

Beneath the Scores: Rethinking Hallucination Evaluation for Video Understanding Models

Shuzhi Gong Fengze Sun Yuansan Liu

The University of Melbourne

{shuzhi, fengze.sun1, yuansan.liu1}@unimelb.edu.au

Abstract

Video understanding is increasingly performed by multi-stage LLM agents that separate temporal grounding, visual observation, and reasoning. Yet these stages are typically evaluated on different benchmarks and distributions, making it difficult to determine where hallucinations originate. We first organize existing benchmarks around these stages and show that their scores provide inconsistent diagnostic signals: stronger stage-level performance does not reliably imply lower downstream hallucination, and even benchmarks targeting the same capability can disagree.

We therefore introduce a causal stage-intervention protocol that overwrites individual stages while holding the downstream task fixed. Across 60,008 runs on three video-agent architectures, we find that grounding is the dominant source of downstream error, with roughly four times the causal impact of corrupting visual observations. Successful grounding depends primarily on locating the correct region rather than precise temporal overlap, explaining why standard mIoU metrics poorly predict downstream reliability. We further find that incorrect evidence is substantially more harmful than missing evidence. Finally, auditing existing benchmarks against these interventions reveals that their scores do not reliably predict causal cascade sensitivity and can fail under distribution shift. These results motivate intervention-based, stage-aware evaluation for trustworthy video agents.

1 Introduction

Video understanding is increasingly moving from monolithic Video LLMs toward agentic systems that decompose long-video reasoning into specialized stages Tang et al. [2025], Maaz et al. [2024], Zhang et al. [2023], Lin et al. [2024], Li et al. [2024b], Wang et al. [2024a], Liu et al. [2025b], Zhang et al. [2026b], Liu et al. [2025a]. A typical video agent first **grounds** the question to relevant moments, then **watches** those moments to extract visual evidence, and finally reasons over the evidence to produce an answer (Figure 1). This decomposition improves flexibility and scalability, but also creates multiple points of failure: an agent may retrieve the wrong moment, misinterpret the right one, or reason incorrectly over otherwise valid evidence. Errors can therefore propagate across stages before appearing in the final answer.

Evaluation, however, has not evolved with this pipeline. Existing hallucination benchmarks largely score final outputs Huang et al. [2026], Lu et al. [2025], Wang et al. [2024b], Rawal et al. [2025a], whereas capability benchmarks separately evaluate grounding Chen et al. [2025a], Xiao et al. [2024a], temporal understanding Wang et al. [2025], or related intermediate behaviors. Because these benchmarks use different datasets, distributions, and protocols, their scores are not directly comparable across stages. A high grounding score, for example, does not establish that grounding errors are unimportant downstream, nor does a poor hallucination score identify which stage caused the failure.

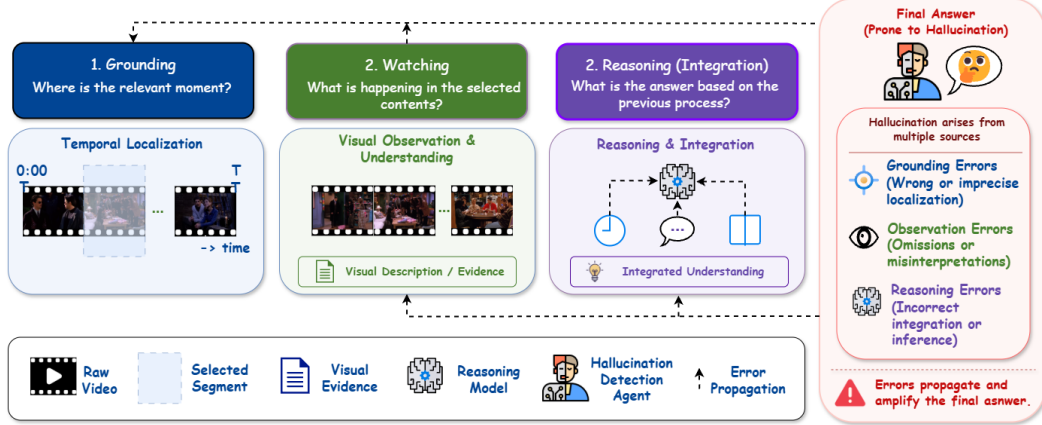


Figure 1: A three-stage view of video understanding. Grounding localizes relevant moments, Watching extracts visual evidence, and Reasoning integrates that evidence into an answer. Downstream failures may originate at any stage and propagate through the pipeline.

This raises a fundamental question: *Can existing benchmark scores actually diagnose failures in multi-stage video-agent pipelines?*

We first examine this question observationally. We organize five established benchmarks around Grounding, Watching, and Reasoning and evaluate representative open-weight video agents under controlled inference settings. The resulting profiles are strikingly inconsistent: stronger upstream scores do not reliably imply better downstream hallucination behavior, benchmarks assigned to the same stage can move in opposite directions under the same inference change, and hallucination metrics can produce anomalous values on agentic architectures. These inconsistencies expose a more fundamental limitation: when stages are evaluated on different data, cross-stage comparisons remain correlational and cannot determine which stage actually changes the final outcome.

We therefore replace cross-benchmark correlation with a **controlled stage-intervention protocol**. Using 2,522 questions over 227 long videos with ground-truth temporal evidence, we overwrite one stage while keeping the downstream task fixed. For Grounding, we substitute oracle, random, off-target, and progressively dilated temporal windows; for Watching, we systematically omit or replace frames within the correct window. Across 60,008 intervened runs over three architecturally distinct agents, this design directly measures how changing a stage changes downstream performance. We further complement answer accuracy with an evidence-level diagnosis that tests whether apparently correct answers are actually supported by the video.

The interventions reveal a consistent causal picture. Grounding dominates downstream error: replacing a length-matched random window with the oracle evidence span improves accuracy by 17–18 points across all three architectures, roughly four times the effect of substantial frame corruption during Watching. Yet deployed grounding modules recover only about one quarter of this available gain. Moreover, being on-target matters far more than being temporally precise: expanding an oracle span by up to $8\times$ incurs essentially no loss, explaining why overlap-based metrics such as mIoU can poorly reflect downstream utility. Finally, wrong evidence is more harmful than missing evidence: agents tolerate severe frame omission but degrade when plausible frames from elsewhere in the video replace the correct evidence. Consistent with these results, our audit finds no existing benchmark score that detectably predicts causally measured cascade sensitivity. Evidence-level diagnosis further reveals that 42–65% of correct answers can be unsupported by the video, while a hallucination metric that discriminates on its native distribution can silently saturate under domain shift.

Our contributions are threefold:

1. We provide a stage-aligned analysis of existing video evaluation, showing that fragmented benchmark scores cannot reliably diagnose where failures arise in multi-stage video agents.
2. We introduce a causal stage-intervention protocol that overwrites individual pipeline stages on a shared item set while holding the downstream task fixed, enabling direct measurement of stage-level effects across agent architectures.

- Using this causal reference, we reveal systematic gaps in current evaluation: grounding dominates downstream error, while deployed grounders capture only a small fraction of the available headroom; on-target evidence matters more than precise temporal localization; incorrect evidence is more damaging than missing evidence; and existing answer- and hallucination-level scores can fail to reflect causal reliability.

2 Preliminary Study

2.1 Pipeline Formalization and Parameters

We abstract modern video agents into three functional stages: **Grounding** ($\mathcal{G}$), which localizes question-relevant temporal evidence; **Watching** ($\mathcal{W}$), which extracts visual information from the selected content; and **Reasoning** ($\mathcal{R}$), which integrates this information with the question to produce the final answer. These stages need not correspond to explicit modules—different agents may implement them through retrieval, frame selection, visual descriptions, latent representations, or tool use—but they provide a common interface for stage-aligned evaluation. The pipeline is sequential, $\mathcal{G} \rightarrow \mathcal{W} \rightarrow \mathcal{R}$, so upstream errors may affect downstream predictions.

For the preliminary study, we vary two inference-time controls shared by the evaluated systems: the number of sampled frames f and the maximum tool-call budget τ . f controls the amount of visual evidence exposed to the model; we use $f \in \{16, 32, 64\}$ for Video-R1, VideoMind, and EVA. τ limits the number of agentic tool invocations and is varied from 4 to 16 for EVA; the other models do not expose an equivalent control. We focus on these parameters because they can be changed at inference time without modifying model weights, enabling controlled comparisons across configurations.

2.2 Models, Benchmarks and Metrics

We evaluate three representative open-weight systems spanning distinct video-agent designs: VIDEO-R1 [Feng et al., 2026], a monolithic Qwen2.5-VL-based reasoner; VIDEOMIND [Liu et al., 2025b], which separates stages with role-specific LoRA modules; and EVA [Zhang et al., 2026b], a multi-agent framework with explicit localization, observation, and answering components. All experiments use publicly released checkpoints without additional fine-tuning. For VIDEOMIND and EVA, intermediate stage outputs are directly exposed; for VIDEO-R1, we obtain them through structured prompting. The same systems are used in the causal intervention study in Section 3.

Stage	Benchmark	Scale	Metric
Grounding	NEXT-GQA [Xiao et al., 2024b]	990 / 5,553	Acc, mIoU
	CG-BENCH [Chen et al., 2025a]	1,219 / 12,129	Acc, mIoU
Watching	LVBENCH (T.G.) [Wang et al., 2025]	72 / 179	Acc
	ARGUS [Rawal et al., 2025a]	500 / 500	Cost _H , Cost _O
Reasoning	ARGUS [Rawal et al., 2025a]	500 / 500	Cost _H , Cost _O
	ELV-HALLUC [Lu et al., 2025]	200 / 3,600	Acc, Diff, SAH
	VideoHalluciner [Wang et al., 2024b]	948 / 1,800	Basic, Halluc, Both

Table 1: Stage-aligned benchmarks and metrics.

We organize existing benchmarks by the pipeline stage they primarily probe, as summarized in Table 1. Grounding is evaluated with NEXT-GQA and CG-BENCH, Watching with the temporal-grounding subset of LVBENCH and ARGUS, and downstream hallucination/reasoning with ARGUS, ELV-HALLUC, and VideoHalluciner. ARGUS spans Watching and Reasoning because it evaluates the faithfulness of generated visual descriptions.

We organize existing benchmarks by the pipeline stage they primarily probe, as summarized in Table 1. Grounding is evaluated with NEXT-GQA and CG-BENCH, Watching with the temporal-grounding subset of LVBENCH and ARGUS, and downstream hallucination/reasoning with ARGUS, ELV-HALLUC, and VideoHalluciner. ARGUS spans Watching and Reasoning because it evaluates the faithfulness of generated visual descriptions.

