RETHINKING HALLUCINATION IN LONG-VIDEO UNDERSTANDING: A UNIFIED BINDING PERSPECTIVE AND A BENCHMARK

Anonymous authors
Paper under double-blind review

ABSTRACT

Hallucination is a major bottleneck in reliable long-video understanding. Unlike short clips, long videos involve expanded context windows, dense temporal dynamics, and complex multi-entity narratives. These factors make long-video hallucination far more complex and diverse. Existing methods tackle this challenge in a piecemeal fashion, relying on task-specific heuristics, such as frame selection or evidence search, to mitigate individual hallucination types. We **rethink** long-video hallucination, arguing that these diverse hallucination stem from a single root cause: incorrect multimodal binding. We establish a unified perspective and a comprehensive taxonomy dividing atomic errors into two families: out-video hallucination (binding to non-existent entities) and in-video hallucination (misbinding valid elements to wrong entities, roles, states, or timing). To address this, we introduce a temporal hypergraph framework for *correct-by-construction* understanding without task-level training, using n-ary relations and temporal validity to eliminate fabrication and misattribution. We further build a taxonomy-labeled benchmark for long video hallucination that not only measures the overall hallucination level of existing models and frameworks, but also diagnoses the source and the bias of hallucination, pointing out directions for improvement. Across many benchmarks, our approach yields consistent gains over strong baselines. This work offers a unified paradigm to tackle long video hallucination, moving beyond fragmented, task-specific heuristics. Code and data are available at https://anonymous.4open.science/r/rrt_echo-874C/

1 INTRODUCTION

Video is the richest modality for capturing the real world. Driven by the progress in vision-language models (VLMs), video understanding has emerged as a crucial capability for general-purpose multimodal intelligence. Whether for embodied agents or video-grounded reasoning, success in these applications requires more than just frame-level perception. A model must not only perceive individual frames, but also understand and reason over the dynamic relations among entities, actions, and events across long temporal spans.

As videos grow from a few seconds to minutes or even hours, however, the difficulty of understanding changes in kind rather than in degree. Short videos typically contain few entities and events, involve limited scene transitions, and require little long-range binding of evidence; a model can therefore extract the correct semantics relatively easily. Long videos introduce several intrinsic challenges(Liu et al., 2024). First, fine-grained evidence is abundant but often transient, making it easy to miss at the perception stage. Second, event-level semantics are rich and relationally complex(Asgarov et al., 2026). While film-level narratives demand reasoning over many interconnected events, models frequently fail to disentangle them (Lu et al., 2026). Third, recurring entities undergo appearance and state transitions across scenes, severely complicating identity tracking and the maintenance of contextual consistency (Ha et al., 2026). These properties make hallucination in long-video understanding not only more frequent but also more diverse and harder to detect.

Historically, the most widely studied form of hallucination is *fabrication*, defined as asserting entities, actions, or events that never appear in the video (Huang et al., 2026). Driven by recent advance-

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.

ments in vision-language models (VLMs), such fabrications have been substantially mitigated. Yet a harder and more subtle class persists in long videos and their downstream reasoning: models often correctly observe what occurs, but misattribute its relations, temporal alignment, states, or attributes. For example, a model might assign an action to another event, swap the roles of two people, or return pre-transition states. Such errors occur even when perception is fully accurate. They arise from incorrect binding during event aggregation and directly lead to downstream reasoning errors. Recent studies (Ma et al., 2026; Wang et al., 2025; Long et al., 2026; Yeo et al., 2026; Yan et al., 2026) have made meaningful progress through customized mechanisms like denser sampling, multi-granularity memory, or agentic retrieval. While these methods demonstrate promise within specialized domains, they often target a particular type of hallucination and may require dedicated post-training. As a result, existing approaches treat these failure modes primarily as localized phenomena, prompting us to examine whether a unified framework could address them from a single underlying cause.

In this work, we revisit long-video hallucination from first principles and identify a single root cause: **a failure of multimodal binding**. Rather than treating fabrication and misattribution as distinct phenomena, we unify them along a single axis based on referent validity. Specifically, (a) *phantom binding* (out-video) attaches a description to a referent that *never existed*, while (b) *mis-binding* (in-video) attaches validly observed content to the *wrong* referent, such as an incorrect entity, event, role, state, or timestamp. This unified perspective naturally consolidates a fragmented problem space into a coherent framework. To address the root cause directly, we represent long videos using a structure inherently designed for multimodal binding: a *temporal hypergraph*. As the video unfolds, the hypergraph registers cross-scene identities and stores n-ary relations with temporal validity. By construction, querying this graph suppresses both hallucination types. The model cannot fabricate facts, as no supporting hyperedge exists; nor can it misattribute valid facts, as scoped graph readoffs isolate correct contexts. Crucially, this construction is task-training-free and model-agnostic, making it directly applicable to any VLM/MLLM frameworks.

To operationalize this binding framework, we introduce a complete taxonomy of long-video hallucinations. At the top level, it distinguishes phantom binding from mis-binding, and subsequently refines the latter into atomic failure modes across event membership, relation roles, entity identity, and temporal binding. Based on this taxonomy, we construct a diagnostic evaluation dataset. Beyond reporting overall accuracy, it systematically pinpoints fine-grained hallucination types and evaluates model-specific error preferences. This reveals that different models exhibit distinct susceptibilities across error categories, turning a single accuracy score into actionable diagnostics for targeted model improvement. Furthermore, we design injection stress tests by inserting irrelevant clips into long videos to evaluate robustness against distractions.

In summary, our contributions are threefold:

- **A unifying perspective and complete taxonomy.** We unify out-video and in-video hallucinations under a single multimodal binding framework down to atomic failure modes, offering a principled characterization of long-video hallucination.
- **A training-free and model-agnostic method.** We represent long videos as temporal hypergraphs that suppress both hallucination types by construction. Operating as a plug-in memory module, our approach ECHO requires no task-level training and seamlessly adapts to arbitrary models. ECHO yields consistent gains over advanced VLMs and SOTA long-video frameworks across multiple benchmarks. Specifically, it surpasses the strongest baselines by 11.86 percentage points on BindScope and 5.57 points on ELV-Halluc. As shown in Figure 5 and Table 2.
- **A taxonomy-labeled diagnostic benchmark.** We introduce BindScope, a human-labeled long-video hallucination dataset that pinpoints fine-grained error types and diagnoses model-specific binding error biases.

2 Related Works

2.1 Long-video Memory and Retrieval

Existing methods organize long-video evidence through entity-centric memories, event structures, and multimodal retrieval Long et al. (2026); Yeo et al. (2026); Ma et al. (2026); Yin et al. (2026); Wang et al. (2025); Yan et al. (2026); Li et al. (2026b), with further advances in compression and

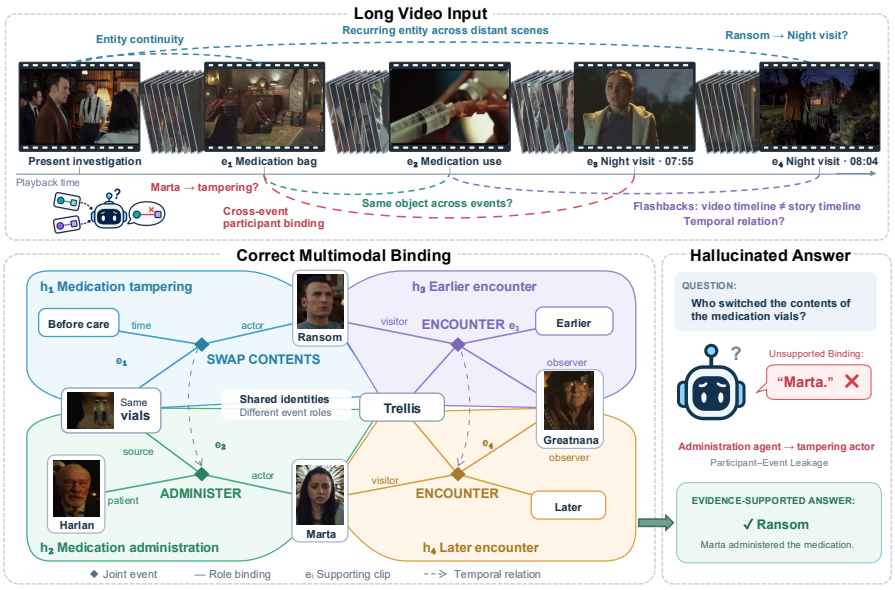

Figure 1: **A binding perspective on long-video hallucination.** In *Knives Out*, recurring entities and flashbacks complicate cross-event bindings. Misbinding Marta's administration role to the tampering event yields an unsupported answer, despite the individual elements being present in the video. Evidence-linked temporal hyperedges jointly bind participants, roles, objects, and times.

evidence recoverability Wang et al. (2026); Li et al. (2026a). Such representations provide grounding that can help curb fabrication of unobserved content. However, retaining valid evidence does not ensure that its joint bindings remain intact. The observed entities, actions, and attributes can still be mis-binded across incompatible events, roles, or times, producing in-video hallucinations. ECHO targets this distinction by explicitly preserving n-ary event bindings and their temporal scope throughout memory construction and retrieval.

