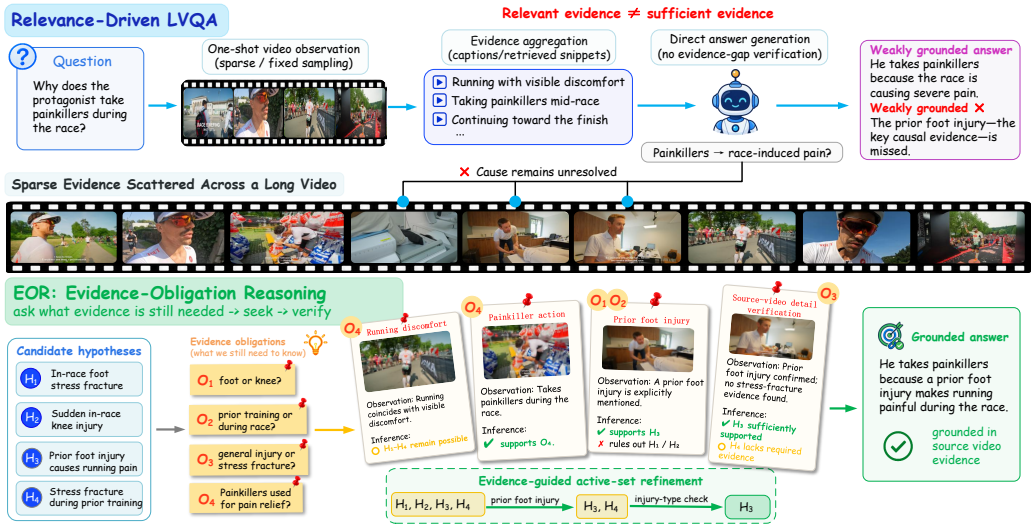


Figure 1: **Motivation of EOR.** Relevant video content is not necessarily sufficient to verify an answer. EOR associates candidate hypotheses with persistent evidence obligations and uses unresolved obligations to drive targeted source-video verification until the answer is sufficiently grounded.

Existing approaches mainly improve how long visual streams are represented or accessed. Sparse memory, textual abstraction, temporal modeling, and adaptive compression reduce the cost of processing extended videos (Song et al., 2024; Zhang et al., 2024; Ren et al., 2024; Shen et al., 2025). More recent systems actively acquire task-relevant evidence rather than processing the entire video uniformly, using iterative retrieval, query-adaptive localization, uncertainty-aware search, or multimodal memory (Wang et al., 2024; Ma et al., 2025; Wang et al., 2025b; Zhi et al., 2025; Yeo et al., 2026). These methods substantially improve efficiency and adaptivity, but evidence acquisition is still largely driven by relevance, uncertainty, or the current reasoning cue. What remains less explicit is a persistent representation of *which evidence is still required* before a particular candidate answer can be considered sufficiently grounded.

Recent hypothesis-verification methods move closer to this objective. Think, Then Verify converts candidate answers into testable hypotheses and uses discriminative clues to guide localized video inspection (Wang et al., 2026). EOR shares this verification-oriented perspective but introduces a different reasoning state: each hypothesis is decomposed into multiple *persistent evidence obligations*, i.e., observable conditions whose verification contributes to determining whether the hypothesis is sufficiently grounded. The state of each obligation is retained across reasoning rounds, so subsequent evidence acquisition is driven by what remains unresolved for the competing hypotheses rather than by a transient verification target.

Persistent evidence obligations provide EOR with a stable reasoning state across iterations. They distinguish missing evidence from explicit contradiction, prevent already verified conditions from being repeatedly rediscovered, and expose which unresolved requirements still separate competing hypotheses. These states serve two complementary roles: they determine what evidence should be sought next, and they provide a basis for allocating a limited verification budget across competing hypotheses. EOR therefore combines obligation-guided source-video verification with an evidence-coverage-aware active set that preserves complementary unresolved requirements while avoiding redundant verification. For open-ended QA, verified evidence can additionally revise an inadequate initial hypothesis space, while revised or newly introduced candidates remain subject to the same verification process. Figure 1 illustrates the motivation behind this shift from relevance-driven retrieval to verification-driven evidence acquisition.

We summarize our contributions as follows:

- We introduce persistent evidence obligations for long-video QA, explicitly representing the verification requirements of competing hypotheses across reasoning rounds.

- We develop obligation-guided verification with evidence-aware hypothesis selection, directing source-video inspection toward unresolved requirements while reducing redundant verification
- We evaluate EOR on both multiple-choice and open-ended long-video QA, examining answer quality, evidence-access efficiency, and the contributions of its major reasoning components through controlled ablations.

2 Related Work

Long-video question answering and representation. Video QA has evolved from short-clip understanding toward reasoning over increasingly extended temporal contexts. ActivityNet-QA and NExT-QA emphasize open-ended answering and temporal-causal reasoning, while EgoSchema, LongVideoBench, Video-MME, MLVU, and LVBench further increase video duration, task diversity, and reasoning complexity (Yu et al., 2019; Xiao et al., 2021; Mangalam et al., 2023; Wu et al., 2024; Fu et al., 2025; Zhou et al., 2025; Wang et al., 2025a). To make such inputs tractable, long-video models reduce visual redundancy through sparse memory, textual abstraction, timestamp-aware modeling, or adaptive compression (Song et al., 2024; Zhang et al., 2024; Ren et al., 2024; Shen et al., 2025). These approaches substantially improve access to long visual contexts, but mainly address how video information is represented or compressed. EOR instead focuses on how the acquired evidence should be organized and verified with respect to competing answer hypotheses.

Adaptive evidence acquisition. Rather than processing an entire video uniformly, recent methods selectively acquire task-relevant evidence. VideoAgent iteratively retrieves visual information under LLM control, while DrVideo converts long videos into coarse documents and augments missing information through retrieval (Wang et al., 2024; Ma et al., 2025). VideoTree further performs query-adaptive coarse-to-fine localization, while VideoINSTA explores informative spatiotemporal reasoning and evidence sufficiency (Wang et al., 2025b; Liao et al., 2024). Other systems introduce uncertainty-aware acquisition, curiosity-driven exploration, multi-agent collaboration, adaptive frame selection, and temporal search (Zhi et al., 2025; Yang et al., 2025; Chen et al., 2025; He et al., 2026; Pan et al., 2026). Recent memory-based agents such as WorldMM additionally support iterative retrieval over multimodal memories (Yeo et al., 2026). EOR follows the selective-access paradigm but makes the search objective more explicit: evidence acquisition is driven by unresolved verification requirements of the current hypotheses rather than relevance or uncertainty alone.