2.3 Results and Limits of Fragmented Evaluation

We first examine whether existing stage-aligned benchmarks provide consistent diagnostic signals for hallucination in video agents.

Table 2 shows that capability scores and hallucination scores do not align consistently across stages. Strong Grounding or Watching performance does not imply lower downstream hallucination: for example, EVA achieves the best CG-Bench accuracy and LVBench score, yet performs poorly on ELV-Halluc and the hallucination subset of VideoHalluciner. Conversely, models with weaker standard capability scores can be substantially more robust on hallucination-oriented evaluation. ARGUS provides a useful intermediate signal by directly evaluating hallucination and omission in visual descriptions, but it still operates on a separate dataset.

Method	Grounding				Watching		W $\leftrightarrow$ H		Hallucination					
	CG-Bench		NeXT-GQA		LVBench*	ARGUS	Cost _H ↓	Cost _O ↓	ELV-Halluc			VideoHalluciner		
	Acc↑	mIoU↑	Acc↑	mIoU↑					Acc↑	Diff↓	SAH↓	Basic↑	Halluc↑	Both↑
Qwen2-VL-7B	31.37	-	74.59	-	54.79	55.52	82.09	8.1	5.8	6.1	76.42	64.25	46.58	
Qwen2.5-VL-7B	32.24	-	72.98	-	58.10	51.19	81.57	14.2	4.5	5.1	74.58	68.83	49.58	
Video-R1	35.32	1.44	75.38	21.14	59.82	66.15	83.58	15.2	-0.3	-0.3	81.08	55.50	42.58	
VideoMind-7B	38.40	7.10	76.79	31.40	54.79	55.63	83.06	11.4	4.2	4.6	76.25	64.58	46.83	
EVA	40.20	4.51	68.99	26.87	60.27	53.10	82.81	0.6	-0.5	-0.5	88.25	45.83	39.08	
Avg.	35.51	4.35	73.75	26.47	57.55	56.32	82.62	9.90	2.74	3.00	79.32	59.80	44.93	

Table 2: Stage-aligned benchmark results under default inference settings.¹

Overall, the benchmark profile can rank individual capabilities, but it does not reveal where a downstream hallucination originates. Detailed per-model comparisons are provided in Appendix A. Additional parameter-sweep results in Appendix A further show that changing the sampled-frame count or tool-call budget produces benchmark- and model-dependent effects, with no consistent improvement in downstream hallucination robustness.

Why Fragmented Evaluation Is Not Enough? The preliminary study exposes two problems. First, stage-level capability scores are only weakly connected to downstream hallucination. Second, conclusions do not reliably transfer across datasets, even within the same nominal stage. The anomalous near-zero or negative ELV-Halluc readings for EVA and VIDEO-R1 make this ambiguity particularly clear: from the score alone, we cannot tell whether the model failed or the metric did.

More fundamentally, all of these comparisons are correlational and are measured on different datasets. They can reveal disagreement, but cannot determine which pipeline stage caused a downstream error or how much correcting that stage would help. Section 3 therefore replaces cross-benchmark comparison with controlled stage interventions on a single item set.

3 Causal Stage Intervention

The preliminary study reveals disagreement between stage-level and hallucination scores, but cannot identify its source because each stage is evaluated on a different dataset. We therefore move to a controlled design in which the *same items and downstream task* are retained while one upstream stage is overwritten with a known value. This allows us to directly measure how errors at each stage affect the final prediction. Overall, we conduct **60,008 intervened agent runs** and **5,784 evidence-level diagnoses** across three architectures.

3.1 Intervention Protocol

Data and grounding interventions. We construct the intervention set from CG-BENCH [Chen et al., 2025a], which provides both ground-truth temporal evidence spans (`clue_intervals`) and a native *Hallucination* question category. We use all **2,522 questions** from the 227 videos containing at least one Hallucination item: 444 Hallucination questions and 2,078 questions from eleven other categories. Because both groups concern the same videos, comparisons are not confounded by video distribution. The median video lasts 26 minutes, whereas the median clue span is only 10 s (0.64% of the video). For each item, we replace the agent’s temporal window while leaving the remaining pipeline unchanged. The conditions are `full` (whole video), `predicted` (the agent’s predicted span), `oracle` (ground-truth clue span), `random` (a length-matched random span), and `off-target` (a length-matched span in the largest evidence-free gap). We additionally dilate the oracle span by $2 \times /4 \times /8 \times /16 \times$ on 300 items to test how precise grounding must be.

Watching interventions and implementation. With grounding fixed to `oracle`, we corrupt the sampled frames within the correct window. *Omission* removes a fraction ρ of frames, whereas *fabrication* replaces them with frames sampled at least 60 s away from the same video while preserving their original timestamp labels, with $\rho \in \{0.25, 0.5, 0.75\}$. Both interventions operate on the sampled frame list shared by all three systems. The intervention is inserted at the corresponding interface

¹Default settings are Video-R1 ($f=16$), VideoMind-7B ($f=32$), and EVA ($\tau=4$, $f=32$). LVBench* denotes the Temporal Grounding subset; ARGUS reports hallucination/omission costs; VideoHalluciner reports Basic/Halluc/Both accuracy.

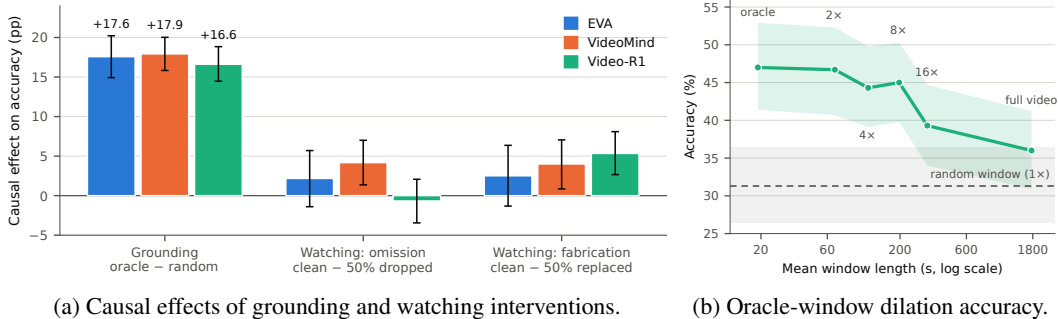


Figure 2: Grounding determines where the model looks, but does not need to be temporally precise. **Left:** replacing a length-matched random window with the oracle span improves accuracy by 17–18 pp across all three architectures, substantially more than corrupting frames within the correct window. **Right:** expanding the oracle window up to 8× causes little degradation; performance drops only once the window becomes very broad.

of each architecture: EVA’s requested windows are restricted to the assigned span, VIDEO-MIND’s answer receives that span instead of the grounder’s prediction, and VIDEO-R1, which has no explicit grounding module, receives it directly through its frame list and serves as the monolithic control. Thus, `predicted` is undefined for VIDEO-R1.

Statistical protocol. We report 95% paired bootstrap intervals clustered by video (227 clusters). Paired binary outcomes are tested with exact McNemar tests and Holm correction within each contrast family; unparseable outputs count as incorrect and are reported separately as abstentions. As a pre-specified sanity check, `oracle` must outperform `random` for every agent, while `off-target` must not outperform `random`. Both conditions hold (Table 3), confirming that the interventions are effectively propagated to the models.

Grounding condition	EVA	VIDEOMIND	VIDEO-R1
<code>oracle</code> (GT clue span)	46.6 [44.2, 48.9]	49.3 [47.0, 51.6]	47.2 [44.9, 49.6]
<code>predicted</code> (agent as deployed)	34.1 [31.8, 36.4]	35.9 [33.6, 38.2]	—
<code>full</code> (whole video)	27.0 [25.1, 28.9]	34.8 [32.5, 37.1]	33.3 [31.1, 35.4]
<code>random</code> (length-matched)	29.0 [27.0, 31.1]	31.4 [29.3, 33.4]	30.6 [28.7, 32.6]
<code>off-target</code> (length-matched)	27.2 [25.3, 29.1]	30.6 [28.4, 33.0]	30.7 [28.6, 32.8]
<i>Headroom decomposition (paired contrasts, pp)</i>			
<code>oracle</code> – <code>random</code> (total headroom)	+17.6 [14.9, 20.2]	+17.9 [15.8, 20.0]	+16.6 [14.5, 18.8]
<code>predicted</code> – <code>random</code> (captured)	+5.0 [2.7, 7.3]	+4.5 [2.6, 6.4]	—
<code>oracle</code> – <code>predicted</code> (left on the table)	+12.5 [10.2, 14.8]	+13.4 [11.3, 15.5]	—

Table 3: Causal grounding intervention on 2,522 paired CG-BENCH items from 227 video clusters. Top: accuracy (%) with clustered 95% bootstrap CIs. Bottom: paired contrasts. All `oracle`–`random` differences are significant at $p < 0.001$. The deployed grounders capture only $\sim 25\%$ of the available `oracle` headroom.

3.2 Analysis and Findings of Causal Stage Intervention

Grounding Dominates Downstream Error. Table 3 and Figure 2 show a clear asymmetry between grounding and watching. Replacing a length-matched random window with the oracle span improves accuracy by 16.6–17.9 pp across all three architectures ($p < 0.001$). By contrast, dropping or replacing half of the frames *within the correct window* changes accuracy by at most about 5 pp. Thus, for long-video understanding, *where* the model looks matters substantially more than moderate corruption of *what* it sees once the relevant moment has been found.

Current grounding modules exploit only a small fraction of this opportunity. Relative to a random window, the oracle span provides roughly 18 pp of headroom, whereas the deployed grounders recover only 5.0 pp for EVA and 4.5 pp for VIDEO-MIND. The remaining gap is 12.5 and 13.4 pp, respectively.

In particular, VIDEOMIND’s predicted grounding performs only marginally better than showing the whole video, while EVA obtains a clearer gain. Across these different architectures, the same pattern emerges: grounding is both the most consequential stage and the stage with the largest remaining room for improvement.

Wrong evidence is more harmful than missing evidence. The watching interventions further show that agents tolerate missing visual evidence better than misleading evidence. Even after dropping 75% of the frames within the correct window, the degradation is small and not consistently significant. Replacing those frames with plausible content from elsewhere in the same video causes a clearer drop in accuracy (Appendix B). This suggests that grounding errors are harmful not simply because they remove useful frames, but because they can supply convincing evidence from the wrong moment.