2.2 Long-video Hallucination benchmarks

Long-video benchmarks evaluate complementary aspects of extended video understanding, including general comprehension, temporal reasoning, and narrative event structure (Fu et al., 2025; Asgarov et al., 2026). Hallucination-focused benchmarks further isolate temporal inconsistency, narrative-prior errors, semantic aggregation hallucination, and failures under multimodal perturbations (Li et al., 2025; Lee et al., 2025; Lu et al., 2026; Dong et al., 2026). These benchmarks reveal important failure modes, but they typically organize errors by observable content type, task setting, or perturbation, leaving the corrupted relation and its originating memory stage underspecified. We instead formulate long-video hallucination as binding failure: *phantom* binding links claims to nonexistent referents, whereas *mis-binding* assigns observed elements to incorrect entities, events, roles, states, or times. Our benchmark labels these atomic binding failures, traces them across memory construction, consolidation, and retrieval.

3 Method

As illustrated in Figure2, we present Evidence-Constrained Hypergraph Organization (**ECHO**), a framework designed to preserve evidence-grounded bindings in long-video memory. We adopt a *temporal hypergraph* as our memory structure because its n-ary hyperedges naturally model complex, multi-entity event structures alongside their temporal validity. Instead of storing observed elements in isolation, ECHO explicitly tracks entity participation, semantic roles, and the temporal scope of attributes and relations. The framework consists of three core components: (1) multimodal recognition-registration-tracing (RRT); (2) ECHO framework for temporal hypergraph construction,

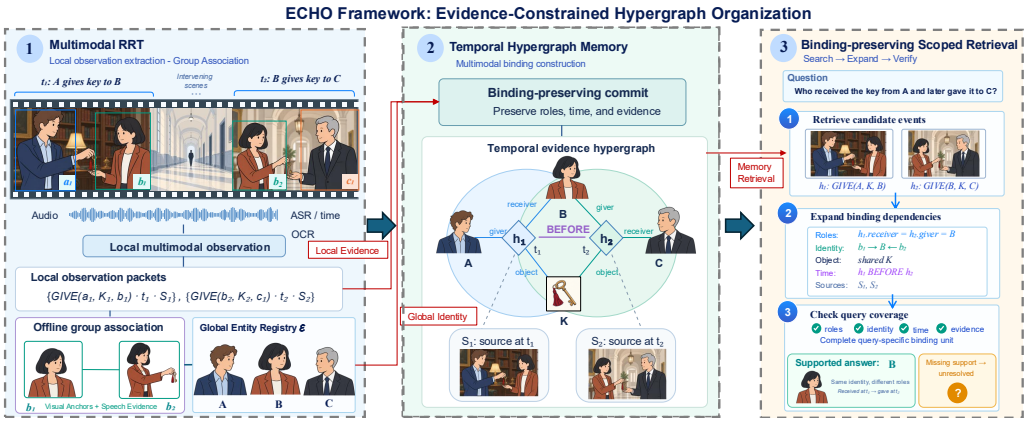

Figure 2: **Overview of ECHO Framework.** (a) Multimodal RRT extracts grounded local observations and proposes revisable identity links. (b) Temporal hypergraphs is organized to store shared entities alongside event-specific roles, temporal scope, and provenance. (c) Binding-preserving scoped retrieval expands candidate events into query-specific binding units while verifying evidence coverage. Example: Entity B acts as receiver in the first event and giver in the second; unverified relations remain unresolved.

and (3) binding-preserving scoped retrieval. Together, these components preserve robust multimodal bindings across complex, multi-scene long video queries.

3.1 Multimodal Recognition-Registration-Tracing (RRT)

To establish grounded observation endpoints without fabricating non-existent events, Multimodal Recognition-Registration-Tracing (RRT) processes video windows through three unified operations. First, *recognition* extracts local instances, predicates, semantic roles, and visual evidence, leaving ambiguous endpoints unresolved rather than making unverified guesses. Second, *registration* decouples local instances $\mathcal{X}$ from global entity commitments $\mathcal{E}$ via a projection:

$$\pi^{(v)} : \mathcal{X} \to \mathcal{E} \cup \{\perp\}, \tag{1}$$

where an unresolved mapping $\perp$ preserves local event structures while withholding unverified global identities. Third, *tracing* maintains revisable candidate links with explicit statuses (accepted, pending, or rejected) and isolates spoken entities from visual participants until explicitly grounded. See **Appendix A.1** for detailed RRT pipeline and implementation rules.

3.2 Evidence-Constrained Hypergraph Organization (ECHO)

ECHO organizes grounded observations into a typed temporal hypergraph $\mathcal{G}^{(v)}$ to preserve multi-entity relationships and temporal validity. An event occurrence is represented as a role-labeled hyperedge:

$$h_e = (\iota_e, p_e, \{(\rho, x_{e,\rho})\}_{\rho \in \mathcal{R}_e}, \tau_e, \mathcal{Z}_e)\,, \tag{2}$$

which binds the predicate p_e, semantic roles $\{(\rho, x_{e,\rho})\}$, time interval τ_e, and evidence $\mathcal{Z}_e$ into a single unit, eliminating predicate leakage and role permutation. Similarly, attribute and state observations are stored as scoped hyperedges:

$$h_a = (x_{\text{owner}}, a, u, \tau_a, \mathcal{Z}_a)\,, \tag{3}$$

where the value u applies strictly to the supported temporal scope τ_a without overwriting earlier states or assuming persistence through unobserved periods. The full pipeline combining RRT and hypergraph synthesis is outlined in Algorithm 1. See **Appendix A.2** for complete graph schema definitions and edge creation protocols.

Algorithm 1 ECHO Framework for Memory Construction

Require: Video V; local multimodal observer f
Ensure: Global entity registry $\mathcal{E}$; evidence-linked temporal hypergraph $\mathcal{G}^{(v)}$

$\mathcal{O} \leftarrow \emptyset$ ▷ Immutable local observations
$\mathcal{E}, \mathcal{G}^{(v)}, \mathcal{U} \leftarrow \emptyset$ ▷ Entities, hypergraph, unresolved links
Stage I: Recognition (Local multimodal observation)
for each clip $c_t \in \textsc{Segment}(V)$ **do**
  $z_t \leftarrow \textsc{AlignEvidence}(c_t)$ ▷ Frames, audio, time, and source references
  $O_t \leftarrow f(z_t)$ ▷ Local instances, utterances, events, and attributes
  $O_t \leftarrow \textsc{ValidateLocalBindings}(O_t, z_t)$ ▷ Roles and evidence remain within clip scope
  $\mathcal{O} \leftarrow \mathcal{O} \cup O_t$
end for
Stage II: Registration & Tracing (Offline global association)
$\mathcal{A} \leftarrow \textsc{SelectVisualAnchors}(\mathcal{O})$ ▷ Instance-specific, multi-view visual anchors
$(\mathcal{E}, \pi, \mathcal{U}) \leftarrow \textsc{AssociateGroups}(\mathcal{O}, \mathcal{A})$ ▷ π: local instance $\rightarrow$ global entity
$\nu \leftarrow \textsc{ResolveNamedReferences}(\mathcal{O}, \pi)$ ▷ Name mention $\rightarrow$ event role $\rightarrow$ entity
Stage III: Temporal Hypergraph Construction
for each local relation $r \in \mathcal{O}$ **do**
  $h \leftarrow \textsc{BindLocal}(r)$ ▷ Predicate, local role fillers, time, and evidence
  if ValidateBinding(h) **then**
    $\mathcal{G}^{(v)} \leftarrow \textsc{Commit}(\mathcal{G}^{(v)}, h)$
  else
    $\mathcal{U} \leftarrow \mathcal{U} \cup \{h\}$
  end if
end for
$\mathcal{G}^{(v)} \leftarrow \textsc{AttachIdentityLinks}(\mathcal{G}^{(v)}, \pi, \mathcal{U})$
$\mathcal{G}^{(v)} \leftarrow \textsc{AttachNamedReferences}(\mathcal{G}^{(v)}, \nu)$
return $\mathcal{E}, \mathcal{G}^{(v)}$

3.3 Binding-Preserving Scoped Retrieval

To deliver verifiable answers in long-horizon reasoning, ECHO retrieves and audits memory through a four-stage scope-preserving pipeline. Given a query q, candidate facts are first retrieved under explicit temporal constraints and accepted identity links. Next, each selected fact f is expanded into a self-contained binding unit:

$$U_q(f) = \text{Closure}_{q,\mathcal{G}^{(v)}}(\{f\}), \tag{4}$$

which collects its predicate, role endpoints, visual evidence, and required dependencies. Retrieval constructs a payload $P_q = \bigcup_{U \in \mathcal{S}_q} U$ under context budget B ($\text{cost}(P_q) \leq B$), where units exceeding B are omitted entirely rather than truncated across roles or dependencies. Furthermore, answer generation is decoupled from support assessment, evaluating candidate claims c via a scope-preserving mapping:

$$\text{Support}(c, P_q) = 1 \iff \exists\, \phi : c \hookrightarrow P_q. \tag{5}$$

Matching identical entities in unrelated events fails this condition. Assuming accurate memory stores, this dependency-closed mechanism guarantees that any invalid role or entity substitution within the queried scope cannot be certified as supported. See **Appendix A.3** for detailed search protocols, closure expansion algorithms, and claim auditing rules.

4 BindScope Benchmark

We introduce **BindScope**, a human-annotated diagnostic benchmark evaluating multimodal binding errors in long videos. Spanning eight fine-grained error types (R0 to R7; Table 1), BindScope contains 1,249 paired units and 312 unpaired questions, totaling 2,810 questions across 77 long videos (1,441 minutes). Rather than evaluating isolated element recognition, BindScope verifies whether observed entities, actions, attributes, and events are correctly assigned to their ground-truth semantic and temporal scopes. Through a task design incorporating mixed question formats and paired-unit evaluations, BindScope precisely diagnoses model hallucination preferences while mitigating shortcut learning and random guessing (hacking). Figure 3 summarizes the dataset statistics.

