STRAND: Benchmarking and Improving Object-Centric Spatio-Temporal Monitoring in Video Large Language Models

Anonymous authors
Paper under double-blind review

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 **STRAND** (*Spatio-Temporal Reasoning Audited by Necessary Decomposition*), 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. Crucially, we score models with *Faithful Accuracy*, an unconditional joint metric that credits a prediction only when the target answer and every prerequisite sub-question are correct, so that a model cannot inflate its score by being selectively consistent on the small subset of targets it happens to answer correctly. To address failure modes exposed by STRAND, we further propose **STRAND-Track**, an object-centric framework that explicitly constructs and reasons over structured object trajectories via chunk-wise state extraction and temporal aggregation. Extensive experiments, including comparisons against end-to-end MLLMs and modular video harnesses on a matched backbone, a configuration matched on frames, calls and tokens simultaneously, and a configuration matched on the answering prompt and its decoding settings, demonstrate that our object-centric framework significantly reduces hallucinated answers and improves spatio-temporal reasoning consistency over state-of-the-art MLLMs. Component ablations isolate cross-chunk identity linking as the step the gain rests on, while an oracle-trajectory reference shows that constructing the trajectories and reasoning over them leave comparable shares of the remaining error, so neither stage alone closes the gap.

1 Introduction

Multimodal large language models (Cui et al., 2024; Tang et al., 2025b; Zohar et al., 2025) have advanced rapidly in video understanding, with gains in semantic reasoning, instruction following and long-context modelling (Wang et al., 2026b; Zhang et al., 2024). However, they still hallucinate plausible but factually incorrect objects, relations and events (Lu et al., 2025; Gao et al., 2025). This matters most in deployment, where a correct answer rarely turns on one salient moment but on a coherent account of what persists and what changes across the video.

Video understanding therefore requires reasoning over a dynamic world in which entities persist, interact and change. Faithful interpretation depends on persistent spatio-temporal monitoring: identifying objects, maintaining their identities, tracking their states, and linking them to temporally ordered events (Yuan et al., 2025; Li et al., 2025b). Because each operation supplies context for the next, a confusion between similar objects propagates into later reasoning while the response stays fluent.

This is precisely the ability that current evaluations leave underdetermined. Work on video hallucination (Wang et al., 2024b; Li et al., 2025a; Rawal et al., 2025) has sharpened attention to unsupported claims, and mitigation methods suppress unlicensed statements (Xu et al., 2025; Bae et al., 2025; Sun et al., 2026). Other benchmarks raise difficulty instead, widening the temporal horizon (Mangalam

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

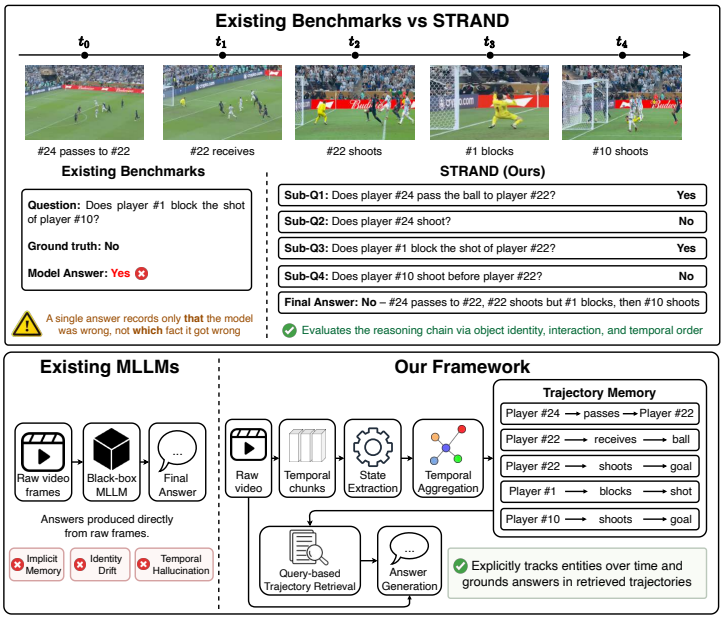


Figure 1: Existing benchmarks and MLLM architectures versus ours. *Top*: a single answer records only that a model is wrong, whereas sub-questions name the facts it rests on. *Bottom*: we answer from object trajectories rather than raw frames alone.

et al., 2023; Wang et al., 2024a), chaining evidence across grounded moments (Chen et al., 2025), or making the query object-centric (Wang et al., 2025a). These make evaluation more demanding but do not reveal the evidence path by which a model reaches its answer, and standard grading compounds the opacity by checking only the final answer (Rawal et al., 2025; Cheng et al., 2026; Li et al., 2026a). No amount of added difficulty repairs that criterion, since a model that reaches the right outcome after misbinding an object or reversing two events earns the credit given to one that reasons faithfully. What is missing is auditable evidence, not harder questions.

We therefore introduce **STRAND**, for *Spatio-Temporal Reasoning Audited by Necessary Decomposition*, which makes the intermediate facts part of the evaluation. Rather than generating question-answer pairs from raw video, it begins from a human-verified world model of each video: objects, their states, their relations, and the moments these change. Compositional target questions are composed from that structure and paired with atomic sub-questions probing the monitoring steps the target depends on, so we can ask not only whether a model answers the target but whether it also reports the supporting facts. We mark a target correct only when the model answers it and all of its sub-questions correctly, and faithful accuracy is the fraction of all targets so marked. The denominator includes every target, so a model cannot raise its score by answering only an easy subset.

Because STRAND credits an answer only when the facts supporting it are correct as well, we design **STRAND-Track**, an answering system that represents those facts explicitly rather than leaving them implicit. STRAND-Track represents a video as object trajectories, each recording how one object's state changes over time: it divides the video into temporal chunks, extracts object states within each, then links observations across chunks so that one object receives one identity. Given a question, it retrieves the trajectories that question depends on and answers from them. Because the representation is built once per video, the same trajectories serve every question about it.

Using STRAND, we evaluate current video models and STRAND-Track. Frontier models frequently answer a target correctly while getting its supporting facts wrong, so their faithful accuracy falls well below their target accuracy. STRAND-Track narrows this gap, and the gain persists under matched-budget controls that equalize the frame budget, the number of model calls, or the token budget in turn, under a control that equalizes all three at once, and under a control that gives the end-to-end baseline the same answering prompt and decoding settings our own answerer uses. Answering from trajectories alone costs the stronger backbone little accuracy, so the trajectories already carry what a question needs, whereas permuting their temporal order costs both backbones more than any component we ablate. The gain therefore rests on the temporal structure of the representation rather than on structured preprocessing alone, and disabling cross-chunk identity linking, which leaves every other

stage intact, is the one intervention that moves the stronger backbone outside a tie. Replacing predicted trajectories with human-verified ones recovers a substantial part of the remaining error but not most of it, so building better trajectories and reasoning more reliably over them are open problems of comparable size.

2 Related Work

Video Multimodal Large Language Models. Multimodal large language models have evolved from vision-language assistants into systems built for long-context temporal reasoning (Cui et al., 2024; Tang et al., 2025b), moving from instruction tuning on short video (Wang et al., 2026b; Zhang et al., 2024) to long-context scaling and temporal grounding (Li et al., 2026b; Wang et al., 2025b; Zohar et al., 2025). Open-weight models such as InternVL3, Qwen3-VL, Qwen3.5-Omni and Gemma-4 (Zhu et al., 2025; Bai et al., 2025; Team, 2026; Team et al., 2026), reasoning-focused video MLLMs such as Cosmos-Reason2, VideoRFT and Video-R1 (Azzolini et al., 2025; Wang et al., 2026a; Feng et al., 2026), and proprietary frontier models such as Gemini 3, GPT-5 and Claude-4.6-Sonnet (Team et al., 2023; Achiam et al., 2023; Liu et al., 2026a) are the systems we evaluate.

Video Hallucination Benchmarks and Mitigation. Existing benchmarks target different failure modes, including object hallucination (Wang et al., 2024b), action and temporal grounding errors (Li et al., 2025a), structural reasoning deficits (Rawal et al., 2025; Cheng et al., 2026), and temporal drift in long videos (Li et al., 2026a), while mitigation spans training-time alignment (Xu et al., 2025; Bae et al., 2025), inference-time decoding (Sun et al., 2026), and retrieval-augmented architectures (Gao et al., 2026; Lu et al., 2025; Gao et al., 2025).

Long-Video and Multi-Hop Reasoning Benchmarks. A parallel line of work raises task difficulty. EgoSchema (Mangalam et al., 2023) and LVBench (Wang et al., 2024a) widen the temporal horizon, MultiHop-EgoQA (Chen et al., 2025) chains evidence across grounded moments, and VideoInfer (Wang et al., 2025a) makes the query object-centric. These raise the bar on what a model must integrate, but they still grade the final answer alone, so a correct response does not show that the model followed the right entities through time. A complementary strategy attacks static bias at the input, pairing a clip with a temporally manipulated counterpart so appearance alone cannot decide the answer (Liu et al., 2024b; Li et al., 2024). That perturbs the video and still grades one answer, whereas STRAND leaves the footage untouched and changes the grading instead. Decomposed evaluation is not new either: AGQA-Decomp grades compositional consistency over automatically generated sub-question graphs (Gandhi et al., 2022) and V-STaR orders spatio-temporal sub-tasks (Cheng et al., 2026), but both decompose programmatic annotation and score consistency conditioned on the model's own correctness, whereas STRAND decomposes human-verified facts under an unconditional metric that cannot be raised by answering fewer targets. Appendix I compares these benchmarks with STRAND axis by axis.

Object-Centric and Structured Video Understanding. VideoRefer (Yuan et al., 2025) and STI (Li et al., 2025b) show that modelling spatio-temporal object interactions improves grounding, and structured pipelines such as SeViLA (Yu et al., 2023), TraveLER (Shang et al., 2024) and VideoMind (Liu et al., 2026b) decompose video reasoning into planning, temporal localisation, evidence verification and answer generation. A related line composes that decomposition programmatically, executing generated code over visual subroutines (Surís et al., 2023), or retrieves a query-conditioned slice of the video before answering (Xu et al., 2023; Ranasinghe et al., 2025). All of these re-read the video per query, since planning, retrieval and execution are driven by the question and nothing survives between questions about one video. STRAND-Track instead builds a query-agnostic object memory once and retrieves from it, which is what lets a single representation serve every question and what makes identity binding, rather than evidence selection, the step the design rests on. A further alternative drops the discrete bottleneck and keeps continuous temporal context inside the backbone (Liu et al., 2024a). The no-chunking control of Section 6.3 measures that regime on our own backbones.

3 Object-Centric Spatio-Temporal Formulation

Video understanding is commonly formulated as a direct mapping $A = F_\theta(V, Q)$, where $V = \{f_t\}_{t=1}^{T}$ is a sequence of frames and Q a question: the model compresses the video into one latent representation and reasons implicitly within it. This works for short clips, but when V spans multi-

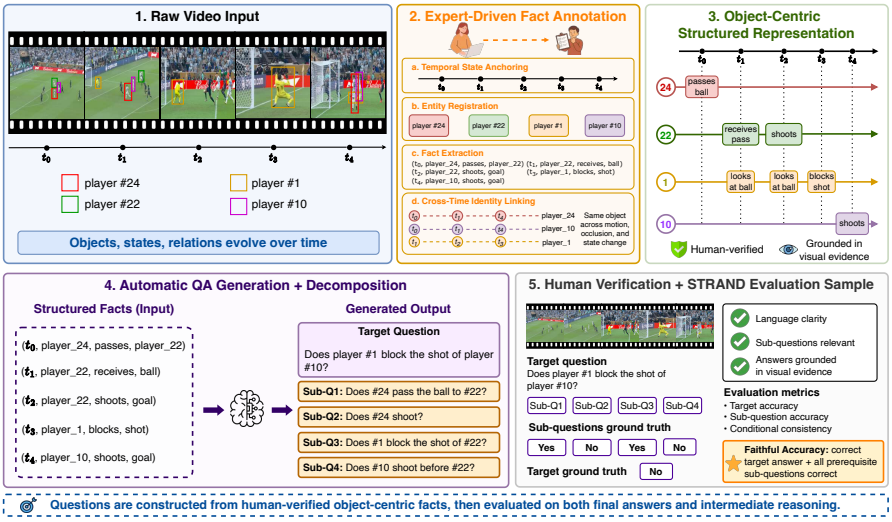


Figure 2: Overview of the STRAND construction pipeline. Annotators anchor timestamps, register entities, extract typed facts, and link identities across time. Target questions and sub-questions are composed from that record and verified.

ple scenes it entangles object identity, temporal dynamics and reasoning signal, leaving no explicit mechanism to preserve entities or track how their states evolve.

We therefore represent a video as a set of persistent entities $\mathcal{O} = \{o_1, \ldots, o_K\}$, where an object is any semantically coherent entity trackable over time. Each o carries a trajectory $\mathcal{S}_o = \{(t, s_t^{(o)}) \mid t \in \mathcal{T}_o\}$ and $\mathcal{S} = \{\mathcal{S}_o\}_{o \in \mathcal{O}}$, with $s_t^{(o)}$ the observable state of o at t, covering unary properties such as `red shirt` and typed relations such as `passing to` o_2 (Appendix H). Identity is thus separated from state, and video answering becomes structured reasoning over trajectories, $A = \Psi_Q(\mathcal{S})$, the abstraction both the benchmark and the framework are built on.

