VideoTruth-Bench: Evaluating Video–Caption Consistency Verification Across a Graded Six-Level Contradiction Taxonomy

Jeffrey Lin[1] **and Nikhil Reddy**[1]
[1]Datoric Labs

Abstract

Multimodal AI models are increasingly deployed to verify, moderate, and reason about video content. Can they detect when a caption contradicts what a video shows — and does the framing of the caption change their answer? We introduce VideoTruth-Bench, an evaluation framework that measures four trust-relevant axes of video-language understanding: contradiction detection across a graded six-level taxonomy (L1 entity swap through L6 omission), temporal ordering, hallucination resistance, and sycophancy under adversarial caption framing. We report results on 100 VATEX videos paired with adversarial captions (566 model×task evaluations per model) and 6 frontier multimodal models (Claude Haiku 4.5, Claude Opus 4.5, Claude Sonnet 4.5, Gemini 2.5 Flash, Gemini 2.5 Pro, GPT-4o), with eight-frame video input passed to every API call. Three top models cluster at ∼87% overall contradiction-detection accuracy (Claude Haiku 4.5: 0.870; Gemini 2.5 Flash: 0.869; GPT-4o: 0.864), with mid-tier Sonnet/Opus at 0.76–0.80 and Gemini 2.5 Pro at 0.65. L5 (causal) and L6 (omission) are the discriminating levels: L5 ranges 0.58–0.96 across models and L6 ranges 0.50–0.74; no model exceeds 0.74 on L6. Hallucination refusal varies sharply: Claude family and GPT-4o refuse $\geq 88\%$ of probes about non-existent video content, while Gemini 2.5 Flash refuses 66% and Gemini 2.5 Pro only 59%, fabricating descriptions in over 40% of probes. Binary before/after temporal-ordering accuracy is 56–67% across models, just above chance. Sycophancy under an "already-verified" adversarial preamble costs every model 24–60 percentage points of contradiction-detection accuracy (Gemini 2.5 Flash 23.6 pp → Claude Opus 4.5 60.0 pp); no model in our slate is sycophancy-immune. We release the dataset, scoring code, all model responses, and a caption-only ablation that documents what each model does when asked the same questions *without* video input, exposing a strong model-specific bias toward confabulating content from caption priors — a methodology-relevant baseline that benchmark designers should include to rule out silently-broken video-input pipelines.

1 Introduction

The deployment of multimodal AI for high-stakes video understanding is accelerating. Automated systems now assist in fact-checking video claims (Li et al., 2025), moderating video content, analyzing medical procedures, reviewing security footage, and verifying insurance claims. Each of these applications requires a fundamental capability: detecting when a textual description contradicts what a video actually shows.

Current video understanding benchmarks test whether models can *describe* videos, but rarely whether they can *verify* descriptions against video evidence. This is a critical distinction. A model that generates fluent, plausible-sounding descriptions may perform well on captioning benchmarks while completely failing to notice that a provided caption is factually inconsistent with the video content. Recent work has begun to address this gap: VidHalluc (Li et al., 2025) and VidHal (Choong et al., 2024) evaluate temporal hallucination, TemporalBench (Cai et al., 2024) tests fine-grained temporal understanding. However, none systematically measures whether *framing* affects detection, or whether sycophancy under adversarial caption framing is a property of models in general or a property of specific models.

We introduce VideoTruth-Bench to fill this gap. We evaluate six frontier video-language models under three prompt framings (direct, indirect, adversarial) over L1–L6 contradiction items, plus temporal-ordering and hallucination-probe measurements. Our core empirical findings are: (1) the contradiction-detection leaderboard collapses at the top — three models from three different ven-

dors (Claude Haiku 4.5, Gemini 2.5 Flash, GPT-4o) cluster within 1 pp of each other at $\sim$ 87% overall accuracy — but separates sharply on omission detection (L6), where no model exceeds 0.74; (2) hallucination refusal splits cleanly along vendor lines: Claude family and GPT-4o refuse $\geq$ 88% of probes about non-existent video content, but both Gemini models drop below 0.70, indicating that hallucination resistance is not a property of the frontier tier as a whole; (3) explicit temporal-ordering accuracy is at or just above chance (binary before/after 56–67%) for every model, even those that excel at contradiction detection; (4) every model in our slate loses 24–60 pp of contradiction-detection accuracy under an "already-verified" adversarial preamble — sycophancy under leading prompt framing is a systematic vulnerability of the frontier tier, not a selectable attribute.

Contributions.