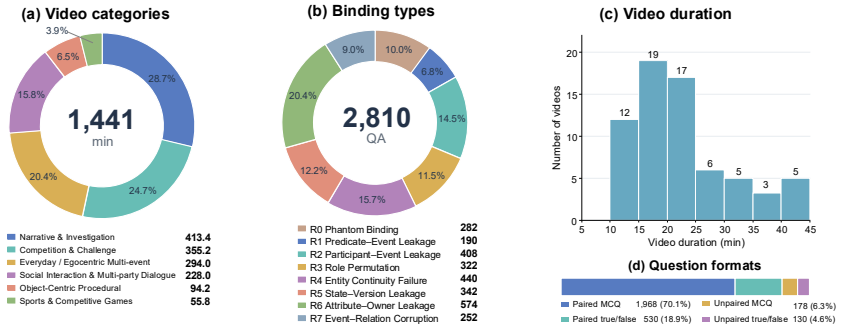

Figure 3: **Statistics of BindScope.** (a) Video categories with measured duration (minutes). (b) Binding error type distribution across 2,810 questions. (c) Video duration distribution (d) Question-format distribution.

4.1 Binding-Centered Task Design

To comprehensively evaluate long-video understanding, BindScope generates questions based on our defined binding error taxonomy, spanning varying difficulty levels and spatial-temporal scopes (e.g., intra-scene single-event binding versus cross-scene entity tracking). Each question contrasts a video-grounded assignment with a controlled alternative that alters a targeted binding while preserving all observed constituents, strictly measuring support for joint assignments rather than the mere presence of individual elements.

Tasks are hierarchically structured into single-event binding, cross-event assignment, and cross-event connection to disentangle localized perception errors (e.g. R6 attribute-owner misbinding) from long-range reasoning (e.g. R1 predicate-event misbinding), as shown in4. See **Appendix B.1** for complete taxonomies and task design protocols.

4.2 Pairwise QA Mechanism

During preliminary evaluations, we observed that modern VLMs frequently bypass genuine video comprehension by exploiting rich pre-training priors (e.g., recalling plot details of well-known movies like *Sherlock Holmes*) or structural shortcuts (e.g., guessing answers based on option length or descriptive features). To accurately diagnose true hallucination preferences and prevent such hacking, we introduce a rigorous pairwise QA mechanism. Each evaluation unit comprises paired questions, $p = (q_{p,1}, q_{p,2})$, which probe the identical underlying binding relation from distinct perspectives using controlled distractors. A model receives credit only if it correctly answers both questions. This strict joint-success requirement fundamentally neutralizes random guessing and prior-based exploits.

Pairwise QA are evaluated independently in isolated contexts without access to paired questions or metadata. We report Question Accuracy (QAcc) and Paired Accuracy (PairAcc):

$$\mathrm{PairAcc} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \prod_{j=1}^{2} \mathbb{1}[\hat{y}_{p,j} = y_{p,j}], \tag{6}$$

4.3 Dataset Construction and Quality Control

BindScope is constructed through a strict multi-stage pipeline. The process begins with automated candidate generation, followed by answer-blind model auditing to verify evidence sufficiency and pair non-redundancy. Finally, human annotators verify every candidate QA one by one to eliminate residual linguistic, visual, and world-knowledge shortcuts, thereby guaranteeing the highest QA quality. See **Appendix B.2** for pipeline specifications and audit protocols.

Table 1: **A binding failure taxonomy of long-video hallucination.**

Family	Type	Definition and example
Out-of-video hallucination: Phantom binding		
Evidence grounding	**R0** Phantom Binding	A binding involves an element absent from the video. *Example:* Alice holds a cup; no phone appears ⇒ Alice holds a **phone**.
In-video hallucination: Misbinding		
Event membership	**R1** Event–Action Misbinding	An action is misbinded to wrong event. *Example:* Alice washes a cup; Bob drops a plate ⇒ Alice **drops** the cup.
	R2 Event–Entity Misbinding	A Entity is misbinded to wrong event. *Example:* Alice cuts a cake; Bob waters flowers ⇒ **Bob** cuts the cake.
Relation roles	**R3** Role Permutation	Correct participants but swapped event roles. *Example:* Alice gives Bob a key ⇒ **Bob gives Alice** the key.
	R6 Entity–Attribute Misbinding	An observed attribute is assigned to the wrong entity. *Example:* Alice wears red; Bob wears blue ⇒ Alice wears **blue**.
Entity identity	**R4** Entity Continuity Failure	Incorrectly splitting or merging identities. *Example:* Alice removes her coat ⇒ The coatless person is **someone else**.
Temporal binding	**R5** State–Timing Misbinding	An observed state is assigned to the wrong time. *Example:* A cup is full at t_1, empty at t_2 ⇒ It is **empty at** t_1.
	R7 Event–Relation Misbinding	Correct events receive an incorrect temporal or causal link. *Example:* Alice enters; then Bob leaves ⇒ Bob leaves **before** Alice enters.

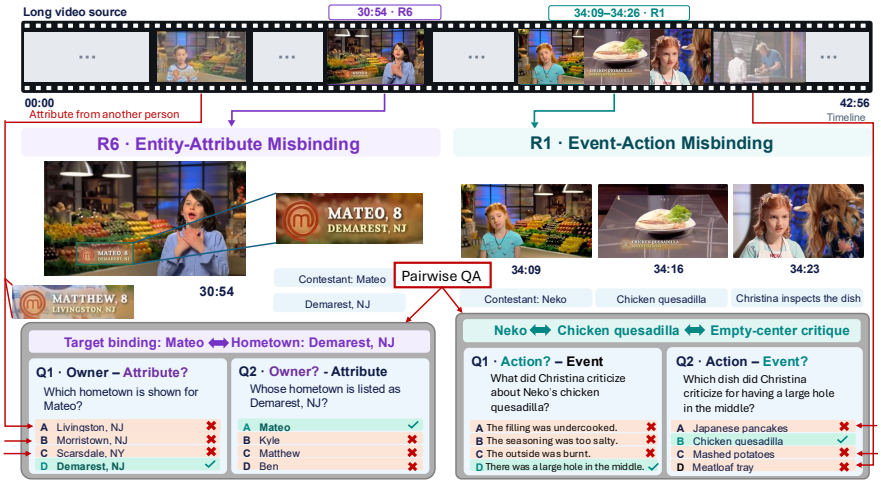

Figure 4: **BindScope QA examples and Pairwise QA Machanism.** This video clip illustrates paired questions targeting R1 and R6 binding hallucinations. Each binding relation is test by a paired QA and both questions must be answered correctly to pass the unit.

5 Experiments

To validate the diagnostic capabilities of BindScope Benchmark and the superiority of **ECHO** in long-video hallucinations elimination, we conduct extensive experiments across two specialized long-video hallucination benchmarks (BindScope and ELV-Halluc). The evaluated baselines include two primary groups:

- **Foundation VLMs:** We evaluated most common VLM models via direct question answering over uniformly sampled video frames, including Qwen3-VL-8B, Qwen3-VL-30B-A3B, Qwen3.8-Omni-Flash, HY-Vision-2.0, and GPT-5.4.
- **Advanced Long-Video Understandig Systems:** We evaluated state-of-the-art RAG-based systems (Caption RAG, Event-Causal RAG (Yan et al., 2026), and MAGIC-Video (Li et al., 2026b)) and Memory-based system on long-video hallucination level (WorldMM (Yeo et al., 2026), M3-

Agent (Long et al., 2026), and ReflectWorld-MM (Ma et al., 2026)). Notably, to ensure a fair comparison, all reproduced long-video frameworks share **Qwen3-VL-8B** as their underlying foundation VLM.

Specifically, our evaluation addresses three core research questions: **RQ1 (Diagnostic Profiling):** What fine-grained binding hallucination patterns does BindScope reveal across existing VLMs and long-video frameworks? **RQ2 (In-Domain Effectiveness):** Does ECHO outperform VLM, RAG-based, and memory-based baselines across all binding categories? **RQ3 (Cross-Benchmark Generalization):** How effectively does ECHO generalize to outperform competitive baselines on external long-video hallucination benchmarks?

5.1 Diagnostic Profiling (RQ1)

We evaluate foundation models and advanced long-video frameworks on our BindScope benchmark, with overall and fine-grained results visualized in Figure 5 (detailed numerical tables are provided in Appendix C). First, long-video hallucination remains a severe bottleneck across existing architectures, persisting heavily in both in-video (mis-binding) and out-video (phantom binding) regimes. Moreover, as shown in Figure 5, even when different models and frameworks achieve comparable overall accuracy, they exhibit markedly distinct hallucination preferences across binding error categories. We have There are many interesting and insightful findings, demonstrating BindScope's diagnostic capacity.

Aggregate Accuracy Masks Binding Type Preferences. Qwen3-VL-32B and HY-Vision-2.0 achieve virtually identical overall accuracies (17.5% and 17.7%, respectively), yet exhibit opposing category-level strengths. HY-Vision-2.0 dominates on R1 (26.4% vs. 19.8%), whereas Qwen3-VL-32B excels at event (relation corruption, R7; 29.9% vs. 21.5%). Similarly, while Qwen3.8-Omni-Flash leads overall foundation models, GPT-5.4 demonstrates superior performance on participant (event binding, R2; 17.5% vs. 13.9%) and role permutation (R3; 28.7% vs. 25.0%). These reversals confirm that aggregate accuracy scores conceal significant category-specific trade-offs and error preferences.

