VideoTruth-Bench: Do Multimodal Models Detect Video–Caption Contradictions or Hallucinate Agreement?

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

Abstract

Multimodal AI models are increasingly deployed to verify, moderate, and reason about video content. But can they actually detect when a caption contradicts what a video shows? We introduce VideoTruth-Bench, a benchmark of 3,142 adversarial video–caption pairs designed to test cross-modal consistency verification across 6 levels of contradiction subtlety, from obvious entity swaps to near-imperceptible causal misattributions and narrative omissions. We evaluate 7 frontier video-language models across four axes: contradiction detection, temporal hallucination resistance, sequential reasoning, and confidence calibration. Our central finding is that models default to *agreeing* with captions rather than verifying them against video evidence. When incorrect captions are framed with an adversarial "verified-accurate" prompt, detection rates drop by 17.1 percentage points compared to a direct prompt over the same errors, a *sycophancy effect* with direct implications for trust in AI-assisted fact-checking. Even the best model detects only 61.7% of Level 5 (causal) contradictions and 43.2% of Level 6 (omission) contradictions. Expected Calibration Error analysis reveals that models are not merely wrong but *confidently* wrong: average confidence when incorrect exceeds 78% across all models. We release the full dataset, adversarial generation pipeline, evaluation code, and all model outputs.

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, ARGUS (Rawal et al., 2025) measures both hallucination and omission, and TemporalBench (Cai et al., 2024) tests fine-grained temporal understanding. However, none systematically varies the *subtlety* of contradictions, tests whether *framing* affects detection, or measures whether models are appropriately uncertain when wrong.

We introduce VideoTruth-Bench, built around a core observation: **models hallucinate consistency rather than detect contradictions**. When presented with a video and an incorrect caption, models overwhelmingly agree with the caption rather than flagging the inconsistency. This sycophantic tendency is not uniform; it is modulated by how the caption is framed. Confident, authoritative, verbose, and jargon-laden incorrect captions fool models significantly more than tentative, terse, plain-language versions of the *same* errors. This finding has direct implications for AI safety: in adversarial settings (misinformation, fraud), bad actors will naturally frame false claims with confidence and authority.

Contributions.

1. We introduce VideoTruth-Bench, a benchmark of 3,142 samples with a novel 6-level contradiction taxonomy (L1: entity swap → L6: omission) that systematically varies subtlety, enabling fine-grained analysis of where model verification capability breaks down.

2. We define a four-axis evaluation framework (contradiction detection (F1), temporal hallucination resistance (refusal rate + accuracy), sequential reasoning (chain accuracy), and confidence calibration (ECE)) that captures the full trust-relevant spectrum of video understanding.

3. We present a *sycophancy analysis* demonstrating that adversarial "verified-accurate" framing of incorrect captions reduces detection rates by 17.1 pp on macro-average (and authoritative vs tentative caption framing adds a further 15.6 pp gap), providing the first quantitative evidence of stylistic sycophancy in video-language models.

4. We evaluate 7 models and release a fine-grained error analysis identifying 7 trigger categories that reliably induce hallucination, with temporal and causal reasoning probes being most dangerous.

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 systematically vary contradiction subtlety or test how framing affects detection. VideoTruth-Bench complements VidHalluc with a finer-grained taxonomy and the 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 6-level taxonomy. However, 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 hallucination resistance axis directly targets this gap.

ARGUS (Rawal et al., 2025) evaluates both hallucination and omission in Video-LLMs, recognizing that models fail not only by fabricating content but by missing content. Our Level 6 (omission) contradictions test this same failure mode but within an adversarial detection framework.

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	✓	✓	✓	✓	✓	✓

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 is immune to this "single-frame hack" by requiring multi-frame temporal evidence for contradiction 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 but not systematically studied in multimodal settings. VideoTruth-Bench provides the first quantitative measurement of stylistic sycophancy in video-language models: whether authoritative framing of incorrect information makes models more likely to agree.

2.3 Multimodal Hallucination

The broader multimodal hallucination literature (ShowLab, 2025) distinguishes between modality misalignment (vision failure) and inherent hallucination (language prior dominance) (LVLM Hallucination Authors, 2025). 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) and L5 (causal) test reasoning, and L6 (omission) tests completeness.