4 STRAND Construction

4.1 Expert-Driven Fact Annotation

To prevent the benchmark from inheriting hallucinatory details from automated generators, we decouple extraction of the world state from generation of the QA pairs, using a closed pool of 50 trained annotators to build explicit factual representations of the videos. Annotators train extensively on the required schema, and because the four annotation steps mix continuous and discrete judgements we report agreement per step. On the typed fact labels, the discrete variable the benchmark is built from, agreement reaches $\kappa = 0.88$ over 500 doubly labelled cells, with $\kappa = 0.85$ for cross-time identity linking, 0.89 F1 for entity registration, and a 0.35s mean offset for timestamp anchoring (Appendix E). Annotators follow a strict protocol to isolate directly observable visual evidence from subjective inferences:

Annotation proceeds in four steps: annotators anchor key timestamps $\mathcal{T}$ at salient interactions, state transitions and occlusions, register the task-relevant entities $\mathcal{O}$ with visually grounded identifiers such as `player_24`, record observable states and relations at each timestamp as typed tuples such as $(t, o_1, \texttt{passing_to}, o_2)$, and trace entities across the timeline to bind identities into trajectories $\mathcal{S}_o$.

4.2 Automatic QA Generation and Decomposition

Given the annotated facts, we use a Large Language Model (LLM) to organize them into the global representation $\mathcal{S}$ and generate reasoning chains. The generator operates only on the symbolic fact tuples and never sees the video, so its errors are ones of phrasing or unsupported inference rather than invented visual detail, and its role is limited to chain selection and surface realization. We use Llama-3.1-70B-Instruct, disjoint from every model family we evaluate, so no evaluated system shares pretraining lineage with the model that phrases its questions.

Target question generation. Rather than naively combining facts, the LLM constructs reasoning chains over the trajectories. Consider the annotated chain from the broadcast football sequence of Figures 1 and 2, in which #24 passes to #22 at t_0, #22 receives at t_1 and shoots at t_2, #1 blocks a shot at t_3, and #10 shoots at t_4. The target is *"Does player #1 block the shot of player #10?"*, with ground truth *No*: two players shoot and #1 blocks once, both attributions describe ordinary play, so only the position of the block distinguishes them, at t_3, after #22's shot and before #10 shoots.

Sub-question decomposition. The LLM decomposes the target into atomic sub-questions probing the facts it draws on: *Does player #24 pass the ball to player #22?* (*Yes*), *Does player #24 shoot?* (*No*), *Does player #1 block the shot of player #22?* (*Yes*), and *Does player #10 shoot before player #22?* (*No*). The third checks the attribution of the block and the fourth the order of the two shooters. Faithful accuracy requires all four, which is stricter than the minimal set the target logically needs. We compose a target under exactly this condition, whenever its answer turns on the order of two events and both orders stay consistent with the events themselves, and 361 of the 382 temporally constrained targets (94.5%) carry an ordering sub-question meeting it. The remaining 595 targets use a second template that drops the ordering requirement and makes the answer turn on a binding instead: which entity a state or typed relation belongs to at a named moment, or whether an entity seen at two separated timestamps is the same one. Both templates share the fact pool, the decomposition step and the verification pass, and Appendix F prints them verbatim.

4.3 Evaluation Metrics

Because STRAND evaluates the final prediction and the internal monitoring of facts, we use a multi-tiered scheme. For a target q with ground truth y and prediction $\hat{y}$, let $S_q = \{s_1, \ldots, s_m\}$ be 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, $A_{\text{target}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\hat{y}_i = y_i)$.
- **Sub-question Accuracy (A_{sub}):** Measures the model's foundational perception ability by computing accuracy across all sub-questions, $A_{\text{sub}} = \sum_{i=1}^{N}\sum_{j=1}^{m_i} \mathbb{I}(\hat{y}^{(j)}_{\text{sub},i} = y^{(j)}_{\text{sub},i}) \big/ \sum_{i=1}^{N} m_i$.
- **Faithful Accuracy (A_{faith}), primary metric:** Credits a question only when the target answer *and every* supporting sub-question are correct. It is computed over all N targets, without conditioning on the model's own correctness:

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

- **Conditional Consistency (A_{cons}), secondary:** The mean proportion of correctly answered sub-questions, taken only over $\mathcal{C} = \{i \mid \hat{y}_i = y_i\}$, the targets the model itself answers correctly: $A_{\text{cons}} = |\mathcal{C}|^{-1} \sum_{i\in\mathcal{C}} m_i^{-1} \sum_{j=1}^{m_i} \mathbb{I}(\hat{y}^{(j)}_{\text{sub},i} = y^{(j)}_{\text{sub},i})$.

Why A_{faith} and not A_{cons}. A_{cons} averages over $\mathcal{C}$, whose membership depends on the model, so two models are scored on different subsets. It is maximised by low-recall selectivity and does not fall when a model misses a target outright. A_{faith} has a fixed denominator, is monotone in both target and sub-question correctness, and is zero for a model that answers nothing, so a model with high A_{target} but low A_{faith} is likely exploiting spurious correlations rather than monitoring the scene. We report A_{cons} alongside it but state all consistency claims in terms of A_{faith}.

What A_{faith} does not certify. It grades separately elicited answers, so it is a property of outputs rather than a certificate about internal computation, and it requires every annotated sub-question rather than a minimal sufficient subset, so a model reaching the target along another valid route goes uncredited. Annotators find an alternative sufficient fact set for 5.8% of targets. It is thus a filter for comprehensive world-state monitoring, not a trace of necessary reasoning (Appendices G and H).

4.4 Statistics and chance baselines

STRAND contains 88 videos averaging 183 seconds, with 977 targets and 2516 sub-questions, or 2.58 per target and 11.1 targets per video. They are selected from 24,500 Wikimedia Commons files collected between January 1 and March 31, 2026, under six rules covering licence, duration, encoding quality, the presence of at least two dynamic entities, the absence of answer-revealing overlays, and

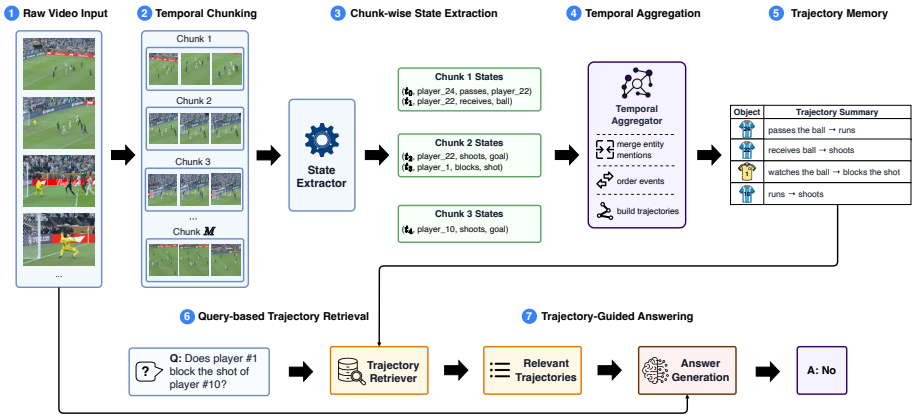


Figure 3: **Overview of STRAND-Track.** Stages 1 to 5 run once per video. Stages 6 and 7 run once per question.

near-duplicate removal. The dynamic-entity rule rejects more files than any other, and Appendix E breaks the rejections down rule by rule. The retained videos span sports, egocentric, surveillance and instructional footage. All items are yes/no, chosen so that sub-questions stay atomic and grading needs no judge, and labels are balanced by construction at 50.9% *Yes* for targets (497 of 977) and 51.6% for sub-questions.

Accuracies should be read against blind references rather than 50%: a majority-class predictor reaches $A_{\text{faith}} = 19.1\%$, a question-only predictor 21.4%, and a caption-only predictor 31.2%. Restricted to ordering sub-questions the question-only predictor falls to 49.3%, so event order has to be read off the video (Appendix H).

4.5 Human Verification

Because LLMs can introduce phrasing ambiguities or hallucinated priors, annotators review every generated QA pair alongside the source video, verifying that (1) the language is clear and natural, (2) the sub-questions are jointly sufficient to answer the target and each probes a fact the target draws on, and (3) all answers are unambiguously grounded in visual evidence. Any sample with unsupported assumptions or unreliable identity bindings is discarded. Annotators also record whether an alternative sufficient fact set exists, finding one for 5.8% of targets (Appendices L and H).

5 STRAND-Track: Object-Centric Trajectory Reasoning

Building on the object-centric formulation, we convert a video into structured object trajectories and answer over them rather than re-reading the full frame sequence per query. Given $V = \{f_t\}_{t=1}^{T}$, we build $V \xrightarrow{E_\phi} \tilde{\mathcal{S}} \xrightarrow{\mathcal{A}_\psi} \hat{\mathcal{S}} = \{\mathcal{S}_o\}_{o \in \mathcal{O}}$ once per video, reuse it across every question about that video, then retrieve and answer per question:

$$(\hat{\mathcal{S}}, Q) \xrightarrow{R_\omega} \mathcal{S}_Q, \qquad A = G_\eta(\mathcal{S}_Q, Q, \bar{V}), \tag{2}$$

where $\bar{V}$ is a fixed uniformly sampled frame budget.

5.1 Chunk-wise state extraction

We partition the video into M disjoint chunks $V = \bigcup_{m=1}^{M} V^{(m)}$, write $\tilde{V}^{(m)}$ for the N_l frames sampled from chunk m, and apply a VLM-based extractor E_ϕ to every chunk independently, so extraction parallelises. Writing $\tilde{\mathcal{I}}_m^{(i)}$ for the sampled timestamps at which object $o_i^{(m)}$ is visible and $s_t^{(i,m)}$ for its state at t,

$$\tilde{\mathcal{S}}^{(m)} = E_\phi(\tilde{V}^{(m)}) = \left\{ \left(o_i^{(m)}, \{(t, s_t^{(i,m)})\}_{t \in \tilde{\mathcal{I}}_m^{(i)}} \right) \right\}_{i=1}^{K_m}, \qquad \tilde{\mathcal{S}} = \bigcup_{m=1}^{M} \tilde{\mathcal{S}}^{(m)}, \tag{3}$$

with K_m the number of objects detected in chunk m. A state is relational rather than purely attributive, so $s_t^{(i,m)}$ holds the predicate the object participates in together with its arguments, which is what the tuples of Figure 3 report.

5.2 Temporal aggregation

Chunk-level observations contain duplicate objects, noisy states and identities fragmented across time. A deterministic, symbolic aggregator consolidates them, $\hat{\mathcal{S}} = \mathcal{A}_\psi(\tilde{\mathcal{S}})$ with $\mathcal{S}_o = \{(t, s_t^{(o)}) \mid t \in \mathcal{T}_o\}$ ordered in time. Here $\psi = (\Delta t_{max}, \tau_{conf})$ collects fixed hyperparameters, not learned weights, and $\mathcal{A}_\psi$ issues no model call in our default configuration. The LLM-summarize row of Table 2 is the ablation that does. Its four steps are similarity scoring over the fields the extractor emits, constraint filtering on temporal proximity and a confidence threshold, bipartite conflict resolution, and trajectory generation in which unmatched observations survive as singletons. Scoring is explicit rather than learned: the free-text `attribute` and `action` strings are case-folded into token sets and the relation arguments taken as a set, and the three Jaccard coefficients are combined as $c(o_i, o_j) = w_{\text{attr}} J_{\text{attr}} + w_{\text{act}} J_{\text{act}} + w_{\text{rel}} J_{\text{rel}}$ with fixed weights $(0.25, 0.25, 0.50)$, relation agreement carrying half the mass because it alone anchors an observation to a second registered entity. Appendix H works the score through an example. The step that binds observations across chunks is the one the design rests on, so Section 6.3 ablates it in isolation. Appendices H and D.3 state each step in full and report the threshold trade-off.

5.3 Retrieval and answering

Not every trajectory bears on a given question, so we retrieve a query-relevant subset $\mathcal{S}_Q = R_\omega(\hat{\mathcal{S}}, Q) \subseteq \hat{\mathcal{S}}$ and answer from it. Operating over structured trajectories rather than raw frames lets this stage address object identities, state transitions and temporal relations directly. $\bar{V}$ is a fixed uniformly sampled frame budget, independent of the question, so among the video-derived inputs only $\mathcal{S}_Q$ varies per query. Holding the prefix constant is deliberate: it makes the prefix cacheable across the questions asked of one video and keeps every configuration on one frame schedule, so a matched-budget control cannot be won or lost on frame selection. The cost is that a brief event may fall between sampled frames, leaving $\mathcal{S}_Q$ naming a fact $\bar{V}$ does not show. Fetching frames at the timestamps the trajectories carry is the natural remedy, excluded here because it would make the visual input query-dependent (Appendix G). For the stronger backbone retrieval acts mainly as a context-length control rather than a source of accuracy, though the weaker one is markedly more sensitive to it, as Appendix D.6 quantifies.

6 Experiments

6.1 Experimental Setup

Benchmarks and metrics. We evaluate on STRAND with the metrics of Section 4.3, and on Video-Hallucer (Wang et al., 2024b), Video-MME (Fu et al., 2025) and EgoSchema (Mangalam et al., 2023) in Appendix D.1.