Entity Continuity Decouples from Relational Binding Capabilities. M3-Agent reaches 32.1% on entity continuity (R4), outperforming Caption RAG (24.7%) and the best foundation model (20.2%), aligning with its persistent face and voice identity modeling. While M3-Agent and MAGIC-Video achieve comparable R4 scores (32.1% vs. 31.8%), their performance diverges sharply on role permutation (R3; 27.8% vs. 17.9%) and event relations (R7; 32.6% vs. 18.5%). Likewise, Event-Causal RAG leads on R4 (37.13%) and R3 (37.91%), yet trails MAGIC-Video on R1 (25.25% vs. 32.1%). These orthogonal profiles demonstrate that success in tracking identity across scenes does not translate into robust temporal or relational binding.

Participant Assignment Presents a Universal Bottleneck Resistant to Model Scaling. Event–entity binding (R2) consistently emerges as the weakest link across all evaluated baselines, with foundation models scoring between 12.2% and 17.5%, and long-video frameworks ranging from 16.7% to 28.07%. Crucially, parameter scaling fails to resolve this bottleneck: scaling from Qwen3-VL-8B to Qwen3-VL-32B yields overall gains, yet provides no improvement on R3 (both 17.1%) and even slightly degrades R2 accuracy (12.2% vs. 12.8%). This persistent fragility highlights that event-level participant and role binding require dedicated architectural designs rather than naive capacity scaling.

5.2 In-domain Effectiveness of ECHO (RQ2)

Figure 5 demonstrates that ECHO achieves superior accuracy across both out-video ($R0$) and in-video ($R1$ to $R7$) hallucination categories. This consistent advantage confirms that long-video hallucinations can be effectively addressed through **a unified binding perspective**. Overall, ECHO outperforms direct video models as well as RAG- and memory-based frameworks. Compared to the strongest baseline in each category, ECHO boosts accuracy by 9.0 to 24.0 percentage points, demonstrating gains across diverse hallucination types rather than a single error category.

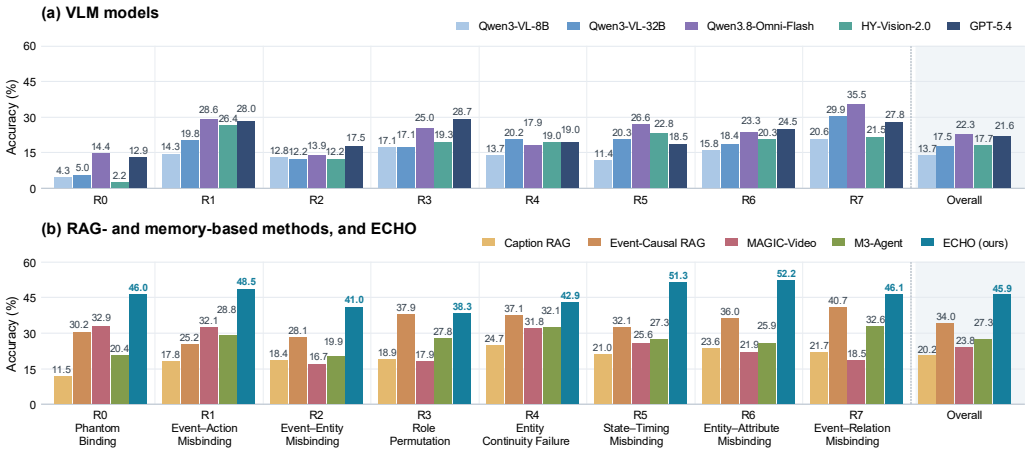

Figure 5: **Binding hallucination evaluation on BindScope.** Overall and type-specific Accuracy (%) across 8 binding categories for (a) VLM fundation models and (b) RAG- and memory-based frameworks, including ECHO.

Table 2: Cross-benchmark evaluation on ELV-Halluc. We report paired accuracy (%) for in-video (In.) and out-of-video (Out.) hallucinations across four content categories. InAcc and OutAcc denote overall accuracy in each setting; Avg. is their unweighted mean. Higher is better.

Model/Method	Visual In. / Out.	Object In. / Out.	Action In. / Out.	Declarative In. / Out.	InAcc.	OutAcc.	AvgAcc.
baseline							
Qwen3-VL-8B	12.56 / 19.23	20.00 / 24.10	17.69 / 21.54	18.46 / 17.18	17.18	20.51	18.85
RAG-based methods							
EC-RAG	11.79 / 15.90	14.62 / 17.44	13.33 / 16.15	11.28 / 13.08	12.76	15.64	14.20
MAGIC-Video	8.72 / 12.82	13.33 / 16.15	15.38 / 15.64	12.05 / 12.31	12.37	14.23	13.30
Memory-based methods							
ReflectWorld-MM	7.55 / 6.60	7.55 / 8.49	7.55 / 6.60	8.49 / 11.32	7.78	8.25	8.01
M3-Agent	17.69 / 21.28	22.56 / 23.08	20.26 / 20.77	**21.79 / 21.79**	20.58	21.73	21.16
WorldMM	14.62 / 19.23	17.95 / 20.77	17.18 / 18.72	15.38 / 15.90	16.28	18.65	17.47
ECHO (Ours)	**19.78 / 33.33**	**28.71 / 39.39**	**21.43 / 33.68**	18.27 / 20.00	**22.08**	**31.36**	**26.72**

The most notable improvements occur in state-timing binding ($R5$) and entity-attribute binding ($R6$). On these categories, ECHO reaches 51.3% and 52.2% accuracy, exceeding the top-performing baselines by 19.23 and 16.16 percentage points, respectively. Additionally, ECHO yields substantial gains on event-action and event-entity binding, outperforming baselines by 16.40 and 12.93 percentage points. These results directly support our formulation: Generating verifiable answers requires strictly preserving event membership, attribute ownership, and temporal validity throughout the video.

5.3 Cross-benchmark Generalization(RQ3)

External validation on ELV-Halluc. To examine whether ECHO's benefits extend beyond BindScope, we evaluate it on ELV-Halluc, which tests in-video and out-of-video hallucinations across four content categories. As shown in Table 2, ECHO achieves 22.08% InAcc and 31.36% OutAcc, yielding an average accuracy of 26.72%. This exceeds M3-Agent, the strongest baseline by average accuracy, by 5.57 percentage points.

ECHO achieves the highest accuracy on visual details, objects, and actions in both settings, with particularly large gains on out-of-video hallucinations. However, improvements are not uniform: the gain in InAcc is smaller (+1.50 points), and declarative content remains below M3-Agent.

AI USE STATEMENT

We used generative AI tools to assist with manuscript drafting and language editing. As part of the BindScope construction pipeline, video-capable models generated candidate questions and performed answer-blind auditing. These model-based checks are distinguished from human verification and do not independently establish ground truth. Multimodal models also serve as components of ECHO and as evaluation baselines, as described in the paper. The authors are responsible for verifying AI-assisted content, including annotations, code, references, and experimental claims, and take responsibility for the final manuscript and artifacts.

ETHICS STATEMENT

BindScope is intended to support the diagnosis and mitigation of hallucinations in long-video understanding. Its source videos may contain identifiable individuals, copyrighted material, and social or cultural biases. Any release of source material will be subject to the applicable permissions and licenses; where redistribution is not permitted, we will provide source references and annotation metadata instead of video files. Entity associations in ECHO are used to maintain correspondence within a video, rather than to identify individuals outside the video context. Model-assisted annotations may contain errors or biases, and benchmark performance should not be interpreted as a guarantee of reliability in real-world applications.

REPRODUCIBILITY STATEMENT

The main text describes ECHO's memory construction and retrieval procedures, the BindScope task design, and the experimental comparisons. Appendix A provides the graph schema, identity association procedures, retrieval and verification algorithms, and implementation settings. Appendix B describes dataset construction, quality control, and evaluation protocols. The anonymous repository linked in the paper provides supporting code and materials. Reproducing results obtained with hosted models may be affected by changes in model versions or service behavior.

REFERENCES

Ali Asgarov, Kaushik Narasimhan, Najibul Haque Sarker, Hani Alomari, Chia-Wei Tang, Anushka Sivakumar, Zaber Ibn Abdul Hakim, Shaurya Mallampati, and Chris Thomas. NEST: Narrative event structures in time for long video understanding. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026. URL https://arxiv.org/abs/2606.19706v2.

Zixuan Dong, Jiafu Tang, Zhide Lei, Zhe Cao, Zijie Zhang, Yanghai Wang, Shihao Li, Xiaodong Wang, Baoyun Peng, and Jiaheng Liu. OmniHalluc-L: Counterfactual benchmarking and modality-perturbation reliability calibration for long-form omni hallucination. *arXiv preprint arXiv:2606.03614*, 2026. doi: 10.48550/arXiv.2606.03614. URL https://arxiv.org/abs/2606.03614.

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. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Hyeonjeong Ha, Jinjin Ge, Bo Feng, Kaixin Ma, and Gargi Chakraborty. NarrativeTrack: Evaluating entity-centric reasoning for narrative understanding. *arXiv preprint arXiv:2601.01095*, 2026. URL https://arxiv.org/abs/2601.01095.

Yiyang Huang, Yitian Zhang, Yizhou Wang, Mingyuan Zhang, Liang Shi, Huimin Zeng, and Yun Fu. Distorted or fabricated? a survey on hallucination in video LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26599–26611. Association for

Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1325. URL https://aclanthology.org/2026.findings-acl.1325/.

Kyuho Lee, Euntae Kim, Jinwoo Choi, and Buru Chang. NOAH: Benchmarking narrative prior driven hallucination and omission in video large language models. *arXiv preprint arXiv:2511.06475*, 2025. doi: 10.48550/arXiv.2511.06475. URL https://arxiv.org/abs/2511.06475.