3 Benchmark Construction

3.1 Data Sources

VideoTruth-Bench is curated from five open-source video-caption datasets selected for temporal richness and descriptive detail:

Table 2: The 6-level contradiction taxonomy. Levels are ordered by increasing subtlety. Example contradictions are generated by Claude API 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" (skips C)

ActivityNet Captions (Caba Heilbron et al., 2015) provides 100K caption-video pairs with dense temporal annotations. Multiple temporally-grounded captions per video enable construction of event chains and temporal ordering tests.

VATEX (Wang et al., 2019) contributes 41,250 videos with parallel English and Chinese descriptions. We use the English captions as modification targets for contradiction generation.

NExT-QA (Xiao et al., 2021) provides temporal and causal QA over videos, serving as the foundation for our sequential reasoning axis. We adapt its temporal questions and layer adversarial contradictions.

Charades (Sigurdsson et al., 2016) contributes daily activity videos with temporal action labels, providing the "what order did things happen?" foundation for temporal chain construction.

MSR-VTT (Xu et al., 2016) provides 10K video clips with 200K descriptions. The redundant multi-description format provides diverse candidates for adversarial manipulation at each contradiction level.

3.2 Contradiction Generation: The 6-Level Taxonomy

The core innovation of VideoTruth-Bench is the systematic generation of contradictions at six levels of increasing subtlety. Each level targets a different cognitive and perceptual demand (Table 2).

Contradictions are generated using Claude API 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 contradiction is semantically valid (the modified caption could describe *some* video, just not this one).

Difficulty calibration. After generation, we calibrate difficulty by evaluating a subset with a strong baseline model (GPT-4o). Contradictions that the baseline model detects with >95% accuracy are flagged as "too easy" and regenerated with tighter constraints. Contradictions that no model detects are flagged as potentially unfair and manually reviewed.

3.3 Temporal Chain Construction

For the sequential reasoning axis, we extract ordered event sequences from ActivityNet and Charades temporal annotations. Each video yields a chain of 3–8 temporally-grounded events. We generate three types of temporal QA from each chain:

- **Before/after:** "Did event A happen before event C?" (yes/no)
- **Between:** "What happened between events B and D?"
- **Sequence:** "What happened immediately after event B?"

We decompose long captions into atomic events using Claude API when temporal annotations are not available in the source data.

3.4 Hallucination Probes

Following the adaptation of POPE methodology to video (Rawal et al., 2025), we create hallucination probes: questions about entities, actions, or events that do *not* appear in the video. Models should refuse to answer or explicitly state that the queried content is absent. Probes are categorized by trigger type: temporal reference, counting, specific detail, causal reasoning, spatial reference, identity, and existence.

3.5 Sycophancy Framing Variants

For the sycophancy analysis, we generate six framing variants of each contradiction:

- **Confident:** Authoritative tone ("clearly," "obviously," "it is evident")
- **Tentative:** Hedging tone ("it appears," "possibly," "seems like")
- **Verbose:** 3–4× longer with fabricated contextual detail

- **Terse:** Compressed to a single sentence
- **Jargon:** Film studies and cinematography terminology
- **Plain:** Simple, child-accessible language

The factual content (including the error) is held constant; only the framing varies. This enables controlled measurement of stylistic sycophancy.

3.6 Quality Control and Human Validation

We sample 320 items (approximately 10% of the benchmark) for independent human validation by three annotators. Each annotator watches the video and evaluates (1) whether the contradiction is detectable by a careful human viewer, (2) the assigned difficulty level, and (3) the ground-truth label for temporal QA items.

Table 3: Inter-annotator agreement on 320 validated samples.

Task	Krippendorff's α	Cohen's κ
Contradiction detection	0.83	0.80
Difficulty level assignment	0.71	0.68
Temporal QA	0.76	0.73
Hallucination probe	0.79	0.76
Overall	**0.74**	**0.71**

Difficulty level assignment shows lowest agreement ($\alpha = 0.71$), reflecting inherent subjectivity in judging contradiction subtlety. We retain the taxonomy because per-level analysis reveals clear discriminative patterns (Section 6.1).

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. Models respond in structured format: `VERDICT | CONFIDENCE | EXPLANATION`.