Implementations. STRAND-Track (Gemini-3.5-Flash) uses `Gemini-3.5-Flash-preview` for E_ϕ, R_ω and G_η. **STRAND-Track (Qwen3-VL-235B)** uses `qwen3-vl-235b-a22b-thinking` for E_ϕ, `qwen3-235b-a22b` for R_ω and `qwen3.5-27b` for G_η. Unless ablated, both use 15-second chunks, 60 frames per chunk, 64 frames for G_η, and identity linking. Appendix F gives every prompt verbatim.

6.2 Main Results

Table 1 shows that STRAND is challenging for frontier MLLMs. No end-to-end model exceeds 68.4% A_{target}, several fall below 50%, and single-frame performance drops below chance for the models audited in Appendix D.5. VideoRFT-7B, Video-R1-7B and VideoMind-7B fall below the majority-class A_{faith} of 19.1%, so without reliable temporal tracking they cannot ground an answer in the facts it rests on. The Gemini-3.5-Flash instantiation reaches 59.4 A_{faith} and 77.1 A_{target}, ahead of Gemini-3.1-Pro by 20.5 and 15.7 points, a margin that sets the scale of the effect rather than isolating its cause.

Temporal ordering remains the bottleneck. Splitting A_{sub} by probe type locates the residual errors: the Gemini instantiation reaches 78.8% on sub-questions about individual facts but 61.4% on those about the order of two events, and Qwen 75.4% against 56.2%. The gap of roughly 18 points holds across both backbones, so even a system that builds explicit trajectories identifies events far more reliably than it orders them (Appendices B and C).

Table 1: STRAND results (%). **Best** and second-best per column. A_{faith} is primary, A_{cons} reference only (Section 4.3). Lat. is per-query latency amortised over $Q = 39.7$, * API-served. Four further models in Appendix H.

Method	STRAND				Lat. (s)
	A_{faith}	A_{target}	A_{sub}	A_{cons}	
STRAND-Track					
STRAND-Track (Gemini-3.5-Flash)	**59.4**	**77.1**	**76.3**	81.3	3.6*
STRAND-Track (Qwen3-VL-235B)	54.2	74.0	72.6	78.1	3.4
Proprietary end-to-end MLLMs					
Gemini-3.1-Pro	38.9	61.4	64.9	69.2	3.5*
Gemini-3.5-Flash	28.7	48.7	61.7	69.9	2.4*
GPT-5	20.0	32.5	43.2	70.5	4.1*
Claude-4.6-Sonnet	30.2	48.3	62.9	73.2	3.3*
Open-weight end-to-end MLLMs					
Gemma-4-27B	35.8	55.5	60.0	72.7	2.7
InternVL3-78B	29.7	49.9	66.1	71.0	3.0
Qwen3-VL-235B Think	47.9	66.8	71.4	77.6	3.2
Qwen3-VL-32B Think	46.1	65.0	70.6	78.1	2.9
Qwen3.5-27B	45.5	68.4	66.7	73.3	2.8
Cosmos-Reason2-8B	41.5	54.7	74.7	**83.7**	2.1
VideoRFT-7B	16.4	38.6	55.1	50.8	2.0
Video-R1-7B	4.3	24.8	19.6	21.9	2.0
Structured pipelines					
VideoMind-7B	18.9	50.3	37.9	43.2	4.4
TraveLER	22.0	50.6	41.8	49.3	4.0
SeViLA	26.1	51.4	59.6	56.8	3.2

Significance is assessed by paired bootstrap over A_{faith}, the primary metric (10,000 resamples, clustered at the video level), and each instantiation improves over the strongest end-to-end model of its family at $p < 0.005$. Against the protocol-matched control of Section 6.3 Gemini gains 8.2 points $[5.4, 11.0]$ and Qwen 3.5 $[1.2, 5.8]$. Disabling identity linking costs 7.8 $[4.5, 11.1]$ and 9.1 $[5.2, 13.0]$. The same tests over A_{target} agree throughout.

Transfer to external benchmarks. The Gemini instantiation leads the strongest baseline on VideoHallucer by 9.3 points and on EgoSchema by 1.6, and trails Gemini-3.1-Pro on Video-MME by 0.5, and the Qwen instantiation improves on its extractor backbone run end-to-end by 11.6 points on VideoHallucer. Since STRAND is built around object-centric trajectories, these check generality rather than claim state of the art (Appendix D.1).

Efficiency. STRAND-Track trades preprocessing for accuracy, but trajectory construction is *amortised*: the M extraction calls are paid once per video and reused across every query on it. With $Q = 39.7$ questions per video and an observed mean of $M = 12.7$ chunks, the $M + 2Q = 92.1$ calls are $2.3\times$ the baseline's 39.7 (Appendix D.7).

6.3 Ablation Study

Matched-budget comparison. A harness brings extra compute as well as a different representation, so we hold the backbone fixed and equalise one resource at a time, matching the baseline up to our budget. On the Qwen backbone, raising the end-to-end baseline from 47.9 A_{faith} to our frame, call and token budgets one at a time leaves us ahead by 3.9, 2.6 and 2.2, and matching all three *simultaneously* still leaves 2.1 points, narrowly above the 2.0-point tie threshold. Sub-question prompting leaves 3.5 and a same-backbone modular harness with no trajectories 4.4, so compute and modularity each explain part but a residual reduces to neither. Capacity runs in the baseline's favour throughout: every Qwen control answers with the 235B backbone against our 27B generator.

Separating the prompt from the representation. Resources are not all a harness changes. Our answerer's prompt asks explicitly for step-by-step reasoning about temporal order, at temperature 0.2, whereas Table 1 uses a generic prompt at temperature 0, and that difference is large: given the G_η prompt and decoding over raw frames, Gemini-3.5-Flash rises from 48.7 to 74.3 A_{target}, most of the way to our 77.1. It buys far less faithfulness, reaching 51.2 A_{faith} against our 59.4, and 50.7

Table 2: Same-backbone controls, ablations and an oracle reference, as A_{faith} / A_{target} (%). Backbone fixed within each column. Δ is the change in A_{faith} from the base row. Protocols in Appendix D.4.

Configuration	Gemini-3.5-Flash A_{faith} / A_{target}	Gemini-3.5-Flash Δ	Qwen3-VL-235B A_{faith} / A_{target}	Qwen3-VL-235B Δ
Base configuration	**59.4** / 77.1	–	**54.2** / 74.0	–
Same-backbone matched controls				
End-to-end, frame-matched (3,303 frames)	33.1 / 52.4	−26.3	50.3 / 68.9	−3.9
End-to-end, call-matched (92.1 calls)	35.6 / 55.0	−23.8	51.6 / 70.1	−2.6
End-to-end, token-matched (966K tokens)	34.8 / 54.1	−24.6	52.0 / 70.4	−2.2
End-to-end, jointly matched	36.4 / 55.8	−23.0	52.1 / 70.5	−2.1
End-to-end, protocol-matched (prompt, decoding)	51.2 / 74.3	−8.2	50.7 / 68.4	−3.5
Sub-question-prompted end-to-end	41.0 / 58.3	−18.4	50.7 / 68.4	−3.5
Generic modular harness, no trajectories	45.3 / 63.7	−14.1	49.8 / 68.1	−4.4
Component ablations				
w/o cross-chunk identity linking	51.6 / 70.2	−7.8	45.1 / 64.8	−9.1
No filter (full timeline)	59.3 / 77.0	−0.1	45.9 / 64.6	−8.3
Aggregator: LLM-summarize	58.9 / 76.7	−0.5	43.2 / 62.7	−11.0
Shuffled trajectory order	47.1 / 65.3	−12.3	37.4 / 54.1	−16.8
Oracle reference				
Human-verified trajectories at G_η	77.9 / 88.6	+18.5	72.4 / 85.1	+18.2

against 54.2 on Qwen, so we rest the representation claim on A_{faith} and note that on A_{target} alone most of the Gemini margin over Table 1 is prompting rather than trajectories.

The mechanism is backbone-dependent. Every figure here carries the metric it is measured in, since Table A4 reports A_{target} and the Δ column of Table 2 reports A_{faith}. For Qwen3-VL-235B every ablated component is load-bearing, an 8-frame answerer, no retrieval filtering and LLM-summarize aggregation costing 10.4, 9.4 and 11.3 points of A_{target}, whereas for Gemini-3.5-Flash the same interventions cost 0.6, 0.1 and 0.4, all ties. A simpler explanation therefore survives for the stronger backbone, that structured preprocessing with a capable model accounts for most of the gain. Shuffling the temporal order costs 11.8 of A_{target} ($77.1 \rightarrow 65.3$), separating temporal order from unordered structure but not object-centric from generic structure. The control that does is disabling cross-chunk identity linking alone, leaving extraction, schema, retrieval and answering untouched. It costs 7.8 points of A_{faith} on Gemini and 9.1 on Qwen, the one intervention moving the strong backbone outside a tie, so we restrict the strong per-stage claim to Qwen. Removing chunking entirely costs 8.7 and 12.8 points of A_{target}. Appendices D.2 and H give the full reading and name the controls we lack.

Where the remaining error sits. Oracle trajectories are worth 18.5 points of A_{faith} on Gemini and 18.2 on Qwen, bounding what better construction can buy, but the oracle row stops at 77.9 and 72.4, leaving 22.1 and 27.6 unreached with gold facts in hand. The larger share of the residual error therefore sits in answering, so we claim a comparable split rather than a construction bottleneck.

Representation quality and answer format. Scored against the human-verified trajectories, both instantiations place event ordering lowest of the five axes, so the deficit precedes answering, and disabling binding cuts identity linking by over 20 points while leaving registration and state untouched (Appendix J). On a contrastive subset of 300 perturbed sub-questions and an open-form subset of 200 probes, scores fall for every system while the ordering of systems is preserved (Appendix K). Figure A1 shows a failure qualitatively.

7 Conclusion

We introduce STRAND, a benchmark for object-centric spatio-temporal monitoring in video language models. By decomposing target queries into sub-questions and scoring them jointly under Faithful Accuracy, it separates faithful reasoning from coincidental correctness, and the framework we build on it aggregates chunk-wise object states into query-relevant trajectories. The design claim rests on same-backbone controls matched on frames, calls, tokens and the answering protocol, the shuffled-trajectory ablation, and an ablation disabling cross-chunk identity linking alone. An oracle reference splits the residual error comparably between building trajectories and reasoning over them, so neither stage alone closes the gap. Appendix G gives the preprocessing cost and the remaining limits. When explicit monitoring stops paying off, as backbones improve, remains open.

Ethics Statement

All 88 videos are drawn from Wikimedia Commons under free content licences, and we collect no footage ourselves. People appear incidentally in public settings. We make no attempt at identification, entities carry visually grounded identifiers such as `player_24` rather than names, and the annotation schema records no biometric or identity attribute. The annotator pool was recruited through Prolific and paid $15 per hour, above local minimum wage, and the protocol was approved by our institutional ethics board. Appendix E gives the full provenance, licensing and compensation record.

STRAND is a diagnostic benchmark for spatio-temporal reasoning. Because it rewards persistent tracking of people across time, it could be repurposed as a development target for surveillance or re-identification systems, which is not its intended use and which the annotation schema is deliberately designed not to support: it contains no coordinates, no appearance embeddings and no cross-video identity links.

Reproducibility Statement

Section 4.3 defines every metric in closed form and Appendix A states the serving configuration. Appendix F gives all seven prompts verbatim, for STRAND-Track and for every baseline, together with the determinism protocol, the two sampling temperatures and their seeds, and the one parser deviation we apply. Appendix D.4 states the protocol for each matched control separately, since a matched comparison is only as strong as its definition, and Appendix J gives the entity-matching rule used to score the constructed trajectories against the human-verified ones.

Two limits on reproducibility are worth stating plainly. Seeding is exact only for the open-weight models we serve ourselves through vLLM. For API-served models we pass the seed but the provider does not guarantee bit-exact reproduction, so those results are repeatable in distribution rather than byte for byte. And re-running the generator reproduces the *candidate* question pool rather than the released benchmark, because the human verification pass of Section 4.5 discards items failing its three checks. We therefore release the verified question set directly. [1].

Use of Large Language Models

In this work we used generative AI tools for one task that requires disclosure, the composition of benchmark items. Llama-3.1-70B-Instruct converts human-annotated symbolic fact tuples into natural-language target questions and their atomic sub-questions (Section 4.2). It operates on the tuples alone and never receives video, and its prompt is printed verbatim in Appendix F. Every item it produced was reviewed by human annotators against the source video and discarded if it failed the three checks of Section 4.5, so no generated item enters the benchmark unverified.

We did not use generative AI to propose or refine the research hypotheses, to design the methodology or the experiments, to develop theoretical models or conceptual frameworks, to formulate or prove mathematical claims, or to interpret results.

Large language models also appear inside the system under study, as the extractor, retriever and answerer of Section 5 and as the evaluated baselines of Section 6. Those are the object of the research rather than an authoring aid, and every model is named with its exact identifier in Section 6.1.

We have reviewed all AI-assisted output and take full responsibility for the content of this paper, including every claim, table and released artefact.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[1] The benchmark, the per-video licence and attribution manifest, the fact annotations and the evaluation code are available at https://github.com/... (the link is hidden now due to double-blind review)

Alisson Azzolini, Junjie Bai, Hannah Brandon, Jiaxin Cao, Prithvijit Chattopadhyay, Huayu Chen, Jinju Chu, Yin Cui, Jenna Diamond, Yifan Ding, et al. Cosmos-reason1: From physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025.

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)*, pp. 13744–13753. IEEE, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Qirui Chen, Shangzhe Di, and Weidi Xie. Grounded multi-hop videoqa in long-form egocentric videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2159–2167, 2025.

Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-star: Benchmarking video-llms on video spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9155–9164, 2026.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 958–979. IEEE, 2024.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 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)*, pp. 24108–24118. IEEE, 2025.

Mona Gandhi, Mustafa Omer Gul, Eva Prakash, Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Measuring compositional consistency for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5046–5055, 2022.

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. *arXiv preprint arXiv:2503.19622*, 2025.

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 e-prints*, pp. arXiv–2601, 2026.

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*, 2026a.

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)*, pp. 13723–13733. IEEE, 2025a.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. In *International Conference on Learning Representations*, volume 2026, pp. 109089–109117, 2026b.

Yun Li, Yiming Zhang, Tao Lin, XiangRui Liu, Wenxiao Cai, Zheng Liu, and Bo Zhao. Sti-bench: Are mllms ready for precise spatial-temporal world understanding? In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5622–5632. IEEE, 2025b.

Jiacheng Liu, Xiaohan Zhao, Xinyi Shang, and Zhiqiang Shen. Dive into Claude Code: the design space of today's and future AI agent systems. *arXiv preprint arXiv:2604.14228*, 2026a.

Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. St-llm: Large language models are effective temporal learners. In *European Conference on Computer Vision*, pp. 1–18, 2024a.

Ye Liu, Kevin Qinghong Lin, Chang-Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for temporal-grounded video reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 57481–57506, 2026b.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024b.

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. *arXiv preprint arXiv:2508.21496*, 2025.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Kanchana Ranasinghe, Xiang Li, Kumara Kahatapitiya, and Michael Ryoo. Understanding long videos with multimodal language models. In *International Conference on Learning Representations*, volume 2025, pp. 11810–11835, 2025.

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)*, pp. 20280–20290. IEEE, 2025.

Chuyi Shang, Amos You, Sanjay Subramanian, Trevor Darrell, and Roei Herzig. Traveler: A modular multi-lmm agent framework for video question-answering. *arXiv preprint arXiv:2404.01476*, 2024.

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, pp. 9251–9259, 2026.

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11888–11898, 2023.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29118–29128, 2025a.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(2):1355–1376, 2025b.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

Qwen Team. Qwen3.5-Omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Haochen Wang, Qirui Chen, Cilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, Weidi Xie, and Stratis Gavves. Object-centric video question answering with visual grounding and referring. *arXiv preprint arXiv:2507.19599*, 2025a.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *Advances in neural information processing systems*, 38:4350–4376, 2026a.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Lvbench: An extreme long video understanding benchmark, 2024a.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, et al. Internvideo2. 5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*, 2025b.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024b.

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, pp. 10467–10475, 2026b.

Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, and Yan Lu. Retrieval-based video language model for efficient long video question answering. *arXiv preprint arXiv:2312.04931*, 2023.

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 *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)*, pp. 1571–1590, 2025.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36:76749–76771, 2023.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18970–18980. IEEE, 2025.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

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

Table A1: Faithful Accuracy (A_{faith}) breakdown across STRAND question categories (%). The four categories partition the 977 target questions.

Model	State Change	Identity Tracking	Action Sequ.	Multi-Obj Interact.
Target count (n_k)	*200*	*277*	*300*	*200*
STRAND-Track (Gemini-3.5-Flash)	**70.00**	**49.82**	**52.00**	**73.00**
Gemini-3.1-Pro	50.00	31.05	32.67	48.00
Qwen3.5-27B	58.00	37.91	38.00	55.00

Table A2: Sub-question error types (%), as a share of each model's incorrect sub-questions. Rows sum to 100.

Model	Obj. Hallucination	State Misattribution	Temporal Order Error
STRAND-Track (Gemini-3.5-Flash)	**12.4**	34.1	53.5
Gemini-3.1-Pro	38.2	36.5	25.3
Qwen3.5-27B	33.7	38.4	27.9

A 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.

B Question-Type Breakdown

To provide a granular view of model performance across diverse reasoning requirements, we disaggregate STRAND evaluation by target question type. As illustrated in Table A1, query types range from unary state tracking, for example identity preservation, to complex multi-object spatial dynamics. The breakdown shows that identity tracking and action sequencing remain the hardest categories for every evaluated model. STRAND-Track provides a consistent uplift across all four, with the largest absolute gains against Gemini-3.1-Pro, averaging 20.8 points.

C Error-Type Breakdown

Complementary to the question-type taxonomy, we categorise failures at the sub-question level. Table A2 reports the composition of sub-question error modes when the target answer is wrong. Baselines that rely on implicit encoding spread their failures across visual hallucination and state misattribution, whereas STRAND-Track suppresses both. Because those modes fall away, temporal order errors become the dominant remaining failure for our strongest instantiation, at 53.5% with a 19.4-point margin over the next category. These are shares of each model's own errors rather than absolute rates, so the shift reflects what is left once the other modes are reduced, and it matches the ordering bottleneck of Section 6.2.

D Extensive results and full ablation studies of STRAND-Track

D.1 Performance on Other Video Benchmarks

As a supplement to the main text, Table A3 details the evaluation results of STRAND-Track on the VideoHallucer, Video-MME, and EgoSchema benchmarks. As shown in the table, STRAND-Track achieves results on par with or better than existing video understanding methods.

D.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 STRAND-Track's performance and reports the results in Table A4.

Table A3: Results on external benchmarks. † marks models we evaluate ourselves under the shared protocol of Appendix A. Unmarked rows reproduce figures reported by the original papers. ‡ marks the one model requiring the parser fallback of Appendix F. Basic and Hallucinated sub-columns are unmarked because Video-LLaVA's yes-bias inflates Basic. VideoHallucer Overall is the paired both-correct score, so Overall $\leq \min(\text{Basic}, \text{Halluc.})$. Video-MME Overall is the mean of the three subset columns.

Method	VideoHallucer			Video-MME (w/o subs)				EgoSchema
	Basic	Halluc.	Overall	Short	Med.	Long	Over.	Acc.
STRAND-Track								
STRAND-Track (Gemini-3.5-Flash)	83.5	**79.9**	**76.8**	**85.8**	**78.3**	66.2	76.8	**78.4**
STRAND-Track (Qwen3-VL-235B)	80.6	75.1	71.4	82.2	76.0	62.7	73.6	76.4
Frontier end-to-end MLLMs								
Gemini-3.1-Pro†	85.2	68.4	63.1	84.1	76.5	**71.3**	**77.3**	76.8
GPT-5†	84.5	72.1	67.5	83.4	75.8	69.4	76.2	75.4
Gemini-3.5-Flash†	81.2	58.5	52.3	79.5	71.0	62.5	71.0	68.5
Qwen3-VL-235B†	81.5	64.2	59.8	80.8	72.6	64.4	72.6	72.1
Earlier reference models								
Gemini-1.5-Pro	83.6	42.3	37.8	81.7	74.3	67.4	74.5	72.2
GPT-4o	75.1	74.2	53.5	80.0	70.3	65.3	71.9	72.0
Open-weight end-to-end MLLMs								
LLaVA-OV	84.1	60.8	48.5	76.7	62.2	60.0	66.3	60.1
Qwen3.5-27B†	76.4	52.3	45.1	75.2	64.8	55.3	65.1	61.4
Video-LLaVA‡	95.1	20.3	17.8	45.3	38.0	36.2	39.8	n/a
VideoChat2	29.7	25.8	7.8	48.3	37.0	33.2	39.5	63.6

Modality & Granularity (Clusters E & F). Removing either text trajectories or visual frames degrades performance, indicating that they are complementary. The frames-only row is an ablation *of STRAND-Track*, not a re-measurement of the end-to-end baseline, and the two must not be read against each other. It keeps the G_η prompt, which asks explicitly for step-by-step reasoning about temporal order, together with G_η's decoding temperature and its uniform 64-frame schedule, and removes only $\hat{\mathcal{S}}$. The Table 1 row for the same backbone instead uses the shared baseline prompt at temperature 0. The gap between them (75.7 against 48.7 for Gemini-3.5-Flash) therefore measures prompt and decoding, not representation. We flag its size because on A_{target} it is larger than any effect we attribute to the representation itself, and we now run the control that separates the two: the protocol-matched row of Table 2 gives an ordinary end-to-end baseline the G_η prompt, its temperature and its 64-frame uniform schedule, with no chunking, extraction or aggregation anywhere in the pipeline. It reaches 51.2 A_{faith} / 74.3 A_{target} / 69.8 A_{sub} on Gemini and 50.7 / 68.4 on Qwen. The frames-only ablation of STRAND-Track sits close to it, at 53.4 A_{faith} / 75.7 A_{target} / 72.1 A_{sub} for Gemini and 51.5 / 66.5 / 71.0 for Qwen. The two are near-identical by construction, differing only in that the ablation runs inside the STRAND-Track harness with an emptied state while the control is a plain single-pass baseline, and the residual difference of 2.2 points on Gemini and 0.8 on Qwen sits at or below our 2.0-point tie threshold, which is the outcome we would expect. The comparison that matters is therefore protocol-matched against base: on A_{target} the prompt recovers most of the gap (74.3 against 77.1), while on A_{faith} a gap of 8.2 points survives on Gemini and 3.5 on Qwen. Readers should take the A_{target} column of Table 1 as a protocol-uncontrolled comparison and the protocol-matched row as the controlled one. For chunk granularity, expanding to 30s dilutes event-level state extraction, while shrinking to 7.5s fragments cross-chunk interactions, and 15s provides the optimal balance.

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

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 A4, Cluster H). Replacing structured JSON states with dense prose captioning (dense captioning ablation) causes performance to drop to 72.4% for Gemini, confirming

Table A4: Full ablation studies on STRAND Target Accuracy (%).

Ablation Focus	Backbone	Chunk (s)	Frames/chunk	Frames@G_η	A_{target} (%)
Reference Base Config					
Base Configuration	STRAND-Track (Gemini-3.5-Flash)	15	60	64	77.1
Base Configuration	STRAND-Track (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 and Identity Linking					
Aggregator: LLM-summarize	Gemini / Qwen	15	60	64	76.7 / 62.7
w/o cross-chunk identity linking	Gemini / Qwen	15	60	64	70.2 / 64.8
Cluster E: Modality Drop-outs at G_η					
Text-only (no frames)	Gemini / Qwen	15	60	64	73.6 / 58.1
Frames-only (no $\hat{S}$)	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
Full context, no chunking, no aggregation	Gemini / Qwen	–	–	64	68.4 / 61.2
Cluster I: Oracle Reference					
Human-verified trajectories at G_η	Gemini / Qwen	15	–	64	88.6 / 85.1

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.

D.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 agreement of the extracted attribute, action and relation-argument fields. 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 A5 presents an ablation study on these hyperparameters using the Qwen3-VL-235B pipeline. Relaxing the temporal constraint to $\Delta t_{max} = 30$s increases false-positive identity matches across distinct events, whereas tightening it to $\Delta t_{max} = 5$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 Faithful Accuracy.

D.4 Same-Backbone Matched Controls

Table 2 holds the backbone fixed and equalises frames, calls and tokens, first one at a time and then all three together. A matched control is only as good as its definition, so we state each one.

Frame-matched. Two frame accountings are available and we state both. STRAND-Track *decodes* 762 unique frames at extraction (60 per chunk over an observed mean of $M = 12.7$ chunks) plus 64 at the answerer, so 826 unique frames per video. Across a full evaluation pass it *presents* $60M + 64Q =$

Table A5: Ablation of identity-linking thresholds and conflict resolution on STRAND. The base row reproduces STRAND-Track (Qwen3-VL-235B) in Table 1.

Configuration	A_{target} (%)	A_{faith} (%)
Base: $\Delta t_{max} = 15$s, $\tau_{conf} = 0.75$, bipartite	**74.0**	**54.2**
Temporal constraint Δt_{max}		
5s (strict)	71.2	50.4
30s (relaxed)	72.8	51.1
Confidence threshold τ_{conf}		
0.90 (high precision)	70.5	48.9
0.50 (high recall)	71.8	51.2
Conflict resolution		
w/o bipartite matching (greedy)	69.4	47.5

$762 + 64 \times 39.7 = 3{,}303$ frame inputs, against $64 \times 39.7 = 2{,}540$ for an end-to-end baseline. We match the baseline on the larger of the two: the frame-matched baseline decodes 3,303 uniformly sampled frames in a single pass, which exceeds what we ever hold in one context and is deliberately generous.

Call-matched. The baseline runs on a 92.1-call schedule matching $M + 2Q$. The video is split into the same 12.7 segments, each is summarised in one call, and each query is answered twice over the concatenated summaries with a majority vote resolving disagreement. Two samples tie on 12.4% of queries for Gemini and 15.6% for Qwen, and a third sample breaks the tie, so the baseline in fact runs slightly above the matched call budget. We leave those extra calls uncharged, which again favours the baseline.

Token-matched. The baseline receives the same 966K input-token budget, spent on denser uniform sampling plus a dense caption of the full video, so the comparison isolates what the tokens encode rather than how many there are.

Jointly matched. The three controls above relax one constraint each and leave the other two at the baseline's default, so none of them faces a fully equipped opponent. The jointly matched configuration applies all three at once. The video is split into the same 12.7 segments, each is summarised in one call over 60 densely sampled frames, and each query is then answered twice over the concatenated summaries with a majority vote, which reaches 826 unique frames, 92.1 calls and the 966K token budget together. The three resources overlap heavily, since extra calls and extra frames both buy coverage of the same timeline, so their effects do not add: the jointly matched baseline reaches 52.1 A_{faith} against 52.0 for the token-matched control alone. The surviving 2.1-point margin is the narrowest we report and sits just above our 2.0-point tie threshold, so we state it as a real but small effect rather than a decisive one.