Aiden Yiliu Li, Nels Numan, and Anthony Steed. Visual agentic memory: Enabling online long video understanding via online indexing, hierarchical memory, and agentic retrieval. *arXiv preprint arXiv:2605.16481*, 2026a. doi: 10.48550/arXiv.2605.16481. URL https://arxiv.org/abs/2605.16481.

Chaoyu Li, Eun Woo Im, and Pooyan Fazli. VidHalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13723–13733, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_VidHalluc_Evaluating_Temporal_Hallucinations_in_Multimodal_Large_Language_Models_for_CVPR_2025_paper.html.

Jiazheng Li, Chi-Hao Wu, Yunze Liu, Kaize Ding, Jundong Li, and Chuxu Zhang. Bridging modalities, spanning time: Structured memory for ultra-long agentic video reasoning. *arXiv preprint arXiv:2605.08271*, 2026b. doi: 10.48550/arXiv.2605.08271. URL https://arxiv.org/abs/2605.08271.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 2024. URL https://aclanthology.org/2024.tacl-1.9/.

Lin Long, Yichen He, Wentao Ye, Yiyuan Pan, Yuan Lin, Hang Li, Junbo Zhao, and Wei Li. Seeing, listening, remembering, and reasoning: A multimodal agent with long-term memory. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ec370e882e8a02be955679ddbd2731b8-Abstract-Conference.html.

Hao Lu, Jiahao Wang, Yaolun Zhang, Ruohui Wang, Xuanyu Zheng, Yepeng Tang, Dahua Lin, and Lewei Lu. ELV-Halluc: Benchmarking semantic aggregation hallucinations in video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32572–32581, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Lu_ELV-Halluc_Benchmarking_Semantic_Aggregation_Hallucinations_in_Video_Understanding_CVPR_2026_paper.html.

Haoran Luo, Haihong E, Guanting Chen, Yandan Zheng, Xiaobao Wu, Yikai Guo, Qika Lin, Yu Feng, Zemin Kuang, Meina Song, Yifan Zhu, and Anh Tuan Luu. HyperGraphRAG: Retrieval-augmented generation via hypergraph-structured knowledge representation. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2503.21322.

Xiaokang Ma, Yifan Sun, Zhihong Jin, Jie Gu, Yudong Luo, Shenyi Shao, Chu Tang, Jingmin Chen, and Li Pu. ReflectWorld-MM: An entity-oriented multimodal memory system for open-ended video streams. *arXiv preprint arXiv:2607.09759*, 2026. doi: 10.48550/arXiv.2607.09759. URL https://arxiv.org/abs/2607.09759v2.

Junxi Wang, Te Sun, Jiayi Zhu, Junxian Li, Haowen Xu, Zichen Wen, Xuming Hu, Zhiyu li, and Linfeng Zhang. StreamMeCo: Long-term agent memory compression for efficient streaming video understanding. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13234–13251. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.647. URL https://aclanthology.org/2026.findings-acl.647/.

Yun Wang, Long Zhang, Jingren Liu, Jiaqi Yan, Zhanjie Zhang, Jiahao Zheng, Ao Ma, Run Ling, Xun Yang, Dapeng Wu, Xiangyu Chen, and Xuelong Li. Video-EM: Event-centric episodic memory for long-form video understanding. *arXiv preprint arXiv:2508.09486*, 2025. doi: 10.48550/arXiv.2508.09486. URL https://arxiv.org/abs/2508.09486v2.

Peizheng Yan, Yu Zhao, Liang Xie, Juntong Qi, Mingming Wang, and Erwei Yin. Event-Causal RAG: A retrieval-augmented generation framework for long video reasoning in complex scenarios. *arXiv preprint arXiv:2605.06185*, 2026. doi: 10.48550/arXiv.2605.06185. URL https://arxiv.org/abs/2605.06185.

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. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yeo_WorldMM_Dynamic_Multimodal_Memory_Agent_for_Long_Video_Reasoning_CVPR_2026_paper.html.

Xinlei Yin, Xiulian Peng, Xiao Li, Zhiwei Xiong, and Yan Lu. Hierarchical long video understanding with audiovisual entity cohesion and agentic search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32882–32891, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yin_Hierarchical_Long_Video_Understanding_with_Audiovisual_Entity_Cohesion_and_Agentic_CVPR_2026_paper.html.

Juwei Yue, Chuanrui Hu, Jiawei Sheng, Zuyi Zhou, Wenyuan Zhang, Tingwen Liu, Li Guo, and Yafeng Deng. HyperMem: Hypergraph memory for long-term conversations. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, 2026. URL https://arxiv.org/abs/2604.08256.

A Evidence-Constrained Hypergraph Organization (ECHO) Framework

ECHO turns a video into a query-independent temporal hypergraph memory (§A.1–§A.2) and answers or verifies statements only through bindings that exist in that memory (§A.3–§A.5). Table 3 summarizes the notation.

Symbol	Meaning
v, $W_k = [s_k, s_k + \Delta)$	video, k-th observation window of length Δ
$\mathcal{X}$, $\mathcal{E}$	local instances, global entities
$\pi^{(v)} : \mathcal{X} \to \mathcal{E} \cup \{\perp\}$	versioned identity projection
$h_e = (\iota_e, p_e, \{(\rho, x_{e,\rho})\}, \tau_e, \mathcal{Z}_e)$	event hyperedge
$h_a = (x_{\text{owner}}, a, u, \tau_a, \mathcal{Z}_a)$	attribute / state hyperedge
$h_s = (\iota_s, \sigma_s, \tau_s, spk_s)$	utterance record (text σ_s, speaker possibly $\perp$)
$\mathcal{G}^{(v)}$	temporal hypergraph memory of v
Q_c	query hypergraph parsed from a statement / caption c
$T_c = [t_0, t_1]$	temporal scope of c
$\text{sim}(\cdot, \cdot)$	cosine similarity of text embeddings

Table 3: Notation.

A.1 Multimodal Recognition–Registration–Tracing (RRT)

Recognition: evidence-scoped local observations. RRT processes a video in query-independent temporal windows $W_k = [s_k, s_k + \Delta)$ ($\Delta = 8$ s, $K = 8$ uniformly sampled frames per window). Within each window, a perception model jointly extracts local instances (people, objects, on-screen text carriers), predicates with directed semantic roles (agent, theme, recipient, instrument, target, source, destination), attributes, readable text, and the frames that support each record. Three constraints keep a window's observation evidence-scoped: (i) events are recorded in active voice, so the agent role is the performer and role permutation is not introduced by surface syntax; (ii) only visible attributes are recorded; values such as *not visible*, *unknown* or *none* are dropped rather than stored as values; and (iii) a role whose filler cannot be resolved to an instance of the same window remains *unresolved*, instead of being filled by a nearby entity. Partial observations are therefore kept without

asserting events that were not observed. Speech transcribed from the same window is shown to the perception model only as a disambiguation hint and is never recorded as visual evidence.

Free-form attribute names are normalized into canonical slots (`jersey_number`, `worn_text`, `team`, `color`, `clothing`, `hair`, `accessory`, `material`, `shape`, `size`, `count`, `state`, `role_cue`); composite values are split so that, e.g., a jersey description yields a separate `jersey_number` value. We call $\mathcal{S}_{\text{id}} = \{\texttt{jersey_number}, \texttt{worn_text}, \texttt{team}\}$ the *identity slots*.

Person grounding. Identity association requires a region that shows the person, not a hand or a fragment. Each recognized person is grounded by a separate call that receives the person's description and one frame, and returns a single box $b = (x_1, y_1, x_2, y_2)$ together with a view label $\in \{\texttt{head_and_body}, \texttt{body}, \texttt{head}, \texttt{partial}, \texttt{hands_or_arms}\}$. One person per call avoids positional mix-ups among several boxes; several people in one frame are handled by one call each, and two boxes of different people with IoU > 0.6 are marked `ambiguous`. If a person is not found in the first frame where it was recorded, up to three fallback frames are tried. Regions whose view is `partial` or `hands_or_arms`, or that are `ambiguous`, are stored but are not *identity-usable*: they are excluded from the galleries used by tracing. At most 8 people per window are grounded.

Registration: local instances before global identities. RRT registers local instances independently from global identity commitments. Spatial regions form local observations, instances group observations within a window, and global entities connect instances across time. Event roles attach strictly to local instances, so no global identity is asserted by an event. Cross-occurrence identity is governed by a versioned projection

$$\pi^{(v)} : \mathcal{X} \to \mathcal{E} \cup \{\perp\}. \tag{7}$$

An unresolved projection ($\perp$) preserves the local event while withholding a global identity assignment.

Tracing: revisable correspondence. Cross-window identity links are *nominated* from multimodal cues and then *verified* against endpoint evidence. Candidate pairs (x_i, x_j), $x_i \in W_k$, $x_j \in W_{k'}$, $k \neq k'$, of the same kind are scheduled in groups; for each group a verifier compares the identity-usable region galleries of the two instances ("do LEFT and RIGHT depict the same individual?") and returns `accepted`, `rejected` or `pending`. Link statuses are revisable when later windows add evidence. Only accepted links enter $\pi^{(v)}$, and only after an *identity-slot veto*: accepted links are applied in a fixed order with union–find, and a link (x_i, x_j) is refused if the merged component would carry two distinct values of a single-valued identity slot,

$$\left|\{\, u : (x, \texttt{jersey_number}, u) \in \mathcal{A},\ x \in C(x_i) \cup C(x_j) \,\}\right| > 1 \ \Rightarrow\ \text{refuse } (x_i, x_j), \tag{8}$$

where $C(\cdot)$ is the current component and $\mathcal{A}$ the attribute observations. Look-alike uniforms can make a gallery comparison say *same* for team-mates; an independently read number is evidence that they differ. A misread number can only fragment an identity, never merge two people. To further prevent mis-binding, RRT decouples entity identity from surface ownership via `partial` relations, and isolates spoken entities from visual participants until they are explicitly grounded.