Why the unified design matters. These effects are difficult to recover from cross-benchmark comparisons. On the shorter DR.V-BENCH slice, for example, predicted grounding does not outperform simply exposing the whole video for either EVA or VIDEOMIND. This is consistent with the dilation result: when the video itself is relatively short, a broad but on-target context can be sufficient. More generally, comparing grounding and hallucination scores across different datasets cannot separate such distribution effects from genuine stage sensitivity. The intervention design avoids this confound by changing one stage while evaluating the same downstream items.

4 Auditing the Benchmark Ecosystem Against the Causal Standard

The intervention protocol provides more than stage-level error attribution. For each item and system configuration, it gives a direct measure of how much a particular stage affects the final prediction. This allows us to revisit the observation in Section 2.3 that existing benchmark scores are only weakly connected across stages, and ask a more concrete question: *do existing benchmark scores predict how sensitive the final answer is to a causal intervention on that stage?*

Cascade sensitivity. For each item, we define cascade sensitivity as the change in answer accuracy when grounding is replaced by the oracle span rather than a length-matched random span. At the system level, we average this quantity over items. Because sensitivity is measured on the same 2,522 items used in Section 3, it serves as a causal reference quantity rather than another benchmark score.

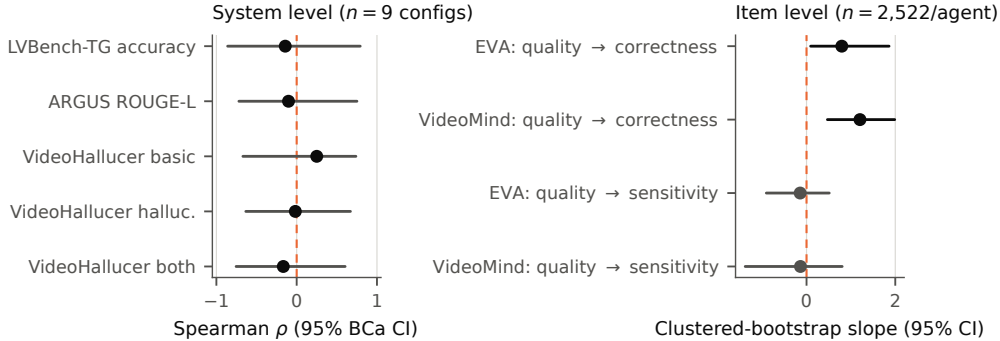


Figure 3: No existing benchmark score detectably predicts causally measured cascade sensitivity. Left: Spearman correlation between five retained stage-aligned benchmark scores (recomputed from per-item outputs of Section 2.3) and causal sensitivity across nine agent $\times$ frame-budget configurations; every 95% BCa interval spans zero. Right: within-agent, per-item clustered regression. Grounding quality (frame-level evidence rate) predicts *whether the agent answers correctly* (top two rows, intervals exclude zero), but does not predict *whether grounding causally determines the answer* (bottom two rows). The same pattern appears for both agents with an explicit grounding stage.

4.1 No benchmark score predicts what causally matters

Figure 3 shows the same pattern at both the item and system levels: conventional benchmark scores track *performance*, but not *causal dependence*.

Item level. Better grounding is associated with higher final-answer accuracy for both EVA and VIDEOMIND, confirming the familiar leaderboard signal that agents tend to answer better when they localize more relevant evidence. However, grounding quality does not predict cascade sensitivity: items with better grounding are no more likely to be those whose answers actually change when grounding is corrected. The regression slopes for correctness are clearly positive, whereas those for causal sensitivity remain near zero (Figure 3). In other words, standard grounding metrics can tell us whether an agent is doing well, but not whether grounding is the stage responsible for that outcome.

System level. The same conclusion holds across agent and frame-budget configurations. None of the retained stage-aligned benchmarks—LVBENCH-TG, ARGUS, or the three VideoHalluciner scores—shows a detectable association with causally measured cascade sensitivity. Despite measuring grounding, visual faithfulness, and hallucination from different perspectives, these scores do not identify which systems are more sensitive to grounding errors.

Taken together, the two analyses expose a fundamental gap between benchmark performance and diagnostic value: **existing scores can rank systems by how well they perform, but they do not reveal how strongly a particular stage causally determines the final answer.**

4.2 Beyond right and wrong: correct answers can still be hallucinated

Final-answer accuracy cannot distinguish a correct answer supported by the video from one reached without sufficient visual evidence. To expose this difference, we apply DR.V, a stage-aware hallucination diagnosis framework that verifies claims in an agent’s answer against sampled visual evidence and attributes failures to perceptive, temporal, or cognitive causes.² We run the diagnosis on 900 outputs from the DR.V-BENCH long-video slice and 4,884 outputs from CG-BENCH.

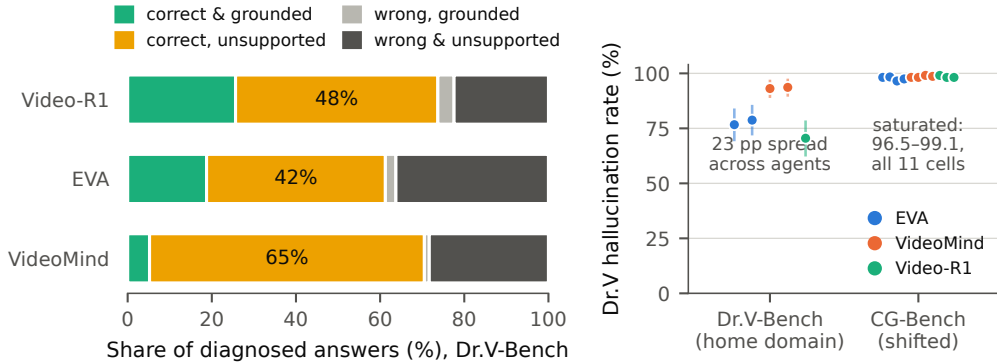


Figure 4: **Left:** evidence-level diagnosis on DR.V-BENCH. A large fraction of answers are *correct but unsupported*: 42–65% of correct answers lack sufficient visual evidence, a failure mode hidden by final-answer accuracy. **Right:** DR.V clearly separates agents on its home domain, but saturates on CG-BENCH across agents and intervention conditions, indicating that the same hallucination metric can lose discriminative power under distribution shift.

Figure 4 reveals a substantial gap between answer correctness and evidential support. Across the three agents, **42–65% of correct answers are not sufficiently supported by the video**. These answers would be counted as successful under a conventional accuracy metric, even though the model did not establish the visual evidence needed to justify them.

This difference is important for reliability evaluation. Agents with similar final-answer accuracy can differ substantially in how much of that accuracy is actually grounded in the video. In our results, DR.V separates the agents by more than 20 percentage points in hallucination rate despite much smaller differences in answer accuracy. Thus, final-answer correctness and evidential support capture different aspects of model behavior.

A correct answer is therefore not necessarily a reliable answer. Evaluation should distinguish whether the model arrives at the right prediction from supported visual evidence, rather than treating all correct outputs as equally successful.

²Reference withheld for double-blind review.

4.3 Metric failure versus model failure

The preliminary study shows that anomalous benchmark scores do not necessarily imply model failure. On ELV-HALLUC, EVA and VIDEO-R1 obtain near-zero or negative scores, far outside the regime observed for the monolithic models on which the benchmark was validated.

Our intervention results show how such failures can arise. DR.V clearly separates agents on its native DR.V-BENCH distribution, but saturates near 100% on CG-BENCH across agents and intervention conditions (Figure 4, right). More importantly, replacing a random grounding window with the oracle span improves answer accuracy by about 17 pp while barely changing the DR.V score. The metric therefore loses discriminative power under the shifted evaluation regime. The likely cause is a mismatch between the metric and the target setting: CG-BENCH produces long free-form answers with many checkable claims, while the diagnosis relies on sparsely sampled frames from long videos, making a positive hallucination verdict almost inevitable. A similar architectural mismatch may explain the anomalous ELV-HALLUC scores for tool-calling agents.

An extreme benchmark score can therefore reflect metric failure rather than model failure.

4.4 Implications for trustworthy evaluation

These results suggest three practical changes to how multi-stage video systems should be evaluated.

(i) Measure whether grounding is on target, not only how tightly it overlaps the annotation. Finding 3 shows that substantial temporal dilation around the correct evidence has little effect on downstream accuracy, even though it can strongly reduce overlap-based measures such as mIoU. Grounding metrics should therefore distinguish between missing the relevant event and selecting a somewhat wider window around it. One useful criterion is whether the annotated evidence is contained within the selected window at a practically useful dilation.

(ii) Check that a metric remains discriminative in the target regime. Before interpreting a hallucination score, the metric should be tested against a controlled manipulation known to change model behavior. An oracle-span intervention provides one inexpensive and reusable probe. If a metric remains nearly constant despite a large and independently verified change in answer accuracy, as in Section 4.3, its numerical value should not be treated as a reliable measure of model behavior in that regime.

(iii) Evaluate evidential support in addition to answer correctness. The 42–65% rate of correct-but-unsupported answers shows that correctness alone is insufficient for assessing reliability. Evaluation should therefore record not only whether an answer is correct, but also whether the visual evidence available to the agent actually supports it. Stage-annotated datasets such as CG-BENCH, which provide clue intervals, make this type of evidence-level evaluation feasible at benchmark scale.

5 Conclusion

Existing video benchmarks can rank capabilities, but they do not reliably diagnose where failures arise or which stages causally determine downstream outcomes in multi-stage video agents. By replacing cross-benchmark comparison with controlled stage interventions, we find that grounding is the dominant downstream bottleneck, that being on-target matters more than precise temporal overlap, and that wrong evidence is substantially more harmful than missing evidence. Our benchmark audit further shows that conventional capability- and hallucination-level scores do not reliably predict causal cascade sensitivity. These results suggest that trustworthy evaluation should move beyond isolated benchmark scores toward intervention-based, evidence-aware evaluation of video agents.

Limitations. First, the CG-BENCH intervention results measure downstream answer *error* rather than hallucination defined as unsupported content. The latter is examined separately with DR.V-BENCH, where the diagnosis instrument retains discriminative power. Second, we evaluate one checkpoint per architecture, so cross-agent dissociations (e.g., VIDEO-R1’s sensitivity to fabrication) cannot separate architecture from model identity. Third, the system-level audit covers nine configurations and its wide intervals support only “no detectable relationship”; the item-level analysis ($n=2,522$ per agent) is the powered version of the claim. Fourth, EVA’s full condition required a forced-answer prompt to suppress tool-denied refusals (flagged per-run; Appendix D), and multi-interval clues (11% of items) are served as their temporal hull.

