Tracking the Truth: Object-Centric Spatio-Temporal Monitoring for Video Large Language Models

Tri Cao*[1], Khoi Le*[1], Thong Nguyen †[1], Cong-Duy Nguyen [2], Quynh Vo [2], Anh-Tuan Luu [2,3], Miao Chunyan [3], See-Kiong Ng [1], Shuicheng Yan [1], Bryan Hooi [1]

[1] National University of Singapore, [2] VinUniversity, [3] Nanyang Technological University.

Email: `thong.nguyen@u.nus.edu`

Abstract

While multimodal large language models (MLLMs) have advanced video understanding, they remain highly prone to hallucinations in dynamic scenes. We argue this stems from a failure in spatio-temporal monitoring, the ability to persistently track object identities, states, and relations over time. Existing benchmarks obscure this deficit by relying on single final-answer evaluations for queries that can often be resolved via local visual cues or statistical priors. To rigorously diagnose this, we introduce **STEMO-Bench** (**S**patio-**TE**mporal **MO**nitoring), a benchmark of human-verified object-centric facts that evaluates intermediate reasoning by decomposing queries into sub-questions, distinguishing genuine temporal understanding from coincidental correctness. To address failure modes exposed by STEMO, we propose **STEMO-Track**, a novel object-centric framework that explicitly constructs and reasons over structured object trajectories via chunk-wise state extraction and temporal aggregation. Extensive experiments demonstrate that our object-centric framework significantly reduces hallucinated answers and improves spatio-temporal reasoning consistency over state-of-the-art MLLMs.

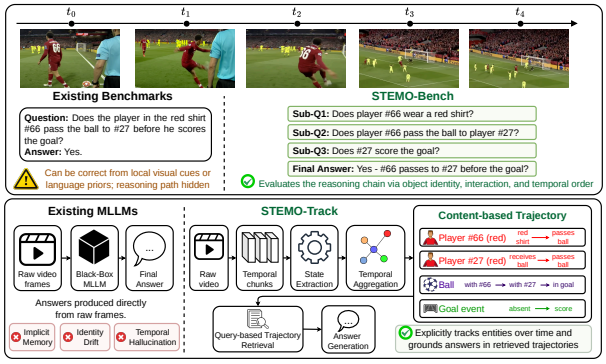

Figure 1: Comparison of existing evaluation benchmarks and MLLM architectures against our proposed STEMO-Bench and STEMO-Track.

1 Introduction

Multimodal large language models (MLLMs) [32, 26, 37] have recently made rapid progress on video understanding, improving semantic reasoning [30], instruction following [35], and long context modeling [13, 28]. Despite these advances, their reliability remains limited, as MLLMs are still prone to hallucinations, producing plausible yet factually incorrect outputs [16, 7]. Understanding and diagnosing these failures is essential for deploying such systems in real world applications.

*Contributes equally, †Corresponding author.

Published at the Second Workshop on Agents in the Wild: Safety, Security, and Beyond (AIWILD) at ICML 2026.

At its core, video understanding involves reasoning over a dynamic world. A video consists of entities whose properties and relations evolve continuously over time. To faithfully interpret such content, a model must perform persistent spatio-temporal monitoring, including identifying objects, maintaining their identities, tracking their state changes, and linking them to temporally ordered events [34, 14]. This process requires consistent reasoning across time, where errors in early steps such as misidentifying an object or missing a state transition can propagate and lead to incorrect conclusions.

However, existing evaluation benchmarks do not fully capture these requirements. Prior work has introduced a range of hallucination benchmarks [29, 12, 23] and mitigation strategies [31, 2, 25, 8], typically by constructing fine-grained question-answer pairs using a combination of MLLMs and human annotators with access to raw video inputs such as frames and subtitles [29, 12]. While effective at exposing certain failure modes, these benchmarks exhibit two key limitations under the spatio-temporal monitoring perspective. First, many questions can be answered by inspecting a single frame or a short segment, without requiring consistent tracking of objects over time. This allows models to rely on local cues rather than performing global temporal reasoning. Second, evaluation is typically based on a single final answer [23, 4, 11], which obscures the reasoning process. A model may arrive at the correct answer through statistical priors or partial observations, without correctly resolving the underlying spatio-temporal dependencies [22, 23, 11]. As a result, current benchmarks provide only a limited view of a model's true understanding capabilities.

To address this gap, we introduce **STEMO-Bench** (**S**patio-**TE**mporal **MO**nitoring), a benchmark designed to explicitly evaluate an MLLM's ability to track object-centric dynamics over time. STEMO-Bench adopts a structure-first data construction pipeline that aligns with the requirements of spatio-temporal monitoring. Instead of generating question-answer pairs directly from raw video inputs, we first construct a human-verified structural representation of each video, explicitly capturing objects, their evolving states, and temporal relations. These structured facts serve as the foundation for generating compositional question-answer pairs that require multi-step reasoning across time. By grounding all questions in explicit human-verified structure, STEMO-Bench ensures that each query directly probes temporal dependencies and object interactions, rather than superficial visual cues.

Beyond dataset construction, STEMO-Bench introduces a fine-grained evaluation protocol that assesses the reasoning process itself. Each query is decomposed into a sequence of sub-questions corresponding to key steps in spatio-temporal monitoring, such as identifying objects, retrieving states at specific timestamps, and verifying temporal transitions. A model is considered consistent only if it correctly resolves these intermediate steps. This design allows us to distinguish genuine understanding from coincidental correctness, providing a more faithful evaluation of hallucination behavior.

Motivated by the failure modes exposed by STEMO, we construct **STEMO-Track**, an object-centric pipeline for video understanding that explicitly models entities and their temporal evolution. Instead of treating a video as an unstructured sequence of frames, we represent it as a set of objects and their associated state trajectories, and perform reasoning over these structured representations. This perspective establishes a strong baseline, aligning the model's internal computation with the requirements of spatio-temporal monitoring.

Concretely, our pipeline first converts a video into a structured representation before performing question answering. We partition the video into temporal chunks and apply a state extractor to identify objects and their states over time, producing a set of local observations. A temporal aggregation module then links these observations across chunks, enforcing consistency over object identities, states, and timestamps to form coherent object trajectories. Given a question, the model retrieves the relevant trajectories and performs reasoning over them to produce the final answer. This design separates representation construction from query-specific reasoning, enabling more reliable modeling of long-horizon spatio-temporal dynamics.

In summary, our contributions are threefold: (1) We introduce the STEMO-Bench alongside a faithful evaluation protocol that explicitly tests intermediate reasoning through atomic sub-questions. (2) We propose STEMO-Track a novel object-centric pipeline that systematically constructs and queries structured object trajectories from videos. (3) We conduct a systematic evaluation of modern MLLMs on target versus intermediate consistency, highlighting key spatio-temporal deficiencies and demonstrating the efficacy of explicit monitoring.

2 Related Work

Video Multimodal Large Language Models. Recent years have seen rapid progress in multimodal large language models (MLLMs), evolving from general vision-language assistants into specialized systems capable of long-context temporal reasoning [32, 26]. Early models focused on instruction tuning and short-video understanding [30, 35], while newer architectures emphasize long-context scaling, temporal grounding, and reasoning [13, 28, 37]. This advancement is driven by strong open-source/open-weight models such as InternVL3, Qwen3-VL, Qwen3.5-Omni, and Gemma-4 [36, 3, 21, 9], alongside reasoning-focused video MLLMs including Cosmos-Reason2, VideoRFT, and Video-R1 [18, 27, 5]. Proprietary frontier models, including Gemini 3, GPT-5, and Claude-4.6-Sonnet, further continue to push multimodal reasoning capabilities [10, 20, 19, 1].