Hypothesis verification and open-ended reasoning. Most closely related to EOR, Think, Then Verify formulates long-video understanding as hypothesis verification, where candidate answers are converted into testable hypotheses and discriminative clues guide localized video inspection (Wang et al., 2026). EOR shares this verification-oriented perspective but represents the reasoning state differently. Instead of organizing verification around a clue-level target, each hypothesis is decomposed into multiple persistent evidence obligations whose satisfied, contradicted, and unresolved states are retained across reasoning rounds. These obligation states directly determine subsequent evidence acquisition, so the distinction is not whether iterative verification is performed, but how unresolved verification requirements are represented and reused. Open-ended video QA further removes the guarantee that the correct answer direction is available beforehand (Yu et al., 2019; Xiao et al., 2021; Maaz et al., 2024). EOR addresses this setting by allowing verified source-video evidence to revise an inadequate hypothesis space while returning newly introduced candidates to the same verification loop.

3 Method

EOR formulates long-video question answering as iterative verification of competing answer hypotheses. Given a long video V and a question q, the framework maintains a reasoning state consisting of candidate hypotheses, their evidence obligations, and a lightweight temporal event graph of the video (Sec. 3.1). Each iteration selects the most informative unresolved obligation, uses it to retrieve candidate regions through the event graph, and verifies them in the source video; the resulting observations update obligation states and hypothesis scores (Sec. 3.2). A budgeted selection mechanism then decides which hypotheses remain worth verifying, and a grounding test decides

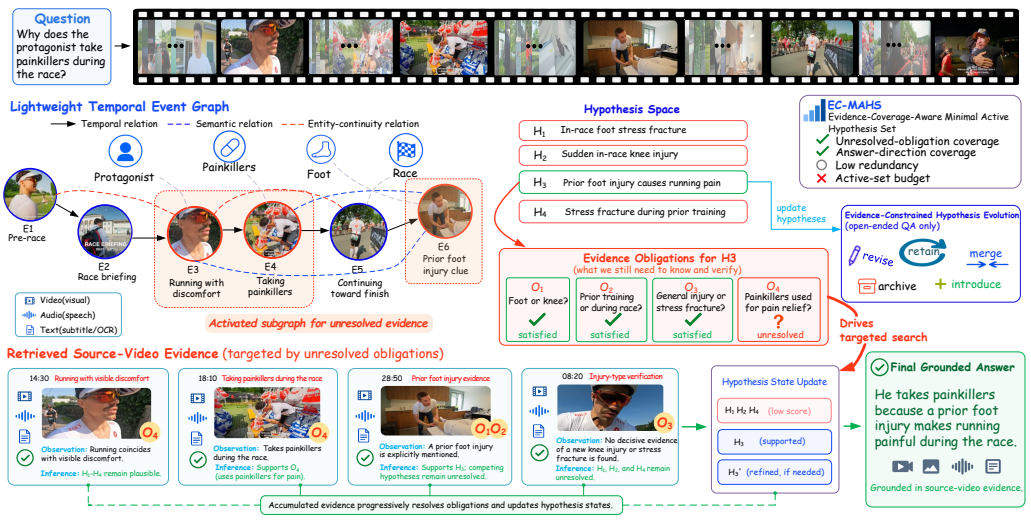


Figure 2: **Overview of EOR.** Persistent evidence obligations guide targeted source-video verification, while EC-MAHS allocates the verification budget across competing hypotheses. For open-ended QA, verified evidence can additionally revise the hypothesis space.

when to stop (Sec. 3.3). For open-ended questions, the hypothesis space itself may evolve under verified evidence (Sec. 3.4). Fig. 2 gives an overview of the framework.

3.1 Persistent Verification State

Hypotheses and obligations. At iteration t, EOR maintains a hypothesis pool $\mathcal{H}^{(t)} = \{h_i^{(t)}\}_{i=1}^{K_t}$, where $h_i^{(t)} = (a_i^{(t)}, \mathcal{O}_i^{(t)}, s_i^{(t)})$ contains a candidate answer, its evidence obligations, and an evidence-based score. Multiple-choice options initialize hypotheses directly, while open-ended QA uses a small candidate pool proposed from the question and coarse video context. Each obligation $o_{ij}^{(t)} = (c_{ij}, z_{ij}^{(t)}, \alpha_{ij})$ specifies an observable condition expected under the hypothesis, with $z_{ij}^{(t)} \in \{\text{unresolved}, \text{satisfied}, \text{contradicted}\}$ and importance weight $\alpha_{ij} > 0$. Obligation states persist across iterations: satisfied conditions are not re-verified, unresolved conditions guide subsequent evidence seeking, and contradicted conditions affect hypothesis scores. Semantically equivalent obligations are grouped into families $\mathcal{F}$, allowing retrieved source evidence to be reused across hypotheses. We omit the superscript (t) when unambiguous.

Temporal event graph. To avoid repeatedly processing the full video, EOR maintains a lightweight temporal graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where each event $n_k = (x_k, \ell_k, [t_k^s, t_k^e], p_k)$ stores a retrieval feature, a coarse semantic cue, its temporal span, and a source-video pointer. Typed temporal, entity-continuity, and semantic edges support local expansion during retrieval. The graph serves only as a navigation layer; all evidential decisions are grounded in the corresponding source-video intervals.

Hypothesis scoring. We summarize each hypothesis by its weighted satisfied and contradicted coverage:

$$C_{i,+}^{(t)} = \frac{\sum_j \alpha_{ij} \mathbb{I}[z_{ij}^{(t)} = \text{satisfied}]}{\sum_j \alpha_{ij}}, \qquad C_{i,-}^{(t)} = \frac{\sum_j \alpha_{ij} \mathbb{I}[z_{ij}^{(t)} = \text{contradicted}]}{\sum_j \alpha_{ij}}, \tag{1}$$
$$s_i^{(t)} = b_i + \beta_+ C_{i,+}^{(t)} - \beta_- C_{i,-}^{(t)}.$$

where $\mathbb{I}[\cdot]$ is the indicator function, b_i is the hypothesis prior, and β_+ and β_- weight supporting and contradictory evidence, respectively. The resulting score ranks hypotheses according to verified evidence, while sufficient grounding is determined separately by the grounding criterion in Sec. 3.3.

3.2 OBLIGATION-GUIDED EVIDENCE SEEKING

This subsection determines what missing evidence to acquire next and where to find it in the video.

What to verify next. Not every unresolved obligation is equally informative. EOR scores each unresolved obligation by

$$\pi_{ij}^{(t)} \;=\; \alpha_{ij}\,\big(1+\lambda_d\,u_{ij}^{(t)}+\lambda_c\,w_{ij}^{(t)}\big), \tag{2}$$

where $u_{ij}^{(t)} \in [0,1]$ measures the discriminative value of c_{ij} among the currently plausible hypotheses: it is high when verifying the condition would support and refute different competing hypotheses, and low when the condition affects them similarly. The support/refute/neutral relation between a condition and each hypothesis is assigned when the obligation is created, while $u_{ij}^{(t)}$ is recomputed over the current plausible pool at each iteration. Its exact computation is provided in Appendix A.3. The binary term $w_{ij}^{(t)}$ indicates whether reliable supporting and refuting observations currently coexist for the obligation family (Appendix A.4). Among the unresolved obligations of active hypotheses, EOR verifies the one with the highest $\pi_{ij}^{(t)}$ next.