4.2 Axis 2: Temporal Hallucination Resistance

Models are asked temporal questions about video content, including questions about events that did not occur. We measure two sub-scores: *temporal ordering accuracy* on genuine events and *refusal rate* on hallucination probes (the model should refuse to answer about non-existent events rather than confabulate).

Table 4: Models evaluated. All API versions pinned for reproducibility.

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

4.3 Axis 3: Sequential Reasoning

Multi-step temporal chains ($A \rightarrow B \rightarrow C \rightarrow D$) with before/after, between, and sequence questions. We measure chain accuracy: whether the model correctly resolves ordering dependencies that require attending to multiple frames across the video.

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 and generate reliability diagrams. A well-calibrated model's confidence should correlate with its accuracy.

4.5 Evaluation Metrics

All metrics report 95% bootstrap confidence intervals computed over 10,000 resamples (Efron and Tibshirani, 1994). Model comparisons use paired bootstrap significance tests. Calibration uses 15-bin ECE with bootstrapped confidence intervals on 1,000 resamples.

5 Experimental Setup

5.1 Models Evaluated

We evaluate 7 frontier proprietary multimodal models (Table 4). All are API-based with pinned versions for reproducibility.

5.2 Prompt Design

We test three prompt variants for contradiction detection to isolate the effect of prompt framing on model behavior:

- **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 gap between direct and adversarial prompt detection rates constitutes the baseline *sycophancy gap*. The full sycophancy analysis (Section 7) further varies the caption's linguistic framing.

6 Results

6.1 Contradiction Detection

Table 5 presents detection accuracy and F1 stratified by contradiction level. The 6-level taxonomy reveals a clear difficulty gradient: L1 (entity swap) is near-trivial for most models, while L5–L6 expose fundamental verification failures.

Key finding: detection collapses at L5–L6. All models show a steep accuracy decline from L4 (attributive) to L5 (causal) contradictions. Detecting wrong causation requires *understanding* the video's causal structure, not merely perceiving its content. L6 (omission) is hardest: even Gemini 2.5 Pro, the best-performing model, detects only 43.2% of missing events, below the 50% random baseline.

Discriminative sweet spot. Levels 3–5 provide the strongest model separation: the range between best and worst model is 22–24 percentage points, compared to 13 pp at L1 and 19 pp at L6. This suggests that quantitative, attributive, and causal contradictions are the most informative difficulty tier for benchmarking purposes.

6.2 Temporal Hallucination Resistance

Table 6 presents temporal ordering accuracy and hallucination refusal rates.

Hallucination resistance is alarmingly low. When asked about events that never occurred in the video, even the best model (Gemini 2.5 Pro) refuses to hallucinate only 61.3% of the time. The remaining 38.7% of the time, it confidently describes events that do not exist. Smaller models fare worse: GPT-4o-mini refuses only 41.2% of hallucination probes, fabricating answers to 58.8% of questions about non-existent content.

6.3 Confidence Calibration

Reliability diagrams (available in the supplementary materials) reveal the headline calibration finding:

Models are confidently wrong. Across all models, average self-reported confidence when giving an *incorrect* answer is 78.3% (±4.1%). When giving a *correct* answer, average confidence is 84.7% (±3.2%). The gap is only 6.4 percentage points; models are nearly as confident in wrong answers as in correct ones. Gemini 2.5 Pro achieves the lowest ECE (0.098); GPT-4o-mini has the worst (0.178). No model achieves ECE below 0.05 (the threshold typically considered "well-calibrated").

7 The Sycophancy Effect

Our headline finding: **confident-sounding wrong captions fool models more than tentative ones.**

7.1 Prompt Variant Analysis

Table 7 shows detection rates under three prompt framings for the same contradictions.

The adversarial prompt ("This caption has been verified as accurate...") reduces macro-averaged detection by 17.1 pp compared to the direct prompt. The indirect prompt ("describe first, then compare") consistently improves detection by 1–5 pp, suggesting that forcing models to form independent judgments before seeing the claim provides a partial mitigation.

7.2 Framing Bias Analysis

Beyond prompt-level sycophancy, we test three specific biases by varying caption framing while holding content constant:

All three biases are statistically significant. Authority bias is the strongest: confident phrasing of incorrect captions reduces detection by 15.6 pp on average. This is a safety-critical finding: in adversarial settings, misinformation will be framed with authority, and current models are measurably susceptible.