Video Hallucination Benchmarks and Mitigation. As video MLLMs advance, growing efforts focus on understanding and reducing their failure modes. Existing benchmarks target different vulnerabilities, including object hallucination [29], action and temporal grounding errors [12], structural reasoning deficits [23, 4], and temporal drift in long videos [11]. Mitigation approaches span training-time alignment [31, 2], inference-time decoding strategies [25], and retrieval-augmented architectures [8, 16, 7].

Object-Centric and Structured Video Understanding. To address the limitations of holistic frame embeddings, recent work explores object-centric and structured video reasoning. Methods such as VideoRefer [34] and STI [14] show that modeling spatio-temporal object interactions improves grounding. Building on this idea, structured pipelines like VideoMind [15], TraveLER [24], and SeViLA [33] decompose video reasoning into modular steps, including planning, temporal localization, evidence verification, and answer generation.

3 Object-Centric Spatio-Temporal Formulation

Video understanding is commonly formulated as a direct mapping:

$$A = F_\theta(V, Q), \tag{1}$$

where $V = \{f_t\}_{t=1}^{T}$ denotes a sequence of video frames and Q is a question. Under this formulation, the model encodes the entire video into a monolithic latent representation and performs reasoning implicitly within this space.

While effective for short clips, this formulation becomes increasingly fragile for long or complex videos. When V spans multiple scenes and events, compressing heterogeneous visual content into a single representation entangles object identity, temporal dynamics, and reasoning signals. Consequently, the model lacks an explicit mechanism to preserve persistent entities or to track how their states evolve over time.

To better characterize the requirements of robust video understanding, we adopt an object-centric spatio-temporal formulation. Instead of viewing a video as an unstructured sequence of frames, we represent it as a set of persistent entities:

$$\mathcal{O} = \{o_1, o_2, \ldots, o_K\}. \tag{2}$$

The notion of an *object* here is abstract and task-dependent, referring to any semantically coherent entity that can be consistently tracked over time (e.g., physical actors, scene elements, or higher-level constructs).

Each object $o \in \mathcal{O}$ is associated with a temporally ordered sequence of states defined over a discrete time domain $\mathcal{T} = \{t_1, \ldots, t_L\}$. We define its trajectory as:

$$\mathcal{S}_o = \{(t, s_t^{(o)}) \mid t \in \mathcal{T}\}, \tag{3}$$

where $s_t^{(o)} \in \mathcal{S}$ denotes the observable state of object o at time t. The state $s_t^{(o)}$ captures observable unary properties (e.g., visual attributes like `red shirt`, actions like `walking`) as well as binary relations or interactions with other entities (e.g., `passing to` o_2).

This formulation explicitly enforces temporal consistency: each object maintains a persistent identity across time, while its trajectory encodes its state evolution. Changes in state are thus disentangled from changes in identity. Aggregating over all objects yields a global spatio-temporal representation of the video:

$$\mathcal{S} = \{\mathcal{S}_o\}_{o \in \mathcal{O}}. \tag{4}$$

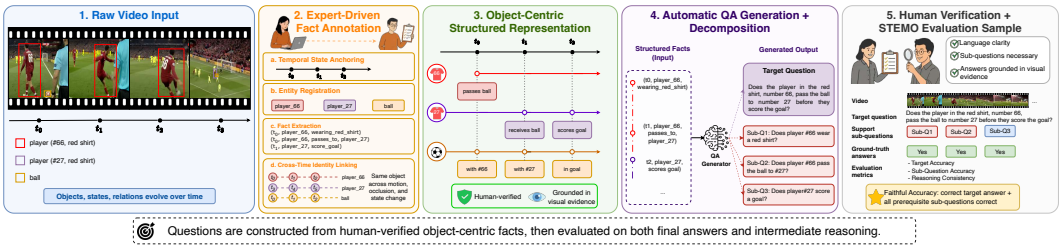

Figure 2: Overview of the STEMO-Bench construction pipeline.

Under this formulation, video answering shifts from implicit pattern matching to structured reasoning over object trajectories:

$$A = \Psi_Q(\mathcal{S}), \tag{5}$$

where Ψ_Q denotes a query-dependent reasoning operator. This perspective provides a principled abstraction for analyzing model behavior and designing evaluation benchmarks that target explicit spatio-temporal monitoring capabilities.

4 STEMO-Bench Construction

From the object-centric perspective, faithful video understanding requires maintaining an internal monitor over objects, states, interactions, and identities across time. However, existing datasets often fail to evaluate this capability. Many question-answer pairs can be solved by inspecting a single frame or relying on language priors, allowing models to bypass temporal reasoning entirely. Furthermore, evaluating models based solely on a final answer obscures whether the model genuinely resolved the underlying temporal dependencies or simply guessed correctly.

To address this gap, we propose **STEMO-Bench**, a **S**patio-**TE**mporal **MO**nitoring benchmark for evaluating faithful video understanding. STEMO-Bench pairs each complex target question with a set of atomic *sub-questions* that test the prerequisite facts necessary to justify the final answer. This design ensures that a model is only credited for a correct answer if its reasoning process is structurally sound, allowing us to distinguish faithful video understanding from shortcut learning. We construct STEMO-Bench using a rigorous, expert-driven pipeline. Readers can refer to Appendix A for illustrated examples from STEMO-Bench.

4.1 Expert-Driven Fact Annotation

To prevent the benchmark from inheriting hallucinatory details from automated generators, we decouple the extraction of the world state from the generation of the QA pairs. We utilize a closed pool of 50 trained annotators to construct explicit factual representations of the videos. The annotators were trained extensively on the required schema, resulting in an inter-annotator agreement score of 0.88 (Cohen's kappa), indicating a strong consensus on object identities and state transitions. Annotators follow a strict protocol to isolate directly observable visual evidence from subjective inferences:

1. **Temporal Anchoring:** Annotators define a discrete set of key timestamps $\mathcal{T}$ that capture salient moments of interaction, state transitions, or occlusions.
2. **Entity Registration:** Annotators identify all task-relevant entities $\mathcal{O}$ and assign them canonical, visually grounded identifiers (e.g., `leftmost_suitcase`, `player_red`).
3. **Fact Extraction:** At each timestamp t, annotators record observable states and relations. These are expressed as structured tuples, such as $(t, o_1, \texttt{wearing_red})$ for unary states, or $(t, o_1, \texttt{passing_to}, o_2)$ for interactions.
4. **Cross-Time Identity Binding:** To establish trajectories $\mathcal{S}_o$, annotators must manually trace entities across the temporal anchors. If an object undergoes severe occlusion or a state change, annotators explicitly link its identity across time.

This yields a human-verified, structured factual representation of the video's dynamics.

4.2 Automatic QA Generation and Decomposition

Given the annotated facts, we utilize a Large Language Model (LLM) to organize them into the global object-centric representation $\mathcal{S}$ and generate reasoning chains.