References

- Mubashara Akhtar, Anka Reuel, Prajna Soni, Sanchit Ahuja, Pawan Sasanka Ammanamanchi, Ruchit Rawal, Vilém Zouhar, Srishti Yadav, Chenxi Whitehouse, Dayeon Ki, et al. When ai benchmarks plateau: A systematic study of benchmark saturation. *arXiv preprint arXiv:2602.16763*, 2026.
- Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, et al. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13787–13805, 2024.
- Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. In *International Conference on Learning Representations*, volume 2025, pages 45647–45682, 2025a.
- Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. In *International Conference on Learning Representations*, volume 2025, pages 45647–45682, 2025b.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.
- Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. Videoagent: A memory-augmented multimodal agent for video understanding. In *European Conference on Computer Vision*, pages 75–92. Springer, 2024.
- Bo Feng, Zhengfeng Lai, Shiyu Li, Zizhen Wang, Simon Wang, Ping Huang, and Meng Cao. Breaking down video llm benchmarks: Knowledge, spatial perception, or true temporal understanding? *arXiv preprint arXiv:2505.14321*, 2025.
- 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, Yuhang 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24108–24118, 2025.
- Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, et al. Video-mme-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*, 2026.
- Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280, 2024.
- Yiyang Huang, Yitian Zhang, Yizhou Wang, Mingyuan Zhang, Liang Shi, Huimin Zeng, and Yun Fu. Distorted or fabricated? a survey on hallucination in video llms. *arXiv preprint arXiv:2604.12944*, 2026.
- Ming Kong, Xianzhou Zeng, Luyuan Chen, Yadong Li, Bo Yan, and Qiang Zhu. Mhbench: Demystifying motion hallucination in videollms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4401–4409, 2025.
- Chaoyu Li, Eun Woo Im, and Pooyan Fazli. Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13723–13733, 2025.
- Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024a.

- Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024b.
- Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 5971–5984, 2024.
- Runtao Liu, Ziyi Liu, Jiaqi Tang, Yue Ma, Renjie Pi, Jipeng Zhang, and Qifeng Chen. Longvideoagent: Multi-agent reasoning with long videos. *arXiv preprint arXiv:2512.20618*, 2025a.
- Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for long video reasoning. *arXiv e-prints*, pages arXiv–2503, 2025b.
- 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.
- Hao Lu, Jiahao Wang, Yaolun Zhang, Ruohui Wang, Xuanyu Zheng, Yepeng Tang, Dahua Lin, and Lewei Lu. Elv-halluc: Benchmarking semantic aggregation hallucinations in video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 32572–32581, 2026.
- Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, 2024.
- 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.
- Patrick Ramos, Ryan Ramos, and Noa Garcia. Data leakage in visual datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6309–6319, 2025.
- Ruchit Rawal, Reza Shirkavand, Heng Huang, Gowthami Somepalli, and Tom Goldstein. Argus: Hallucination and omission evaluation in video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20280–20290, 2025a.
- Ruchit Rawal, Reza Shirkavand, Heng Huang, Gowthami Somepalli, and Tom Goldstein. Argus: Hallucination and omission evaluation in video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20280–20290, 2025b.
- Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024.
- Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024.
- 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*, 2025.
- Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967, 2025.
- Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer, 2024a.

- Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohalluciner: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024b.
- Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohalluciner: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024c.
- Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214, 2024a.
- Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214, 2024b.
- Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words. *Advances in Neural Information Processing Systems*, 37:57240–57261, 2024.
- Xiaocui Yang, Wenfang Wu, Shi Feng, Ming Wang, Daling Wang, Yang Li, Qi Sun, Yifei Zhang, Xiaoming Fu, and Soujanya Poria. Mm-instructeval: Zero-shot evaluation of (multimodal) large language models on multimodal reasoning tasks. *Information Fusion*, 122:103204, 2025.
- Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pages 543–553, 2023.
- Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, Zhiyu Tan, Hao Li, Xingjun Ma, and Jingjing Chen. Eventhallusion: Diagnosing event hallucinations in video llms. *arXiv preprint arXiv:2409.16597*, 2024.
- Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. Deep video discovery: Agentic search with tool use for long-form video understanding. *Advances in Neural Information Processing Systems*, 38:89863–89895, 2026a.
- Yaolun Zhang, Ruohui Wang, Jiahao Wang, Yepeng Tang, Xuanyu Zheng, Haonan Duan, Hao Lu, Hanming Deng, and Lewei Lu. Eva: Efficient reinforcement learning for end-to-end video agent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12289–12299, 2026b.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?

Answer: [\[Yes\]](#)

Justification: Contributions are correctly reflected in abstract and introduction.

Guidelines:

- The answer [\[N/A\]](#) means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [\[No\]](#) or [\[N/A\]](#) answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: Limitations and corresponding future works are discussed.

Guidelines:

- The answer [\[N/A\]](#) means that the paper has no limitation while the answer [\[No\]](#) means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate “Limitations” section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren’t acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[Yes\]](#)

Justification: Proof or analysis of the proposed theoretical result are included in the paper.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Model details are disclosed

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Code and instructions will be released upon acceptance

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://neurips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: All details are provided.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports standard metrics including error bars.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: All details are provided

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: the research conform with the NeurIPS Code of Ethics.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Discussed in conclusion

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: the paper poses no such risks

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: existing assets are properly credited

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: the paper does not release new assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: the paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [\[Yes\]](#)

Justification: the core research in this paper involves LLMs as important components.

Guidelines:

- The answer [\[N/A\]](#) means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

A Additional Analysis of the Preliminary Study

This appendix expands the observational results reported in Section 2.3. Table 2 remains in the main paper; here we provide additional interpretation of the benchmark profiles and parameter sweeps.

A.1 Detailed Benchmark Comparisons

Grounding. The two Grounding benchmarks distinguish temporal localization capability, but their rankings do not translate consistently to hallucination robustness. VIDEOMIND achieves the strongest grounding mIoU, while EVA obtains the best CG-Bench accuracy. However, EVA performs substantially worse on hallucination-oriented metrics, indicating that stronger measured grounding does not necessarily correspond to lower downstream hallucination.

Watching. A similar disconnect appears for visual observation. EVA obtains the highest LVBench accuracy but weak hallucination scores, whereas Qwen2.5-VL-7B performs slightly worse on LVBench yet much better on ELV-Halluc and VideoHalluciner. ARGUS is more directly related to hallucination because it measures fabricated and omitted content in generated visual descriptions, but it still provides only an observational signal on its own distribution.

Hallucination metrics. ELV-Halluc and VideoHalluciner show some common model-level tendencies, but their absolute behavior differs substantially. In particular, ELV-Halluc returns near-zero or negative values for EVA and VIDEO-R1, well outside its usual operating regime. These values are therefore difficult to interpret without an independent reference for whether the underlying model behavior actually changed.

A.2 Detailed Parameter-Sweep Analysis

Frame count. For EVA, increasing the number of sampled frames raises CG-Bench long-video accuracy from 36.93% to 41.00%, but decreases NeXT-GQA mIoU. Thus, two Grounding benchmarks react in opposite directions to the same change. ARGUS varies only mildly for EVA and VIDEOMIND, while VideoHalluciner is largely flat for EVA, changes slightly for VIDEOMIND, and degrades more substantially for VIDEO-R1.

Tool-call budget. Increasing EVA’s tool-call budget likewise produces no monotonic benefit. CG-Bench accuracy peaks at an intermediate budget, NeXT-GQA changes only marginally, and hallucination-oriented metrics show small or inconsistent responses. Hence, additional reasoning steps do not reliably improve hallucination robustness.

A.3 Detailed Parameter-Sweep Analysis

Frame count. For EVA, increasing the number of sampled frames raises CG-Bench long-video accuracy from 36.93% to 41.00%, but decreases NeXT-GQA mIoU. Thus, two Grounding benchmarks react in opposite directions to the same change. ARGUS varies only mildly for EVA and VIDEOMIND, while VideoHalluciner is largely flat for EVA, changes slightly for VIDEOMIND, and degrades more substantially for VIDEO-R1.

Tool-call budget. Increasing EVA’s tool-call budget likewise produces no monotonic benefit. CG-Bench accuracy peaks at an intermediate budget, NeXT-GQA changes only marginally, and hallucination-oriented metrics show small or inconsistent responses. Hence, additional reasoning steps do not reliably improve hallucination robustness.

A.4 Inference-Parameter Sweeps

Figure 5 shows that the same inference change can lead to different conclusions across benchmarks. For EVA, increasing the sampled-frame count improves CG-Bench long-video accuracy but decreases NeXT-GQA mIoU, even though both evaluate Grounding. Watching and hallucination metrics are similarly model-dependent.

Figure 6 shows the same instability for agentic reasoning: increasing EVA’s tool-call budget does not monotonically improve either Grounding or hallucination metrics. More frames or more tool calls therefore cannot be interpreted as uniformly improving hallucination robustness.

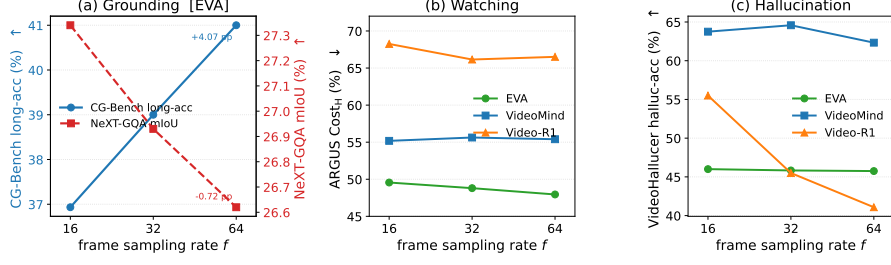


Figure 5: Effect of frame sampling rate on Grounding, Watching, and Hallucination benchmarks. The same increase in frame count can improve one benchmark while degrading another, revealing benchmark-dependent responses to identical inference changes.

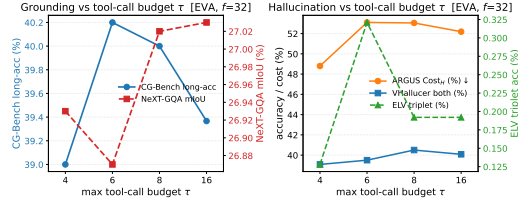


Figure 6: Effect of maximum tool-call budget τ for EVA at fixed frame count $f = 32$. Larger tool budgets do not yield monotonic gains in either Grounding or hallucination metrics.