Where to look. For the selected obligation, EOR encodes the obligation-conditioned query $\phi(q, a_i^{(t)}, c_{ij})$ into the retrieval space of the event graph and activates

$$\mathcal{N}_{ij}^{(t)} = \Big\{ n_k \in \mathcal{N} \;\Big|\; \cos\Big(\phi(q, a_i^{(t)}, c_{ij}), x_k\Big) \geq \theta_{ij}^{(t)} \Big\}, \tag{3}$$

where $\theta_{ij}^{(t)}$ is determined from the similarity-score distribution for obligation o_{ij}. Its quantile level is progressively relaxed when the same obligation remains unresolved, thereby expanding the search range, while a per-round event budget prevents excessive video access. If no event satisfies the threshold, the highest-scoring event is retained as a fallback. The exact relaxation schedule and budget are specified in the Appendix A.5. Activated nodes are then expanded along the typed edges of the temporal event graph $\mathcal{G}$:

$$\widetilde{\mathcal{N}}_{ij}^{(t)} = \mathcal{N}_{ij}^{(t)} \cup \big\{\, n_m \in \mathcal{N} \;\big|\; \exists\, n_k \in \mathcal{N}_{ij}^{(t)},\ (n_k, n_m, r) \in \mathcal{E},\ r \in \rho(c_{ij}) \,\big\}, \tag{4}$$

where $\rho(c_{ij})$ selects the edge types relevant to the obligation (e.g., temporal-order conditions expand along temporal edges; entity-state conditions along entity-continuity edges). The relevant relation type is assigned together with the obligation when the condition is generated, using a fixed vocabulary of temporal, entity-continuity, and semantic relations. Expansion is limited to one hop and at most η additional nodes. In this view, question-only retrieval corresponds to $\mathrm{Retrieve}(q)$, whereas EOR performs $\mathrm{Retrieve}(q, h_i, o_{ij})$; the retrieval target therefore changes as the verification state evolves.

Verifying in the source video. Activated events only nominate candidate regions and are not treated as evidence. EOR inspects the corresponding source-video intervals conditioned on (q, c_{ij}), and the verifier returns *support*, *refute*, or *uncertain* with confidence γ. Only predictions with $\gamma \geq \gamma_h$, where γ_h is the reliability threshold, update the obligation state: reliable support yields *satisfied*, reliable refutation yields *contradicted*, while conflicting or insufficient evidence leaves it *unresolved*. Thus, all state updates remain grounded in source-video observations.

Cross-hypothesis evidence reuse. An inspected source interval can be reused for a semantically similar obligation of another hypothesis when $\mathrm{sim}(\psi(c_{ij}), \psi(c_{i'j'})) \geq \tau_{\mathrm{prop}}$. The verifier re-evaluates the cached interval for the new condition rather than propagating the previous verification label, reducing repeated video access while preserving obligation-specific decisions.

3.3 EVIDENCE-AWARE HYPOTHESIS MANAGEMENT

Active-set selection. Verifying every live hypothesis at every iteration would exhaust the source-video and model budget. EOR therefore maintains a live pool $\mathcal{H}_{\mathrm{live}}^{(t)}$ of all non-archived hypotheses and restricts active-set selection to currently plausible hypotheses, defined as those whose scores are within a margin ϵ_{pl} of the current leader, i.e., $s_i^{(t)} \geq s_{\mathrm{top}}^{(t)} - \epsilon_{\mathrm{pl}}$. The resulting plausible pool is

denoted by $\mathcal{H}_{\mathrm{pl}}^{(t)}$. We evaluate an active set from two complementary aspects: unresolved-obligation coverage and answer-direction coverage. Let $\mathcal{F}_{\mathrm{hi}}^{(t)}$ denote the obligation families containing at least one high-priority unresolved obligation with $\pi_{ij}^{(t)} \geq \tau_\pi$, and let $D_1, \ldots, D_Q$ denote the answer directions in $\mathcal{H}_{\mathrm{pl}}^{(t)}$. We define

$$\begin{aligned} \mathrm{Cov}_{\mathrm{obl}}(\mathcal{A}) &= \frac{\left|\left\{F \in \mathcal{F}_{\mathrm{hi}}^{(t)} : \exists h \in \mathcal{A}, \exists o \in \mathcal{O}(h) \cap F, z(o) = \text{unresolved}\right\}\right|}{\left|\mathcal{F}_{\mathrm{hi}}^{(t)}\right|}, \\ \mathrm{Cov}_{\mathrm{dir}}(\mathcal{A}) &= \frac{|\{D_q : D_q \cap \mathcal{A} \neq \emptyset\}|}{Q}. \end{aligned} \tag{5}$$

When $\mathcal{F}_{\mathrm{hi}}^{(t)}$ is empty, we set $\mathrm{Cov}_{\mathrm{obl}}(\mathcal{A}) = 1$. Each plausible option forms one answer direction in multiple-choice QA, while semantically equivalent open-ended candidates share a direction.

Based on these coverage measures, EOR selects the smallest active set that preserves sufficient obligation and answer-direction coverage while retaining the current leading hypothesis:

$$\mathcal{A}^{(t)} = \arg\min_{\mathcal{A} \subseteq \mathcal{H}_{\mathrm{pl}}^{(t)}} |\mathcal{A}| \quad \text{s.t.} \quad \mathrm{Cov}_{\mathrm{obl}}(\mathcal{A}) \geq \tau_{\mathrm{obl}}, \quad \mathrm{Cov}_{\mathrm{dir}}(\mathcal{A}) \geq \tau_{\mathrm{dir}}, \quad h_{\mathrm{top}}^{(t)} \in \mathcal{A}. \tag{6}$$

We refer to this minimum-cardinality objective as the *Evidence-Coverage-Aware Minimal Active Hypothesis Set* (EC-MAHS).

In practice, we approximate the ideal solution of Eq. 6 with a greedy coverage heuristic. Starting from the leading hypothesis, the algorithm repeatedly adds the candidate with the largest combined marginal gain in obligation and answer-direction coverage. It stops when both coverage thresholds are reached, the active-set budget K_{act} is exhausted, or no remaining hypothesis provides positive coverage gain. Thus, a lower-scoring hypothesis may remain active when it uniquely covers a discriminative unresolved family, whereas redundant hypotheses stop consuming verification calls.