Target Question Generation. Instead of naively combining facts, the LLM constructs depth-bounded reasoning chains over the object trajectories. A valid chain $C = (f_1, \ldots, f_k)$ spans multiple facts connected by temporal order, shared entities, or identity links. To ensure complexity, chains must contain $k \geq 2$ facts, feature at least one multi-object interaction, and maintain temporal monotonicity ($t_1 \leq t_2 \leq \cdots \leq t_k$). The LLM then expresses this causal chain as a natural language target question.

For example, consider an annotated chain involving two players:

- f_1: $(t_1, \texttt{player_66}, \texttt{wearing_red_shirt})$
- f_2: $(t_1, \texttt{player_66}, \texttt{passes_to}, \texttt{player_27})$
- f_3: $(t_2, \texttt{player_27}, \texttt{scores_goal})$

The LLM generates the compositional target question: *"Does the player in the red shirt, number 66, pass the ball to number 27 before they score a goal?"*

Sub-Question Decomposition. To verify the reasoning process, the LLM decomposes the target question into atomic, supporting sub-questions mapped directly to the facts f_i in the chain. Answering these sub-questions correctly is logically prerequisite to answering the target query. For the example above, the sub-questions are:

1. *Does player number 66 wear a red shirt?* (Probing f_1)
2. *Does player number 66 pass the ball to player number 27?* (Probing f_2)
3. *Does player number 27 score a goal?* (Probing f_3)

4.3 Human Verification

Even with structured inputs, LLMs can introduce phrasing ambiguities or hallucinated priors. To ensure benchmark quality, annotators review the generated QA pairs alongside the source video. They verify that (1) the language is clear and natural, (2) the sub-questions are jointly necessary and sufficient to answer the target query, and (3) all answers are unambiguously grounded in visual evidence. Any sample featuring unsupported assumptions or unreliable identity bindings is discarded.

4.4 Evaluation Metrics

Because STEMO evaluates both the final prediction and the internal monitoring of facts, we introduce a multi-tiered evaluation scheme. For a target question q with ground-truth answer y and prediction $\hat{y}$, let $S_q = \{s_1, \ldots, s_m\}$ denote its supporting sub-questions with answers $y^{(j)}_{\text{sub}}$ and predictions $\hat{y}^{(j)}_{\text{sub}}$.

- **Target Accuracy (A_{target}):** Standard performance on the complex target questions:

$$A_{\text{target}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = y_i) \tag{6}$$

- **Sub-question Accuracy (A_{sub}):** Measures the model's foundational perception ability by computing accuracy across all sub-questions:

$$A_{\text{sub}} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{m_i} \mathbb{I}(\hat{y}^{(j)}_{\text{sub},i} = y^{(j)}_{\text{sub},i})}{\sum_{i=1}^{N} m_i} \tag{7}$$

- **Consistency (A_{cons}):** Evaluates whether the model maintains coherent intermediate reasoning when the final target prediction is correct. For each correctly answered target question, consistency is defined as the proportion of correctly answered sub-questions:

$$A_{\text{cons}} = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \frac{1}{m_i} \sum_{j=1}^{m_i} \mathbb{I}(\hat{y}^{(j)}_{\text{sub},i} = y^{(j)}_{\text{sub},i}) \tag{8}$$

  where $\mathcal{C} = \{i \mid \hat{y}_i = y_i\}$ is the set of target questions answered correctly by the model.

A model with high A_{target} but low A_{cons} is likely relying on spurious correlations rather than maintaining a robust spatio-temporal monitor.

4.5 Dataset Statistics

STEMO-Bench contains 88 curated videos across sports, egocentric, surveillance, and instructional domains, with 977 target questions and 2516 supporting sub-questions (3.25 per target on average). The videos average 183 seconds, requiring models to reason over long temporal horizons rather than short localized interactions.

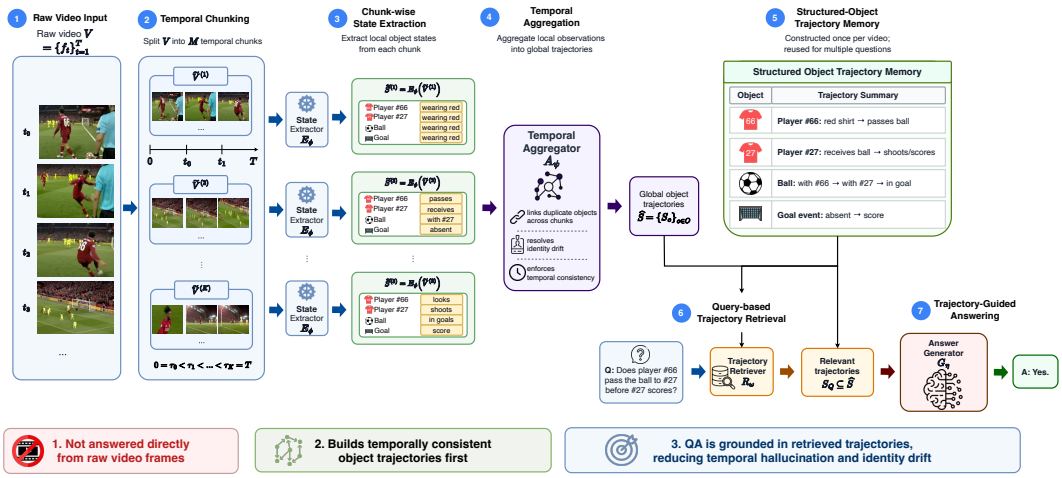

Figure 3: STEMO-Track divides a video into temporal chunks, extracts object states, aggregates them into consistent trajectories, and retrieves relevant trajectories to generate the final answer.

5 STEMO-Track

5.1 Overview

Building on the object-centric spatio-temporal formulation, we propose **STEMO-Track**, an object-centric video reasoning framework that first converts a video into structured object trajectories and then performs question answering over this representation. Rather than reasoning directly over raw frame sequences for every query, STEMO-Track separates video representation construction from query-specific reasoning. An overview of the proposed pipeline is shown in Figure 3.

Formally, given a video $V = \{f_t\}_{t=1}^{T}$, we first construct a set of temporally consistent object trajectories:

$$V \xrightarrow{E_\phi} \tilde{\mathcal{S}} \xrightarrow{\mathcal{A}_\psi} \hat{\mathcal{S}},$$

where E_ϕ is a VLM-based state extractor that produces local object-state observations from video chunks, and $\mathcal{A}_\psi$ is a VLM-based temporal aggregator that consolidates these observations across time. The resulting representation is:

$$\hat{\mathcal{S}} = \{\mathcal{S}_o\}_{o \in \mathcal{O}},$$

where each $\mathcal{S}_o$ denotes the trajectory of object o. Importantly, this trajectory construction is performed only once per video and can be reused across different questions.

Given a question Q, we retrieve the relevant trajectories and generate the final answer:

$$(\hat{\mathcal{S}}, Q) \xrightarrow{R_\omega} \mathcal{S}_Q \xrightarrow{G_\eta} A,$$

where R_ω selects the query-relevant subset of trajectories and G_η is a VLM-based answer generator that performs trajectory-guided answering.

This decomposition enables efficient reuse of video representations while allowing each question to reason over the relevant object identities, states, and temporal transitions.