Protocol-matched (prompt and decoding). The raw MLLM answers each question in a single pass over 64 uniformly sampled frames, exactly as in Table 1, but with the G_η prompt in place of the shared baseline prompt and with G_η's decoding settings ($T = 0.2$, `top_p` $= 0.95$, seed 42) in place of greedy decoding. No chunking, extraction, aggregation or retrieval is present. This control exists because the G_η prompt asks explicitly for step-by-step reasoning about temporal order, which is an intervention on the model independent of the representation, and Appendix D.2 shows it is worth more on A_{target} than the representation itself.

Sub-question-prompted end-to-end. The raw MLLM receives the target together with its generated sub-questions, answers all of them, and the target answer is read from its own response. Gold sub-question answers are never shown. On the Qwen backbone this control and the protocol-matched one land on the same figures to the precision we report (50.7 / 68.4). They are distinct configurations that happen to coincide, and the Gemini column separates them clearly (41.0 / 58.3 against 51.2 / 74.3).

Model capacity in the controls. Every same-backbone control answers with the full backbone in a single pass: `Gemini-3.5-Flash-preview` in the Gemini column and `qwen3-vl-235b-a22b-thinking` in the Qwen column. STRAND-Track's Qwen instantiation uses the 235B model only for extraction and retrieval and answers with the 27B `qwen3.5-27b` generator, so the controls hold roughly a ninefold parameter advantage at the reasoning stage. The asymmetry runs against us and we do not correct for it.

Table A6: Single-frame solvability and shortcut audit. Full-video columns reproduce Table 1. The last two rows receive no visual input at all rather than a single frame.

Model	Full video		Restricted input	
	A_{target}	A_{faith}	A_{target}	A_{faith}
Single centre frame				
Gemini-3.1-Pro	61.4	38.9	31.4	14.1
Gemini-3.5-Flash	48.7	28.7	29.8	12.5
GPT-5	32.5	20.0	28.2	10.2
Qwen3.5-27B	68.4	45.5	32.7	13.8
No visual input				
Question-only (Llama-3.1-70B)	–	–	52.6	21.4
Caption-only	–	–	59.8	31.2

Generic modular harness. A same-backbone harness on the same call budget, in which a planner proposes temporal windows, a retriever selects window-level captions, and the same G_η answers from them. It is modular and structured but maintains no persistent entity identity, which is the single property it lacks relative to STRAND-Track.

Without cross-chunk identity linking. Only step 3 of the aggregator is disabled. Chunk-wise extraction, the state schema, retrieval and answering are unchanged and chunk-local identifiers are retained verbatim, so an entity appearing in three chunks becomes three trajectories. Nothing else changes, which is what makes this an isolation rather than a bundled intervention.

Oracle trajectories. The human-verified trajectories replace the predicted $\hat{\mathcal{S}}$ at the input to R_ω, leaving retrieval and answering unchanged. This is an oracle-*input* reference rather than a mathematical upper bound, since an answerer given gold trajectories can still misread them, and the residual 22.1 points at $A_{\text{faith}} = 77.9$ are exactly that. The two residuals should be compared directly. Oracle trajectories are worth 18.5 points on Gemini ($59.4 \rightarrow 77.9$) and 18.2 on Qwen ($54.2 \rightarrow 72.4$), which bounds the gain available from better construction. What remains above the oracle row is 22.1 points on Gemini and 27.6 on Qwen, which bounds the gain available from better reasoning over correct facts. Since $22.1 > 18.5$ and $27.6 > 18.2$, the larger share of the remaining error sits in answering on both backbones, and we state the conclusion as a comparable split rather than as a construction bottleneck.

D.5 Single-Frame Solvability and Shortcut Audit

To verify that STRAND rigorously resists spatial and language priors, we conduct a single-frame solvability audit. We restrict the visual input of frontier MLLMs (Gemini-3.1-Pro, Gemini-3.5-Flash, GPT-5 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 falls to between 28.2% and 32.7%, below the 50% a coin would give on a binary task, and A_{faith} drops below 15% (see Table A6). This confirms that STRAND queries inherently demand multi-step spatio-temporal reasoning and cannot be resolved through single-frame shortcut heuristics, at least for the models audited.

D.6 Trajectory Retrieval Analysis

Query-based retrieval R_ω is the component whose contribution the ablations support least (Cluster C), so we characterise it directly rather than only through downstream accuracy. Because STRAND ships human-verified trajectories, retrieval can be scored against ground truth: for each target we take the annotated facts in its reasoning chain as the gold trajectory set, and measure the precision and recall of the trajectories $\mathcal{S}_Q$ that R_ω returns. This separates "retrieval selects the right evidence" from "retrieval improves the answer", which are distinct claims, and only the first is about the retriever.

Retrieval substantially reduces timeline clutter. The mean number of trajectory triplets per video is 145.2 before filtering and 24.6 per query after it. Against the gold chain facts, R_ω reaches 88.4% precision and 92.1% recall. On downstream impact, an error analysis finds that in 12.4% of queries the no-filter variant fails specifically because timeline clutter misleads the generator while the filtered variant succeeds. An atomic intervention in which single retrieved triplets are dropped shows that

removing one gold triplet takes target accuracy on those queries to 51.2%, indistinguishable from chance on a binary task, so the answerer does consume the retrieved evidence rather than ignoring it.

D.7 Efficiency and Inference Cost

STRAND-Track performs more model calls than a single-pass MLLM, and the accuracy–cost trade-off should be legible to a reader deciding whether to adopt it. The accounting separates two components. The *one-off* cost is trajectory construction, namely M extraction calls plus symbolic aggregation, paid once per video however many questions are asked. The *per-query* cost is one retrieval call plus one answering call. With Q questions per video the total is $M + 2Q$ calls against Q for an end-to-end baseline, so the amortised preprocessing cost is M/Q per question and falls as more questions are asked of the same video. On STRAND, with $Q = 39.7$ and M averaging 12.7 across the benchmark, this is a $2.3\times$ call overhead, 92.1 against 39.7. The mean video is 183 seconds and chunks are 15 seconds, which gives 12.2 in expectation. The observed mean is higher because a trailing partial chunk is retained and extracted rather than discarded, so a video is never truncated at its final chunk boundary.

Table A8 tabulates preprocessing wall-clock per video, per-query wall-clock, model calls, and input and output tokens for STRAND-Track and every measured baseline. Our method incurs higher token cost during extraction, while per-query latency stays close to an end-to-end call, at 2.9 seconds against 2.8 for the Qwen instantiation, because the per-query calls operate on retrieved trajectories and a fixed frame budget rather than on the full timeline.

Open-weight models are measured on $8\times$ H100 80GB GPUs via vLLM. Latencies are medians over 10 runs (IQR $< 5\%$), excluding decoding and sampling, and preprocessing runs at a concurrency of 4. API-served models, marked * in Table A8, were measured on August 10, 2026, and their latencies include network overhead, so they are not commensurable with local measurements. Qwen3.5-9B, Qwen3-VL-8B, EFS and CLIP-Retrieval are omitted because they do not appear in Table 1. The **Lat.** column of Table 1 is the per-query figure plus preprocessing divided by $Q = 39.7$, which is what a reader adopting the framework pays per question.

E Data Provenance, Licensing, and Annotation Ethics

Source. All 88 videos come from Wikimedia Commons. We collect no footage ourselves and use no material from other datasets or from video-sharing platforms.

Selection and curation. Candidates were harvested from the Commons video category tree between January 1 and March 31, 2026, giving a pool of $24{,}500$ files. Six rules were applied in order: *(i)* the file must carry a free content licence permitting redistribution, *(ii)* its duration must fall between 60 and 600 seconds, *(iii)* it must decode cleanly at a minimum of 480p and 24 fps, *(iv)* it must contain at least two entities that are both re-identifiable and dynamic in the sense of Appendix H, *(v)* it must carry no on-screen text, scoreboard or commentary overlay that states an answer directly, and *(vi)* it must not be a near-duplicate, re-encode or excerpt of an already accepted file, tested by perceptual hash over uniformly sampled frames. Table A7 gives the rejection count per rule. Rule *(iv)* is the dominant filter, rejecting more files than any other: most freely licensed Commons footage is single-subject or largely static, and such videos cannot support an identity-binding or ordering target. We did not stratify by domain, so the resulting mix of sports, egocentric, surveillance and instructional footage reflects what survives the rules rather than a designed quota, and the per-domain caveat of Appendix G applies for that reason.

Agreement, measured per annotation step. The four annotation steps produce different kinds of judgement, so a single agreement figure would be ill-defined and we report one statistic per step. Double annotation covers 20 of the 88 videos (22.7%), drawn at random, and yields 500 doubly labelled cells spread over 15 distinct annotator pairs, with no pair covering more than 3 videos, and a pair enters the per-pair statistics only if it shares at least 20 cells. *Timestamp anchoring* is continuous and is reported as an offset rather than as a κ: over 200 matched anchors the mean absolute offset is 0.35 seconds and 92.5% of anchors agree within $t \pm 1.0$s. *Entity registration* is a set-matching task with no fixed label inventory, so chance correction does not apply and we report F1, which is 0.89 over 50 registered entities. *Typed fact extraction* is the discrete labelling the benchmark is built from, and this is the variable the headline $\kappa = 0.88$ is computed on, with observed agreement $p_o = 0.90$

Table A7: Video curation. Rules are applied in order, so each count is of files rejected by that rule among those surviving the previous ones.

Stage	Files
Candidate pool (Wikimedia Commons, Jan 1 – Mar 31, 2026)	24,500
(i) rejected: licence does not permit redistribution	4,200
(ii) rejected: duration outside 60–600s	8,600
(iii) rejected: encoding or resolution below threshold	2,100
(iv) rejected: fewer than two dynamic, re-identifiable entities	8,700
(v) rejected: on-screen overlay reveals an answer	350
(vi) rejected: near-duplicate of an accepted file	462
Retained	**88**

against chance agreement $p_e = 0.17$ over the typed relation and state inventory. *Cross-time identity linking* is a binary same-entity decision over candidate pairs, giving $\kappa = 0.85$ over 150 candidate pairs with $p_o = 0.92$ and $p_e = 0.46$. The higher p_e reflects the binary label space and is why the same observed agreement yields a lower κ than fact extraction. Because fact labels are anchored to timestamps, the fact-level κ inherits the anchoring tolerance: widening it to $t \pm 2.0$s raises κ to 0.91 and narrowing it to $t \pm 0.5$s lowers it to 0.82, so the reported figure sits in the middle of that range rather than at a favourable end of it.

Licensing. Commons material carries free content licences, namely CC BY-SA, CC BY, CC0 or public domain. Licences vary per file, so we record the licence, the author and the source page for every video and ship that manifest with the release. We release the fact annotations and QA pairs under CC-BY-NC 4.0.

Redistribution. Because every source licence permits it, we redistribute the video files themselves alongside the annotations, with per-file attribution as those licences require. This removes the link-rot problem that affects benchmarks distributing URLs alone.

Personal data. The footage is material its uploaders published on Commons under a free licence, so it involves no recording of participants by us and no private data. People appear incidentally in public settings. We do not attempt identification, we assign entities visually grounded identifiers such as `player_24` rather than names, and the annotation schema records no biometric or identity attribute.

Annotators. The 50-annotator pool was recruited through Prolific and paid $15 per hour, above local minimum wage. The protocol and working conditions were approved by our institutional ethics board (protocol number withheld for anonymity and supplied on acceptance). Since the study collects no new footage and annotates only freely licensed public material, the review concerned the annotation task rather than filming.

Intended use. STRAND is a diagnostic benchmark for spatio-temporal reasoning in video language models. It is not intended for training surveillance or identity-tracking systems.

F Prompts

We give verbatim every prompt used across STRAND-Track and the evaluated baselines. To connect the structured output of E_ϕ to the symbolic tuples (t, o_1, s, o_2) of Section 3, the JSON fields map as follows: `timestamp` is t, `object_id` is o_1, `action` or `relation_type` is the predicate s, and `target_id` is o_2.