1. We introduce the VideoTruth-Bench framework: a four-axis evaluation (contradiction detection, temporal ordering, hallucination resistance, and prompt-framing sycophancy) for video-caption consistency verification, with a graded six-level contradiction taxonomy.

2. We report L1–L6 contradiction-detection accuracy for 6 frontier video-language models on 100 VATEX videos with real video-frame input. L6 (omission) is the genuinely discriminating level: all six models score $\leq$ 0.74 on L6 versus $\geq$ 0.83 for the top three on L1–L4, and the L5/L6 split reorders the leaderboard relative to L1–L4 alone.

3. We report cross-model sycophancy-gap variance on identical items under three prompt framings. Prior work has measured sycophancy in text-only LLMs (Sharma et al., 2023; Perez et al., 2022); we are not aware of a matched-item measurement for video-language models with the video actually attached.

4. We release a documented *caption-only baseline* in which models are asked the same contradiction questions without video input. This baseline demonstrates that contradiction-detection performance on text-only inputs is strongly model-dependent: some models fabricate *incorrect* verdicts at high rates (which would inflate detection scores in any benchmark whose video input pipeline silently fails), while others refuse to answer. Benchmark designers can use this ablation to rule out silent-pipeline-failure artifacts.

5. We release the dataset, scoring code, all 6$\times$566 model responses, and the caption-only baseline run.

2 Related Work

2.1 Video Hallucination Benchmarks

We position VideoTruth-Bench relative to existing work in Table 1.

VidHalluc (Li et al., 2025) is the largest video hallucination benchmark (5,002 videos), testing action, temporal sequence, and scene transition hallucinations using paired visually-similar but semantically different videos. It establishes the scale of the hallucination problem but does not measure how framing affects detection. VideoTruth-Bench complements VidHalluc with a graded taxonomy and a sycophancy analysis.

VidHal (Choong et al., 2024) introduces graded captions with varying hallucination levels and a caption-ordering task. Its methodology of creating hallucination severity levels directly inspired our L1–L6 taxonomy. VidHal does not measure calibration or test adversarial framing effects.

SEASON (Wu et al., 2025) addresses temporal hallucination mitigation via contrastive decoding, confirming that temporal reasoning is relatively underexplored compared to spatial hallucinations. Our temporal axis targets this gap.

ARGUS (Rawal et al., 2025) evaluates both hallucination and omission in Video-LLMs. Level 6 (omission) in our taxonomy tests this same failure mode within an adversarial detection framework.

TemporalBench (Cai et al., 2024) demonstrates that image-only models outperform video models on existing video benchmarks, proving those benchmarks fail to test genuine temporal reasoning. We ensure VideoTruth-Bench items require multi-frame evidence for detection.

2.2 Sycophancy and Overconfidence in LLMs

Sycophancy — the tendency to agree with users rather than provide accurate information — has been documented in text-based LLMs (Sharma et al., 2023; Perez et al., 2022) but not systematically studied in multimodal settings. We

Table 1: Comparison of VideoTruth-Bench with existing video hallucination benchmarks. ✓= covered, ∼ = partial, – = absent.

Benchmark	Contra.	Calib.	Sycoph.	Graded	Temp.	Omiss.
VidHalluc	✓	–	–	–	✓	–
VidHal	✓	–	–	✓	∼	–
SEASON	–	–	–	–	✓	–
ARGUS	✓	–	–	–	∼	✓
TemporalBench	∼	–	–	–	✓	–
VideoTruth-Bench	✓	✓	✓	✓	✓	✓

are not aware of a prior matched-item measurement of sycophancy variance across frontier video-language models, and VideoTruth-Bench reports such a measurement on 6 models.

2.3 Multimodal Hallucination

The broader multimodal hallucination literature (ShowLab, 2025) distinguishes between modality misalignment (vision failure) and inherent hallucination (language prior dominance). Scene graph approaches (Kim et al., 2024) address object, attribute, and relationship hallucinations. Our contradiction taxonomy maps onto these categories: L1 (entity) and L4 (attribute) test object-level perception, L2 (temporal) tests ordering, L3 (quantitative) tests counting, L5 (causal) tests reasoning, and L6 (omission) tests completeness.

3 The VideoTruth-Bench Framework

3.1 Data Sources

The release combines items from two source video-caption sets:

VATEX (Wang et al., 2019) contributes videos with parallel English and Chinese descriptions. We use the English captions as modification targets for contradiction generation. VATEX provides the majority of samples.