5.2 Chunk-wise State Extraction

To handle long videos efficiently, STEMO-Track partitions the video into M disjoint temporal chunks:

$$V = \bigcup_{m=1}^{M} V^{(m)}, \quad V^{(m)} = \{f_t\}_{t \in \mathcal{I}_m},$$

where $\{\mathcal{I}_m\}_{m=1}^{M}$ forms a partition of $\{1, \ldots, T\}$.

For each chunk, we sample a subset of frames:

$$\tilde{V}^{(m)} = \{f_{t_j}^{(m)}\}_{j=1}^{N_l}.$$

A VLM-based state extractor E_ϕ processes each sampled chunk independently to identify visible objects and estimate their states:

$$\tilde{\mathcal{S}}^{(m)} = E_\phi(\tilde{V}^{(m)}),$$

where

$$\tilde{\mathcal{S}}^{(m)} = \left\{ \left(o_i^{(m)}, \{(t, s_t^{(o_i^{(m)})})\}_{t \in \mathcal{I}_m} \right) \right\}_{i=1}^{K_m}.$$

Here, $o_i^{(m)}$ denotes an object instance detected in chunk m, K_m is the number of detected objects in that chunk, and $s_t^{(o_i^{(m)})}$ denotes the estimated state of object $o_i^{(m)}$ at time t. This stage produces local object-state observations within each chunk. Since each chunk is processed independently, state extraction can be parallelized:

$$\tilde{\mathcal{S}} = \bigcup_{m=1}^{M} \tilde{\mathcal{S}}^{(m)}.$$

5.3 Temporal Aggregation

Local observations extracted from individual video chunks ($\tilde{\mathcal{S}}$) often contain duplicate objects, noisy state predictions, or fragmented identities across time. To construct a coherent, global set of object trajectories ($\hat{\mathcal{S}}$), STEMO-Track applies a VLM-based temporal aggregation module:

$$\hat{\mathcal{S}} = \mathcal{A}_\psi(\tilde{\mathcal{S}}).$$

This aggregator transforms the unlinked, noisy inputs ($\tilde{\mathcal{S}}$) into structured global trajectories ($\hat{\mathcal{S}}$) through a four-step identity-linking algorithm:

1. **Similarity Scoring:** For any two object observations o_i and o_j across different chunks, we compute a matching confidence score $c(o_i, o_j)$ based on visual attribute similarity and spatial proximity.
2. **Constraint Filtering:** Candidate links are only formed if they satisfy a temporal proximity constraint ($|t_i - t_j| < \Delta t_{max}$) and exceed a minimum confidence threshold ($c > \tau_{conf}$).
3. **Conflict Resolution:** In ambiguous cases where multiple candidate links satisfy the filtering criteria, we apply a bipartite matching protocol prioritized by temporal adjacency to ensure one-to-one identity mapping.
4. **Trajectory Generation:** Once linked, the observations are merged to form the final global set of trajectories. This global set is simply the collection of all individual object trajectories, defined as $\hat{\mathcal{S}} = \{\mathcal{S}_o \mid o \in \mathcal{O}\}$, where $\mathcal{O}$ represents the set of all unique objects found in the video.

Within this global set $\hat{\mathcal{S}}$, the specific trajectory $\mathcal{S}_o$ for a single, unique object o is defined as a temporal sequence of its states:

$$\mathcal{S}_o = \{(t, s_t^{(o)}) \mid t \in \mathcal{T}_o\}, \quad \forall o \in \mathcal{O},$$

where $\mathcal{T}_o$ denotes the timestamps at which object o is observed or inferred. Together, chunk-wise state extraction and temporal aggregation transform the raw video into a structured trajectory representation that is constructed once per video and reused for downstream tasks.

5.4 Query-based Trajectory Retrieval

Given a question Q, not all object trajectories are necessary for answering it. We therefore retrieve a query-relevant subset:

$$\mathcal{S}_Q = R_\omega(\hat{\mathcal{S}}, Q),$$

where $\mathcal{S}_Q \subseteq \hat{\mathcal{S}}$. The trajectory retriever selects the relevant objects, states, and temporal intervals needed to answer the question. By operating over structured trajectories rather than raw frames, this step can explicitly access object identities, state transitions, and temporal relations.

5.5 Trajectory-Guided Answering

Finally, a VLM-based answer generator G_η reasons over the retrieved trajectories and the question:

$$A = G_\eta(\mathcal{S}_Q, Q).$$

The answer generator produces the final response by composing information from the selected object trajectories, such as retrieving states at specific timestamps, comparing state transitions, or reasoning over interactions among multiple objects. This final stage performs query-specific reasoning only after the video has been converted into object trajectories.

6 Experiments

6.1 Experimental Setup

Benchmarks and metrics. We evaluate our approach on four benchmarks: STEMO-Bench, VideoHallucer [29], Video-MME [6], and EgoSchema [17]. In the main paper, we focus on STEMO-Bench using the metrics defined in Section 4.4, including Target Accuracy (A_{target}), Sub-question Accuracy (A_{sub}), and Consistency (A_{cons}). Results on the remaining benchmarks are provided in Appendix E.1.

Implementations. We instantiate two STEMO-Track versions to validate generalizability. **STEMO-Track (Gemini-3-Flash)** uses `gemini-3-flash-preview` for the state extractor E_ϕ, text filter R_ω, and answer generator G_η, with deterministic temporal concatenation. **STEMO-Track (Qwen)** uses `qwen/qwen3-vl-235b-a22b-thinking` for E_ϕ, `qwen/qwen3-235b-a22b` for R_ω, and `qwen/qwen3.5-27b` for G_η. Unless ablated, both pipelines use 15-second chunks, 60 frames per chunk, 64 frames for G_η, JSON chunk-prompt, answerer prompt, and identity linking.

Baselines. We compare against proprietary end-to-end MLLMs, open-weight end-to-end MLLMs, budget-aware frame-selection pipelines using Qwen3-VL-8B, and structured pipelines including VideoMind-7B, TraveLER, and SeViLA. Baselines receive a uniform 64-frame visual budget unless constrained by sequence length.

Table 1: STEMO-Bench results. Values are percentages. Best results are in **bold**; second-best results are underlined. Cons. denotes consistency (A_{cons}).

Method	STEMO-Bench		
	A_{target}	A_{sub}	A_{cons}
STEMO-Track			
STEMO-Track (Gemini-3-Flash)	**77.1**	**76.3**	**81.3**
STEMO-Track (Qwen3)	74.0	72.6	78.1
Proprietary end-to-end MLLMs			
Gemini-3.1-Pro	61.4	64.9	69.2
Gemini-3-Flash	48.7	61.7	69.9
Gemma-4-27B	55.5	60.0	72.7
GPT-5	32.5	43.2	70.5
Claude-4.6-Sonnet	48.3	62.9	73.2
Open-weight end-to-end MLLMs			
InternVL3-78B	50.0	66.1	71.0
Qwen3-VL-32B Instruct	59.8	70.0	78.1
Qwen3-VL-32B Think	65.0	70.6	78.1
Qwen3.5-27B	68.4	66.7	73.3
Qwen3.5-9B	50.5	49.0	57.9
Cosmos-Reason2-8B	54.7	74.7	83.7
Qwen3-VL-8B Think	61.3	70.1	77.4
Qwen3-VL-8B Instruct	34.3	59.8	77.4
VideoRFT-7B	38.6	55.1	50.8
Video-R1-7B	24.8	19.6	21.9
Budget-aware pipelines			
AKS	27.6	53.9	75.6
EFS	24.5	50.2	73.1
CLIP-Retrieval	26.2	52.9	72.8
Structured pipelines			
VideoMind-7B	50.3	37.9	43.2
TraveLER	50.6	41.8	49.3
SeViLA	51.4	59.6	56.8

6.2 Main Results

Table 1 shows that STEMO-Bench is challenging for frontier MLLMs, with end-to-end models generally lagging behind STEMO-Track. Under the single-frame setting, models drop to near-random performance (Appendix E.5), suggesting that shortcut cues are insufficient. STEMO-Track (Gemini-3-Flash) achieves 77.1% A_{target}, 76.3% A_{sub}, and 81.3% A_{cons}, outperforming Gemini-3.1-Pro by 15.7, 11.4, and 12.1 points, respectively. These results indicate that explicit trajectory construction is a key factor behind the gains, further supported by budget-matched comparisons (Appendix E.4).