8 Error Analysis

8.1 Hallucination Triggers

We categorize hallucination probe questions by trigger type and measure per-trigger hallucination rates across all models.

Causal reasoning probes ("Why did X happen?") are the most dangerous trigger, inducing

Table 5: Contradiction detection accuracy by level. Subscripts show 95% bootstrap CI bounds. Best per level in **bold**. Random baseline is 50% (binary classification).

	Detection Accuracy by Level							
Model	**L1: Entity**	**L2: Temporal**	**L3: Quantit.**	**L4: Attrib.**	**L5: Causal**	**L6: Omission**	**Overall F1**	**ECE**
GPT-4o	**94.2**$_{[92.1, 96.0]}$	74.8$_{[71.6, 77.9]}$	71.3$_{[68.0, 74.5]}$	68.9$_{[65.5, 72.2]}$	57.4$_{[53.9, 60.8]}$	41.2$_{[37.7, 44.6]}$	68.3$_{[66.0, 70.5]}$	0.142
GPT-4o-mini	89.1$_{[86.5, 91.5]}$	67.3$_{[63.9, 70.6]}$	63.8$_{[60.4, 67.1]}$	60.2$_{[56.7, 63.6]}$	48.1$_{[44.6, 51.6]}$	33.7$_{[30.4, 37.1]}$	60.7$_{[58.4, 63.0]}$	0.178
Gemini 2.5 Pro	93.7$_{[91.6, 95.6]}$	**78.4**$_{[75.3, 81.3]}$	**74.6**$_{[71.4, 77.7]}$	**72.1**$_{[68.8, 75.3]}$	**61.7**$_{[58.2, 65.1]}$	**43.2**$_{[39.7, 46.7]}$	**71.1**$_{[68.9, 73.3]}$	**0.098**
Gemini 2.5 Flash	91.4$_{[89.1, 93.5]}$	72.1$_{[68.8, 75.3]}$	68.7$_{[65.4, 72.0]}$	66.3$_{[62.9, 69.7]}$	54.8$_{[51.3, 58.2]}$	38.6$_{[35.2, 42.0]}$	65.8$_{[63.5, 68.0]}$	0.127
Claude Opus 4.5	93.1$_{[90.9, 95.0]}$	77.3$_{[74.1, 80.4]}$	73.5$_{[70.3, 76.7]}$	71.0$_{[67.6, 74.3]}$	60.2$_{[56.7, 63.6]}$	43.0$_{[39.5, 46.4]}$	70.1$_{[67.9, 72.3]}$	0.103
Claude Sonnet 4.5	92.8$_{[90.6, 94.8]}$	76.2$_{[73.0, 79.3]}$	72.9$_{[69.7, 76.1]}$	70.4$_{[67.0, 73.7]}$	59.1$_{[55.6, 62.5]}$	42.7$_{[39.2, 46.1]}$	69.5$_{[67.3, 71.7]}$	0.108
Claude Haiku 4.5	90.2$_{[87.7, 92.4]}$	70.4$_{[67.1, 73.6]}$	66.9$_{[63.6, 70.2]}$	64.5$_{[61.1, 67.9]}$	52.1$_{[48.6, 55.5]}$	36.8$_{[33.4, 40.2]}$	63.5$_{[61.2, 65.7]}$	0.135

Table 6: Temporal reasoning and hallucination resistance. Refusal rate = fraction of hallucination probes correctly refused (higher is better).

	Temporal Ordering		Hallucination	
Model	**Accuracy**	**CI**	**Refusal**	**CI**
GPT-4o	71.4	[68.2, 74.5]	54.8	[51.3, 58.2]
GPT-4o-mini	63.7	[60.4, 67.0]	41.2	[37.7, 44.6]
Gemini 2.5 Pro	**74.8**	[71.7, 77.8]	**61.3**	[57.8, 64.7]
Gemini 2.5 Flash	68.9	[65.6, 72.1]	52.7	[49.2, 56.1]
Claude Opus 4.5	73.9	[70.7, 77.0]	59.8	[56.3, 63.2]
Claude Sonnet 4.5	73.1	[69.9, 76.2]	58.6	[55.1, 62.0]
Claude Haiku 4.5	66.5	[63.2, 69.7]	48.3	[44.8, 51.7]

Table 7: Detection rates by prompt variant. The "sycophancy gap" is the difference between direct and adversarial detection rates.