When to stop. EOR does not terminate merely because one hypothesis currently leads. The leading hypothesis $h_{\mathrm{top}}^{(t)}$ is considered sufficiently grounded when all of the following conditions hold, where the subscript top denotes quantities associated with the current leading hypothesis:

$$\begin{aligned} &C_{\mathrm{top},+}^{(t)} \geq \tau_+ \ \wedge \ \sum_j \alpha_{\mathrm{top},j} \mathbb{I}\left[z_{\mathrm{top},j}^{(t)} = \text{satisfied}\right] \geq \nu, \\ &C_{\mathrm{top},-}^{(t)} \leq \tau_- \ \wedge \ s_{\mathrm{top}}^{(t)} - \max_{h \in \mathcal{H}_{\mathrm{live}}^{(t)} \setminus \{h_{\mathrm{top}}^{(t)}\}} s_h^{(t)} \geq \delta. \end{aligned} \tag{7}$$

Here, τ_+ and τ_- are the minimum satisfied-coverage and maximum contradicted-coverage thresholds, ν requires sufficient absolute verified evidence, and δ enforces a minimum score margin. If the grounding criterion is not met, multiple-choice QA stops after r stagnant iterations or budget exhaustion, whereas open-ended QA triggers hypothesis evolution while budget remains (Sec. 3.4); fallback is used only when no further verification or evolution is possible.

Final grounding check. Local verification evaluates evidence against individual obligations. Before answering, EOR jointly re-inspects the decisive source-video evidence of h_{top} and checks whether these observations jointly support the leading hypothesis without temporal or evidential inconsistencies. If the check fails, the obligations implicated in the inconsistency are reset to *unresolved*, and verification resumes while budget remains. When no further verification or hypothesis evolution is possible, EOR returns the live hypothesis with the highest evidence-based score s_i as a fallback prediction. The frequency of this fallback is reported in the experiments.

3.4 Evidence-Constrained Hypothesis Evolution

The preceding verification process assumes that the live pool contains a candidate capable of explaining the correct answer. This is reasonable for multiple-choice QA, but not for open-ended QA, where the initial pool may contain incorrect entities or relations, or miss the correct explanation

altogether. Re-ranking a fixed candidate set cannot recover from such failures. Let $\mathcal{B}^{(\leq t)}$ denote the source-grounded evidence accumulated up to iteration t. For open-ended QA, EOR triggers hypothesis evolution when the leading hypothesis fails to satisfy the grounding criterion for r consecutive iterations, or when every currently plausible hypothesis contains at least one high-priority contradicted obligation:

$$\mathcal{H}^{(t+1)} = \text{Evolve}\left(\mathcal{H}^{(t)}, \mathcal{B}^{(\leq t)}, q\right). \tag{8}$$

The evolution step may retain supported hypotheses, revise or merge partially correct ones, archive contradicted or redundant ones, or introduce a new candidate when the current pool cannot explain the verified evidence. Previously verified states are inherited only for unchanged obligations; modified or newly introduced obligations are initialized as *unresolved*. An archived hypothesis may be reactivated if later source-video evidence invalidates the observation that caused its rejection. Every revised or newly introduced hypothesis is decomposed into evidence obligations and returned to the same verification loop. Thus, verified evidence may modify the candidate space, but hypothesis generation alone never certifies an answer.

4 Experiments

We evaluate EOR around three questions. First, does verification-driven evidence acquisition improve answer quality over relevance-based retrieval under a comparable video-access budget? Second, do persistent evidence obligations provide a more effective reasoning state than question-driven, hypothesis-driven, or transient verification strategies? Third, does evidence-aware hypothesis management improve the use of a limited verification budget, and can hypothesis evolution benefit open-ended QA when the initial candidate space is inadequate? We study these questions on EgoSchema and ActivityNet-QA, focusing on both answer quality and the behavior of the evidence-seeking process.

4.1 Experimental Setup

Datasets. We evaluate EOR on two complementary benchmarks. **EgoSchema** is our primary long-form multiple-choice benchmark, where answer options naturally instantiate competing hypotheses and many questions require temporally distributed evidence. **ActivityNet-QA** evaluates open-ended QA and is used to test whether verified evidence can revise an initially incomplete hypothesis space rather than only re-rank fixed candidates.

Baselines. We use both literature-level and controlled comparisons. Table 1 and 2 compare EOR with published methods on EgoSchema and ActivityNet-QA, respectively. We prioritize closely matched Qwen2.5-VL-7B training-free settings, while larger-model systems are included only as references when their backbones or video-access strategies differ substantially. For controlled analysis, Table 3 compares Question Retrieval, Hypothesis Retrieval, Transient Verification, EOR without EC-MAHS, and the full EOR under the same verification budget.

Metrics. We report accuracy on EgoSchema, and both accuracy and QA score on ActivityNet-QA. For controlled analysis, we additionally report the average number of source-video verification calls (*Avg. Verify Calls*) and *Verified Obligation Coverage*, defined as the final satisfied-obligation coverage of the selected hypothesis. Additional efficiency statistics are provided in the appendix.

Implementation details. All controlled variants use the same event representation, retrieval encoder, verifier, preprocessing pipeline, and source-video access budget. The maximum source-video verification budget is 12 calls per question. For open-ended QA, the number of initial hypotheses is selected adaptively. Full implementation details, prompts, thresholds, and stopping criteria are provided in the appendix.

4.2 Main Results

Comparison on EgoSchema. Table 1 places EOR in the context of published results on the EgoSchema 500-question subset. The table separates comparable 7B training-free methods from

Table 1: Comparison on the EgoSchema 500-question subset. Published results use different backbone and video-access settings; we therefore separate comparable 7B training-free methods from larger-model system references.

Method	Backbone	Visual Access	Acc. (%) ↑
Comparable 7B training-free methods			
Qwen2.5-VL-7B (Bai et al., 2025)	Qwen2.5-VL-7B	32 frames	59.4
Explore-Then-Select (Shi et al., 2025)	Qwen2.5-VL-7B	128→32	60.6
A.I.R. (Zou et al., 2026)	Qwen2.5-VL-7B	≤32 frames	62.4
Large-model system references			
DrVideo (Ma et al., 2025)	GPT-4	Iterative	66.4
VideoHV-Agent (Wang et al., 2026)	GPT-4o	1 fps	81.0
Ours			
EOR	Qwen2.5-VL-7B	Adaptive	70.2

larger-model system references so that differences in backbone capacity are visible rather than implicitly treated as method-only gains. Among the comparable 7B methods, the reported accuracies range from 59.4% to 62.4%, whereas the larger GPT-4/GPT-4o systems report 66.4% and 81.0%, respectively. EOR obtains an accuracy of 70.2% under its adaptive evidence-access strategy. The controlled study in Sec. 4.3 further isolates the contribution of EOR's reasoning mechanism under a fixed verification budget.