To ensure rigorous evaluation of these gains, we report statistical significance for STEMO-Bench target accuracy using paired bootstrap tests (10,000 resamples). All reported improvements for our pipelines over their respective baselines are statistically significant ($p < 0.005$). For example, Pipeline (Gemini-3-Flash) significantly outperforms the strongest open-weight model (Qwen3.5-27B) with a 95% confidence interval of $[+6.8\%, +10.6\%]$ on Target Accuracy.

Furthermore, our framework transfers successfully to established video benchmarks (detailed results are provided in the Appendix E.1). As a brief overview, Pipeline (Gemini-3-Flash) achieves top overall results on VideoHallucer, with strong performance on Video-MME and EgoSchema, while Pipeline (Qwen3-VL-235B) offers a highly competitive open-weight alternative.

6.3 Ablation Study

We ablate key components of STEMO-Bench to evaluate their impact on target accuracy. While we focus on Clusters A–D below, the full analysis for Clusters A–H is provided in Appendix E.2.

Visual Budget (Clusters A & B) Answerer Budget: Lowering the budget to 8 frames barely impacts Gemini-3-Flash, as trajectory memory encodes key temporal cues, but severely degrades Qwen3-VL, which relies heavily on raw visual grounding.

Table 2: Ablation studies on STEMO-Bench Target Accuracy (%).

Ablation Focus	Pipeline	Chunk (s)	Frames/chunk	Frames@G_η	A_{target} (%)
Reference Base Config					
Base Configuration	Pipeline (Gemini-3-Flash)	15	60	64	77.1
Base Configuration	Pipeline (Qwen3-VL-235B)	15	60	64	74.0
Cluster A: Visual Budget at the Answerer					
Frames@G_η = 32	Gemini / Qwen	15	60	32	76.3 / 66.4
Frames@G_η = 16	Gemini / Qwen	15	60	16	76.5 / 65.5
Frames@G_η = 8	Gemini / Qwen	15	60	8	76.5 / 63.6
Cluster B: Visual Budget at the Extractor					
Frames/chunk = 30	Gemini / Qwen	15	30	64	76.6 / 65.2
Frames/chunk = 15	Gemini / Qwen	15	15	64	77.0 / 64.9
Frames/chunk = 8	Gemini / Qwen	15	8	64	74.9 / 62.6
Cluster C: Stage B Retrieval Ablations					
No filter (full timeline)	Gemini / Qwen	15	60	64	77.0 / 64.6
Random filter (top-k)	Gemini / Qwen	15	60	64	76.5 / 68.0
Cluster D: Temporal Aggregation					
Aggregator: LLM-summarize	Gemini / Qwen	15	60	64	76.7 / 62.7

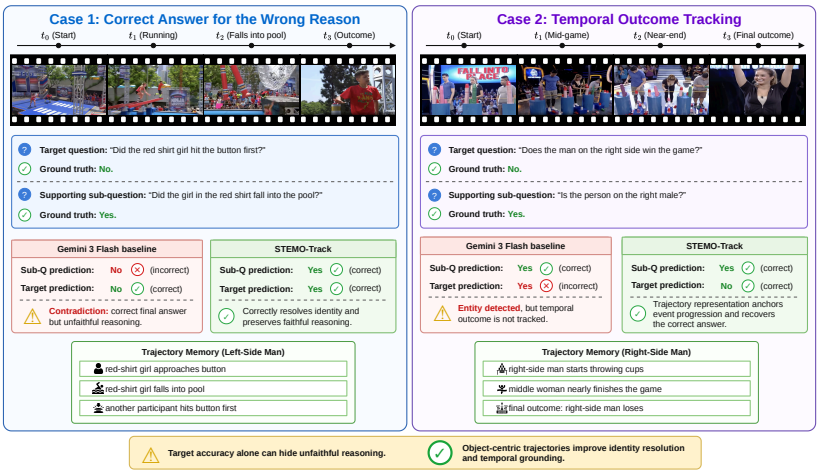

Figure 4: Qualitative analysis of target and sub-question behavior.

Extractor Density: Lowering density to 8 frames/chunk degrades both models, proving that sufficient sampling density is vital for robust state extraction.

Retrieval & Aggregation (Clusters C & D) Retrieval Filtering: Skipping filtering harms Qwen but not Gemini, indicating that open-weight models struggle more with timeline clutter. **Aggregation Method:** Deterministic concatenation outperforms LLM summarization by preventing the omission of low-salience, critical details.

6.4 Qualitative Analysis

STEMO's paired evaluation exposes shortcut reasoning hidden by target-only scoring. Baselines often pass target questions while failing prerequisite sub-questions. In Figure 4, Gemini-3-Flash answers a target query but fails the associated identity sub-question, whereas our method preserves identity binding. Similarly, our trajectory representation corrects baseline failures in temporal tracking. These cases prove target accuracy alone overestimates true video understanding.

7 Conclusion

In this work, we introduce STEMO-Bench, a benchmark for evaluating object-centric spatio-temporal monitoring in video Large Language Models. By decomposing target queries into supporting sub-questions, STEMO-Bench distinguishes faithful reasoning from coincidental correctness. To address the limitations of existing baselines, we further propose STEMO-Track, an object-centric framework that extracts chunk-wise object states and aggregates them into query-relevant trajectories. Experiments show that explicit trajectory construction substantially improves spatio-temporal reasoning consistency and reduces hallucinated predictions. Future work will focus on improving fine-grained visual attribute binding within object trajectories to further enhance the robustness of video MLLMs.

References

[1] Anthropic. Introducing claude sonnet 4.6. Anthropic, 2026.

[2] Kyungho Bae, Jinhyung Kim, Sihaeng Lee, Soonyoung Lee, Gunhee Lee, and Jinwoo Choi. Mash-vlm: Mitigating action-scene hallucination in video-llms through disentangled spatial-temporal representations. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13744–13753, 2025.

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[4] Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-star: Benchmarking video-llms on video spatio-temporal reasoning. *arXiv preprint arXiv:2503.11495*, 2025.

[5] Kaituo Feng, Kaixiong Gong, Bohao Li, et al. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*, 2025.

[6] 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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24108–24118, 2025.

[7] Hongcheng Gao, Jiashu Qu, Jingyi Tang, Baolong Bi, Yue Liu, Hongyu Chen, Li Liang, Li Su, and Qingming Huang. Exploring hallucination of large multimodal models in video understanding: Benchmark, analysis and mitigation, 2025.