A.5 Interpretation

These observations demonstrate instability rather than attribution. Because each benchmark uses a different dataset and protocol, cross-benchmark differences conflate model capability with dataset effects. They therefore motivate, but cannot replace, the controlled intervention study in Section 3.

B Full Intervention Ladders

Tables 4 and 5 report the complete watching-corruption and dilation ladders. All intervals are 95% bootstrap CIs clustered by video. At the strongest corruption rate ($\rho=0.75$), the paired clean–corrupted contrasts are: omission +5.7 pp [−0.3, +11.6] (EVA) and +1.7 pp [−3.0, +6.2] (VIDEO-R1); fabrication +6.7 pp [+1.0, +12.6] (EVA) and +4.3 pp [0.0, +8.9] (VIDEO-R1). Even discarding three quarters of the evidence frames inside a correct window is not reliably detectable, whereas replacing them with plausible frames from elsewhere in the same video is.

Corruption	ρ	EVA	VIDEOMIND	VIDEO-R1
none (clean, oracle window)	0	43.2 [39.3, 47.0]	49.5 [45.5, 53.5]	46.5 [42.6, 50.4]
omission (frames dropped)	0.25	46.7 [41.1, 52.2]	50.0 [44.3, 55.6]	48.0 [42.2, 53.9]
	0.50	41.0 [37.2, 44.8]	45.3 [41.4, 49.3]	47.2 [43.3, 50.9]
	0.75	40.3 [35.3, 45.4]	45.7 [40.0, 51.3]	43.7 [38.1, 49.3]
fabrication (frames replaced)	0.25	42.7 [37.2, 48.4]	50.3 [44.9, 55.7]	45.0 [39.7, 50.6]
	0.50	40.7 [36.9, 44.5]	45.5 [41.4, 49.5]	41.2 [37.3, 45.0]
	0.75	39.3 [34.5, 44.3]	43.0 [37.6, 48.5]	41.0 [35.5, 46.6]

Table 4: Watching-corruption ladder: accuracy (%) with grounding pinned to the oracle span and a fraction ρ of in-window frames corrupted ($n=600$ at $\rho \in \{0, 0.5\}$, $n=300$ otherwise). Fabricated frames are sampled ≥ 60 s away in the same video and keep their original timestamp labels.

Oracle window dilation	EVA	VIDEOMIND	VIDEO-R1
2×	45.7 [40.3, 51.2]	50.3 [44.6, 55.9]	46.7 [40.7, 52.4]
4×	46.7 [41.6, 51.7]	46.7 [41.1, 52.3]	44.3 [39.0, 49.8]
8×	43.3 [38.0, 48.6]	46.0 [40.0, 52.0]	45.0 [39.7, 50.3]
16×	38.7 [33.6, 43.7]	45.3 [39.6, 51.0]	39.3 [34.0, 44.8]

Table 5: Dilation ladder for all three agents ($n=300$): accuracy (%) when the oracle window is expanded about its centre. The plateau through $8\times$ replicates across architectures; only $16\times$ (≈ 5 min around a 10 s clue) degrades.

C Audit Details

Table 6 lists the nine agent $\times$ frame-budget configurations used in the system-level audit (Section 4.1), with their causally measured cascade sensitivity (oracle – random, set S, $n=400$ items). The retained benchmark scores were recomputed from the per-item outputs of the preliminary study; the recomputation reproduces Table 2 exactly (e.g., VIDEOMIND at $f=32$ gives VideoHallucur 76.25/64.58/46.83).

Configuration	Sens. (pp)	95% CI	Configuration	Sens. (pp)	95% CI
EVA $f=16$	16.8	[11.0, 22.5]	VIDEOMIND $f=64$	13.8	[9.2, 18.4]
EVA $f=32$	17.8	[11.9, 23.5]	VIDEO-R1 $f=16$	11.5	[6.8, 16.3]
EVA $f=64$	15.0	[10.0, 20.1]	VIDEO-R1 $f=32$	16.3	[11.9, 20.7]
VIDEOMIND $f=16$	15.8	[10.5, 21.0]	VIDEO-R1 $f=64$	15.5	[10.5, 20.5]
VIDEOMIND $f=32$	16.5	[11.6, 21.4]			

Table 6: The nine configurations of the system-level audit and their causal cascade sensitivity. Across these systems, the Spearman correlations between each retained benchmark score and sensitivity are: LVBench-TG -0.14 $[-0.86, +0.79]$, ARGUS ROUGE-L -0.10 $[-0.72, +0.75]$, VideoHallucur basic $+0.25$ $[-0.67, +0.74]$, halluc -0.02 $[-0.63, +0.67]$, both -0.17 $[-0.75, +0.60]$ (95% BCa); every interval spans zero.

D Protocol Details

EVA under the full condition. Denied its frame-selection tool, EVA initially refused to answer 20% of items; a single forced-answer follow-up prompt reduced refusals to 4%, and every nudged run is flagged `forced_answer` in the released logs. Without this control, every EVA `oracle-full` contrast would be inflated by a format artefact rather than a grounding effect.

Multi-interval clues. 11% of items carry multiple ground-truth clue intervals; since all three agents consume one contiguous window, these are served as their temporal hull.

Abstention. Unparseable output counts as wrong and is additionally reported as abstention; abstention rates are $\leq 1.7\%$ in every grounding condition (max: EVA `off-target`).

Dr.V diagnosis coverage. Diagnosis failure rates are 3.3–16.7% on DR.V-BENCH and 1.1–3.4% on CG-BENCH; rates and per-level (perceptive/temporal/cognitive) attributions are included in the released reports. On CG-BENCH the graded verdict variant is as flat as the binary one (0.746–0.768 across all cells).

Compute and reproduction. The campaign comprises 60,008 intervened runs and 5,784 diagnoses. Item manifests and span generation are seeded, so all three agents receive byte-identical spans and a re-run reproduces every number. Each run row records its condition, requested and effective span, frame timestamps, frame-level evidence rate, span IoU, parse path, and latency. Only the 227 required videos (64 GB) of the gated 411 GB CG-BENCH release were fetched, by reading each remote archive’s central directory over HTTP range requests.

E Related Work

E.1 Video Understanding Pipelines

Video understanding has evolved from monolithic vision-language models that directly answer questions from uniformly sampled video frames to increasingly agentic systems that decompose reasoning into multiple stages Tang et al. [2025], Maaz et al. [2024], Zhang et al. [2023], Lin et al. [2024], Li et al. [2024b], Wang et al. [2024a]. Early video-language models typically process an entire video through a single forward pass and generate an answer without exposing intermediate reasoning, making their decision process difficult to interpret or diagnose Maaz et al. [2024], Zhang et al. [2023], Lin et al. [2024], Li et al. [2024b]. To improve temporal grounding, long-video understanding, and reasoning efficiency, recent approaches increasingly decompose video understanding into multiple functional stages, including temporal localization, visual observation, memory retrieval, and answer synthesis Song et al. [2024], Ren et al. [2024], Huang et al. [2024], Fan et al. [2024], Wang et al. [2024a], Liu et al. [2025b]. Representative systems include retrieval-based pipelines, multi-agent frameworks, and role-specialized architectures, where different modules are responsible for locating relevant moments, extracting visual evidence, and integrating observations into a final response Wang et al. [2024a], Fan et al. [2024], Liu et al. [2025b], Zhang et al. [2026a], Liu et al. [2025a], Zhang et al. [2026b]. Such modular designs substantially improve scalability, interpretability, and long-video reasoning capability, and have rapidly become a common paradigm for modern video understanding systems Tang et al. [2025], Wang et al. [2024a], Zhang et al. [2026a], Liu et al. [2025a], Zhang et al. [2026b]. However, while video understanding has evolved into an inherently multi-stage process, its evaluation largely remains end-to-end Mangalam et al. [2023], Li et al. [2024a], Fu et al. [2025]. Existing studies primarily assess only the correctness of the final prediction, providing limited insight into how errors emerge and propagate across the underlying reasoning pipeline.

E.2 Video Hallucination Evaluation

Hallucination has recently emerged as one of the central challenges in video understanding, motivating a growing number of benchmark datasets and evaluation protocols Huang et al. [2026], Wang et al. [2024c], Zhang et al. [2024], Kong et al. [2025], Li et al. [2025], Rawal et al. [2025b]. Existing benchmarks characterize hallucination from different perspectives, including visual fabrication, temporal inconsistency, semantic aggregation, descriptive omission, and event-level reasoning errors Wang et al. [2024c], Yang et al. [2024], Li et al. [2025], Kong et al. [2025], Rawal et al. [2025b], Zhang et al. [2024], Lu et al. [2026]. Some benchmarks evaluate whether generated answers are faithfully grounded in visual evidence, while others measure caption faithfulness or consistency between local observations and global event understanding Wang et al. [2024c], Yang et al. [2024], Rawal et al. [2025b], Zhang et al. [2024], Lu et al. [2026]. Together, these benchmarks have substantially advanced the systematic evaluation of hallucination in video-language models and enabled quantitative comparison across different architectures Huang et al. [2026], Li et al. [2025], Rawal et al. [2025b], Lu et al. [2026]. Despite their different definitions and evaluation protocols, most existing benchmarks share a common assumption: hallucination is evaluated solely from the model’s final output Wang et al. [2024c], Li et al. [2025], Rawal et al. [2025b], Lu et al. [2026]. Consequently, they determine whether an answer is hallucinated, but provide little evidence about where the hallucination originates. For modern multi-stage video agents, however, hallucinated responses may arise from inaccurate temporal localization, incomplete visual observation, or incorrect reasoning over otherwise correct evidence. End-to-end evaluation therefore conflates failures from multiple reasoning stages, making targeted diagnosis and model improvement considerably more difficult.