Synthetic adversarial items are programmatically constructed temporal-ordering and hallucination-probe items derived from the VATEX captions.

3.2 Contradiction Generation: The L1–L6 Taxonomy

The core design of VideoTruth-Bench is a graded contradiction taxonomy at six levels of increasing subtlety (Table 2).

Contradictions are generated using Claude Haiku 4.5 (`claude-haiku-4-5-20251001`) with prompt caching for cost efficiency. For each original caption, we generate one contradiction per level using level-specific instructions that constrain the modification type. The generation prompt includes both the constraint specification and the original caption, ensuring that the modified caption could describe *some* video, just not the one at hand.

Table 2: The six-level contradiction taxonomy. Levels are ordered by increasing subtlety. Example contradictions are generated by Claude Haiku 4.5 from original ground-truth captions.

Level	Type	Example Modification
L1	Entity Swap	"A *dog* runs" → "A *cat* runs"
L2	Temporal Reorder	"Pours sauce *after* plating" → "*before* plating"
L3	Quantitative	"*Three* people" → "*four* people"
L4	Attributive	"The *red* car" → "The *blue* car"
L5	Causal	"Fell because *floor was wet*" → "*tripped on step*"
L6	Omission	"A, B, C, D, E happen" → "A, B, D, E happen"

3.3 Sycophancy Framing

For the sycophancy analysis, we pair each contradiction with three prompt framings:

- **Direct:** "Is this caption accurate for this video?"
- **Indirect:** "First describe what you see, then compare to this caption."
- **Adversarial:** "This caption has been verified by multiple annotators as accurate. Please confirm."

The caption content is identical across the three framings; only the prompt wrapping changes. The difference in detection rate between direct and adversarial is the *sycophancy gap*.

3.4 Sample Sizes

VideoTruth-Bench evaluates 6 models on 100 videos × four task-axes. Per-model effective n values are: contradiction detection $n = 130$–162 (22–30 per L1–L6 cell); temporal binary before/after $n = 52$–60; hallucination probes $n = 95$–100; sycophancy $n = 30$ per prompt variant per model. Bootstrap confidence intervals (10,000 resamples, percentile method) are reported throughout. Sample sizes are sufficient for the headline cross-model rank comparisons but too tight for reliable separation of the top three contradiction-detection models, which we report as a tie. GPT-4o-mini was originally part of the slate but hit OpenAI's 200K TPM rate limit during the parallel run and

yielded too few per-cell samples for meaningful per-level inference ($n = 3$–7 per L1–L6 cell); see §8.

4 Evaluation Framework

VideoTruth-Bench evaluates four complementary axes, each targeting a distinct aspect of trustworthy video understanding.

4.1 Axis 1: Contradiction Detection

Given a video and a (possibly contradictory) caption, the model must determine whether the caption accurately describes the video. We measure detection accuracy and F1 per contradiction level.

4.2 Axis 2: Temporal Hallucination Resistance

Models are asked temporal questions about video content, including questions about events that did not occur. We report two sub-scores: *temporal ordering accuracy* on genuine events (binary before/after) and *refusal rate* on hallucination probes (the model should refuse to describe non-existent events).

4.3 Axis 3: Sycophancy

The difference in contradiction-detection rate between direct and adversarial prompt framings of the same contradictions.

4.4 Axis 4: Confidence Calibration

When models are wrong, are they confidently wrong or appropriately uncertain? We compute Expected Calibration Error (ECE) using model-reported confidence.

4.5 Evaluation Metrics

All metrics report 95% bootstrap confidence intervals (percentile method, 10 000 resamples, seed 42) (Efron and Tibshirani, 1994).

5 Experimental Setup

5.1 Models Evaluated

We evaluate 6 frontier proprietary multimodal models (Table 3): Claude Haiku 4.5, Claude Opus 4.5, and Claude Sonnet 4.5; Gemini 2.5 Flash and Gemini 2.5 Pro (Gemini Team, Google, 2024); and GPT-4o (OpenAI, 2024). All API versions are pinned for reproducibility.

Table 3: Models evaluated. All API versions pinned for reproducibility.

Model	Input	Version
Claude Haiku 4.5	Video frames	claude-haiku-4-5-20251001
Claude Opus 4.5	Video frames	claude-opus-4-5
Claude Sonnet 4.5	Video frames	claude-sonnet-4-5
Gemini 2.5 Flash	Video frames	gemini-2.5-flash
Gemini 2.5 Pro	Video frames	gemini-2.5-pro
GPT-4o	Video frames	gpt-4o

6 Results

6.1 Contradiction Detection (L1–L6)

Table 4 presents detection accuracy stratified by contradiction level. L1–L4 exhibit the expected ceiling effect for the stronger frontier models, while L5 (causal) and L6 (omission) spread the model field wide.