Chunk-wise state extractor E_ϕ (temperature 0.0, max tokens 1024)

```
Models: Gemini-3.5-Flash-preview or qwen/qwen3-vl-235b-a22b-thinking
SYSTEM: You are an expert video annotator. Identify and track distinct, dynamic entities
     across the provided frames.
USER: Analyse the provided video chunk. Extract all visible, dynamic objects and their
     states. Output strictly in this JSON schema:
{
  "objects": [
    {
```

```
      "object_id": "string (e.g. player_10)",
      "timestamp": "float (seconds)",
      "attribute": "string (e.g. wearing red)",
      "action": "string (e.g. running)",
      "relations": [{"target_id": "string", "relation_type": "string"}]
    }
  ]
}
```

Aggregator $\mathcal{A}_\psi$, LLM-summarize variant (temperature 0.0, max tokens 2048)

```
Used only in the Cluster D ablation. The default aggregator is symbolic and issues no model
     call.
Models: Gemini-3.5-Flash-preview or qwen/qwen3.5-27b
SYSTEM: You are an expert data aggregator.
USER: You are given concatenated chunk-level object state observations extracted from a
     video: {chunk_tuples}. Merge co-referent entities across time to form consolidated
     global trajectories. Resolve redundancies but DO NOT drop any state transitions or low
     -salience details. Output strictly in the exact same JSON schema as the input.
```

Query-based retriever R_ω (temperature 0.0, max tokens 1024, text only)

```
Models: Gemini-3.5-Flash-preview or qwen/qwen3-235b-a22b
SYSTEM: You are a trajectory retrieval agent.
USER: Given the following global object trajectories: {global_trajectories} and the target
     question: {target_question}, select the subset of trajectories required to answer the
     question. You must select and return whole trajectories exactly as they appear, not
     summarise or rewrite them. Limit your selection so that the total number of output
     triplets does not exceed 30. Output the selected subset in the exact same JSON format
     as provided.
```

Trajectory-guided answerer G_η (temperature 0.2, max tokens 512)

```
Models: Gemini-3.5-Flash-preview or qwen/qwen3.5-27b
SYSTEM: You are a temporal reasoning assistant. You will be given a set of extracted object
     trajectories and a few sampled visual frames.
USER: Using the structured trajectories {retrieved_trajectories} and the sampled frames,
     answer the question: {target_question}. First reason step by step about the temporal
     order of events, then give a final Yes or No answer.
```

QA generation and sub-question decomposition, temporally constrained template

```
 (temperature 0.7, max tokens 2048, text only)
Model: Llama-3.1-70B-Instruct
SYSTEM: You are an expert benchmark composer.
USER: You are given a sequence of human-verified fact tuples representing a video: {
     annotated_facts}. Do not use any external visual context or priors.
Task 1: Compose a compositional target question based ONLY on these facts. The target
     question must turn on the order of two events, where both temporal orders stay
     consistent with the events themselves.
Task 2: Decompose the target question into atomic sub-questions. These must be jointly
     sufficient to answer the target. Each sub-question must probe exactly one fact the
     target draws on. Do not generate any sub-question whose fact the target does not turn
     on.
Keep all entities as visually grounded identifiers (e.g. player_24). Format every item as a
     binary Yes or No question.
```

Baseline evaluation, all evaluated MLLMs (temperature 0.0, max tokens 256)

```
SYSTEM: You are a video QA assistant.
USER: Answer the following question based on the provided video frames: {target_question}.
     Reason step by step, then provide a final binary answer strictly as "Yes" or "No".
```

Prose extractor, Cluster G Variants B and C (temperature 0.0, max tokens 1024)

```
Identical to the $E_\phi$ prompt above except for the requested output format.
SYSTEM: You are an expert video annotator. Identify and track distinct, dynamic entities
     across the provided frames.
USER: Analyse the provided video chunk. Extract all visible, dynamic objects and their
     states. Output your observations as a dense prose paragraph rather than structured
     JSON. Record timestamps, attributes, actions and interactions for grounded entities
     such as player_24.
```

The template above produced the 382 temporally constrained targets. A second template, identical except that Task 1 reads "Compose a compositional target question based ONLY on these facts" with no ordering requirement, produced the remaining 595. Both share Task 2 verbatim. We print the temporal template because it is the one the ordering analysis of Section 6.2 rests on.

The baseline prompt is identical for every evaluated model, rendered through each model's own chat template with no per-model wording changes. One deviation is worth recording: Video-LLaVA does not reliably emit a terminating Yes or No, so for that model the extraction parser falls back to the first boolean token in the output. This is the parsing behaviour referenced in the footnote to Table A3.

Determinism protocol. Two prompts run at non-zero temperature and we state how each is controlled. The trajectory-guided answerer G_η runs at $T = 0.2$ with $\texttt{top_p} = 0.95$ and a fixed seed of 42. We take a single generation pass per query rather than a multi-sample vote, and those single-pass outputs are what produce the accuracies in Table 1. The QA-generation and decomposition prompt runs at $T = 0.7$ with a fixed seed of 1337, executed once over each video's fact tuples.

Two caveats bound what this buys. First, seeding is exact only for the open-weight models we serve ourselves through vLLM, where we control the sampler. For API-served models we pass the seed but providers do not guarantee bit-exact reproduction, so the Gemini instantiation is repeatable in distribution rather than byte for byte. Second, re-running the generator reproduces the *candidate* question pool, not the released benchmark, because the human verification pass of Section 4.5 discards items that fail its three checks. We therefore release the verified question set directly rather than asking users to regenerate it.

G Limitations

While STRAND-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, STRAND-Track explicitly relies on chunk-wise state extraction, object-centric trajectory construction, and temporal aggregation prior to query resolution. This design is associated with improved reasoning performance throughout our experiments, but it 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 framework.

The preprocessing cost is quantified in Appendix D.7: 92.1 calls per video against 39.7 for a single-pass baseline, a $2.3\times$ overhead that amortises as more questions are asked of the same video but is paid in full on the first one.

Beyond efficiency, we record 7 further limitations that bound what our results support.

Scale, and what it licenses. At 88 videos and 977 targets, aggregate comparisons are well resolved but per-domain claims are not (Section 4.3). We report domain breakdowns as descriptive only, and we do not claim that the observed ordering of models is stable within any single domain.

Faithfulness is observable, not certified. Faithful Accuracy scores the joint correctness of separately elicited answers. This is a property of what a system outputs, not proof that its internal computation followed the evidence path the sub-questions describe. A model could answer every sub-question correctly and reach the target by an unrelated route, and STRAND would credit it. The metric raises the bar a shortcut has to clear rather than eliminating shortcuts.

Co-design of benchmark and method. STRAND is organised around object-centric trajectories, and STRAND-Track constructs exactly that representation, so STRAND-Track is plausibly better matched to this benchmark than a generic method would be. The external-benchmark results (Appendix D.1) are the check on this, and we regard the generality claim as supported only to the extent those results hold up. A benchmark built around a non-trajectory representation would be a stronger test and we have not run one.

Binary answer format. Both targets and sub-questions are yes/no items, chosen for atomicity and judge-free grading (Section 4.4). This makes the benchmark vulnerable to answer priors in a way an open-ended format would not be, and many sub-questions are leading in the sense that they presuppose the entity they ask about. We control for this with label balancing and prior-only baselines, but the format remains a real restriction, and the contrastive and open-form subsets of Appendix K

bound the effect and show it does not reorder the systems, but they cover 300 and 200 items against 2,516 in the main set. Rebuilding STRAND in open form at full scale remains the most valuable single addition we can identify.

A query-agnostic frame budget. The answerer sees a fixed uniform sample of the video rather than frames chosen for the question (Section 5). This is what makes the visual prefix cacheable and what keeps the matched-budget controls free of a frame-selection confound, but a brief localised event can fall between sampled frames, so a retrieved trajectory may name a fact the frames do not show. Since retrieved trajectories carry timestamps, fetching frames at those timestamps is the obvious remedy and we have not evaluated it. We expect it to help most on the configurations where the visual budget is already binding, which Cluster A of Table A4 identifies as the Qwen instantiation rather than the Gemini one.

Baseline families we do not run. Our controls hold the backbone fixed, which rules in end-to-end, sub-question-prompted, protocol-matched and generic-modular comparisons but rules out two families that cannot be instantiated on the same backbone. Programmatic visual reasoning composes the decomposition as executable code over visual subroutines (Surís et al., 2023), and architectures that keep continuous temporal context avoid a discrete intermediate representation altogether (Liu et al., 2024a). The no-chunking control of Table A4 approximates the second on our own backbones, but neither family is evaluated in its original form, and a programmatic baseline in particular would test whether an executable decomposition reaches the same facts our trajectories do.

Margin under the strictest control. The jointly matched configuration of Appendix D.4 leaves 2.1 points on the Qwen backbone, barely clear of the 2.0-point threshold below which we report differences as ties. The design claim therefore survives the strictest resource control we can construct, but with a margin small enough that a different allocation of the same budget, or a stronger baseline harness, could plausibly close it. We do not claim more than the number supports.

H Extended discussion

This section collects material moved out of the main paper for length. Each item is referenced from the section it supports. Appendix I expands the benchmark positioning of Section 2, Appendix J scores the constructed trajectories against the human-verified ones, and Appendix K reports the contrastive and open-form probe sets. Appendices B and C break performance down by question category and by error type, Appendix G states what these results do not support, and Appendix L shows worked items from the benchmark.

H.1 Efficiency measurement protocol

Open-weight models are measured on $8\times$ H100 80GB GPUs via vLLM. Latencies are medians over 10 runs (IQR $< 5\%$), excluding decoding and sampling, and preprocessing runs at a concurrency of 4. API-served models, marked * in Table A8, were measured on August 10, 2026, and their latencies include network overhead, so they are not commensurable with local measurements. Gemma-4-27B, GPT-5, Claude-4.6-Sonnet, InternVL3-78B, Qwen3.5-9B, Qwen3-VL-8B, VideoRFT-7B, Video-R1-7B, EFS and CLIP-Retrieval are omitted because closed-API rate limits preclude commensurable profiling, or because we lack the local infrastructure to serve them at scale.

H.2 Which comparison supports the design claim

The structured-pipeline baselines run on substantially smaller backbones than either instantiation. VideoMind-7B employs a single Qwen2-VL-7B backbone (7B parameters) across its pipeline, TraveLER employs a LLaVA-1.5-13B (13B parameters) visual extractor and a Vicuna-13B (13B parameters) text planner, and SeViLA chains a localizer and an answerer that share a single BLIP-2 Flan-T5-XL backbone (3B parameters). By contrast, our Qwen instantiation allocates a 235B-parameter model for extraction (E_ϕ), a 235B-parameter model for retrieval (R_ω), and a 27B-parameter model for answer generation (G_η). Because the structural design of these baselines cannot be separated from the capacity of the models beneath them, and the STRAND-Track open-weight instantiation employs up to $78\times$ more parameters at a single stage than SeViLA does in total, these rows do not isolate the contribution of the harness, and we do not rest the design claim on them.

H.3 Conditional consistency is not comparable across models

Table 1 contains a direct illustration of the metric pathology discussed in Section 4.3. Cosmos-Reason2-8B reports the highest A_{cons} in the table (83.7) while answering 22.4 points fewer target questions correctly than the Gemini-3.5-Flash instantiation (54.7 vs. 77.1). This is not evidence that Cosmos-Reason2-8B monitors the scene better. It follows from A_{cons} being averaged over a model-dependent subset, so that a model answering fewer targets is graded only on the targets it found easiest. This is why we do not claim a consistency improvement on the basis of the A_{cons} column. The corresponding claim under the unconditional metric is made in terms of A_{faith}.

H.4 What the controls settle, and what they do not

Three counterfactuals were needed to separate the object-centric representation from extra compute, and Table 2 reports all three. A sub-question-prompted baseline tests whether decomposition alone recovers the gap without any explicit representation, and it does not, leaving 3.5 points on the matched backbone. A generic modular harness separates modularity from object trajectories and leaves 4.4. Disabling cross-chunk identity linking, while leaving chunk-wise structure intact, costs 7.8 points on Gemini-3.5-Flash and 9.1 on Qwen3-VL-235B. That third result is the one the mechanism claim turns on, because identity drift motivates the whole design and it is the only intervention moving the strong backbone well outside a tie.

Three things remain unsettled. Joint correctness of separately elicited answers is an observable property of outputs, not a certificate about internal computation. The generic harness is one instantiation of "structured but not object-centric", and a different planner might close more of the gap. And the jointly matched control clears our own tie threshold by 0.1 points, which is a real margin under the criterion we stated but not a comfortable one.

H.5 Two instantiations of the aggregator

(a) Deterministic concatenation with identity linking (default for all headline numbers): the four steps above are executed as symbolic operations over the extracted tuples, with no additional model call. "Concatenation" refers to the union of chunk-level observations *after* the linking steps have merged co-referent entities, not to a raw string concatenation of chunk outputs. *(b) LLM-summarize*: an LLM receives the concatenated chunk-level tuples and is asked to emit a consolidated trajectory set in the same schema, performing merging implicitly in natural language. Variant (b) is the Cluster D ablation and is consistently worse, which we attribute to the summariser dropping low-salience states that later prove decisive. Variant (a) cannot drop a state in this way because it never re-generates one.

Occlusion, re-entry, and look-alikes. The temporal proximity constraint $|t_i - t_j| < \Delta t_{max}$ bounds how long a trajectory may be interrupted before a re-appearance is treated as a new entity. This makes the failure modes explicit rather than silent: a long occlusion beyond Δt_{max} fragments one entity into two (a recall error), while visually similar entities within the window may be merged (a precision error), and the bipartite matching step exists to prevent the second failure from cascading into a many-to-one collapse. Section D.3 quantifies this trade-off over Δt_{max} and τ_{conf}. Extraction is *query-agnostic*: E_ϕ never sees the question, which is what permits the representation to be built once and reused, but also means that an entity no annotator would consider salient may be omitted before any query is posed.

H.6 What retrieval contributes

We are explicit that R_ω is the component with the weakest empirical support. Removing it entirely ("no filter, full timeline") costs the Gemini pipeline 0.1 points, which is a tie, and only the Qwen pipeline degrades meaningfully. The reading we adopt is that retrieval is a *context-length* mechanism rather than a reasoning mechanism: it matters when the answerer cannot attend reliably over the full trajectory set, and stops mattering once it can. We therefore do not claim retrieval as a source of the method's accuracy, and we quantify its behaviour directly, in terms of retrieved-trajectory precision and recall against the human-verified trajectories, in Appendix D.6.