E.3 Benchmark Diagnosis

Beyond proposing new benchmarks, an emerging line of research has begun to investigate the reliability and limitations of existing evaluation benchmarks themselves Chen et al. [2024], Yang et al. [2025], Alzahrani et al. [2024], Feng et al. [2025], Akhtar et al. [2026]. Previous studies have examined issues including benchmark saturation, dataset bias, inconsistent model rankings across benchmarks, shortcut learning, and the sensitivity of evaluation results to inference configurations Chen et al. [2024], Yang et al. [2025], Alzahrani et al. [2024], Ramos et al. [2025], Akhtar et al. [2026], Feng et al. [2025], Fu et al. [2026]. These findings suggest that benchmark performance does not always faithfully reflect a model’s underlying capability, motivating more reliable evaluation methodologies

and benchmark diagnosis Chen et al. [2024], Alzahrani et al. [2024], Akhtar et al. [2026], Feng et al. [2025], Fu et al. [2026]. Our work is closely related to this direction but differs in both objective and methodology. Rather than introducing another hallucination benchmark or comparing benchmark difficulty, we investigate whether existing benchmarks remain informative when the evaluation target shifts from monolithic video-language models to modern multi-stage video agents. Specifically, we organize representative benchmarks according to the functional stages of video understanding and study how effectively they diagnose hallucinations throughout the reasoning pipeline. Our analysis reveals that benchmarks developed independently for individual capabilities can produce inconsistent, incomplete, or even misleading signals when collectively evaluating modern agentic architectures. Methodologically, prior benchmark-diagnosis studies remain observational: they compare scores across datasets, models, or inference settings, and thus inherit the confounds of the distributions they compare. We instead adopt the standard of *controlled intervention*—overwriting a single pipeline stage with a known value while holding the downstream task fixed on one stage-annotated item set—which turns stage attribution from a correlational conjecture into a measured causal quantity, and provides a reference standard against which the predictive value of existing benchmark scores can itself be audited.

F EVA’s Sweep on CG-Bench

Method	Size	long-acc.	mIoU	rec.@IoU	acc.@IoU	Averaged Token Usage		
						Visual	Total	Vis. Frac.
EVA ($mt=4, nf=16$)	7B	36.93	4.41	4.97	2.63	8.1K	9.2K	85.8%
EVA ($mt=4, nf=32$)	7B	39.00	4.47	4.90	2.57	13.2K	14.7K	88.3%
EVA ($mt=4, nf=64$)	7B	41.00	4.83	5.35	2.69	16.5K	18.5K	88.3%
EVA ($mt=6, nf=16$)	7B	38.17	4.58	5.28	2.93	8.2K	9.5K	84.5%
EVA ($mt=6, nf=32$)	7B	40.20	4.51	4.99	2.59	13.3K	14.9K	87.8%
EVA ($mt=6, nf=64$)	7B	41.70	4.87	<u>5.39</u>	2.85	16.6K	18.7K	88.2%
EVA ($mt=8, nf=16$)	7B	37.53	4.51	5.28	3.17	8.1K	9.4K	84.4%
EVA ($mt=8, nf=32$)	7B	40.00	4.64	5.13	2.71	13.3K	14.9K	87.9%
EVA ($mt=8, nf=64$)	7B	40.70	4.77	5.38	2.69	16.6K	18.6K	88.2%
EVA ($mt=10, nf=16$)	7B	37.77	4.50	5.14	2.83	8.2K	9.5K	84.6%
EVA ($mt=10, nf=32$)	7B	39.87	4.52	4.87	2.58	13.2K	14.8K	87.8%
EVA ($mt=10, nf=64$)	7B	<u>41.03</u>	4.79	5.32	2.61	16.6K	18.7K	88.1%
EVA ($mt=16, nf=16$)	7B	37.83	4.49	5.21	2.73	8.2K	9.5K	84.5%
EVA ($mt=16, nf=32$)	7B	39.37	4.64	5.21	2.57	13.3K	14.9K	87.8%
EVA ($mt=16, nf=64$)	7B	40.80	4.77	5.27	2.64	16.8K	18.9K	88.2%

Table 7: Performance of EVA variants on CG-Bench [Chen et al., 2025b]. mt denotes the maximum number of tool-call turns and nf the number of frames per tool call.

G Eva’s Sweep on Next-GQA

H LVBench and ARGUS Performance

I Per-Sample Cascade Analysis: VideoMind Grounding Quality on CGBench

Table 11 reports how VIDEOMIND-7B’s per-sample grounding quality (measured as IoU between the Grounder’s predicted span and the ground-truth temporal span) correlates with final answer correctness across $N = 3,000$ CGBench questions. The threshold $\text{IoU} \geq 0.1$ is used to distinguish samples where the Grounder at least partially localized the relevant segment from those where it missed entirely; mean grounding IoU is 0.065, confirming that accurate temporal localization is rare and constitutes the primary bottleneck.

Method	Size	IoU			IoP			Long Acc	Acc@GQA	Averaged Token Usage		
		R@0.3	R@0.5	mIoU	R@0.3	R@0.5	mIoP			Visual	Total	Vis. Frac.
EVA ($mt=4, nf=16$)	7B	35.47	17.46	27.34	37.44	20.17	29.83	69.43	14.84	3.2K	4.0K	79.9%
EVA ($mt=4, nf=32$)	7B	34.51	16.60	26.93	36.10	19.20	29.21	68.99	14.26	5.9K	7.0K	84.8%
EVA ($mt=4, nf=64$)	7B	33.46	16.35	26.62	34.84	18.59	28.69	68.95	13.80	10.1K	11.6K	86.7%
EVA ($mt=6, nf=16$)	7B	35.37	17.25	27.29	37.04	20.06	29.82	69.25	14.76	3.2K	4.0K	80.0%
EVA ($mt=6, nf=32$)	7B	34.55	16.39	26.87	36.16	19.14	29.22	69.27	13.92	5.9K	6.9K	84.7%
EVA ($mt=6, nf=64$)	7B	33.38	16.25	26.52	34.91	18.62	28.62	68.87	13.58	10.1K	11.7K	86.6%
EVA ($mt=8, nf=16$)	7B	35.37	17.40	27.27	<u>37.15</u>	20.08	29.68	68.97	14.72	3.2K	4.0K	79.8%
EVA ($mt=8, nf=32$)	7B	34.63	16.67	27.02	36.33	19.31	29.30	68.85	14.19	5.8K	6.9K	84.6%
EVA ($mt=8, nf=64$)	7B	33.37	16.27	26.61	34.88	18.55	28.73	69.31	13.42	10.1K	11.7K	86.7%
EVA ($mt=10, nf=16$)	7B	35.66	17.55	27.39	37.44	20.29	29.83	69.16	15.01	3.2K	4.0K	79.8%
EVA ($mt=10, nf=32$)	7B	34.23	16.73	26.90	35.95	19.24	29.17	69.37	14.09	5.9K	6.9K	84.6%
EVA ($mt=10, nf=64$)	7B	33.17	16.21	26.52	34.67	18.59	28.65	68.32	13.61	10.2K	11.7K	86.7%
EVA ($mt=16, nf=16$)	7B	35.24	<u>17.51</u>	27.31	37.04	<u>20.23</u>	29.70	<u>69.43</u>	<u>14.99</u>	3.2K	4.0K	79.9%
EVA ($mt=16, nf=32$)	7B	34.53	16.86	27.03	36.12	19.50	29.36	69.90	14.28	5.9K	6.9K	84.6%
EVA ($mt=16, nf=64$)	7B	33.38	16.18	26.56	34.88	18.55	28.63	68.89	13.54	10.1K	11.6K	86.7%

Table 8: Ablation study on EVA sweep configurations on NExT-GQA [Xiao et al., 2024a]. mt denotes the maximum number of tool-call turns and nf the number of frames per tool call.

Model	Entity Recognition	Event Understanding	Key Info. Retrieval	Reasoning	Summarization	Overall Avg.
VideoMind-7B	62.90	57.61	62.50	29.73	54.14	54.79
VideoMind-2B	61.29	58.70	58.33	37.84	50.00	55.25
Video-R1	58.06	65.22	66.67	45.95	42.86	59.82
MARC-3B	38.71	30.43	58.33	29.73	14.29	35.16
EVA	66.13	56.52	70.83	54.05	57.14	60.27
# of samples	62	92	24	37	14	total: 219

Table 9: Performance comparison on different LVBench categories. Metric: Accuracy (%).

J Performance on ELV-Halluc Dataset

K Performance on VideoHalluc

These tables report per-category performance on VideoHalluc [Wang et al., 2024b] for Video-R1, VideoMind-7B, and EVA under all evaluated configurations. nf denotes the number of input frames; for EVA, mt additionally denotes the maximum number of tool-call turns. Three metrics are reported: *Basic Acc* (accuracy on basic questions), *Halluc Acc* (accuracy on hallucination questions), and *Both Acc* (both questions simultaneously correct).

Model	Param.		Cost _H ↓	Cost _O ↓
	n _{frame}	n _{max-tool}		
Video-R1	16	–	0.6825	0.8384
	32	–	0.6615	0.8358
	64	–	0.6651	0.8328
VideoMind-7B	16	–	0.5518	0.8299
	32	–	0.5563	0.8306
	64	–	0.5540	0.8424
EVA	16	4	0.4956	0.8254
	32	4	0.4880	0.8157
	64	4	0.4795	0.8141
	16	6	0.5165	0.8262
	32	6	0.5310	0.8281
	64	6	0.5305	0.8159
	16	8	0.5292	0.8190
	32	8	0.5305	0.8156
	64	8	0.5232	0.8107
	16	10	0.5235	0.8254
	32	10	0.5251	0.8191
	64	10	0.5390	0.8136
	16	16	0.5253	0.8184
	32	16	0.5219	0.8199
	64	16	0.5150	0.8223

Table 10: Cost comparison under different frame numbers (n_{frame}) and maximum tool budgets (max_{tool}). Lower is better.

Category	N	Mean IoU	Acc (IoU<0.1)↑	Acc (IoU≥0.1)↑	Δ↑
2D Spatial Perception	180	0.052	33.1	46.2	+13.1
Entity Cognition	300	0.046	27.5	23.8	−3.7
Entity Perception	600	0.071	31.0	38.1	+7.0
Event Cognition	300	0.087	40.3	51.4	+11.1
Event Perception	600	0.069	30.6	43.0	+12.4
Hallucination	240	0.054	38.8	59.0	+20.2
Scene Cognition	15	0.176	20.0	60.0	+40.0
Scene Perception	180	0.047	47.4	73.1	+25.7
Text Cognition	45	0.081	30.6	55.6	+25.0
Text Perception	300	0.065	44.4	54.0	+9.6
Time Cognition	60	0.171	25.6	14.3	−11.4
Time Perception	180	0.031	33.8	50.0	+16.2
Overall	3000	0.065	34.6	44.9	+10.3

Table 11: Per-sample cascade analysis for VIDEOMIND-7B on CGBench (default configuration, $nf = 32$). For each CGBench question the Grounder’s predicted temporal span is compared against the ground-truth span to compute IoU. “Acc (IoU<0.1)” and “Acc (IoU≥0.1)” are final answer accuracies when grounding fails and succeeds, respectively. Δ is the cascade lift in percentage points. Scene, Hallucination, and Text categories show the strongest cascade (>20pp); Time Cognition reverses, suggesting temporal-reasoning questions are impaired by narrow window grounding. $r = 0.072$, $p < 0.001$ overall.