Open-ended performance on ActivityNet-QA. Table 2 evaluates EOR on open-ended video question answering. The published Qwen2.5-VL-7B baselines provide a compact reference for direct prediction, retrieval-based evidence selection, and query-adaptive selection. EOR obtains an accuracy of 57.2% and a QA score of 3.21. The same table also includes EOR without hypothesis evolution, allowing the open-ended extension to be assessed without introducing a separate ablation table.

Table 2: Results on ActivityNet-QA. We prioritize Qwen2.5-VL-7B training-free baselines with closely matched evaluation settings. The fixed-pool variant disables evidence-constrained hypothesis evolution.

Method	Backbone	Acc. (%) ↑	Score ↑
Qwen2.5-VL-7B (Bai et al., 2025)	Qwen2.5-VL-7B	52.1	2.96
Retrieval (Shi et al., 2025)	Qwen2.5-VL-7B	52.7	2.98
Explore-Then-Select (Shi et al., 2025)	Qwen2.5-VL-7B	54.3	3.07
EOR w/o Hypothesis Evolution	Qwen2.5-VL-7B	55.6	3.12
EOR	Qwen2.5-VL-7B	57.2	3.21

Effect of hypothesis evolution. The last two rows of Table 2 isolate the effect of hypothesis evolution. Keeping the initial candidate pool fixed yields 55.6% accuracy and a QA score of 3.12, whereas enabling evidence-constrained hypothesis evolution improves them to 57.2% and 3.21, respectively. This result suggests that verified evidence can refine an inadequate candidate space rather than being used only to re-rank fixed hypotheses.

4.3 Ablation and Analysis

Persistent verification state. The comparison between *Transient Verification* and *EOR w/o EC-MAHS* isolates the effect of retaining evidence requirements across reasoning rounds. Transient Verification regenerates verification targets at each iteration without preserving previously resolved states, whereas EOR w/o EC-MAHS maintains these obligation states while excluding evidence-aware active-hypothesis management. Accuracy improves from 67.0% to 69.2%, while Verified Obligation Coverage increases from 56.8% to 68.5% and the average number of verification calls increases from 6.8 to 7.9. This directly tests whether persistent obligations provide useful reasoning memory beyond transient verification cues.

From relevance-driven to obligation-driven retrieval. Question Retrieval obtains an accuracy of 62.4%, while conditioning retrieval on the current hypothesis increases it to 64.6%. Adding explicit verification targets further improves accuracy to 67.0% for Transient Verification, while persistent obligations yield 69.2% with EOR w/o EC-MAHS and 70.2% with full EOR. Verified Obligation Coverage follows the same controlled progression, increasing from 41.5% to 48.2%, 56.8%, 68.5%, and 76.3%, respectively. These results indicate a progressive shift from relevance-based retrieval toward persistent, obligation-driven evidence acquisition.

Table 3: Ablation study on EgoSchema under the same verification budget. Verified Obligation Coverage (Verif. Obl. Cov.) is the final satisfied-obligation coverage of the selected hypothesis (Section 4.1). Avg. Verify Calls counts source-video verification calls per question.

Variant	EgoSchema Acc. (%) ↑	Verif. Obl. Cov. (%) ↑	Avg. Verify Calls ↓
Question Retrieval	62.4	41.5	3.2
Hypothesis Retrieval	64.6	48.2	4.1
Transient Verification	67.0	56.8	6.8
EOR w/o EC-MAHS	69.2	68.5	7.9
EOR	70.2	76.3	6.1

Evidence-aware hypothesis selection. Removing EC-MAHS decreases accuracy from 70.2% to 69.2% and Verified Obligation Coverage from 76.3% to 68.5%, while increasing the average number of verification calls from 6.1 to 7.9. Unlike score-only hypothesis selection, EC-MAHS retains lower-ranked hypotheses when they uniquely cover unresolved evidence requirements, while redundant candidates no longer consume the same verification budget.

Reasoning under limited verification budgets. We further examine how different reasoning strategies behave when source-video access is increasingly restricted.

Figure 3 compares Question Retrieval, Transient Verification, and EOR under the same source-video verification budgets. With a budget of 4 calls, EOR achieves 68.2% accuracy, compared with 64.6% for Transient Verification and 61.4% for Question Retrieval. As the budget increases, EOR reaches 70.2% accuracy at 12 calls, while the two baselines remain at 67.0% and 62.4%, respectively.

The budget analysis complements Table 3: rather than testing only whether individual reasoning mechanisms improve final accuracy, it examines whether the proposed verification state makes more effective use of a constrained evidence-acquisition budget. This is particularly important for long-video QA, where repeatedly inspecting additional source-video regions can substantially increase inference cost.

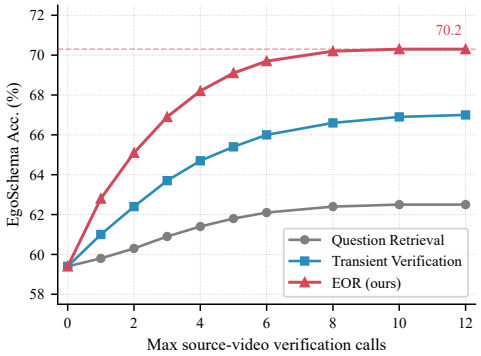


Figure 3: EgoSchema accuracy under different verification budgets for three reasoning strategies.

5 Conclusions

We introduced EOR, a verification-driven framework for long-video question answering that represents unresolved evidence requirements as persistent evidence obligations. Rather than repeatedly retrieving video content from question relevance or transient reasoning cues, EOR uses the evolving verification state to determine what evidence should be inspected next and allocates a limited verification budget across complementary hypotheses. Experiments on EgoSchema and ActivityNet-QA show that EOR achieves 70.2% accuracy on EgoSchema and 57.2% accuracy with a QA score of 3.21 on ActivityNet-QA, while controlled analyses examine the contributions of persistent obligation states, obligation-guided evidence seeking, and evidence-aware hypothesis management.

AI USE STATEMENT

Generative AI tools were used **for** language translation, literature search, and code debugging. All research decisions, implementations, experiments, analyses, and conclusions were independently reviewed and validated by the authors. The authors take full responsibility for the content of this work.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. LVAgent: Long video understanding by multi-round dynamical collaboration of mllm agents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20237–20246, 2025.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025.

Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. FrameThinker: Learning to think with long videos via multi-turn frame spotlighting. In *International Conference on Learning Representations*, 2026.

Ruotong Liao, Max Erler, Huiyu Wang, Guangyao Zhai, Gengyuan Zhang, Yunpu Ma, and Volker Tresp. VideoINSTA: Zero-shot long video understanding via informative spatial-temporal reasoning with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6577–6602. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.384.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. DrVideo: Document retrieval based long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18936–18946, 2025. doi: 10.1109/CVPR52734.2025.01764.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.679.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46212–46244, 2023.