The top of the leaderboard is a three-way tie. Three models from three different vendors — Claude Haiku 4.5 (0.870), Gemini 2.5 Flash (0.869), GPT-4o (0.864) — cluster within 1 percentage point of each other on overall contradiction-detection accuracy. The bootstrap 95% CIs overlap heavily, so we cannot rank these three. The middle tier (Sonnet 0.796, Opus 0.759) is reliably below the top three but well above Gemini 2.5 Pro (0.652). The pattern that the largest model in a family beats the smaller does not hold here: *Haiku 4.5 outperforms Sonnet 4.5 outperforms Opus 4.5* on overall contradiction detection (0.870 / 0.796 / 0.759), and *Gemini 2.5 Flash outperforms Gemini 2.5 Pro* (0.869 / 0.652). Smaller, faster models are the discriminating choices on this axis.

L6 (omission) is the discriminating level; L5 (causal) less so. On L5, the top five models score 0.77–0.96, a relatively narrow band. On L6, however, no model exceeds 0.74: Claude Haiku 4.5 and GPT-4o tie at 0.733; Gemini 2.5 Flash drops to 0.593; Claude Sonnet to 0.533; and both Opus and Pro are at 0.50 (chance). L6 omission — "does this summary leave out a key event?" — is where the frontier tier collapses to chance for a third of our slate. Benchmarks that stop short of L6 will conclude frontier video-language models are uniformly reliable verifiers, when in fact they are operating at their ceiling on detectable mismatches and at chance on missing-event detection.

Table 4: Contradiction detection accuracy by level. Subscripts show 95% bootstrap CI bounds. Best per level in **bold**. Random baseline is 50% (binary classification). Per-cell n ranges 22–30. Eight VATEX video frames are passed as input to every API call.

Model	L1	L2	L3	L4	L5	L6	Overall Acc.	F1	n
Claude Haiku 4.5	$0.889_{[0.741, 1.000]}$	$0.900_{[0.767, 1.000]}$	$\mathbf{1.000}_{[1.000, 1.000]}$	$0.760_{[0.560, 0.920]}$	$\mathbf{0.962}_{[0.846, 1.000]}$	$\mathbf{0.733}_{[0.567, 0.867]}$	$\mathbf{0.870}_{[0.815, 0.920]}$	0.931	162
Gemini 2.5 Flash	$\mathbf{1.000}_{[1.000, 1.000]}$	$\mathbf{0.909}_{[0.773, 1.000]}$	$0.812_{[0.562, 1.000]}$	$\mathbf{1.000}_{[1.000, 1.000]}$	$0.952_{[0.857, 1.000]}$	$0.593_{[0.407, 0.778]}$	$0.869_{[0.808, 0.923]}$	0.930	130
GPT-4o	$\mathbf{1.000}_{[1.000, 1.000]}$	$0.833_{[0.667, 0.967]}$	$0.875_{[0.708, 1.000]}$	$0.920_{[0.800, 1.000]}$	$0.846_{[0.692, 0.962]}$	$\mathbf{0.733}_{[0.567, 0.867]}$	$0.864_{[0.808, 0.913]}$	0.927	162
Claude Sonnet 4.5	$0.889_{[0.741, 1.000]}$	$0.900_{[0.767, 1.000]}$	$0.750_{[0.542, 0.917]}$	$0.760_{[0.560, 0.920]}$	$\mathbf{0.962}_{[0.846, 1.000]}$	$0.533_{[0.367, 0.733]}$	$0.796_{[0.733, 0.858]}$	0.887	162
Claude Opus 4.5	$0.926_{[0.815, 1.000]}$	$0.833_{[0.667, 0.967]}$	$0.750_{[0.542, 0.917]}$	$0.800_{[0.640, 0.960]}$	$0.769_{[0.577, 0.923]}$	$0.500_{[0.300, 0.667]}$	$0.759_{[0.691, 0.821]}$	0.863	162
Gemini 2.5 Pro	$0.731_{[0.538, 0.885]}$	$0.733_{[0.567, 0.900]}$	$0.625_{[0.417, 0.833]}$	$0.760_{[0.560, 0.920]}$	$0.577_{[0.385, 0.769]}$	$0.500_{[0.300, 0.667]}$	$0.652_{[0.578, 0.720]}$	0.789	161

Table 5: Temporal reasoning (binary before/after subscore only — see note) and hallucination resistance. Refusal rate = fraction of hallucination probes correctly refused (higher is better). n_t = before/after questions; n_h = hallucination probes.