Audio registration. Speech is transcribed into utterance records $h_s = (\iota_s, \sigma_s, \tau_s, spk_s)$. Utterances are nodes of their own: an utterance is not a participant of any visual event, a name mentioned in speech is not resolved to a visual role, and speech content is not converted into a world fact. The speaker spk_s stays $\perp$ unless it is grounded.

A.2 ECHO: Temporal Hypergraph Memory Construction

Joint event binding. Observations and accepted bindings are organized in a typed temporal hypergraph $\mathcal{G}^{(v)}$ whose nodes include local instances, global entities, evidence records, utterances and attribute values. An event occurrence is stored as a role-labeled hyperedge (Yue et al., 2026; Luo et al., 2025)

$$h_e = (\iota_e,\, p_e,\, \{(\rho, x_{e,\rho})\}_{\rho \in \mathcal{R}_e},\, \tau_e,\, \mathcal{Z}_e)\,, \tag{9}$$

where ι_e identifies the occurrence, p_e is its predicate, $x_{e,\rho}$ the local instance filling role ρ, τ_e the observed time interval, and $\mathcal{Z}_e$ the supporting evidence (frames, regions). The hyperedge binds the predicate and all of its roles in one unit. Sharing an object across two events does not mix their roles: a person administering medication does not fill the actor role of a medication-tampering event merely because both events involve the same vial. This joint representation prevents predicate–event leakage, participant–event leakage and role permutation.

Scoped attribute states. An attribute or state observation is

$$h_a = (x_{\text{owner}},\, a,\, u,\, \tau_a,\, \mathcal{Z}_a)\,, \tag{10}$$

with canonical slot a, value u and supported temporal scope τ_a. Repeated observations of the same attribute are stored separately and never overwrite earlier values, and an attribute observed over τ_a is not assumed to persist through unobserved periods. Identity slots are the exception that tracing licenses: once instances are merged into an entity ϵ, the identity values read on any of them describe ϵ,

$$\mathrm{Id}(\epsilon) = \big\{(a,u) :\ (x,a,u,\cdot,\cdot) \in \mathcal{G}^{(v)},\ \pi^{(v)}(x) = \epsilon,\ a \in \mathcal{S}_{\mathrm{id}}\big\}, \tag{11}$$

which is what lets a number legible in one window identify the same player in windows where it is not legible. Event–relation hyperedges connect distinct event occurrences with the relation type and its evidence; process links preserve the observed order of related events.

Cross-modal co-temporal hyperedges. To give every event access to the speech that accompanies it without asserting who said what, each utterance is linked to the visual events it overlaps in time. For an utterance h_s with $\tau_s = [a_s, b_s]$ and an event h_e with $\tau_e = [a_e, b_e]$ and margin m ($m = 1.5\,\mathrm{s}$),

$$h_{se} = \big(h_s,\, h_e,\, r,\, g,\, \Sigma_{se}\big), \quad r = \begin{cases} \texttt{during} & [a_s,b_s] \cap [a_e,b_e] \neq \emptyset, \\ \texttt{just_before} & 0 < a_e - b_s \le m, \\ \texttt{just_after} & 0 < a_s - b_e \le m, \end{cases} \tag{12}$$

with gap $g \ge 0$ and Σ_{se} the content terms shared by the utterance text and the event's predicate and participant descriptions. The edge records temporal co-occurrence of two independently observed records and nothing more: speaker and referent remain unresolved.

Construction. Algorithm 2 summarizes memory construction.

Figure 6 shows the interface used to inspect ECHO's constructed memory alongside the source video.

A.3 Binding-Preserving Scoped Retrieval

To deliver verifiable answers without cross-event mis-binding, retrieval filters scope-aware candidate facts, expands them into dependency-closed binding units, audits claim support during answer generation, and guarantees binding soundness.

Scope-filtered retrieval. Given a question q, ECHO extracts entity cues, target relations and explicit temporal constraints. A pre-built index retrieves candidate facts and utterances; search expansion is restricted to relevant local observations and accepted identity links.

Dependency-closed expansion. Retrieving isolated facts risks exposing compatible-looking elements while losing the relations that bind them. Each selected fact f is expanded into a self-contained binding unit

$$U_q(f) = \mathrm{Closure}_{q,\mathcal{G}^{(v)}}(\{f\}), \tag{13}$$

which collects the event occurrence, its role endpoints, visual evidence, and the identity, ownership, reference and co-temporal dependencies it requires; unrelated entity background is left out. Retrieval selects complete units under an input budget B,

$$P_q = \bigcup_{U \in \mathcal{S}_q} U, \qquad \mathrm{cost}(P_q) \le B. \tag{14}$$

A unit that exceeds B is omitted entirely rather than truncated across its roles or dependencies, and a structural check verifies that the delivered payload preserves all selected endpoints, bindings and evidence links.

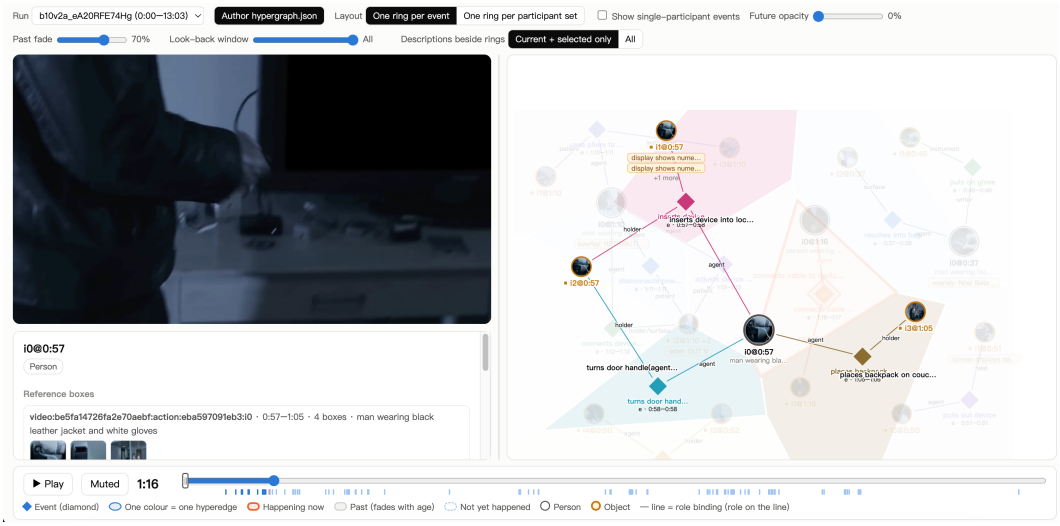

Figure 6: **ECHO's hypergraph inspection interface.** The left panel displays the source video and details of the selected entity, including linked local instances. The right panel visualizes event hyperedges and role-labeled connections to participants. Colored regions distinguish hyperedges, while opacity indicates temporal context relative to playback. The interface supports inspection of entity associations and event bindings against video evidence.

Algorithm 2 ECHO memory construction (RRT + hypergraph synthesis)

Require: video v with audio; window length Δ; frames per window K
Ensure: temporal hypergraph memory $\mathcal{G}^{(v)}$

$\mathcal{U} \leftarrow \mathrm{ASR}(v)$ ▷ utterances h_s, speaker $\perp$
for all windows W_k **do**
  $F_k \leftarrow K$ frames sampled in W_k; $\mathcal{U}_k \leftarrow \{h_s \in \mathcal{U} : \tau_s \cap W_k \neq \emptyset\}$
  $O_k \leftarrow \textsc{Recognize}(F_k,\ \text{hint} = \mathcal{U}_k)$ ▷ instances, events, attributes, text
  normalize attributes to canonical slots; drop unseen values; leave unresolvable roles unresolved
  for all person instances $x \in O_k$ **do**
    $(b_x, view_x) \leftarrow \textsc{GroundOne}(x, F_k)$ ▷ one person per call, fallback frames
  end for
  mark boxes of different people with IoU > 0.6 `ambiguous`
  register O_k: regions $\rightarrow$ observations $\rightarrow$ local instances; events/attributes as h_e, h_a on local instances
end for
$\mathcal{P} \leftarrow \textsc{Nominate}(\{O_k\})$ ▷ cross-window candidate pairs
$L \leftarrow \textsc{Verify}(\mathcal{P})$ on identity-usable galleries ▷ accepted / rejected / pending
$L_{\mathrm{acc}} \leftarrow$ accepted links of L that pass the identity-slot veto (Eq. 8)
$\pi^{(v)} \leftarrow$ connected components of L_{acc}; unlinked instances map to singleton or $\perp$
attach $\mathrm{Id}(\epsilon)$ to each entity (Eq. 11)
add utterance nodes and co-temporal edges h_{se} (Eq. 12)
return $\mathcal{G}^{(v)}$

Answer generation and support assessment. The reader receives q and P_q and is instructed to respect event roles, identity links and temporal scopes. Answer generation is decoupled from support assessment. For a candidate claim c,

$$\mathrm{Support}(c, P_q) = 1 \iff \exists\, \phi : c \hookrightarrow P_q, \tag{15}$$

where ϕ maps the claim's predicate, role assignments, entity links, time constraints and evidence dependencies onto P_q. The same entity or action in an unrelated event does not satisfy this condition.