Junwen Pan, Qizhe Zhang, Rui Zhang, Ming Lu, Xin Wan, Yuan Zhang, Chang Liu, and Qi She. TimeSearch-R: Adaptive temporal search for long-form video understanding via self-verification reinforcement learning. In *International Conference on Learning Representations*, 2026.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14313–14323, 2024. doi: 10.1109/CVPR52733.2024.01357.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. LongVU: Spatiotemporal adaptive compression for long video-language understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 54582–54599. PMLR, 2025.

Yumeng Shi, Quanyu Long, and Wenya Wang. Static or dynamic: Towards query-adaptive token selection for video question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10759–10771, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.545. URL https://aclanthology.org/2025.emnlp-main.545/.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. MovieChat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18221–18232, 2024. doi: 10.1109/CVPR52733.2024.01725.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, and Jie Tang. LVBench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22958–22967, 2025a.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *Computer Vision – ECCV 2024*, pp. 58–76. Springer, 2024. doi: 10.1007/978-3-031-72989-8_4.

Zheng Wang, Haoran Chen, Haoxuan Qin, Zhipeng Wei, Tianwen Qian, and Cong Bai. Think, then verify: A hypothesis-verification multi-agent framework for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33784–33793, 2026.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3272–3283, 2025b.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In *Advances in Neural Information Processing Systems*, volume 37, pp. 28828–28857, 2024. doi: 10.52202/079017-0907.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9777–9786, 2021.

Zeyuan Yang, Delin Chen, Xueyang Yu, Maohao Shen, and Chuang Gan. VCA: Video curious agent for long video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20168–20179, 2025.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. WorldMM: Dynamic multimodal memory agent for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25599–25609, 2026.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9127–9134, 2019. doi: 10.1609/aaai.v33i01.33019127.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple LLM framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21715–21737. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.1209.

Zhuo Zhi, Qiangqiang Wu, Minghe Shen, Wenbo Li, Yinchuan Li, Kun Shao, and Kaiwen Zhou. VideoAgent2: Enhancing the llm-based agent system for long-form video understanding by uncertainty-aware cot. *arXiv preprint arXiv:2504.04471*, 2025.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025. doi: 10.1109/CVPR52734.2025.01278.

Yuanhao Zou, Shengji Jin, Andong Deng, Youpeng Zhao, Jun Wang, and Chen Chen. A.I.R.: Enabling adaptive, iterative, and reasoning-based frame selection for video question answering. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=SZVpOKw0YD.

A EOR Instantiation and Implementation Details

This appendix specifies how the reasoning state introduced in the main paper is instantiated and updated. The temporal event graph is used only to navigate the video; an event cue or graph edge never changes an obligation state by itself. All state changes require an observation produced from the corresponding source-video interval. Unless otherwise stated, decoding is deterministic and all configuration values are selected on the development split and then fixed for test evaluation.

Algorithm 1: Evidence-Obligation Reasoning (EOR)

Input: video V, question q, optional answer choices $\mathcal{C}$, source-video verification budget B
Output: grounded answer a^*

$G \leftarrow \mathrm{BuildEventGraph}(V)$
$\mathcal{H} \leftarrow \mathrm{InitializeHypotheses}(q, G, \mathcal{C})$
Decompose every $h_i \in \mathcal{H}$ into persistent obligations O_i.
Initialize every obligation as unresolved and evidence cache $\mathcal{B} \leftarrow \varnothing$.
Construct and cache obligation families.
while the verification budget is not exhausted **do**
  Update evidence-based scores and the plausible pool.
  $\mathcal{A} \leftarrow \mathrm{ECMAHS}(\mathcal{H})$.
  if the leader satisfies the grounding criterion **then**
    Jointly verify its decisive source-video evidence.
    if the evidence is globally consistent **then return** its answer.
    else reset only the implicated obligations to unresolved.
    end if
  end if
  if no unresolved obligation remains in $\mathcal{A}$ **then break**.
  Select the highest-priority unresolved obligation o^* in $\mathcal{A}$.
  Form $\phi(q, h, o^*)$ and retrieve candidate events from G.
  Apply relation-aware one-hop expansion and merge intervals.
  Verify the selected intervals in the source video.
  Update obligation states, family conflict state, and $\mathcal{B}$.
  Re-evaluate cached intervals for compatible obligations.
  if open-ended QA and evolution is triggered **then**
    $\mathcal{H} \leftarrow \mathrm{Evolve}(\mathcal{H}, \mathcal{B}, q)$.
    Preserve states only for unchanged obligations; initialize modified or new obligations as unresolved.
    Incrementally assign new obligations to cached families.
  end if
end while
return the live hypothesis with the largest evidence-based score.

Algorithm 1 gives the complete inference loop. The following subsections retain the numbering of the original appendix wherever the main paper refers to a specific subsection, in particular Appendices A.3–A.5.

A.1 Temporal Event Graph Construction

Event segmentation. Given a video V, we partition its timeline into localized events using an off-the-shelf shot boundary detector. Segments longer than 30 seconds are subdivided. When reliable boundaries are unavailable, we use fixed windows of 20 seconds with 5 seconds of overlap; the 25% overlap prevents actions from being cut at window boundaries. Every event retains its source timestamps and a pointer to the original video.

For event n_k, we sample 4 frames uniformly from its temporal span and use Qwen2.5-VL-7B to obtain a coarse visual cue. If aligned speech, subtitle, or OCR text is available in the input, it is appended to the cue; otherwise, the corresponding field is omitted. No absent modality is replaced with generated content. The resulting cue ℓ_k is encoded by a frozen text retrieval encoder (BGE-large-en-v1.5) to obtain the retrieval feature $x_k \in \mathbb{R}^{1024}$. The same encoder is used for obligation-conditioned queries (Eq. 3), obligation families (Appendix A.4), and entity matching below, so that all cosine similarities are computed in a single shared embedding space.

Graph edges. The event graph is $G = (\mathcal{N}, \mathcal{E})$. Temporal edges connect consecutive events. Entity-continuity edges connect events whose cues refer to the same entity: we extract entity spans (persons and salient objects) from each cue and link two events when their entity-span embeddings have cosine similarity at least 0.8 under the same retrieval encoder. Semantic edges connect each event to its $K_{\text{sem}} = 3$ nearest neighbors under cosine similarity, keeping the graph sparse so that one-hop expansion (Eq. 4, at most $\eta = 5$ additional nodes) stays within the per-round event budget $B_{\text{evt}} = 3$ (Appendix A.5). The graph is an index and navigation layer only; final evidence always comes from the source-video interval referenced by an activated node.

A.2 Hypothesis and Obligation Initialization

Multiple-choice QA. Each answer option initializes one hypothesis and all hypotheses use the same prior, $b_i = 0$. Consequently, their relative ordering is determined by verified evidence rather than answer-position priors.