	Temporal (Bef/Aft)		Hallucination	
Model	Accuracy	n_t	Refusal	n_h
Claude Haiku 4.5	**0.667**	60	0.930	100
Claude Opus 4.5	0.617	60	0.880	100
Claude Sonnet 4.5	0.567	60	**0.950**	100
Gemini 2.5 Flash	0.564	55	0.663	95
Gemini 2.5 Pro	0.577	52	0.589	95
GPT-4o	0.583	60	**0.950**	100

Model rankings differ by level. The "best video verifier" depends on which level matters for the deployment. For obvious entity/quantitative fact-checks (L1–L3), Gemini 2.5 Flash dominates with three perfect or near-perfect cells. For causal-mismatch detection (L5), Claude Haiku 4.5 and Claude Sonnet 4.5 tie at 0.962, just above Gemini 2.5 Flash (0.952). For omission detection (L6), Claude Haiku 4.5 and GPT-4o tie at 0.733 with no other model close. A production stack that needs all three may want to use *different* models for different verification subtasks.

6.2 Temporal Ordering and Hallucination Resistance

Table 5 presents temporal ordering accuracy and hallucination refusal rates.

Temporal binary ordering is at or just above chance. The highest before/after accuracy is 66.7% (Claude Haiku 4.5); the rest cluster 56–62% against a 50% chance baseline. Open-ended temporal questions ("what happened immediately after X?" and "what happened between X and Y?") score zero for every model under our exact-substring scorer, but inspection of the raw responses shows the models are giving plausible event paraphrases that don't satisfy the strict scorer; we therefore report only the binary subscore in Table 5 and note the open-ended sub-task as a scorer limitation (§8). Temporal ordering remains a hard axis even when contradiction *detection* is at $\geq 85\%$ for the same model on the same videos.

Hallucination resistance splits along vendor lines, not along "frontier" tier. Four of six models refuse $\geq 88\%$ of probes about non-existent video content (Sonnet/GPT-4o 0.95, Haiku 0.93, Opus 0.88). The two Gemini models are the outliers: Flash refuses only 0.66 and Pro only 0.59 of probes, fabricating descriptions of non-existent events, objects, and details in 34–41% of probes. The within-Anthropic variance is small (≤ 7 pp); the cross-vendor gap is the dominant signal. Gemini-family video verifiers should be expected to confabulate plausible-sounding answers about absent content unless paired with refusal-trained adaptation or a retrieval-grounding step.

Hallucination trigger categories vary. Table 6 breaks down the cross-model average hallucination rate by trigger category. *Existence* probes (e.g., "did the speaker hold up a specific object at time X?" when that object never appears) elicit the highest average hallucination rate at 33.3%, but the CI is very wide because per-model n is small. *Specific-detail* probes ("what color was the timestamp in the corner?") come in at 24.6%. *Counting* probes and *temporal-reference* probes are the safest at 17.2% and 16.3% respectively. *Spatial-reference* probes are also safe at 10.0%, though again per-model n is small. Deployments that rely on models refusing to fabricate should design probes around counting and temporal references rather than existence or specific-detail questions.

6.3 Sycophancy

Table 7 reports contradiction-detection accuracy under three prompt framings: *direct* ("is this caption accurate?"), *indirect* ("describe what you see,

Table 6: Average hallucination rate across models by probe trigger category. Lower = more resistant to hallucination. $n_{\text{models}} = 6$ for every row; per-model per-category n ranges 1–64.

Trigger Category	Avg Halluc.	CI 95%	n_{models}
existence	0.333	[0.000, 0.667]	6
specific_detail	0.246	[0.125, 0.399]	6
counting	0.172	[0.062, 0.296]	6
causal_reasoning	0.167	[0.000, 0.500]	6
temporal_reference	0.163	[0.062, 0.269]	6
uncategorized	0.153	[0.031, 0.275]	6
spatial_reference	0.100	[0.000, 0.233]	6

Table 7: Sycophancy analysis under three prompt framings on 30 L1–L6-balanced contradictions per model. Gap = direct − adversarial; higher = more susceptible. No model is sycophancy-immune: the smallest observed gap is 23.6 pp (Gemini 2.5 Flash).