Claims are audited as supported, explicitly contradicted, or insufficiently supported; absence from memory is not treated as proof of falsity.

Binding soundness. Assuming stored observations and accepted correspondences are accurate, a role or entity substitution that has no valid mapping within the queried scope cannot be certified as supported; dependency-closed retrieval preserves every relation needed to evaluate this condition.

A.4 Statement Verification as Hypergraph Matching

Hallucination benchmarks such as ELV-Halluc ask whether a timestamped caption c with scope $T_c = [t_0, t_1]$ is fully correct. We instantiate Eq. equation 15 by matching the caption, written as a hypergraph, against the memory.

Query hypergraph. A language model parses c (without judging it) into $Q_c = (\mathcal{N}_c, \mathcal{H}_c)$: entities $\mathcal{N}_c$ with their surface forms, relation hyperedges $q = (p_q, \{(\rho, n_{q,\rho})\})$, attribute hyperedges $q = (n_{\text{owner}}, a, u)$, and statements $q = (\sigma_q, spk_q)$. Each $q \in \mathcal{H}_c$ is one *claim*. A parse without any claim is treated as a failed parse and repeated, never as a caption without content.

Memory view. Each memory node n (a global entity, or a local instance with $\pi^{(v)} = \perp$) is rendered as text from its descriptions, the text it carries and $\mathrm{Id}(n)$; each hyperedge h as its predicate (or slot and value) with its role fillers; each utterance as σ_s. All texts are embedded once per video ($\mathbf{e}(\cdot)$, Qwen3-Embedding-8B).

Binding score. For a relation or attribute claim q and a memory hyperedge h of the same type (event for relations; attribute/text for attributes),

$$s(q,h) = \min\Big(\mathrm{sim}\big(\mathbf{e}(p_q), \mathbf{e}(p_h)\big),\ \min_{n \in \mathrm{args}(q)}\ \max_{x \in \mathrm{fill}(h)} \mathrm{sim}\big(\mathbf{e}(n), \mathbf{e}(\pi(x))\big)\Big), \tag{16}$$

where p_h is the predicate (for attributes, slot and value) and $\mathrm{fill}(h)$ the role fillers of h. The score is the *weakest* constituent: a hyperedge that shares the predicate but not the participants, or the participants but not the predicate, cannot score as a binding. For a statement claim, $s(q, h_s) = \mathrm{sim}(\mathbf{e}(\sigma_q), \mathbf{e}(\sigma_s))$. We also keep the content-only score $s_{\mathrm{c}}(q,h) = \mathrm{sim}(\mathbf{e}(p_q), \mathbf{e}(p_h))$.

Scope partition. A record is *in scope* if its interval overlaps T_c within a tolerance δ ($\delta = 1\,\mathrm{s}$): $\mathcal{H}_c^{\mathrm{in}} = \{h : \tau_h \cap [t_0 - \delta, t_1 + \delta] \neq \emptyset\}$, $\mathcal{H}_c^{\mathrm{out}} = \mathcal{G}^{(v)} \setminus \mathcal{H}_c^{\mathrm{in}}$. For each claim we keep three candidate sets, each the top-k records by score:

$$\begin{aligned} \mathcal{C}_{\mathrm{in}}^{\mathrm{bind}}(q) &= \mathrm{top}_k\,\{s(q,h) : h \in \mathcal{H}_c^{\mathrm{in}}\}, \\ \mathcal{C}_{\mathrm{in}}^{\mathrm{cont}}(q) &= \mathrm{top}_k\,\{s_{\mathrm{c}}(q,h) : h \in \mathcal{H}_c^{\mathrm{in}}\}, \\ \mathcal{C}_{\mathrm{out}}^{\mathrm{bind}}(q) &= \mathrm{top}_k\,\{s(q,h) : h \in \mathcal{H}_c^{\mathrm{out}}\}. \end{aligned} \tag{17}$$

$\mathcal{C}_{\mathrm{in}}^{\mathrm{bind}}$ looks for the claimed binding where the caption places it; $\mathcal{C}_{\mathrm{in}}^{\mathrm{cont}}$ finds the same content bound to *someone else* in scope (the source of an in-video swap); $\mathcal{C}_{\mathrm{out}}^{\mathrm{bind}}$ finds the binding in *another temporal event*. Records below a relevance floor θ are dropped ($k = 2$, $\theta = 0.5$, at most 5 records per claim and 10 claims per caption).

A.5 Structural Verdicts

Embedding similarity alone cannot decide whether two descriptions denote the same participant ("John" vs. "man in an apron") or the same value. The reader model therefore answers only local, structural questions: for every claim q and every candidate record h it assigns

$$\ell(q,h) = \big(\ell_{\mathrm{cont}}, \ell_{\mathrm{part}}\big) \in \{\texttt{same}, \texttt{conflict}, \texttt{different}\} \times \{\texttt{same}, \texttt{different}, \texttt{unclear}\}, \tag{18}$$

where `conflict` means the same kind of action or property with an incompatible value, and $\ell_{\mathrm{part}} = \texttt{different}$ requires positive identity evidence (a different number, name, team or printed text, or an appearance contradicting the claim's own description); a name against an appearance is `unclear`. The model is not told, and does not decide, whether a record lies in or out of scope, and

Verdict	Condition (exists a candidate h with ...)	Decisive
CONFLICT	$h \in \mathcal{H}_c^{\text{in}}$, $\ell = (\texttt{conflict}, \texttt{same})$	yes
SUPPORTED	$h \in \mathcal{H}_c^{\text{in}}$, $\ell_{\text{cont}} = \texttt{same}$, $\ell_{\text{part}} \neq \texttt{different}$	–
MISBOUND	$h \in \mathcal{H}_c^{\text{in}}$, $\ell = (\texttt{same}, \texttt{different})$, ≥ 2 participants in scope	no†
MISPLACED	$h \in \mathcal{H}_c^{\text{out}}$, $\ell = (\texttt{same}, \texttt{same})$, $s(q,h) \geq \max_{h' \in \mathcal{C}_{\text{in}}(q)} s(q,h') + \gamma$	yes
UNKNOWN	otherwise	–

Table 4: Claim verdicts, checked in the listed order (CONFLICT first; MISBOUND and MISPLACED only when the claim is not SUPPORTED). $\gamma = 0.03$. †Reported but not decisive in our runs: the participant label is unreliable when a name meets an appearance description.

Algorithm 3 Binding-closed statement verification

Require: caption c with scope T_c; memory $\mathcal{G}^{(v)}$ with embeddings
Ensure: verdict $\hat{y}(c)$ and per-claim evidence
1: $Q_c \leftarrow \text{PARSE}(c)$; **if** $\mathcal{H}_c = \emptyset$ **then** re-parse (bounded retries)
2: **for all** claims $q \in \mathcal{H}_c$ **do**
3: compute $s(q,h)$, $s_{\text{c}}(q,h)$ for all type-compatible h (Eq. 16)
4: $\mathcal{C}(q) \leftarrow \mathcal{C}_{\text{in}}^{\text{bind}} \cup \mathcal{C}_{\text{in}}^{\text{cont}} \cup \mathcal{C}_{\text{out}}^{\text{bind}}$ (Eq. 17), dropping records below θ
5: **end for**
6: $\{\ell(q,h)\} \leftarrow \text{READER}(\{(q, \mathcal{C}(q))\})$ ▷ labels only, Eq. 18
7: **for all** claims q **do**
8: $\text{verdict}(q) \leftarrow$ first matching row of Table 4, with cited record ids
9: **end for**
10: **return** $\hat{y}(c)$ by Eq. equation 19, with the cited records of each verdict

it does not judge the caption. Code derives the claim verdict from the labels and the record positions (Table 4).

The caption is judged *not fully correct* if any claim has a decisive verdict,

$$\hat{y}(c) = \texttt{False} \iff \exists\, q \in \mathcal{H}_c : \ \text{verdict}(q) \in \{\text{CONFLICT}, \text{MISPLACED}\}, \tag{19}$$

and $\texttt{True}$ otherwise; UNKNOWN never makes a caption false (absence is not falsity). The margin γ in MISPLACED requires that another temporal event carries the claim *clearly better* than anything in scope; without it, content that recurs across a video (the same cook stirring, repeated dunks) rejects true captions. Algorithm 3 summarizes verification.

A.6 READING WITH FRAMES AND MEMORY EVIDENCE

Memory verdicts cover only the details the memory recorded; a caption detail never observed yields UNKNOWN. The final reader therefore combines the caption's frames with memory evidence. Given K frames sampled uniformly in T_c and, for each claim, its in-scope and out-of-scope binding evidence (the best record of $\mathcal{C}_{\text{in}}^{\text{bind}}$ with $s \geq 0.75$, and the best record of $\mathcal{C}_{\text{out}}^{\text{bind}}$ when it also exceeds the in-scope score by 0.05), including the utterances matched by statement claims, the vision-language reader answers whether c is fully correct. We score the answer by

$$P_{\text{True}}(c) = \frac{\exp z_{\texttt{True}}}{\exp z_{\texttt{True}} + \exp z_{\texttt{False}}}, \tag{20}$$

from the logits of the first answer token, and predict $\texttt{True}$ iff $P_{\text{True}}(c) \geq \lambda$.