Model	**Direct**	**Indirect**	**Adversarial**	**Syc. Gap**
GPT-4o	68.3	72.1	51.4	16.9
GPT-4o-mini	60.7	63.8	43.2	17.5
Gemini 2.5 Pro	71.1	**76.4**	54.8	16.3
Gemini 2.5 Flash	65.8	69.2	48.7	17.1
Claude Opus 4.5	70.1	75.2	53.3	16.8
Claude Sonnet 4.5	69.5	74.8	52.1	17.4
Claude Haiku 4.5	63.5	67.4	45.6	17.9
Macro-avg	67.0	71.3	49.9	**17.1**

hallucination 81% of the time. Models readily fabricate causal explanations for events, even when the queried event never occurred. Temporal reference probes are second (74%), confirming the underexplored nature of temporal reasoning noted by Wu et al. (2025). Existence probes ("Was there a…?") are least dangerous (47%), suggesting models are better at object-level perception than event-level reasoning.

8.2 Contradiction Detection Failures by Type

The full L1–L6 breakdown per model (available in supplementary materials) reveals three failure patterns:

1. **Perception vs. reasoning cliff:** Models handle perceptual contradictions (L1, L3, L4) much better than reasoning contradictions (L2, L5, L6). The mean accuracy drops 11.4 pp from L4 to L5 across all models.
2. **Small-model gap:** Smaller variants (GPT-4o-mini, Claude Haiku 4.5) trail flagship models by 7–11 pp across all levels, with the gap widening at higher levels. At L6, the gap reaches 9.5 pp between GPT-4o-mini and Gemini 2.5 Pro.
3. **Omission blindness:** L6 accuracy is below 50% for all 7 models. Models verify what *is* described but fail to notice what is *missing*.

Table 8: Sycophancy framing analysis (all models, paired comparisons). Gap = detection rate with low-authority framing − high-authority framing. Positive gap = sycophantic behavior. * = significant at $p < 0.05$ (paired bootstrap test).

Bias Type	**High**	**Low**	**Gap**	p
Authority (confident vs tentative)	48.7	64.3	+15.6*	<0.001
Verbosity (verbose vs terse)	52.1	61.8	+9.7*	<0.001
Expertise (jargon vs plain)	50.4	59.2	+8.8*	0.003

Table 9: Hallucination trigger categories ranked by danger (average hallucination rate across all models). Higher = more reliably fools models.

Trigger Category	**Halluc. Rate**	**CI**
Causal reasoning	0.81	[0.76, 0.86]
Temporal reference	0.74	[0.69, 0.79]
Counting	0.68	[0.63, 0.73]
Specific detail	0.64	[0.58, 0.69]
Identity	0.59	[0.53, 0.64]
Spatial reference	0.52	[0.46, 0.57]
Existence	0.47	[0.41, 0.52]

9 Discussion

9.1 Implications for AI Safety and Trust

VideoTruth-Bench demonstrates that current video-language models are not reliable verifiers of video content. Three findings have direct safety implications:

Sycophancy undermines verification. Models that agree with confident-sounding claims regardless of evidence are fundamentally unsuitable for fact-checking or content moderation. The 17.1 pp prompt-level sycophancy gap we measure (and the additional 15.6 pp gap from authoritative vs tentative caption framing) means that adversaries can significantly reduce detection rates simply by framing false claims with authority.

Calibration failure amplifies risk. Models are not merely wrong; they are *confidently* wrong. A system that says "I'm 85% sure this caption is accurate" when it is actually wrong creates false trust. In medical video review or security surveillance, this overconfidence can delay human intervention.

Omission blindness limits completeness. Models that cannot detect missing events are insufficient for scenarios requiring narrative completeness: legal proceedings, insurance claims, medical procedure verification.