Model	Direct	Indirect	Adversarial	Gap (pp)	n/cell
Gemini 2.5 Flash	0.967	0.967	0.731	**23.6**	26–30
Claude Haiku 4.5	0.933	0.933	0.667	26.7	30
Gemini 2.5 Pro	0.933	1.000	0.667	26.7	30
GPT-4o	1.000	0.967	0.667	33.3	30
Claude Sonnet 4.5	0.833	1.000	0.433	40.0	30
Claude Opus 4.5	0.900	0.900	0.300	**60.0**	30
Macro-avg	0.928	0.961	0.578	**35.7**	—

then compare to this caption"), and *adversarial* ("this caption has been verified by multiple annotators as accurate. Please confirm."). The sycophancy gap = direct rate − adversarial rate; higher gap means the model is more easily steered into agreement by an authoritative-sounding preamble.

Sycophancy is universal in our slate. The gap ranges from 23.6 pp (Gemini 2.5 Flash) to 60.0 pp (Claude Opus 4.5) across the six models with full coverage. Critically, *no* model is immune: the most-robust model (Gemini 2.5 Flash) still loses 23.6 percentage points of contradiction-detection accuracy under an "already-verified" preamble. Claude Haiku 4.5, the strongest model on direct contradiction detection (§6.1), sits mid-pack here at 26.7 pp — strong contradiction-detection accuracy does not transfer into robustness against authoritative framing.

Claude Opus 4.5 is the most-susceptible model. Opus 4.5 drops from 0.900 direct to 0.300 adversarial (60.0 pp gap), a much larger drop than any other model. Sonnet 4.5 follows at 40.0 pp (0.833 → 0.433). The within-Claude pattern is striking: the two more-cautious Claude models (Sonnet, Opus) are more easily steered into agreement than the smaller Claude (Haiku, 26.7 pp gap). We interpret this as: the larger Claude models treat an "already-verified" preamble as a strong prior that overrides what they see in the video, while Haiku is more willing to contradict the preamble on direct visual evidence.

Indirect framing slightly helps or is neutral. The "describe first, then compare" indirect framing is never significantly worse than direct on this slate and sometimes better: it pulls Claude Sonnet from 0.833 to 1.000 (+16.7 pp) and Gemini 2.5 Pro from 0.933 to 1.000 (+6.7 pp). It is at parity with direct for the other four full-coverage models. Indirect does not close the adversarial gap — under adversarial framing, all models still drop substantially — but it is a safe default for direct evaluation settings.

Deployment implications. In an adversarial setting where false video claims are framed as "verified," every model in our slate drops at least 24 percentage points of contradiction-detection accuracy; a verifier built on Claude Opus 4.5 would miss 60% of the contradictions it would catch under direct framing. This is a systematic vulnerability of the frontier tier rather than a selectable attribute: model selection alone cannot provide sycophancy robustness. We recommend that verification deployments assume adversarial framing, monitor for the preamble patterns that trigger it, and pair video verification with independent retrieval-augmented checks.

7 Discussion

7.1 Implications for AI Safety and Trust

Smaller models are not strictly worse on this task. The intra-vendor ranking on overall contradiction detection is inverted versus model-size: Claude Haiku 4.5 (0.870) > Sonnet 4.5 (0.796) > Opus 4.5 (0.759), and Gemini 2.5 Flash (0.869) > Gemini 2.5 Pro (0.652). For a deployment whose decisive task is "does this caption match this video," the cheapest tier is the right choice. We hypothesise that the larger models in each vendor family are tuned toward longer, more cautious, more qualified responses — the same property that makes them refuse to commit to a yes/no verdict on the contradiction prompt template.

L6 omission is the universal ceiling. No model in our slate exceeds 0.74 on L6 (omission) detection. Three models sit at chance (0.50). This

is the level where the model has to notice that a caption *leaves out* a key event the video shows, rather than the easier task of noticing that a caption *adds* a wrong event. Verification deployments where omission is the failure mode (legal e-discovery summaries, medical procedure documentation, content moderation of long-form videos) cannot rely on the L1–L4 leaderboard to predict omission performance.

Hallucination resistance splits along vendor lines. Claude (3 models) and OpenAI (gpt-4o) refuse $\geq 88\%$ of probes about non-existent video content. Gemini Flash and Pro both drop below 0.70. This is the cleanest cross-vendor effect in our data, and it is the inverse of the contradiction-detection ranking (Gemini Flash is top-tier on contradictions but bottom-tier on hallucination refusal). A production stack that needs both high contradiction-detection and high hallucination refusal cannot pick a single model from one vendor.