H.7 What the answer generator sees

Here $\bar{V}$ is a uniformly sampled frame budget over the whole video, fixed in size and independent of the question, which is why G_η is a VLM rather than a text-only model. Among the video-derived

inputs only $\mathcal{S}_Q$ depends on the query, so the reusable part of the computation stays the trajectory memory. Section 6.3 reports a text-only variant in which G_η sees $\mathcal{S}_Q$ and Q alone.

H.8 On the uniqueness of a decomposition

Verification establishes that the annotated sub-questions form *a* sufficient reasoning path, not the only one, so a model could in principle reach the target answer by another route. The necessity check bounds this. Annotators discard any sub-question whose fact the target does not turn on, so each retained sub-question corresponds to a fact on the evidence path the target requires. We also probe facts rather than inference steps, since facts are route-invariant in a way that intermediate deductions are not, so an alternative route still has to resolve them. To quantify the residual risk, annotators record whether an alternative sufficient fact set exists for each target and find one for 5.8% of targets. In a separate elicitation study, three annotators independently select minimal fact sets for 100 sampled targets and agree exactly on 92.0% of them, with Fleiss' $\kappa = 0.87$. These figures indicate that the sub-questions probe route-invariant facts rather than one arbitrary reasoning path.

H.9 Answer Format and Chance Baselines

Format. Target questions and sub-questions are both posed as binary verification queries. We adopt this format deliberately. A sub-question must be *atomic* to serve as a prerequisite check, and it must be automatically gradable without a judge model whose own errors would contaminate the faithfulness measurement. Both requirements argue against open-ended generation at the sub-question level. The cost is that binary items are vulnerable to answer-prior effects, which we control for explicitly rather than leaving the format implicit.

Label balance. We balance the yes/no distribution at construction time by requiring the generator to emit, for each positive chain, a minimally perturbed negative counterpart, in which a single fact is replaced by one the annotators record as false at that timestamp. The resulting distribution is 50.9% *Yes* at the target level (497 of 977) and 51.6% *Yes* at the sub-question level (1,298 of 2,516). Raw and balanced accuracy therefore differ by less than 0.3 points for every model, and we report raw accuracy throughout.

Chance and prior-only baselines. Accuracies on STRAND should be read against three references rather than against 50%. A majority-class predictor reaches $A_{\text{target}} = 50.9\%$, $A_{\text{sub}} = 51.6\%$ and $A_{\text{faith}} = 19.1\%$. A question-only predictor, Llama-3.1-70B-Instruct given the question text but no video, reaches 52.6%, 53.8% and 21.4%, so language priors alone do not solve the task. Restricted to the ordering sub-questions it falls to 49.3%, no better than guessing on a binary task. Knowing which events occur therefore carries no information about the order in which they occur, and the relation has to be read off the video. A caption-only predictor, additionally given a detailed caption of the video, reaches 59.8%, 61.5% and 31.2%. Captions supply partial global context, and the margin that remains between them and systems with access to the frames is what the benchmark attributes to direct visual perception.

H.10 Dataset Statistics

STRAND contains 88 curated videos across sports, egocentric, surveillance, and instructional domains, with 977 target questions and 2516 supporting sub-questions (2.58 per target on average). The videos average 183 seconds, requiring models to reason over long temporal horizons rather than short localized interactions. Video sources, licensing, and the annotator compensation protocol are documented in Appendix E.

On scale and evaluation reliability. At 88 videos, STRAND is small by the standards of modern video benchmarks, so we state what this scale supports. The unit of evaluation is the question, and with 977 targets the standard error near 70% target accuracy is 1.5 points, so aggregate differences above 10 points sit well outside sampling noise. The scale does not support strong per-domain claims. At roughly 22 videos per domain the intervals are wide, and we report domain breakdowns as descriptive rather than as evidence of differential capability. Questions drawn from one video share an annotated world state and visual style, so they are not independent. We therefore cluster the bootstrap at the video level rather than resampling questions independently, using 10,000 resamples, which widens the 95% intervals relative to a naive question-level bootstrap. Under this scheme, ablation differences below 2.0 points are indistinguishable from zero, and we report them as ties.

H.11 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 $c(o_i, o_j)$ from the fields the extractor actually emits, namely agreement of the `attribute` string, agreement of the `action` string, and overlap in the relation arguments each observation participates in. The schema carries no coordinates, so no spatial term enters the score. Because E_ϕ emits those fields as free text, agreement is computed symbolically rather than semantically. Each string is lowercased, stripped of punctuation and split on whitespace into a token set, after which a fixed stop-list of eight schema-neutral words is removed. The relation field contributes the set of `target_id` values the observation participates in, which are canonical identifiers and need no normalisation. Each field yields a Jaccard coefficient $J(X, Y) = |X \cap Y|/|X \cup Y|$, with $J = 0$ by convention when both sets are empty, and the three are combined linearly,

$$c(o_i, o_j) = w_{\text{attr}} J_{\text{attr}} + w_{\text{act}} J_{\text{act}} + w_{\text{rel}} J_{\text{rel}}, \qquad (w_{\text{attr}}, w_{\text{act}}, w_{\text{rel}}) = (0.25, 0.25, 0.50). \quad (4)$$

   The weights are fixed once before any evaluation and never tuned per video or per backbone. Relation agreement carries half the mass because a shared typed relation anchors an observation to a second registered entity, which neither an attribute nor an action string does on its own. A worked example: `{red, shirt, number_24}` against `{red, shirt, striped, number_24}` gives $J_{\text{attr}} = 0.667$, `{running, forward}` against `{running}` gives $J_{\text{act}} = 0.500$, and a shared single relation argument gives $J_{\text{rel}} = 1.000$, so $c = 0.25(0.667) + 0.25(0.500) + 0.50(1.000) = 0.792$, which clears $\tau_{conf} = 0.75$. Had the relation argument been absent from one of the two observations, J_{rel} would fall to 0 and c to 0.292, below threshold, which is the sense in which the relation field dominates the decision. Malformed extractor output is handled before scoring: a chunk whose JSON fails to parse after one reformat retry is dropped, which occurs for 15 of the 1,118 chunks in the benchmark (1.3%), and its objects simply do not enter the aggregation.
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 and sorted by timestamp to form the final global set of trajectories. An observation that matches nothing is retained as a singleton trajectory rather than discarded. 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.

H.12 Efficiency accounting

All figures in Table A8 cover a complete evaluation pass over one video. Model calls follow each pipeline's structure. Single-pass end-to-end models issue Q calls, and STRAND-Track issues $M = 12.7$ extraction calls once per video plus two calls per query, one for text retrieval and one for vision-language generation, giving $M + 2Q = 92.1$. TraveLER issues $3Q = 119.1$, SeViLA and AKS each issue $2Q = 79.4$, and VideoMind issues $1 + Q = 40.7$. For input tokens we adopt a uniform rate of 256 tokens per frame across all systems, so that the comparison reflects structural cost rather than a resolution choice. An end-to-end baseline processes $2{,}540$ frames per video, or 39.7×64, for roughly $654{,}415$ input tokens including prompts. STRAND-Track processes $3{,}303$ frames, namely 762 at extraction plus 39.7×64 at generation, for roughly $966{,}093$ tokens once the retrieved state is counted. Per-query latency is correspondingly higher for STRAND-Track than for the matched open-weight baseline, at 2.9 seconds against 2.8, so the framework buys its 5.6-point gain in target accuracy over Qwen3.5-27B with more wall-clock time, more calls, and a larger token budget.

H.13 What counts as an object, and what a state contains

Because the abstraction above is deliberately generic, we make three commitments explicit. *(i) Entity selection.* $\mathcal{O}$ is not the set of all visible objects. An entity is registered only if it is *re-identifiable* (a human annotator can point to the same entity at two separated timestamps using only visual evidence) and *dynamic* (it participates in at least one state change or interaction over the video). Static

Table A8: Efficiency and cost per video ($Q = 39.7$ queries). Measurement protocol in Appendix D.7.

System	Latency (s)		Total cost per video ($Q = 39.7$)		
	Pre-proc.	Per-query	Calls	Input tokens	Output tokens
STRAND-Track (Gemini-3.5-Flash)*	14.2	3.2	92.1	970,336	18,100
STRAND-Track (Qwen3-VL-235B)	18.4	2.9	92.1	966,093	17,450
Gemini-3.1-Pro*	0.0	3.5	39.7	654,415	1,985
Gemini-3.5-Flash*	0.0	2.4	39.7	654,415	1,985
GPT-5*	0.0	4.1	39.7	654,415	1,985
Claude-4.6-Sonnet*	0.0	3.3	39.7	654,415	1,985
Gemma-4-27B	0.0	2.7	39.7	654,415	1,985
InternVL3-78B	0.0	3.0	39.7	654,415	1,985
Qwen3-VL-235B Think	0.0	3.2	39.7	654,415	4,760
VideoRFT-7B	0.0	2.0	39.7	654,415	3,180
Video-R1-7B	0.0	2.0	39.7	654,415	3,180
Qwen3.5-27B	0.0	2.8	39.7	654,415	1,985
Qwen3-VL-32B Instruct	0.0	2.9	39.7	654,415	1,985
Cosmos-Reason2-8B	0.0	2.1	39.7	654,415	1,985
VideoMind-7B	6.5	4.2	40.7	710,000	14,000
TraveLER	8.2	3.8	119.1	820,000	12,000
SeViLA	5.1	3.1	79.4	690,000	9,500
AKS	3.0	2.5	79.4	680,000	8,000

background elements are excluded, which keeps $|\mathcal{O}|$ tractable and keeps the annotation burden on the entities that carry temporal content. *(ii) State schema.* A state is a symbolic tuple, not a coordinate. Each $s_t^{(o)}$ is drawn from a three-part schema, `attribute` (appearance and persistent properties), `action` (the intransitive activity the entity is engaged in), and `relation` (a typed, directed link to another registered entity). We deliberately do not annotate bounding boxes: the failure mode we target is symbolic binding drift, not localisation error, and requiring boxes would restrict the benchmark to entities a detector can localise. *(iii) Uncertainty and partial observability.* $\mathcal{T}_o$ is the set of timestamps at which o is *observed*, and is not required to be contiguous. When an entity is occluded, leaves the frame, and re-enters, annotators record the gap explicitly and then assert an identity link across it only when re-identification is supported by visual evidence (e.g., a jersey number). Otherwise the two segments remain distinct entities. Facts that an annotator can infer but not observe are never recorded, which is what makes "the model should have known this" a well-posed claim.

H.14 Budget-aware frame-selection baselines

These three pipelines share a premise with STRAND-Track, that a fixed visual budget should be spent where the evidence is, but they spend it by selecting frames rather than by building a representation, and all three select per query. *AKS* is adaptive keyframe sampling (Tang et al., 2025a), which recursively partitions the timeline and allocates the frame budget to the partitions scoring highest for query relevance while keeping coverage of the whole video. *EFS* is the embedding-based frame selector we implement as a simpler reference point: frames are scored by cosine similarity between a CLIP image embedding and the question embedding, and the top-k are kept in temporal order, with no coverage constraint. *CLIP-Retrieval* retrieves the same budget by CLIP similarity (Radford et al., 2021) against a set of query-derived text prompts rather than the raw question, which is the standard retrieval-style frame selector. All three then answer in a single pass from the selected frames. The three budget-aware pipelines evaluated with a Qwen3-VL-8B backbone reach the following STRAND scores, in the order A_{faith}, A_{target}, A_{sub}, A_{cons}. AKS reaches 18.9, 27.6, 53.9 and 75.6. EFS reaches 15.9, 24.5, 50.2 and 73.1. CLIP-Retrieval reaches 16.9, 26.2, 52.9 and 72.8. All three fall below the majority-class A_{faith} of 19.1%.