Models	LLM Size	Visual Details			Object			Action			Declarative Content			Avg Acc $\uparrow$	Avg Diff. $\downarrow$	SAH Ratio $\downarrow$
		In.	Out.	Diff.	In.	Out.	Diff.	In.	Out.	Diff.	In.	Out.	Diff.			
Existing Results from ELV-Halluc																
InternVL3-1B	0.5B	8	11	3	8.7	11	2.3	8.7	12.5	3.8	11.3	8.3	-3	9.9	1.5	1.6
InternVL3-2B	1.5B	7	15.5	8.5	8.7	17.2	8.5	7.2	10.5	3.3	10	13	3	11.1	5.8	6.3
SmolVLM-2.2B	1.7B	0	0	0	3	5	2	0	0	0	0	0	0	1	0.5	0.5
Qwen2.5VL-3B	3B	2.2	10.5	8.3	7.7	13.8	6.1	5	8	3	6	6	0	7.4	4.3	4.5
LLaVA-Video-7B	7B	3.7	3.7	0	4.5	2.5	-2	3.7	3.2	-0.5	4	4	0	3.6	-0.6	-0.6
Video-chatgpt-7B	7B	2	2.5	0.5	2.5	1.7	-0.7	1.2	1.2	0	2.2	3.2	1	2.0	0.2	0.1
LLaVA-OV-7B	7B	8	13.2	5.2	9.5	13.7	4.2	8.7	10.7	2	7.7	7.5	-0.2	9.9	2.8	3.0
Qwen2.5VL-7B	7B	10.2	26	15.8	17.5	30.7	13.2	13	20.7	7.7	16.8	10.5	-6.3	18.1	7.6	8.8
InternVL3-8B	7B	12.5	19.5	7.0	14.5	19.5	5.0	13.5	20.5	7.0	12.8	17.7	4.9	16.3	5.9	6.8
InternVL3-14B	14B	17.5	24.5	7.0	22.8	24.5	1.7	16.3	17.7	1.4	15.2	15.5	0.3	19.2	2.6	3.1
Qwen2.5VL-32B	32B	16.5	24.5	8.0	21.7	24.5	2.8	17.2	15.0	-2.2	15.2	7.2	-8.0	17.7	0.1	0.2
InternVL3-38B	32B	25.3	29	3.7	24.2	28	3.8	24	30	6	24.5	24.2	-0.3	26.1	3.3	4.3
Qwen2.5VL-72B	72B	24	35.5	11.5	35.7	41.5	5.8	27.8	32.3	4.5	32.3	27	-5.3	32.0	4.1	5.8
InternVL3-78B	72B	25	31.2	6.2	32	36.5	4.5	28.5	31.2	2.7	24.2	26.5	2.3	29.3	3.9	5.4
GPT-4o	/	7.7	8.3	0.6	8	8.7	0.7	8.7	10.2	1.5	8.5	9.5	1	8.7	0.9	1.0
Gemini2.5-Flash	/	47	58	11	56.5	58.8	2.3	50.5	53.2	2.7	48.7	52	3.3	53.1	4.8	9.8
VideoMind																
VideoMind-2B (f=16)	2B	9.0	11.0	2.1	8.2	9.7	1.5	10.5	11.8	1.3	9.5	7.9	-1.5	9.7	0.8	0.9
VideoMind-2B (f=32)	2B	9.7	11.5	1.8	7.7	9.2	1.5	9.5	11.0	1.5	8.7	8.7	0.0	9.5	1.2	1.3
VideoMind-7B (f=16)	7B	12.6	20.0	7.4	18.2	24.6	6.4	13.1	15.1	2.1	13.8	12.6	-1.3	16.2	3.7	4.3
VideoMind-7B (f=32)	7B	6.9	14.9	8.0	12.1	19.5	7.4	10.0	10.5	0.5	8.2	9.0	0.8	11.4	4.2	4.6
VideoMind-2B (f=64)	2B	13.8	13.6	-0.3	12.3	12.3	0.0	11.3	13.6	2.3	9.7	9.0	-0.8	12.0	0.3	0.4
VideoMind-7B (f=64)	7B	6.2	10.0	3.8	9.0	16.2	7.2	6.7	10.3	3.6	5.6	7.2	1.5	8.9	4.0	4.3
VideoMind-2B (f=128)	2B	21.3	20.5	-0.8	17.7	19.7	2.1	21.0	20.8	-0.3	21.0	19.7	-1.3	20.2	-0.1	-0.1
VideoMind-7B (f=128)	7B	3.6	8.2	4.6	6.9	10.3	3.3	4.6	8.5	3.8	5.6	5.4	-0.3	6.6	2.9	3.0
VideoMind-2B (f=256)	2B	25.6	29.5	3.8	25.1	24.1	-1.0	27.7	28.2	0.5	26.2	24.6	-1.5	26.4	0.4	0.6
VideoMind-7B (f=256)	7B	2.8	5.1	2.3	4.4	8.5	4.1	3.3	5.4	2.1	4.1	4.4	0.3	4.7	2.2	2.3
Video-R1 (frame-count sweep)																
Video-R1-7B-16F	7B	12.1	15.9	3.8	14.9	15.4	0.5	15.1	13.1	-2.1	19.2	15.9	-3.3	15.2	-0.3	-0.3
Video-R1-7B-32F	7B	9.2	15.9	6.7	10.3	17.7	7.4	13.3	16.2	2.8	13.3	13.3	0.0	13.7	4.2	4.8
Video-R1-7B-64F	7B	6.9	13.8	6.9	10.0	15.9	5.9	10.0	14.9	4.9	12.3	8.2	-4.1	11.5	3.4	3.8
EVA (parameter sweep)																
EVA (mt4, 16F)	7B	0.5	0.8	0.3	0.0	1.8	1.8	0.8	0.3	-0.5	0.5	0.0	-0.5	0.6	0.3	0.3
EVA (mt4, 32F)	7B	0.3	0.3	0.0	1.5	1.0	-0.5	0.5	0.0	-0.5	1.0	0.0	-1.0	0.6	-0.5	-0.5
EVA (mt4, 64F)	7B	0.8	0.5	-0.3	0.5	0.5	0.0	0.5	1.0	0.5	0.5	0.0	-0.5	0.5	-0.1	-0.1
EVA (mt4, 128F)	7B	0.3	0.3	0.0	0.5	0.5	0.0	0.5	0.3	-0.3	0.5	0.0	-0.5	0.4	-0.2	-0.2
EVA (mt6, 16F)	7B	0.0	1.0	1.0	0.5	1.5	1.0	0.3	0.3	0.0	0.8	0.0	-0.8	0.5	0.3	0.3
EVA (mt6, 32F)	7B	0.5	0.8	0.3	1.3	0.3	-1.0	0.5	0.3	-0.3	0.5	0.3	-0.3	0.5	-0.3	-0.3
EVA (mt6, 64F)	7B	0.3	0.8	0.5	0.3	0.8	0.5	0.3	0.5	0.3	0.8	0.0	-0.8	0.4	0.1	0.1
EVA (mt6, 128F)	7B	0.3	0.8	0.5	0.8	0.8	0.0	0.5	0.8	0.3	0.0	0.3	0.3	0.5	0.3	0.3
EVA (mt8, 16F)	7B	0.0	1.0	1.0	1.3	1.3	0.0	0.5	0.8	0.3	0.5	0.0	-0.5	0.7	0.2	0.2
EVA (mt8, 32F)	7B	0.3	0.3	0.0	0.8	0.5	-0.3	0.5	0.3	-0.3	0.8	0.0	-0.8	0.4	-0.3	-0.3
EVA (mt8, 64F)	7B	0.3	0.5	0.3	1.0	0.8	-0.3	0.3	0.8	0.5	0.3	0.0	-0.3	0.5	0.1	0.1
EVA (mt8, 128F)	7B	0.3	0.0	-0.3	0.5	0.3	-0.3	0.3	0.5	0.3	0.0	0.0	0.0	0.2	-0.1	-0.1
EVA (mt10, 16F)	7B	0.3	1.5	1.3	1.0	1.3	0.3	0.5	0.8	0.3	0.8	0.0	-0.8	0.8	0.3	0.3
EVA (mt10, 32F)	7B	0.3	0.8	0.5	1.3	1.0	-0.3	0.3	0.5	0.3	0.8	0.3	-0.5	0.6	0.0	0.0
EVA (mt10, 64F)	7B	0.3	1.0	0.8	0.5	0.8	0.3	0.0	1.0	1.0	0.5	0.0	-0.5	0.5	0.4	0.4
EVA (mt10, 128F)	7B	0.0	0.0	0.0	0.3	0.3	0.0	0.0	0.3	0.3	0.5	0.0	-0.5	0.2	-0.1	-0.1
EVA (mt16, 16F)	7B	0.5	1.3	0.8	1.0	1.0	0.0	0.8	0.5	-0.3	0.3	0.3	0.0	0.7	0.1	0.1
EVA (mt16, 32F)	7B	0.3	0.5	0.3	0.5	1.0	0.5	0.5	0.5	0.0	0.5	0.3	-0.3	0.5	0.1	0.1
EVA (mt16, 64F)	7B	0.8	1.0	0.3	1.0	1.0	0.0	0.0	0.8	0.8	0.5	0.0	-0.5	0.6	0.1	0.1

Table 12: Performance on ELV-Halluc. “In.” and “Out.” denote accuracies on in-video and out-video hallucination samples, respectively. “Diff.” denotes the difference between the two accuracies. Lower SAH Ratio indicates better robustness against semantic aggregation hallucinations.