A.7 EVALUATION PROTOCOL ON ELV-HALLUC

Each ELV-Halluc item is a triplet of captions for the same segment: the ground truth c^{gt}, an in-video hallucination c^{in} (a detail replaced by one that appears elsewhere in the same video) and an out-of-video hallucination c^{out} (a detail that does not appear in the video), in four aspects (visual details, objects, actions, declarative content). A pair counts as correct only if both of its captions are judged correctly,

$$\text{Acc}_{\text{in}} = \frac{1}{N} \sum_i \mathbb{1}\big[\hat{y}(c_i^{\text{gt}}) = \texttt{True} \wedge \hat{y}(c_i^{\text{in}}) = \texttt{False}\big], \tag{21}$$

and $\mathrm{Acc}_{\mathrm{out}}$ analogously. Semantic aggregation hallucination is summarized by

$$\mathrm{SAH} = \frac{\mathrm{Acc}_{\mathrm{out}} - \mathrm{Acc}_{\mathrm{in}}}{100 - \mathrm{Acc}_{\mathrm{in}}} \times 100. \tag{22}$$

The decision threshold λ of Eq. equation 20 is chosen on a held-out development split of videos and then fixed for all remaining videos; we additionally report the threshold-free paired win rate $\Pr[P_{\mathrm{True}}(c^{\mathrm{gt}}) > P_{\mathrm{True}}(c^{\mathrm{hal}})]$. Captions whose parse remains empty after retries are excluded from scoring and counted.

A.8 Implementation Details

Component	Setting
Perception / grounding / identity verifier	Qwen3-VL-8B-Instruct (vLLM, bf16), JSON-object decoding
Window	$\Delta = 8$ s, $K = 8$ frames; ≤ 8 people, 10 objects, 10 events, 8 text entries
Grounding	one person per call, box in $[0, 1000]^2$ + view label; ≤ 3 fallback frames; IoU 0.6
Identity veto	single-valued slot `jersey_number`, union–find in link order
ASR	utterance-level transcription of the original audio track
Co-temporal alignment	margin $m = 1.5$ s
Caption parser / reader	Qwen3-VL-8B-Instruct
Embeddings	Qwen3-Embedding-8B, cosine similarity
Matching	$\delta = 1$ s, $k = 2$, $\theta = 0.5$, $\gamma = 0.03$

Table 5: Settings used for ELV-Halluc.

Structured outputs. All structured outputs (observations, grounding, identity decisions, parses, labels) are produced as free-form JSON objects and validated afterwards; an empty or invalid result is retried rather than accepted.

B Datasets

B.1 Binding-centered task design

Controlled binding alternatives. Each question contrasts a video-grounded assignment with a controlled alternative that alters a targeted binding while preserving all observed constituents. For instance, temporal and identity tasks maintain observed endpoints while modifying their correspondence or relative ordering. BindScope thus strictly measures support for joint assignments rather than the mere presence of individual constituents.

Paired-unit evaluation. Each evaluation unit p consists of two independently answered binary-choice questions:

$$p = (q_{p,1}, q_{p,2}). \tag{23}$$

These two questions probe the same binding mechanism through distinct, independently evidenced assignments and controlled distractors, avoiding simple paraphrases or logical negations. A model must answer both questions correctly to pass the unit, preventing shortcuts based on entity bias, attribute salience, or option positioning.

Hierarchical scope distinction. BindScope organizes tasks across three distinct levels: single-event binding, cross-event assignment, and cross-event connection. This hierarchy disentangles localized perception errors from long-range reasoning. For instance, role permutation (R3) exchanges complete role assignments within a single event, whereas identity tracking (R4) requires cross-scene entity correspondence while excluding local identification shortcuts like visible name tags.

B.2 Dataset Construction and Quality Control

Automated candidate generation. A video-capable annotation model processes multimodal source videos to propose candidate paired units alongside their binding labels, supporting time in-

tervals, binding tuples, and error descriptions. Legacy benchmark questions serve strictly as scene-search hints without inheriting original ground-truth answers or distractors. To prevent hallucinated ground truths, unsupported binding categories within a given video are explicitly skipped rather than artificially filled.

Answer-blind model auditing. Candidates first undergo programmatic validation for option formatting and timestamp bounds. Next, a separate, answer-blind auditing model re-evaluates the candidates with access to the video, questions, and options, but without proposed answers or evidence rationales. This auditor independently solves the questions to verify type consistency, answer uniqueness, evidence sufficiency, distractor plausibility, and pair non-redundancy. Candidates passing both stages form the model-verified candidate pool.

Shortcut filtering and human verification. To ensure diagnostic rigor, candidate questions undergo strict filtering for linguistic, visual, and world-knowledge shortcuts. Role (R3) and identity (R4) questions receive targeted scrutiny to prevent models from exploiting social-role priors or local identifiers such as name tags. Furthermore, text-only evaluations using independent LLMs flag residual language shortcuts for manual review. Option order is randomized and answer distributions are balanced across the dataset.

B.3 Prompts for Answerability Auditing

We use a two-stage procedure to audit question answerability. A video-conditioned respondent produces a short free-form answer without access to the answer options or reference answer. A separate text-based judge then assesses whether this answer semantically entails the reference answer in the context of the question. The judge receives no video input; its decision measures semantic agreement rather than independently verifying the reference answer against the video.

Respondent prompt.

```
Watch the video carefully. Answer the question below using ONLY what
    you actually see and hear in the video.
Answer with a short phrase only (a name, an object, an action, or a
    brief description). Do NOT explain, and do NOT list options.

Question: {question}
```

Judge prompt.

```
You are provided with a question, a ground truth answer, and an answer
    from an agent model. Your task is to determine whether the ground
    truth answer can be logically inferred from the agent's answer, in
    the context of the question.
Do not directly compare the surface forms of the agent answer and the
    ground truth answer. Instead, assess whether the meaning expressed
    by the agent answer supports or implies the ground truth answer. If
    the ground truth can be reasonably derived from the agent answer,
    return "Yes". If it cannot, return "No".

Important notes:
- Do not require exact wording or matching structure.
- Semantic inference is sufficient, as long as the agent answer entails
    or implies the meaning of the ground truth answer, given the
    question.
- Only return "Yes" or "No", with no additional explanation or
    formatting.

Input fields:
- question: the question asked
- ground_truth_answer: the correct answer
- agent_answer: the model's answer to be evaluated
```

Table 6: Binding evaluation on BindScope. All entries report paired accuracy (%), where a pair is correct only if both questions are answered correctly. **Overall** is computed over all paired units in R0–R7.

Method	R0 ↑	R1 ↑	R2 ↑	R3 ↑	R4 ↑	R5 ↑	R6 ↑	R7 ↑	Overall ↑
Base model									
Qwen3-VL-8B	4.3	14.3	12.8	17.1	13.7	11.4	15.8	20.6	13.7
Qwen3-VL-32B	5.0	19.8	12.2	17.1	20.2	20.3	18.4	29.9	17.5
Qwen3.8-Omni-Flash	14.4	28.6	13.9	25.0	17.9	26.6	23.3	35.5	22.3
HY-Vision-2.0	2.2	26.4	12.2	19.3	19.0	22.8	20.3	21.5	17.7
GPT-5.4	12.9	28.0	17.5	28.7	19.0	18.5	24.5	27.8	21.6
RAG-based methods									
Caption RAG	11.5	17.8	18.4	18.9	24.7	21.0	23.6	21.7	20.2
Event-Causal RAG	30.2	25.3	28.1	37.9	37.1	32.1	36.0	40.7	34.0
MAGIC-Video	32.8	32.1	16.7	17.9	31.8	25.6	21.9	18.5	23.8
Memory-based methods									
M3-Agent	17.65	28.8	19.9	27.8	32.1	27.3	25.9	32.6	27.3
ECHO (ours)	46.0	48.5	41.0	38.3	42.9	51.3	52.2	46.1	**45.9**

```
Now evaluate the following input:
Input:
- question: {question}
- ground_truth_answer: {ground_truth_answer}
- agent_answer: {agent_answer}
Output ('Yes' or 'No'):
```

B.4 Evaluation Protocol

Questions are evaluated independently in isolated contexts without access to paired questions or annotation metadata. We report both individual Question Accuracy (QAcc) and Paired Accuracy (PairAcc):

$$\mathrm{PairAcc} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \prod_{j=1}^{2} \mathbb{1}[\hat{y}_{p,j} = y_{p,j}], \tag{24}$$

where a pair p earns credit only if both binary-choice questions ($j \in \{1, 2\}$) are correctly answered (random guess baselines: 25% for $\mathrm{PairAcc}$ vs. 50% for QAcc). Under independent uniform guessing, the chance-level PairAcc is 6.25% for four-option multiple-choice pairs and 25% for true/false pairs. Overall PairAcc is computed over all paired units rather than averaged equally across binding types. Unpaired questions contribute only to QAcc.

C Results

Table 6 provides the numerical results underlying the BindScope comparison in the main text. We report paired accuracy for phantom binding (R0) and seven in-video misbinding types (R1–R7), grouped by foundation models, RAG-based methods, and memory-based methods. Overall accuracy is computed over all paired units rather than averaged equally across types; unpaired questions are excluded from this metric. ECHO achieves the highest accuracy across all eight types, with an overall score of 45.9%, compared with 34.0% for the strongest evaluated baseline, Event-Causal RAG.

D Conclusion

We introduced a unified binding perspective on long-video hallucination, distinguishing phantom bindings to absent elements from misbindings among observed elements. Based on this perspective, we developed BindScope, a diagnostic benchmark covering eight binding-error types, and ECHO,

a training-free framework that preserves evidence-grounded bindings through temporal hypergraph memory and scoped retrieval. Our experiments reveal distinct binding-specific weaknesses across existing models and frameworks, while demonstrating ECHO's effectiveness on BindScope and ELV-Halluc. Future work can use these diagnostic profiles to develop targeted interventions for different hallucination types and investigate how perception, identity association, and memory retrieval contribute to each failure mode.