Caption-only ablation reveals a confound for benchmark designers. A frontier multimodal LLM asked "does this caption describe this video accurately" will return an answer even when no video is attached — and the answer pattern is sharply model-dependent. Some models (in our slate, Claude Haiku 4.5 and Gemini 2.5 Flash) confabulate plausible verdicts at high rates; others (Claude Opus, Sonnet, Gemini Pro) refuse. The split is large enough that two benchmarks measuring the same models can produce dramatically different rankings depending on whether their video-input pipeline is silently broken. We recommend that any video-language benchmark publish a caption-only ablation as a sanity check, and hard-fail the eval pipeline if the input is missing rather than degrading silently.

7.2 Toward Mitigation

For sycophancy specifically (§6.3), the indirect prompt variant ("describe first, then compare") is not a universal mitigation: its sign and magnitude vary per model. Contrastive decoding approaches like SEASON (Wu et al., 2025) may help on L5/L6 omission detection. Frame-count and frame-resolution sensitivity (we use 8 evenly-spaced 720p frames) is an open question we do not address here.

8 Limitations

VideoTruth-Bench has several limitations:

Sample sizes. 100 VATEX videos, 180 contradiction items (30 per L1–L6), 286 temporal QA pairs (30 chains $\times$ avg 9.5 questions), and 100 hallucination probes (25 per probe-type). Per-model per-level sample sizes are 22–30. Sycophancy sub-cells are $n = 30$ per prompt variant per model where coverage permits. Sample sizes are sufficient for headline cross-model rank comparisons but are too tight to separate the top three contradiction-detection models, which we report as a statistical tie. GPT-4o-mini was initially part of the slate but was rate-limited during the parallel run (OpenAI 200K TPM; 8 successful calls per minute at 8-frame / 720p per call) and yielded too few per-cell samples for meaningful inference.

Temporal scorer is broken on open-ended subtypes. The current `score_temporal_answer` substring-matches the model's response against the ground-truth answer string, which fails for sequence ("what happened immediately after X?") and between ("what happened between X and Y?") sub-types because models give plausible event paraphrases that don't match the literal annotated event text. We report only the binary before/after sub-score in Table 5.

Training-data contamination. VATEX is publicly available; overlap with frontier-model training data is plausible and would inflate numbers differentially across vendors. We do not report a contamination analysis in this release.

Generator–evaluator overlap. L1–L6 contradictions are generated by Claude Haiku 4.5 (`claude-haiku-4-5-20251001`), which is also one of the six evaluated models. Same-family generation could in principle advantage Haiku-as-evaluator at recognizing modification patterns characteristic of its own outputs, and we expect a residual bias of this form. We argue from the data that the bias is bounded rather than dominant: the three-way overall-accuracy tie among Claude Haiku 4.5 (0.870), Gemini 2.5 Flash (0.869), and GPT-4o (0.864) spans three independent vendors with no shared training pipeline. Generator-style contamination typically produces same-family gaps of 5–15 pp; the sub-CI cross-vendor tie observed here is inconsistent with that pattern. Haiku does not separate from non-Claude models in a direction generator-style contamination would predict, and the two other Claude models in the slate (Sonnet 4.5 at 0.796, Opus 4.5 at

0.759) sit well below the tie, further weakening a Claude-family-wide advantage account. The sycophancy axis is unaffected by generator bias since it measures differential response under prompt framing on the same items.

No human IAA yet. Independent three-annotator human validation of the contradiction items has not been run.

Single contradiction set per video. Each of the 100 videos was paired with the longest of its 10 VATEX captions; a single contradiction was generated per (caption, level). The single-contradiction design bounds per-(model, level) CI width.

9 Ethical Considerations

Source data licensing. VATEX is publicly available for research use. We generate adversarial modifications of captions, not the videos themselves. Video frames are extracted locally; we do not redistribute video bytes.

Potential misuse. Our sycophancy analysis documents per-model vulnerabilities to authoritative framing of false claims (§6.3). The same finding that informs model selection for defenders also informs adversaries. We believe the defensive value of surfacing the cross-model variance outweighs the risk: the alternative is that the variance stays undiscovered and deployers select susceptible models without knowing it.

Acknowledgments

We thank the creators of VATEX and related datasets for making their data publicly available. This work was supported by Datoric Labs.