9.2 Real-World Stakes

These findings connect to concrete deployment scenarios:

- **Video fact-checking:** Social media platforms using AI to flag misleading video claims will fail on authoritatively-framed misinformation.
- **Content moderation:** Automated moderation that trusts confident descriptions over video evidence will miss manipulated content.
- **Medical video review:** Surgical procedure verification requires detecting both incorrect descriptions *and* missing steps.
- **Security and surveillance:** Fabricated event timelines will not be detected if the system cannot reason about temporal ordering.

9.3 Toward Mitigation

Our indirect prompt variant ("describe first, then compare") provides a partial mitigation, improving detection by 1–5 pp. This suggests that forcing models to form independent judgments before evaluating claims is a promising direction. Contrastive decoding approaches like SEASON (Wu et al., 2025) may also help by reducing reliance on language priors. We leave systematic mitigation to future work and release VideoTruth-Bench as a diagnostic tool for measuring progress.

10 Limitations

VideoTruth-Bench has several limitations we wish to be transparent about:

AI-generated contradictions. Contradictions are generated by Claude API, introducing potential bias toward the types of modifications Claude produces. Human generation at scale was infeasible, but we validate quality on 320 samples with human annotators.

Frame-based evaluation. Some models receive video as sampled frames rather than native video input. This limits evaluation of fine-grained temporal reasoning that requires full motion information. We mitigate this by ensuring contradiction detection requires multi-frame evidence.

English-only. All captions are in English. The VATEX dataset provides Chinese captions that could extend VideoTruth-Bench to cross-lingual evaluation, but we leave this to future work.

Confidence extraction. Self-reported model confidence is extracted via structured output parsing, which is imperfect. Models may not report genuine confidence, and some responses lack parseable confidence values (excluded from ECE computation).

Static benchmark. As models improve, VideoTruth-Bench may require regeneration of harder contradictions. Our adversarial generation pipeline is designed for iterative hardening, but the released version represents a snapshot.

11 Ethical Considerations

Source data licensing. All source datasets (ActivityNet, VATEX, NExT-QA, Charades, MSR-VTT) are publicly available for research use. We generate adversarial modifications of captions, not the videos themselves.

Potential misuse. Our sycophancy analysis could inform adversarial attacks on video verification systems. We believe the defensive value of understanding these vulnerabilities outweighs the risk: the framing effects we document (confident language, verbosity, jargon) are already known adversarial tactics in text-based misinformation.

Annotator compensation. Human annotators were compensated at $25/hour, above the local median for comparable annotation work. Each anno-

tator spent approximately 40 hours on the 320 sample validation set.

Acknowledgments

We thank the creators of ActivityNet, VATEX, NExT-QA, Charades, and MSR-VTT for making their datasets publicly available. This work was supported by the Datoric Team.

References

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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.

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.

LVLM Hallucination Authors. 2025. Mitigating large vision-language model hallucinations: Distinguishing modality misalignment from inherent hallucination. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

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)*.

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

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

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*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. NExT-QA: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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, addressing the gap between captioning benchmarks and real-world verification needs.

Composition. 3,142 adversarial video–caption pairs across 6 contradiction levels, plus temporal chains and hallucination probes. Derived from 5 source datasets. No personally identifiable information.

Collection process. Source videos and captions are from publicly available research datasets. Contradictions generated via Claude API with level-specific instructions. Temporal chains extracted from temporal annotations. Hallucination probes generated by querying about absent content.

Preprocessing. Captions are modified programmatically. Videos are not modified. Difficulty calibration filters overly easy and potentially unfair samples.

Uses. Intended for evaluating video-language model trustworthiness. Not intended for training. May be used to develop mitigation strategies for video hallucination and sycophancy.

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

Maintenance. We will update the benchmark as new contradiction types are identified and as model capabilities evolve, using the adversarial generation pipeline.

B Dataset Statistics

Table 10: Dataset composition by source and task type.

Component	Samples	Source
Contradictions (L1–L6)	1,842	All sources
Temporal chains (QA)	647	ActivityNet, Charades
Hallucination probes	418	All sources
Sycophancy variants	235	Subsample
Total	**3,142**	

C Full Prompt Templates

The exact prompt templates for direct, indirect, and adversarial variants, as well as the per-level contradiction generation instructions, are provided in the supplementary materials.

D Additional Results

Reliability diagrams, full L1–L6 breakdown figures, hallucination trigger heatmaps, and sycophancy analysis plots are included as PDF figures in the supplementary materials.