[8] Yuansheng Gao, Jinman Zhao, Tong Zhang, Xingguo Xu, Han Bao, Zonghui Wang, and Wenzhi Chen. Mitigating hallucinations in video large language models via spatiotemporal-semantic contrastive decoding. *arXiv preprint arXiv:2601.22574*, 2026.

[9] Google. Gemma 4: Our most capable open models to date. Google Blog, 2026.

[10] Google DeepMind. Gemini 3.1 pro model card. Google DeepMind Model Card, 2026.

[11] Baiqi Li, Kangyi Zhao, Ce Zhang, Chancharik Mitra, Jean de Dieu Nyandwi, and Gedas Bertasius. Timeblind: A spatio-temporal compositionality benchmark for video llms. *arXiv preprint arXiv:2602.00288*, 2026.

[12] Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13723–13733, 2025.

[13] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. Videochat-flash: Hierarchical compression for long-context video modeling. In *The Fourteenth International Conference on Learning Representations*, 2026.

[14] Yun Li, Yiming Zhang, Tao Lin, XiangRui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. Stibench: Are mllms ready for precise spatial-temporal world understanding? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5622–5632, 2025.

[15] Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for temporal-grounded video reasoning. *arXiv preprint arXiv:2503.13444*, 2025.

[16] Hao Lu, Jiahao Wang, Yaolun Zhang, Ruohui Wang, Xuanyu Zheng, Yepeng Tang, Dahua Lin, and Lewei Lu. Elv-halluc: Benchmarking semantic aggregation hallucinations in long video understanding, 2025.

[17] Karttikeya Mangalam, Raiymbek Akshkulakov, and Jitendra Malik. Egoschema: a diagnostic benchmark for very long-form video language understanding. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[18] NVIDIA. Cosmos-reason2. NVIDIA Cosmos Documentation, 2026. Reasoning vision-language model for Physical AI and robotics.

[19] OpenAI. Gpt-4o system card. OpenAI, 2024.

[20] OpenAI. Introducing gpt-5. OpenAI, 2025.

[21] Qwen Team. Qwen3.5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

[22] Kanchana Ranasinghe, Xiang Li, Kumara Kahatapitiya, and Michael S Ryoo. Understanding long videos with multimodal language models. *arXiv preprint arXiv:2403.16998*, 2024.

[23] Ruchit Rawal, Reza Shirkavand, Heng Huang, Gowthami Somepalli, and Tom Goldstein. Argus: Hallucination and omission evaluation in video-llms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20280–20290, 2025.

[24] Chuyi Shang, Amos You, Sanjay Subramanian, Trevor Darrell, and Roei Herzig. Traveler: A modular multi-lmm agent framework for video question-answering. In *Proceedings of EMNLP*, 2024.

[25] Yiming Sun, Mi Zhang, Feifei Li, Geng Hong, and Min Yang. Smartsight: Mitigating hallucination in video-llms without compromising video understanding via temporal attention collapse. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 9251–9259, 2026.

[26] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Pinxin Liu, Mingqian Feng, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(2):1355–1376, 2026.

[27] Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *arXiv preprint arXiv:2505.12434*, 2025.

[28] Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2.5: Empowering video mllms with long and rich context modeling, 2025.

[29] Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models, 2024.

[30] Zikang Wang, Boyu Chen, Zhengrong Yue, Yi Wang, Yu Qiao, Limin Wang, and Yali Wang. Videochat-a1: Thinking with long videos by chain-of-shot reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 10467–10475, 2026.

[31] Xinhao Xu, Hui Chen, Mengyao Lyu, Sicheng Zhao, Yizhe Xiong, Zijia Lin, Jungong Han, and Guiguang Ding. Mitigating hallucinations in multi-modal large language models via image token attention-guided decoding. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1571–1590, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[32] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 12 2024.

[33] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. In *Advances in Neural Information Processing Systems*, 2023.

[34] Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, et al. Videorefer suite: Advancing spatial-temporal object understanding with video llm. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18970–18980, 2025.

[35] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun MA, Ziwei Liu, and Chunyuan Li. LLaVA-video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025.

[36] Jinguo Zhu, Weiyun Wang, Zhe Chen, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

[37] Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, Serena Yeung-Levy, and Xide Xia. Apollo: An exploration of video understanding in large multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18891–18901, 2025.

A STEMO-Bench Samples

Figure 5: **Example of the STEMO-Bench dataset.** To rigorously evaluate model faithfulness, STEMO-Bench requires models to correctly answer all underlying supporting sub-questions in addition to the complex target question. This ensures the model accurately tracks temporal action sequences—from passing (t_0) to scoring (t_2)—and grounds specific visual attributes (e.g., "yellow shirt", "number 10") rather than merely guessing the final answer.

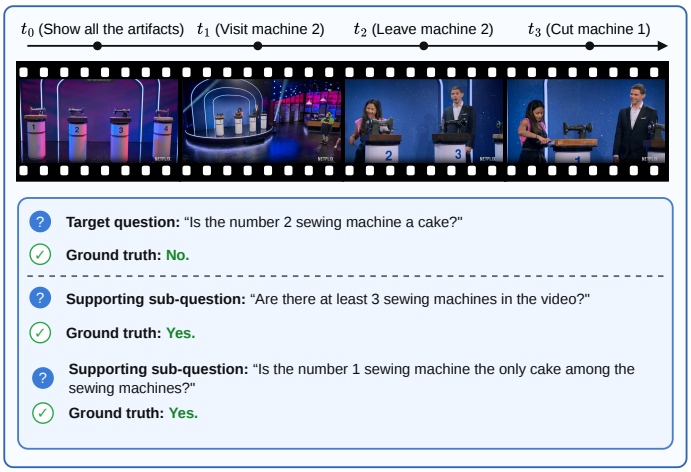

Figure 6: **STEMO-Bench evaluation on multi-step object interactions.** Building on our faithfulness criteria, models must resolve all supporting sub-questions to demonstrate true comprehension. In this scenario, the model must accurately follow a complex sequence—from the initial display of all artifacts (t_0) to cutting machine 1 (t_3)—while distinguishing specific visual attributes ("number 1", "number 2", "cake") to prevent shortcut learning.

Figures 5 through 8 present detailed examples from the STEMO-Bench. These samples showcase the complexity of the target questions and the corresponding step-by-step sub-questions required to verify a model's temporal and visual grounding capabilities.

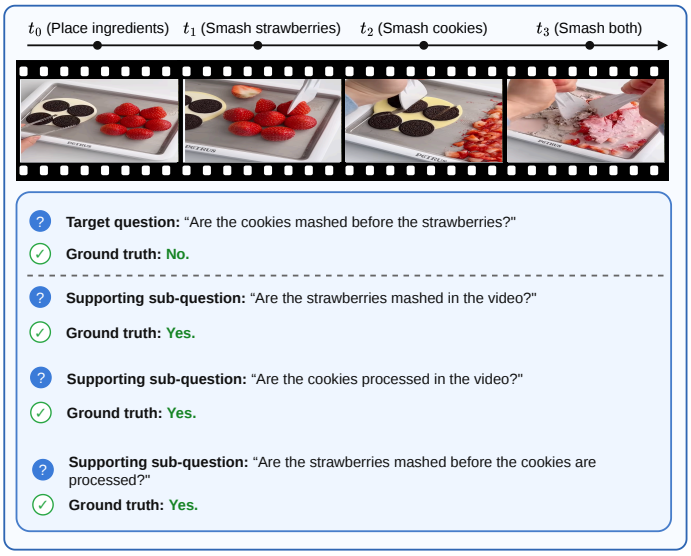

Figure 7: **STEMO-Bench evaluation of temporal ordering.** This example further illustrates how mandatory sub-questions enforce reasoning faithfulness. To succeed, the model must track the precise sequence of events—from placing ingredients (t_0) to smashing both items (t_3)—and ground specific visual objects ("strawberries", "cookies"). This verifies that the model understands the correct temporal ordering of interactions rather than relying on educated guesses.