H.15 Backbone dependence of the mechanism

The mechanism is backbone-dependent, and we do not overclaim it. Read column-wise, the Gemini and Qwen ablations tell different stories, and the difference is the most informative result in this table. For Qwen3-VL-235B every component ablated in Table 2 is load-bearing: dropping the answerer to 8 frames costs 10.4 points of A_{target}, removing retrieval filtering costs 9.4, and LLM-

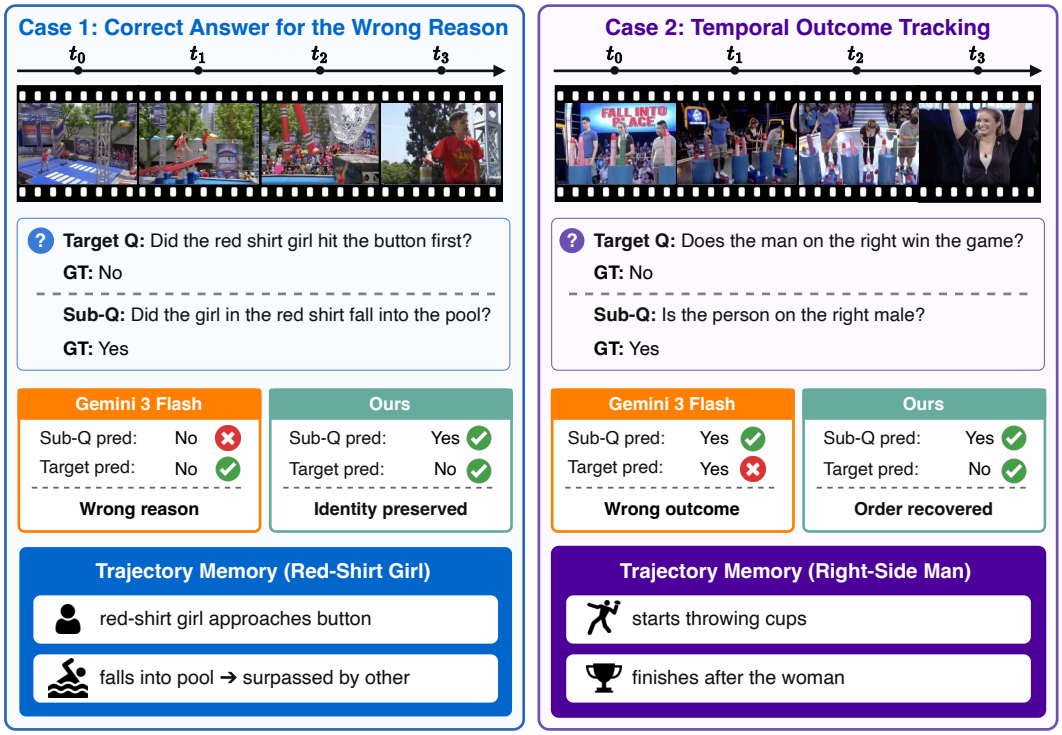


Figure A1: Qualitative analysis of target and sub-question behavior.

summarize aggregation costs 11.3. For Gemini-3.5-Flash the same three interventions cost 0.6, 0.1, and 0.4 points respectively, all within the confidence interval of a tie. The corresponding A_{faith} drops, read from the Δ column of Table 2, are 8.3 and 11.0 for Qwen on retrieval and aggregation and 0.1 and 0.5 for Gemini. Every figure below is labelled with the metric it belongs to, since the two are not interchangeable. The natural alternative hypothesis is that for a sufficiently strong backbone the benefit comes from performing *structured preprocessing with a capable model at all*, rather than from object-centric trajectories or retrieval specifically, and the ablations above do not refute it. Two results do bear on it: text-only answering still reaches 73.6 for Gemini without any frames at the answerer (Appendix D.2, Cluster E), which shows the extracted representation carries most of the information a question needs rather than merely reorganising the answerer's visual input. Shuffling the temporal order of that same representation costs 11.8 points of A_{target} (77.1 $\rightarrow$ 65.3, Appendix D.2, Cluster H), which shows the ordering of the structure, not just its existence, is what the answerer consumes. The shuffle ablation separates temporal order from unordered structured preprocessing. What separates object-centric structure from generic structure is disabling cross-chunk identity linking alone, since the ablated system still performs chunk-wise structured preprocessing with the same capable model and differs only in whether observations are bound to persistent identities. It costs 7.8 points of A_{faith} on Gemini and 9.1 on Qwen, the one intervention moving the strong backbone well outside a tie. We accordingly restrict the strong form of the claim, that each pipeline stage contributes, to the Qwen setting, and state the Gemini claim as the weaker one that temporal ordering of the representation is necessary.

I Benchmark Comparison

STRAND is not the first benchmark to use long videos, object-centric queries or annotated evidence. Table A9 separates the axes. EgoSchema and LVBench widen the temporal horizon but grade one answer. MultiHop-EgoQA annotates the temporal windows supporting an answer and evaluates localisation of that evidence, which is closest to our aim, but the evidence is graded as a span rather than elicited from the model as an independently answerable question. VideoInfer makes queries object-centric with grounding and referring annotations, again scored against the final response. VideoHal-

Table A9: Where STRAND sits relative to long-video, multi-hop, object-centric and hallucination benchmarks. *Evidence annotated*: the supporting evidence exists in the annotation. *Evidence elicited*: the model is asked for it and graded on it. *Joint scoring*: a target is credited only when the answer and all of its prerequisites are correct.

Benchmark	Avg. len. (s)	Object-centric	Evid. annot.	Evid. elicited	Joint scoring
EgoSchema (Mangalam et al., 2023)	180	✗	✗	✗	✗
LVBench (Wang et al., 2024a)	4,101	✗	✗	✗	✗
MultiHop-EgoQA (Chen et al., 2025)	574	✗	✓	✗	✗
VideoInfer (Wang et al., 2025a)	34	✓	✓	✗	✗
VideoHallucer (Wang et al., 2024b)	81	✗	✗	✗	✗
V-STaR (Cheng et al., 2026)	128	✓	✓	✗	✗
STRAND (ours)	183	✓	✓	✓	✓

Table A10: Quality of the constructed trajectories against the human-verified annotations (%). Registration and relation are F1, the others are accuracies. The last two rows disable only cross-chunk identity binding.

Configuration	Entity reg.	Identity link.	State	Relation	Event order
STRAND-Track (Gemini-3.5-Flash)	**84.2**	**79.6**	**81.3**	**74.8**	**71.5**
STRAND-Track (Qwen3-VL-235B)	80.7	73.1	77.9	70.2	65.3
w/o identity linking (Gemini)	84.2	58.3	81.3	71.4	68.9
w/o identity linking (Qwen)	80.7	51.4	77.9	66.8	62.1

lucer and V-STaR probe hallucination and spatio-temporal reasoning with paired or structured items, without a joint criterion over a target and its prerequisites.

What STRAND adds is the combination in the last two columns. The supporting facts are elicited from the model as separate binary questions, and the target is credited only when the answer and every prerequisite are correct on the same item. That combination is what makes "right answer, wrong facts" a scoreable outcome rather than an invisible one.

J Quality of the Intermediate Representation

Every accuracy in the main paper grades a final answer, so none of them says whether the trajectories STRAND-Track builds are correct. Because STRAND ships human-verified trajectories, we can score the representation directly.

Matching rule. A predicted entity matches a gold entity when their canonical identifiers agree after normalisation and their observation windows overlap. Unmatched predictions count as false positives and unmatched gold entities as false negatives. *Entity registration* is the F1 of that matching. *Identity linking* is the fraction of gold cross-chunk co-reference pairs ending up in one predicted trajectory, measured only over entities crossing at least one chunk boundary, which is the subset the linker is responsible for. *State accuracy* is the fraction of matched observations whose `attribute` and `action` fields both agree with the gold state. *Relation* is the F1 over typed directed links (o_1, s, o_2) at matched timestamps. *Event order* is pairwise accuracy over all gold event pairs separated in time.

What the numbers localise. Table A10 places event ordering lowest on both backbones, 12.7 and 15.4 points below entity registration. The same deficit appears at the answer level, where sub-questions about the order of two events trail sub-questions about individual facts by roughly 18 points (Section 6.2). Because it is present in the representation before any answering happens, it is not introduced by the answerer, and improving G_η alone would not remove it. Disabling identity linking leaves entity registration and state accuracy untouched, by construction, and cuts identity-linking accuracy by 21.3 and 21.7 points, with a smaller knock-on effect on relations and event order. This is the representational counterpart of the 7.8 and 9.1-point drops in A_{faith} in Table 2, and it is what licenses attributing those drops to cross-chunk binding rather than to a side effect.

Table A11: Sub-question accuracy (%) on the standard, contrastive and open-form probe sets. Absolute scores fall on both controls and the ordering of systems is preserved.

Model	Standard	Contrastive	Open-form
STRAND-Track (Gemini-3.5-Flash)	**76.3**	**70.1**	**61.7**
STRAND-Track (Qwen3-VL-235B)	72.6	65.4	57.3
Gemini-3.1-Pro	64.9	54.8	48.2
Qwen3.5-27B	66.7	55.9	47.6
Cosmos-Reason2-8B	74.7	58.2	44.1
Majority class / most frequent value	51.6	50.0	12.4
Question-only (no video)	53.8	50.6	18.9

K Contrastive and Open-Form Probes

Binary sub-questions can be leading. *"Does player #24 wear a red shirt?"* presupposes both the entity and the attribute, so a model exploiting a yes-bias may score well without monitoring anything. Section 4.4 reports the label balance and prior-only baselines that bound this. Two further probe sets bound it directly.

Contrastive subset. For 300 sub-questions we construct a minimally perturbed counterpart in which exactly one element, the entity, the predicate or the argument, is replaced by a value the annotators record as *false* at that timestamp, with the rest of the sentence unchanged. A model answering from presupposition answers both members of a pair the same way, whereas a model monitoring the scene separates them. We score the subset as ordinary sub-question accuracy.

Open-form subset. For 200 probes we replace the binary form with an open question over an annotated attribute or relation argument, such as *"What colour is player #24's shirt?"*. Grading is normalised exact match against the annotated value, with a small synonym list per attribute type fixed before evaluation and released with the benchmark. No judge model is used.

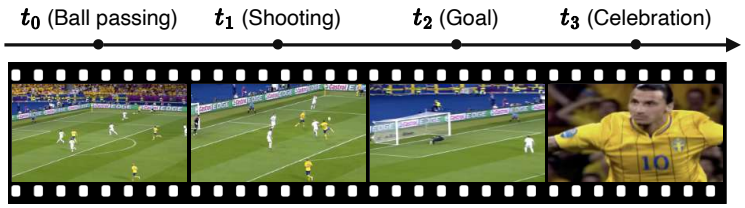
Reading the controls. Table A11 reports both. Every system loses on both probe sets, the expected consequence of removing presupposition and of replacing a two-way choice with an open vocabulary. Two details matter. The question-only baseline collapses to chance on the contrastive subset (50.6%), confirming the perturbation removes the prior rather than merely making the item harder. And Cosmos-Reason2-8B, which has the second-highest standard sub-question accuracy in Table 1, falls furthest on both controls, at 16.5 and 30.6 points, so part of its apparent perception strength was format-dependent. The ordering of STRAND-Track against the end-to-end baselines is unchanged on all three probe sets, so the reported advantage is not an artefact of the binary format.

L STRAND Samples

Figures A2 through A5 present detailed examples from STRAND. 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.

L.1 Additional open-weight baselines

Four further open-weight end-to-end models were evaluated on STRAND and are omitted from Table 1 for space, in the order A_{faith}, A_{target}, A_{sub}, A_{cons}. Qwen3-VL-32B Instruct reaches 42.0, 59.8, 70.0 and 78.1. Qwen3.5-9B reaches 25.6, 50.5, 49.0 and 57.9. Qwen3-VL-8B Think reaches 42.5, 61.3, 70.1 and 77.4. Qwen3-VL-8B Instruct reaches 22.5, 34.3, 59.8 and 77.4. None changes the ordering of the table or the position of any bolded or underlined value.

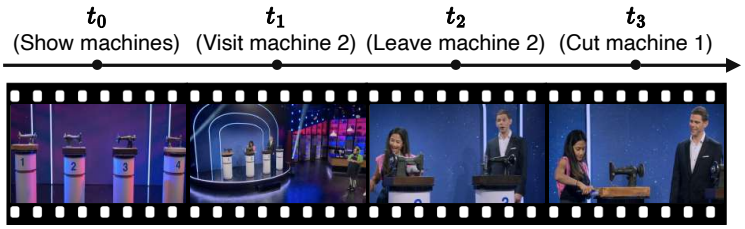


Target Q: Does player #20 score a goal?
GT: No

Sub-Q: Is there a goal in the video?
GT: Yes

Sub-Q: Does player #10 score a goal?
GT: Yes

Figure A2: **Example of the STRAND dataset.** To rigorously evaluate model faithfulness, STRAND 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.



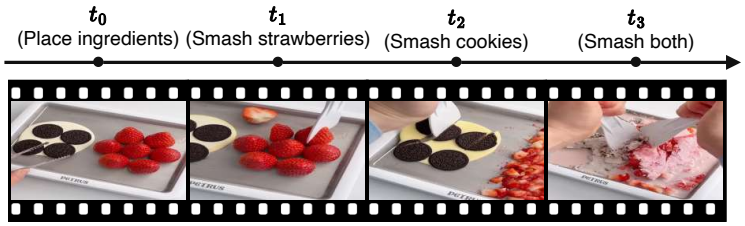
Target Q: Is sewing machine 2 a cake?
GT: No

Sub-Q: Is sewing machine 1 a cake?
GT: Yes

Sub-Q: Is sewing machine 3 a cake?
GT: No

Sub-Q: Is sewing machine 4 a cake?
GT: No

Figure A3: **STRAND 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.



Target Q: Are the cookies smashed before the strawberries?

GT: No

Sub-Q: Are the strawberries smashed in the video?

GT: Yes

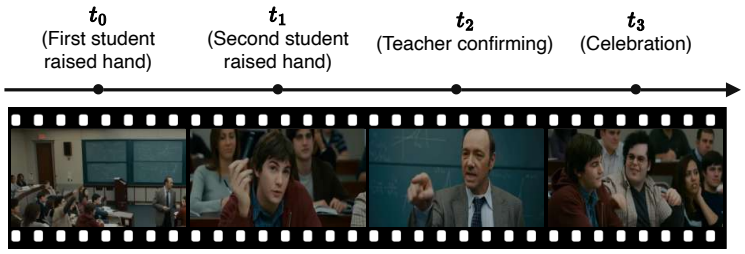
Sub-Q: Are the cookies smashed in the video?

GT: Yes

Sub-Q: Are the cookies smashed after the strawberries?

GT: Yes

Figure A4: **STRAND 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.



Target Q: Did the student who raised hand second later celebrate with his friend?

GT: No

Sub-Q: Did the student in tan jacket raise hand first?

GT: Yes

Sub-Q: Did the student in red jacket raise hand second?

GT: Yes

Sub-Q: Did the student in red jacket celebrate with the student in tan jacket?

GT: No

Figure A5: **STRAND evaluation of actor disambiguation over time.** This scenario highlights how STRAND 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.