Method	Size	Overall	Obj-Rel	Temporal	Sem-Det	Ext-Fact	Ext-NF	Fact-Det	Interact
Video-R1 ($nf=16$)	7B	81.08	72.00	55.11	93.00	91.50	89.50	98.00	69.35
Video-R1 ($nf=32$)	7B	71.67	28.00	51.70	84.00	91.50	92.00	94.00	67.74
Video-R1 ($nf=64$)	7B	64.08	4.50	60.80	52.00	91.00	93.00	94.00	70.16
VideoMind ($nf=16$)	7B	74.75	81.00	55.68	89.00	89.50	89.50	34.00	54.03
VideoMind ($nf=32$)	7B	76.25	82.50	56.25	92.00	90.00	90.00	38.00	55.65
VideoMind ($nf=64$)	7B	79.00	82.00	61.93	88.50	92.50	92.50	60.00	54.84
EVA ($mt=4$, $nf=16$)	7B	88.50	84.00	91.48	93.00	92.00	93.00	98.00	63.71
EVA ($mt=4$, $nf=32$)	7B	88.25	84.50	94.89	92.00	94.00	89.00	98.00	60.48
EVA ($mt=4$, $nf=64$)	7B	88.00	82.00	95.45	91.50	92.00	92.50	99.00	58.87
EVA ($mt=4$, $nf=128$)	7B	87.58	81.50	94.32	93.50	91.50	90.50	98.00	58.87
EVA ($mt=6$, $nf=16$)	7B	88.42	85.50	93.18	92.50	92.50	92.50	97.00	59.68
EVA ($mt=6$, $nf=32$)	7B	87.17	82.50	92.05	91.50	91.50	89.50	98.00	61.29
EVA ($mt=6$, $nf=64$)	7B	88.00	83.00	93.18	92.50	93.50	90.50	98.00	60.48
EVA ($mt=6$, $nf=128$)	7B	88.25	80.50	94.32	93.00	92.00	92.50	99.00	62.90
EVA ($mt=8$, $nf=16$)	7B	88.08	82.50	93.18	93.00	92.50	91.00	97.00	62.90
EVA ($mt=8$, $nf=32$)	7B	88.42	85.00	96.02	92.00	90.50	91.00	100.00	60.48
EVA ($mt=8$, $nf=64$)	7B	88.83	83.00	95.45	92.50	92.50	92.50	99.00	62.90
EVA ($mt=8$, $nf=128$)	7B	87.92	83.50	92.61	94.00	90.00	92.50	98.00	59.68
EVA ($mt=10$, $nf=16$)	7B	88.25	81.00	93.18	94.00	91.50	92.50	99.00	62.90
EVA ($mt=10$, $nf=32$)	7B	87.67	82.00	94.89	91.50	92.50	91.00	99.00	58.06
EVA ($mt=10$, $nf=64$)	7B	88.00	83.00	93.75	93.50	92.50	90.50	98.00	59.68
EVA ($mt=10$, $nf=128$)	7B	88.33	82.00	94.89	93.00	93.00	91.00	99.00	61.29
EVA ($mt=16$, $nf=16$)	7B	88.42	79.50	94.89	94.00	93.50	92.00	98.00	62.90
EVA ($mt=16$, $nf=32$)	7B	87.17	81.00	92.05	92.00	91.50	91.00	100.00	58.87
EVA ($mt=16$, $nf=64$)	7B	87.67	79.50	92.61	93.50	91.50	91.00	100.00	62.90
EVA ($mt=16$, $nf=128$)	7B	87.83	82.50	94.32	91.50	92.50	91.50	97.00	60.48

Table 13: VideoHalluciner [Wang et al., 2024b] **Basic Accuracy (%)** per category. Obj-Rel = Object Relation, Sem-Det = Semantic Detail, Ext-Fact = External Factual, Ext-NF = External Non-Factual, Fact-Det = Fact Detection, Interact = Interaction.

Method	Size	Overall	Obj-Rel	Temporal	Sem-Det	Ext-Fact	Ext-NF	Fact-Det	Interact
Video-R1 ($nf=16$)	7B	55.50	68.50	90.34	63.50	14.00	55.50	50.00	43.55
Video-R1 ($nf=32$)	7B	45.50	24.00	86.36	54.50	14.00	55.00	49.00	40.32
Video-R1 ($nf=64$)	7B	41.08	5.00	85.23	37.50	14.50	54.00	64.00	45.97
VideoMind ($nf=16$)	7B	63.75	76.50	88.64	75.00	26.00	63.50	41.00	69.35
VideoMind ($nf=32$)	7B	64.58	76.50	88.64	74.50	33.00	67.00	43.00	59.68
VideoMind ($nf=64$)	7B	62.33	77.00	88.64	71.50	28.00	66.00	41.00	53.23
EVA ($mt=4$, $nf=16$)	7B	46.00	72.50	43.75	59.00	24.50	52.00	28.00	25.00
EVA ($mt=4$, $nf=32$)	7B	45.83	72.50	42.61	56.00	20.00	53.00	40.00	25.81
EVA ($mt=4$, $nf=64$)	7B	45.75	73.50	37.50	58.50	22.00	56.50	30.00	25.81
EVA ($mt=4$, $nf=128$)	7B	46.17	74.50	43.75	59.00	19.00	55.50	31.00	24.19
EVA ($mt=6$, $nf=16$)	7B	46.92	73.50	44.32	57.00	22.50	50.50	36.00	33.87
EVA ($mt=6$, $nf=32$)	7B	46.17	72.50	44.32	58.00	20.50	51.00	37.00	28.23
EVA ($mt=6$, $nf=64$)	7B	45.50	74.50	42.05	58.00	18.50	52.50	31.00	27.42
EVA ($mt=6$, $nf=128$)	7B	46.25	75.50	40.91	57.00	22.00	51.50	35.00	29.03
EVA ($mt=8$, $nf=16$)	7B	46.33	75.50	44.32	59.50	20.00	50.50	32.00	28.23
EVA ($mt=8$, $nf=32$)	7B	46.67	70.00	44.32	57.50	24.00	55.50	34.00	27.42
EVA ($mt=8$, $nf=64$)	7B	45.50	75.00	39.20	59.50	19.00	53.00	34.00	24.19
EVA ($mt=8$, $nf=128$)	7B	46.33	75.00	44.89	56.00	19.50	55.50	28.00	29.84
EVA ($mt=10$, $nf=16$)	7B	46.33	73.50	43.75	59.00	22.00	49.00	34.00	30.65
EVA ($mt=10$, $nf=32$)	7B	46.17	76.00	40.34	59.50	22.50	48.50	40.00	24.19
EVA ($mt=10$, $nf=64$)	7B	47.17	75.00	43.75	60.00	20.50	55.50	30.00	29.84
EVA ($mt=10$, $nf=128$)	7B	47.33	74.50	45.45	59.50	19.00	54.00	36.00	30.65
EVA ($mt=16$, $nf=16$)	7B	45.58	72.00	43.18	55.50	20.50	51.00	32.00	33.06
EVA ($mt=16$, $nf=32$)	7B	46.67	74.00	43.75	59.00	21.50	52.00	40.00	24.19
EVA ($mt=16$, $nf=64$)	7B	47.33	72.00	42.61	59.00	22.00	57.00	37.00	29.03
EVA ($mt=16$, $nf=128$)	7B	47.25	75.50	43.18	58.50	19.00	54.50	38.00	30.65

Table 14: VideoHalluciner [Wang et al., 2024b] **Hallucination Accuracy (%)** per category.

Method	Size	Overall	Obj-Rel	Temporal	Sem-Det	Ext-Fact	Ext-NF	Fact-Det	Interact
Video-R1 ($nf=16$)	7B	42.58	52.00	48.86	58.00	12.00	48.00	49.00	29.03
Video-R1 ($nf=32$)	7B	34.08	20.00	43.18	46.50	12.50	49.50	47.00	23.39
Video-R1 ($nf=64$)	7B	29.17	3.50	51.14	19.00	11.50	47.50	59.00	30.65
VideoMind ($nf=16$)	7B	44.58	61.00	46.59	64.50	21.00	54.50	11.00	32.26
VideoMind ($nf=32$)	7B	46.83	61.50	49.43	67.00	27.50	58.00	13.00	27.42
VideoMind ($nf=64$)	7B	46.58	62.50	53.98	60.50	24.00	59.50	22.00	23.39
EVA ($mt=4, nf=16$)	7B	40.67	61.00	42.05	52.50	21.00	47.50	28.00	17.74
EVA ($mt=4, nf=32$)	7B	39.08	60.00	42.05	50.00	17.00	44.00	39.00	11.29
EVA ($mt=4, nf=64$)	7B	39.67	61.00	36.93	53.50	20.00	49.00	30.00	11.29
EVA ($mt=4, nf=128$)	7B	39.42	59.00	41.48	54.50	16.00	47.50	29.00	13.71
EVA ($mt=6, nf=16$)	7B	40.42	60.50	43.75	51.50	19.50	45.50	35.00	15.32
EVA ($mt=6, nf=32$)	7B	39.50	59.00	42.61	52.00	16.50	43.00	37.00	16.94
EVA ($mt=6, nf=64$)	7B	39.17	61.50	40.91	52.50	14.50	45.50	31.00	15.32
EVA ($mt=6, nf=128$)	7B	39.00	59.00	39.77	52.00	17.50	44.50	35.00	13.71
EVA ($mt=8, nf=16$)	7B	39.75	62.00	41.48	54.00	16.50	43.00	32.00	16.94
EVA ($mt=8, nf=32$)	7B	40.50	59.00	44.32	51.50	19.00	49.00	34.00	13.71
EVA ($mt=8, nf=64$)	7B	39.42	61.00	37.50	54.50	15.50	48.00	34.00	12.10
EVA ($mt=8, nf=128$)	7B	40.08	62.00	43.18	52.00	16.00	49.00	28.00	15.32
EVA ($mt=10, nf=16$)	7B	40.00	60.00	41.48	54.50	18.00	42.50	34.00	18.55
EVA ($mt=10, nf=32$)	7B	39.92	62.00	40.34	54.50	18.00	42.00	40.00	12.10
EVA ($mt=10, nf=64$)	7B	39.75	60.00	42.05	54.50	17.00	47.50	30.00	12.10
EVA ($mt=10, nf=128$)	7B	40.67	60.50	43.75	54.50	14.00	46.50	35.00	20.16
EVA ($mt=16, nf=16$)	7B	39.17	56.50	41.48	51.00	18.00	45.50	31.00	19.35
EVA ($mt=16, nf=32$)	7B	40.08	57.00	43.75	53.00	18.00	45.00	40.00	14.52
EVA ($mt=16, nf=64$)	7B	40.58	58.00	41.48	54.50	17.50	49.00	37.00	15.32
EVA ($mt=16, nf=128$)	7B	40.17	62.00	42.05	52.50	15.50	47.00	36.00	14.52

Table 15: VideoHalluciner [Wang et al., 2024b] **Both Accuracy (%)** per category — both basic and hallucination questions answered correctly.