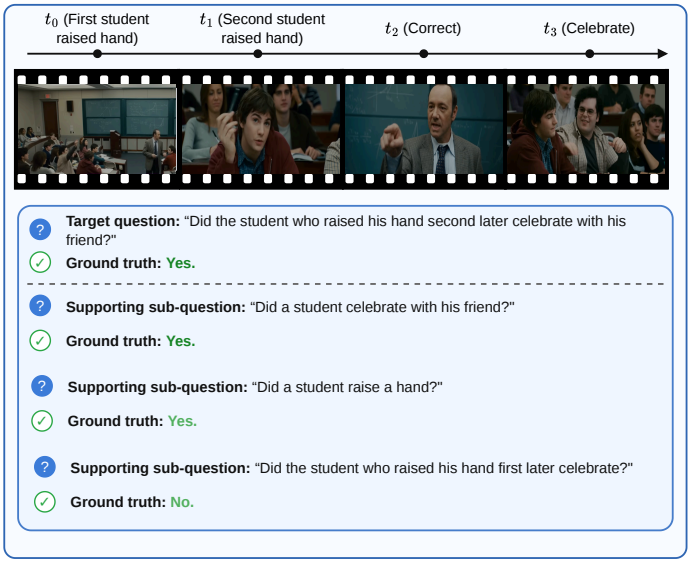

Figure 8: **STEMO-Bench evaluation of actor disambiguation over time.** This scenario highlights how STEMO-Bench tests complex temporal tracking across multiple subjects. To accurately answer the target question, the model must resolve the chronological sequence of events—from the first student raising his hand (t_0) and the second student raising his hand (t_1) to the eventual celebration (t_3). Mandatory supporting sub-questions force the model to explicitly differentiate between the two actors (e.g., verifying that the second student celebrated, while the first did not) to confirm precise spatiotemporal grounding rather than relying on spurious correlations.

B Implementation Details

The full pipeline configuration is defined with 15-second chunks, 60 frames per chunk sampled for E_ϕ, and 64 frames evaluated at G_η. Inference was conducted utilizing vLLM serving for all open-weight

models and native API caching for Gemini to handle long-context trajectory retrieval efficiently. Identity-linking across temporal boundaries is enabled by default in our structured JSON-based pipelines but explicitly disabled in unstructured prose variants.

C Question-Type Breakdown

To provide a granular view of model performance across diverse reasoning requirements, we disaggregate STEMO-Bench evaluation by target question type. As illustrated in Table 3, query types range from unary state tracking (e.g., identity preservation) to complex multi-object spatial dynamics. Our reference pipelines demonstrate substantial consistency advantages, particularly on long-horizon queries requiring sustained identity tracking.

Table 3: Consistency@All (A_{cons}) breakdown across STEMO-Bench question categories (%).

Model	State Change	Identity Tracking	Action Sequ.	Multi-Obj Interact.
STEMO-Track (Gemini-3-Flash)	82.3	78.4	84.1	79.5
Gemini-3.1-Pro	68.1	59.5	71.2	64.3
Qwen3-VL-32B Think	76.2	64.1	77.5	70.0

D Error-Type Breakdown

Complementary to the question-type taxonomy, we categorize models' failures at the sub-question level. Table 4 details the frequency of distinct sub-question error modes when the final target question is incorrect. Models reliant on standard implicit encoding exhibit uniformly high visual hallucination rates and temporal drift. In contrast, our structured trajectory approach minimizes localized visual hallucinations by leveraging the robust state extractor.

Table 4: Percentage of sub-question failure modes when Target Question is failed (%).

Model	Obj. Hallucination	State Misattribution	Temporal Order Error
STEMO-Track (Gemini-3-Flash)	15.2	42.1	42.7
Gemini-3.1-Pro	33.4	38.6	28.0
Qwen3-VL-32B Think	29.8	35.1	35.1

E Extensive results and full ablation studies of STEMO-Track

E.1 Performance on Other Video Benchmarks

As a supplement to the main text, Table 5 details the evaluation results of our proposed method on the VideoHallucer, Video-MME, and EgoSchema benchmarks. As shown in the Table, STEMO-Track achieves on-par or better results than existing video understanding methods.

E.2 Additional Ablation Study

To complement the primary analysis, this section details further ablations evaluating the impact of input modalities and temporal chunking granularity on STEMO-Track's performance and reports the results in Table 6.

Modality & Granularity (Clusters E & F). Removing either text trajectories or visual frames degrades performance, proving their complementary nature. For chunk granularity, expanding to 30s dilutes event-level state extraction, while shrinking to 7.5s fragments cross-chunk interactions; 15s provides the optimal balance.

Representation Form (Cluster G). Swapping structured JSON for prose summaries significantly degrades accuracy. Prose inherently obscures precise entity tracking, timestamps, and identity links, highlighting the necessity of structured state monitoring.

Table 5: VideoHallucer, Video-MME without subtitles, and EgoSchema results. Values are percentages. EgoSchema reports subset accuracy where available.

Method	VideoHallucer			Video-MME (w/o subs)				EgoSchema
	Basic	Halluc.	Overall	Short	Med.	Long	Over.	Acc.
STEMO-Track								
STEMO-Track (Gemini-3-Flash)	83.5	79.9	82.0	85.8	78.3	66.2	76.8	78.4
STEMO-Track (Qwen3-VL-235B)	80.6	75.1	77.9	82.2	76.0	62.7	74.1	76.4
Proprietary end-to-end MLLMs								
Gemini-1.5-Pro	83.6	42.3	37.8	81.7	74.3	67.4	75.0	72.2
GPT-4o	75.1	74.2	53.5	80.0	70.3	65.3	71.9	72.0
Gemini-3-Flash	66.8	45.4	58.1	31.6	24.6	17.1	24.4	37.0
Open-weight end-to-end MLLMs								
VideoChat2	29.7	25.8	7.8	48.3	37.0	33.2	39.5	63.6
Video-LLaVA	95.1	20.3	17.8	45.3	38.0	36.2	39.9	—
LLaVA-OV	84.1	60.8	48.5	76.7	62.2	60.0	66.3	60.1
Qwen3.5-27B	36.1	26.8	32.3	28.3	24.6	25.2	26.0	22.0

Table 6: Full ablation studies on STEMO-Bench Target Accuracy (%).