Open-ended QA. Candidates are generated sequentially from the question and coarse event cues. Generation continues until at least $K_{\min} = 3$ candidates are present and no new answer direction is produced, or until $K_{\max} = 6$ candidates are reached. A candidate is considered a new direction only when its maximum similarity to existing candidates is below $\tau_{\text{ans}} = 0.75$. This realizes the adaptive initialization stated in the main paper and avoids fixing a single K_0 for every question.

Obligation decomposition. Each hypothesis is decomposed into $m_i \in [3, 5]$ observable conditions. An obligation must be directly checkable in the source video and must concern an event, state, entity, or temporal relation rather than merely restate the answer. Importance labels minor, major, and decisive are mapped to weights 1, 2, and 3, respectively. Every obligation is initialized as unresolved and is assigned one or more relation types from {temporal, entity-continuity, semantic}.

A.3 Exact Discriminative Value

For each condition c_{ij} and live hypothesis h_p, the relation $r(c_{ij}, h_p) \in \{\text{support}, \text{refute}, \text{neutral}\}$ is assigned once when the obligation is created. The label describes the logical effect that verifying the condition would have on the hypothesis; it does not claim that the condition is present in the video. Relations are cached and are computed only for new hypotheses introduced by evolution.

Let $\mathcal{H}_{\text{pl}}^{(t)} = \{h_1, \ldots, h_M\}$ be the plausible pool. The discriminative value used in the main-paper priority score is

$$u_{ij}^{(t)} = \begin{cases} \dfrac{2}{M(M-1)} \displaystyle\sum_{p<q} \mathbb{I}[r(c_{ij}, h_p) \neq r(c_{ij}, h_q)], & M \geq 2, \\ 0, & M < 2. \end{cases} \tag{9}$$

The cached relation labels are fixed, but $u_{ij}^{(t)}$ is recomputed because the plausible pool can change across iterations.

A.4 Obligation Families and Conflict State

Semantically equivalent obligations are grouped into families so that one underlying evidence requirement is not repeatedly sought. Conditions are embedded with the retrieval text encoder. During initialization, an obligation is assigned to the closest existing family when its similarity to the family centroid is at least $\tau_{\text{fam}} = 0.80$; otherwise, it starts a new family. Families are then cached rather than reconstructed at every iteration. Obligations created by hypothesis evolution are assigned incrementally using the same rule.

For family F, let $\mathcal{B}^{(t)}_{F,+}$ and $\mathcal{B}^{(t)}_{F,-}$ contain its reliable supporting and refuting observations. The conflict indicator is

$$w^{(t)}_{ij} = \mathbb{I}\Big[|\mathcal{B}^{(t)}_{F,+}| > 0 \ \wedge \ |\mathcal{B}^{(t)}_{F,-}| > 0\Big]. \tag{10}$$

Support-refute coexistence therefore remains an unresolved conflict; one observation never silently overwrites another.

A.5 Adaptive Obligation-Conditioned Retrieval

For selected obligation o_{ij}, the retrieval query is the ordered text triple

$$\begin{aligned}\phi(q, a_i, c_{ij}) = \texttt{Question:}\ \ q \,\|\, \texttt{Candidate:}\ \ a_i \\ \|\, \texttt{Required evidence:}\ \ c_{ij},\end{aligned} \tag{11}$$

where $\|$ denotes concatenation with field delimiters. Let $f^{(t)}_{ij}$ count completed retrieval rounds after which the obligation remained unresolved. Its quantile level is

$$\kappa^{(t)}_{ij} = \max\Big(\kappa_{\min}, \kappa_0 - \Delta\kappa\, f^{(t)}_{ij}\Big), \qquad \kappa_0 = 0.90, \quad \kappa_{\min} = 0.60, \quad \Delta\kappa = 0.10. \tag{12}$$

The activation threshold is the corresponding quantile of all event similarities. Events above it are ranked by similarity and at most $B_{\text{evt}} = 3$ are retained. If none passes the threshold, the highest-scoring event is retained. Repeated failure therefore broadens the search without degenerating into unrestricted full-video inspection.

A.6 Hyperparameters and Stopping Criteria

All thresholds below are kept fixed across both benchmarks and all controlled variants; no per-dataset tuning is performed. Values are summarized in Table 4.

Hypothesis scoring (Eq. 1). The hypothesis prior is uniform, $b_i = 0$ for all hypotheses, since multiple-choice options are exchangeable and open-ended candidates are generated without a ranking signal. Supporting and contradictory evidence are weighted by $\beta_+ = 1.0$ and $\beta_- = 1.5$, respectively: refuting a decisive obligation is penalized more strongly than confirming one is rewarded, which prevents a hypothesis with a large amount of weakly supportive evidence from outscoring a competitor that carries a single reliable contradiction. A hypothesis enters the plausible pool $H^{(t)}_{\text{pl}}$ when its score is within $\epsilon = 0.10$ of the current leader; given that $s^{(t)}_i \in [-\beta_-, \beta_+]$ under $b_i = 0$, this margin retains genuinely competitive hypotheses while excluding clearly dominated ones.

Obligation prioritization (Eq. 2). We set $\lambda_d = 1.0$ and $\lambda_c = 0.5$, so that $\pi^{(t)}_{ij} = \alpha_{ij}\big(1 + u^{(t)}_{ij} + 0.5\, w^{(t)}_{ij}\big)$ with importance weight $\alpha_{ij} \in \{1, 2, 3\}$ (Appendix A.2), discriminative value $u^{(t)}_{ij} \in [0, 1]$ (Appendix A.3), and binary family-conflict indicator $w^{(t)}_{ij} \in \{0, 1\}$ (Appendix A.4). An obligation counts as *high-priority* for family coverage in Eq. 5 when $\pi^{(t)}_{ij} \geq \tau_\pi$ with $\tau_\pi = 3.0$: every decisive obligation ($\alpha_{ij} = 3$) qualifies, a major obligation ($\alpha_{ij} = 2$) qualifies when $u^{(t)}_{ij} \geq 0.5$ or when supporting and refuting observations coexist in its family, and minor obligations never dominate the verification schedule on their own.

Active-set selection (Eq. 5–6). The greedy EC-MAHS approximation stops adding hypotheses once $\mathrm{Cov}_{\mathrm{obl}}(A) \geq \tau_{\mathrm{obl}} = 0.8$ and $\mathrm{Cov}_{\mathrm{dir}}(A) \geq \tau_{\mathrm{dir}} = 0.5$, the active set exceeds $K_{\mathrm{act}} = 3$, or no candidate yields positive coverage gain. Requiring answer-direction coverage is inexpensive—one member per direction suffices—and ensures that the verification budget is always spent on discriminating between competing answer directions rather than on redundant verification of the leading direction.

Graph expansion (Eq. 3–4). One-hop expansion along obligation-relevant edge types adds at most $\eta = 5$ nodes per retrieval round, keeping the candidate region compact relative to the per-round event budget B_{evt} (Appendix A.5).