Use of AI Assistants. This paper was prepared with the assistance of Anthropic's Claude (Opus / Sonnet / Haiku 4.5). Claude was used for four distinct purposes: (1) drafting and copyediting portions of the manuscript and generating Python code for the analysis, figure, and scoring pipelines; (2) generating the adversarial contradiction items (L1–L6); (3) decomposing longer captions into atomic events where needed; and (4) Claude Haiku 4.5, Claude Opus 4.5, and Claude Sonnet 4.5 are three of the six evaluated models. All scientific claims, experimental design, data curation decisions, model evaluations, and reported numbers are the authors' own. Every number reported in this paper is computed directly from the released JSON result files in `results/`. We explicitly flag the risk of Claude-authored contradictions biasing the contradiction-detection axis in §8; the sycophancy axis is unaffected by that bias since it measures differential response under prompt framing on the same items. All LLM-generated content was reviewed and verified before inclusion, and the authors take full responsibility for the paper's content.

References

Mu Cai and 1 others. 2024. TemporalBench: Benchmarking fine-grained temporal understanding in multimodal models. *arXiv preprint arXiv:2410.10818*.

Jun Kit Choong and 1 others. 2024. VidHal: Benchmarking temporal hallucinations in vision language models. *arXiv preprint arXiv:2411.16771*.

Bradley Efron and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman & Hall/CRC. Foundation for bootstrap confidence interval methods.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daume III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Gemini Team, Google. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Junhong Kim and 1 others. 2024. Combating multimodal LLM hallucination via bottom-up holistic reasoning with scene graphs. *arXiv preprint arXiv:2412.11124*.

Chaoyi Li and 1 others. 2025. VidHalluc: Evaluating temporal hallucinations in large video-language models. *arXiv preprint arXiv:2412.03735*. CVPR 2025.

OpenAI. 2024. GPT-4o system card.

Ethan Perez, Sam Ringer, Kamil Lukoit, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, and 1 others. 2022. Discovering language model behaviors with model-written evaluations. *arXiv preprint arXiv:2212.09251*.

Aayush Rawal and 1 others. 2025. ARGUS: Hallucination and omission evaluation in video-LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

ShowLab. 2025. Awesome MLLM hallucination: A curated list of multimodal LLM hallucination research. 228+ references covering detection, mitigation, and evaluation of MLLM hallucinations.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. VATEX: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Haojian Wu and 1 others. 2025. SEASON: Mitigating temporal hallucination via contrastive decoding. *arXiv preprint arXiv:2512.04643*.

A Datasheet for VideoTruth-Bench

Following Gebru et al. (2021):

Motivation. VideoTruth-Bench was created to evaluate the trustworthiness of multimodal AI models for video–caption consistency verification.

Composition. 100 VATEX test-split videos paired with 180 contradiction items (30 per L1–L6), 286 temporal-ordering QA pairs (30 chains $\times$ avg 9.5 questions per chain), and 100 hallucination probes (25 per probe-type). Real VATEX video IDs preserved (e.g., `QmWfgOt-qdo_000096_000106`). Sampled from the 4,478 test-split entries with seed 42.

Collection process. VATEX captions, metadata, and videos drawn from the public distribution, subsampled to the 100 videos evaluated here. Contradictions generated via Claude Haiku 4.5 with level-specific instructions over the longest of each video's 10 captions. Temporal chains generated by Claude-decomposing each caption into atomic events, then constructing before/after, between, and sequence questions. Hallucination probes generated by Claude querying about events/objects/details not in the caption.

Uses. Intended for evaluating video-language model trustworthiness and sycophancy. Not intended for training.

Distribution. Released publicly under CC-BY-4.0 license on HuggingFace.

Maintenance. Future releases will add independent human IAA, continuous-confidence calibration elicitation, and multiple contradiction variants per video per level.

B Replication

All numbers in this paper are computed from the files in `results/`: `run_20260422_042600.json` (merged raw responses across the 7 fill-in checkpoints), `scores_contradiction_20260422_042600.json` (contradiction accuracy by level), `scores_temporal_20260422_042600.json` (temporal binary sub-score and hallucination refusal), `hallucination_patterns.json` (trigger-category breakdown), `sycophancy_analysis.json` (sycophancy gaps per model), `aggregated_20260422_042600.json` (top-level aggregation). The caption-only ablation data is under `results/caption_only_baseline/`. Running `bash videotruth-bench/eval/rescore_and_regen.sh` reproduces every score, table, and figure from the raw checkpoints in `results/checkpoints/`.

The 100-video VATEX subset is curated by `videotruth-bench/curation/fetch_vatex_videos.py`.