Ablation Focus	Pipeline	Chunk (s)	Frames/chunk	Frames@G_η	A_{target} (%)
Reference Base Config					
Base Configuration	Pipeline (Gemini-3-Flash)	15	60	64	77.1
Base Configuration	Pipeline (Qwen3-VL-235B)	15	60	64	74.0
Cluster A: Visual Budget at the Answerer					
Frames@G_η = 32	Gemini / Qwen	15	60	32	76.3 / 66.4
Frames@G_η = 16	Gemini / Qwen	15	60	16	76.5 / 65.5
Frames@G_η = 8	Gemini / Qwen	15	60	8	76.5 / 63.6
Cluster B: Visual Budget at the Extractor					
Frames/chunk = 30	Gemini / Qwen	15	30	64	76.6 / 65.2
Frames/chunk = 15	Gemini / Qwen	15	15	64	77.0 / 64.9
Frames/chunk = 8	Gemini / Qwen	15	8	64	74.9 / 62.6
Cluster C: Stage B Retrieval Ablations					
No filter (full timeline)	Gemini / Qwen	15	60	64	77.0 / 64.6
Random filter (top-k)	Gemini / Qwen	15	60	64	76.5 / 68.0
Cluster D: Temporal Aggregation					
Aggregator: LLM-summarize	Gemini / Qwen	15	60	64	76.7 / 62.7
Cluster E: Modality Drop-outs at G_η					
Text-only (no frames)	Gemini / Qwen	15	60	64	73.6 / 58.1
Frames-only (no $\hat{\mathcal{S}}$)	Gemini / Qwen	15	60	64	75.7 / 66.5
Cluster F: Chunk Granularity					
30s chunks	Gemini / Qwen	30	60	64	75.8 / 69.5
7.5s chunks	Gemini / Qwen	7.5	60	64	76.4 / 71.2
Cluster G: Form Representation					
Variant B: prose extractor, no R_ω	Gemini / Qwen	15	60	64	74.2 / 64.3
Variant C: prose extractor + prose R_ω	Gemini / Qwen	15	60	64	75.1 / 67.2
Cluster H: Trajectory Alternatives					
Dense captioning (prose state)	Gemini / Qwen	15	60	64	72.4 / 61.2
Shuffled-trajectory (no temporal order)	Gemini / Qwen	15	60	64	65.3 / 54.1

Dense Captioning and Shuffled-Trajectory Ablations (Cluster H). To verify the necessity of explicitly structured and temporally ordered trajectories, we introduce dense captioning and shuffled-trajectory ablations (Table 2, Cluster H). Replacing structured JSON states with dense prose captioning (dense captioning ablation) causes performance to drop to 72.4% for Gemini, confirming that unstructured text obscures temporal alignments and multi-object relations. Furthermore, shuffling the temporal order of extracted states (shuffled-trajectory ablation) yields a severe performance drop to 65.3%, underscoring that maintaining the correct temporal sequence is paramount for faithful object-centric reasoning.

Table 7: Budget-matched comparisons evaluating Consistency@All (%). All configurations utilize the Qwen3-VL backbone to isolate architectural contributions from model scale.

Constraint	End-to-End Baseline	Our Pipeline	Δ
Equal total frame budget (64 frames)	46.7	76.0	+29.3
Equal number of model calls (Single pass)	42.5	64.9	+22.4
Equal token budget	49.3	72.8	+23.5

E.3 Identity-Linking and Temporal Aggregation Ablation

Our temporal aggregation module relies on specific hyperparameters to build coherent object trajectories from chunked observations. We define a temporal proximity constraint Δt_{max} to restrict the maximum temporal gap between candidate observations, alongside a matching confidence threshold τ_{conf} based on visual attribute and spatial similarity. In ambiguous cases where multiple candidates exceed the threshold, our conflict resolution applies a bipartite matching algorithm prioritized by temporal adjacency to ensure strict one-to-one identity mapping.

Table 8 presents an ablation study on these hyperparameters using the Qwen3-VL-235B pipeline. Relaxing the temporal constraint ($\Delta t_{max} > 30$s) increases false-positive identity matches across distinct events, whereas an overly strict constraint ($\Delta t_{max} < 5$s) excessively fragments trajectories. Similarly, adjusting the confidence threshold τ_{conf} demonstrates a trade-off between trajectory completeness and identity precision. The base configuration ($\Delta t_{max} = 15$s, $\tau_{conf} = 0.75$) in conjunction with bipartite matching yields the optimal Consistency@All.

Table 8: Ablation of identity-linking thresholds and resolution rules on STEMO-Bench Consistency@All (%).

Configuration	A_{target} (%)	Cons.@All (%)
Base: $\Delta t_{max} = 15$s, $\tau_{conf} = 0.75$, w/ Bipartite Matching	74.0	76.0
Temporal Constraint (Δt_{max})		
$\Delta t_{max} = 5$s (Strict)	71.2	70.4
$\Delta t_{max} = 30$s (Relaxed)	72.8	73.1
Confidence Threshold (τ_{conf})		
$\tau_{conf} = 0.90$ (High Precision)	70.5	68.9
$\tau_{conf} = 0.50$ (High Recall)	71.8	71.2
Conflict Resolution Protocol		
w/o Bipartite Matching (Greedy distance-based)	69.4	67.5

E.4 Budget-Matched Comparisons

To ensure our performance gains stem from the object-centric architecture rather than disparate resource allocation, we conduct budget-matched comparisons against end-to-end MLLMs. We evaluate configurations maintaining an **equal total frame budget** (64 frames globally), an **equal number of model calls** (single-pass extraction and reasoning), an **equal token budget** (truncating trajectory context to match holistic frame embeddings), and an **equal backbone** (using Qwen3-VL for both the baseline and our pipeline). As detailed in Table 7, across all matched settings, explicit trajectory construction consistently yields superior Consistency@All, confirming that the structural prior of spatio-temporal monitoring drives the improvements, rather than brute-force scaling.

E.5 Single-Frame Solvability and Shortcut Audit

To verify that STEMO-Bench rigorously resists spatial and language priors, we conduct a single-frame solvability audit. We restrict the visual input of frontier MLLMs (e.g., Gemini-1.5-Pro, GPT-4o, and Qwen3.5-27B) to a single center frame of the video, forcing them to rely purely on static cues, statistical priors, and linguistic shortcuts. Under this extreme constraint, target accuracy across all tested models plummets to near random chance (e.g., $\sim 28 - 32\%$), and Consistency@All drops below 15% (see Table 9). This confirms that STEMO-Bench queries inherently demand multi-step spatio-temporal reasoning and cannot be resolved through single-frame shortcut heuristics.

Table 9: Single-frame solvability and shortcut audit. Performance metrics when models are restricted to reasoning over a single center frame compared to full video context.

Model	**Full Video**		**Single Center Frame**	
	A_{target} (%)	Cons.@All (%)	A_{target} (%)	Cons.@All (%)
Gemini-1.5-Pro	37.8	34.2	31.4	12.1
GPT-4o	53.5	48.1	32.7	14.3
Qwen3.5-27B	68.4	69.4	28.5	10.8

F Limitation

While STEMO-Track achieves critical gains in interpretability and spatio-temporal consistency, these advantages necessitate increased preprocessing overhead. Unlike conventional end-to-end MLLMs that directly ingest sampled frames, our architecture explicitly relies on a structured pipeline of chunk-wise state extraction, object-centric trajectory construction, and temporal aggregation prior to query resolution. This deliberate design fundamentally improves reasoning capabilities, but inherently introduces computational latency, particularly when processing extended videos or dense object-tracking scenarios. Consequently, the current framework trades strict real-time applicability for enhanced accuracy. We leave it for future work to improve the efficiency of the method.