Source-video verification. The verifier returns *support*, *refute*, or *uncertain* together with a confidence γ. Only predictions with $\gamma \geq \gamma_h = 0.7$ update obligation states; lower-confidence outputs leave the obligation unresolved. A cached source interval is reused for an obligation of another hypothesis only when the obligation embeddings satisfy $\mathrm{sim}(\psi(c_{ij}), \psi(c_{i'j'})) \geq \tau_{\mathrm{prop}}$ with $\tau_{\mathrm{prop}} = 0.8$, and the verifier always re-evaluates the cached interval against the new condition rather than inheriting the previous label.

Grounding criterion (Eq. 7). The leading hypothesis is considered sufficiently grounded when its satisfied coverage reaches $\tau_+ = 0.8$, its contradicted coverage stays below $\tau_- = 0.2$, the accumulated satisfied evidence satisfies $\sum_j \alpha_{\mathrm{top},j}\, \mathbb{I}[z^{(t)}(o_{\mathrm{top},j}) = \textit{satisfied}] \geq \nu$ with $\nu = 4$ (roughly half of the total obligation weight of a typical hypothesis with $m_i \in [3, 5]$ obligations of weights $\{1, 2, 3\}$, or equivalently two decisive obligations), and its score exceeds every other live hypothesis by at least $\delta = 0.15$. When the criterion is not met, multiple-choice QA terminates after $r = 3$ stagnant iterations or upon budget exhaustion; for open-ended QA the same $r = 3$ consecutive failures trigger hypothesis evolution (Eq. 8). The choice $r = 3$ matches the relaxation schedule in Appendix A.5, under which the retrieval quantile reaches its floor $\kappa_{\min}$ after exactly three unsuccessful rounds, so stagnation is declared only after the search range has been fully expanded.

Table 4: Hyperparameters used in all experiments. Equation numbers refer to the main text.

Symbol	Value	Role
β_+	1.0	weight of satisfied coverage in $s_i^{(t)}$ (Eq. 1)
β_-	1.5	weight of contradicted coverage in $s_i^{(t)}$ (Eq. 1)
b_i	0	uniform hypothesis prior (Eq. 1)
ϵ	0.10	plausible-pool score margin (Sec. 3.3)
λ_d	1.0	weight of discriminative value $u_{ij}^{(t)}$ (Eq. 2)
λ_c	0.5	weight of family-conflict indicator $w_{ij}^{(t)}$ (Eq. 2)
τ_π	3.0	high-priority obligation threshold (Eq. 5)
τ_{obl}	0.8	obligation-family coverage threshold (Eq. 6)
τ_{dir}	0.5	answer-direction coverage threshold (Eq. 6)
K_{act}	3	active-set budget (Sec. 3.3)
η	5	max. one-hop expansion nodes (Eq. 4)
γ_h	0.7	verifier reliability threshold (Sec. 3.2)
τ_{prop}	0.8	cross-hypothesis evidence-reuse similarity (Sec. 3.2)
τ_+	0.8	min. satisfied coverage for grounding (Eq. 7)
τ_-	0.2	max. contradicted coverage for grounding (Eq. 7)
ν	4	min. weighted satisfied evidence (Eq. 7)
δ	0.15	min. score margin over runner-up (Eq. 7)
r	3	stagnant iterations before stop/evolution (Sec. 3.3–3.4)

B Prompts and Output Schemas

The templates below are fixed across examples. Braced fields are populated at runtime. Each prompt requires machine-readable JSON; a response that cannot be parsed is retried once with the same

content and a schema-only repair instruction. The repair step may change formatting but may not add evidence.

B.1 Initial Open-Ended Hypothesis Generation

```
Question: {question}
Coarse event cues: {event_cues}
Existing candidates: {candidate_list}

Generate one candidate answer that introduces a semantically distinct
answer direction. Do not repeat or paraphrase an existing candidate.
Treat the candidate as unverified.

Return JSON:
{"candidate": str, "rationale": str, "is_new_direction": bool}
```

B.2 Evidence-Obligation Decomposition

```
Question: {question}
Candidate answer: {candidate}

Decompose the candidate into independently observable conditions that
must be checked before the answer can be accepted. Each condition must
be verifiable in the source video and must not merely restate the answer.

For each obligation return JSON fields:
{"condition": str,
 "importance": "minor" | "major" | "decisive",
 "relation_types": ["temporal" | "entity-continuity" | "semantic"]}
```

B.3 Condition-Hypothesis Relation Assignment

```
Condition: {condition}
Hypothesis: {hypothesis}

If the condition were verified, what logical effect would it have on
the hypothesis? Do not decide whether the video contains the condition.

Return JSON:
{"relation": "support" | "refute" | "neutral"}
```

B.4 Source-Video Verification

```
Question: {question}
Observable condition: {condition}
Source-video interval: {video_interval}

Inspect only the supplied interval. Report explicit support, explicit
counter-evidence, or insufficient evidence. Absence of visible support
is not counter-evidence. Confidence is a reliability score, not a
probability that the final answer is correct.

Return JSON:
{"observation": str,
 "role": "support" | "refute" | "uncertain",
 "confidence": float,
 "time_span": [float, float]}
```

B.5 HYPOTHESIS EVOLUTION

```
Question: {question}
Current hypotheses and obligation states: {hypothesis_state}
Reliable source-grounded observations: {verified_evidence}

Repair the candidate space only when the verified evidence cannot be
adequately explained by the current hypotheses. Allowed operations are
retain, revise, merge, archive, and introduce. Every change must cite
the observation identifiers that motivate it. Do not mark a generated
candidate as verified.

Return JSON:
{"operations": [{"type": str, "inputs": [str], "candidate": str,
                 "evidence_ids": [str]}]}
```

B.6 FINAL GROUNDING CHECK

```
Question: {question}
Leading hypothesis: {hypothesis}
Decisive source-video evidence: {evidence_intervals}

Judge whether the supplied evidence jointly supports the hypothesis
without temporal, entity, or support-refute inconsistency. Identify the
obligations implicated by any inconsistency.

Return JSON:
{"grounded": bool, "reason": str,
 "inconsistent_obligation_ids": [str]}
```

B.7 ACTIVITYNET-QA EVALUATION

ActivityNet-QA predictions are evaluated by DeepSeek-V3 with temperature 0. The correctness field produces the reported accuracy and the integer rating produces the reported mean QA score.

```
Question: {question}
Reference answer: {reference_answer}
Predicted answer: {predicted_answer}

Evaluate semantic correctness, allowing concise paraphrases but not
unsupported extra claims. Return a binary correctness label and an
integer quality score from 0 to 5, where 0 is fully incorrect and 5 is
fully correct and complete.

Return JSON:
{"correct": 0 | 1, "score": 0 | 1 | 2 | 3 | 4 | 5,
 "reason": str}
```