When Seeing Is Not Believing—A Benchmark for Search-Grounded Video Misinformation Detection

Tao Yu[1,2,3*§♠], Yujia Yang[2*], Shenghua Chai[1*†], Jinshuai Zhang[1*†], Haopeng Jin[1†]
Hao Wang[1†], Minghui Zhang[1†], Zhongtian Luo[1†], Yuchen Long[1†], Xinlong Chen[1,2]
Jiabing Yang[1,2], Zhaolu Kang[5], Yuxuan Zhou[4], Siyi Cao[1†], Yuqing Wu[1†]
Peifan Wang[1†], Tianyang Dang[1†], Minghui Guo[1†], Hanyu Wang[1†], Hai Huang[1†]
Zhengyu Man[1†], Xingyu Yang[1†], Xinya Sun[1†], Xinming Wang[1,2]
Hongzhu Yi[2‡], Zheqi He[3‡], Xi Yang[3], Yan Huang[1,2‡], Liang Wang[1,2]
[1]CASIA [2]UCAS [3]BAAI [4]Tsinghua University [5]Peking University
yutao2025@ia.ac.cn

Abstract

Video misinformation increasingly operates at the semantic and evidential level: authentic footage may be selectively edited, temporally reordered, spliced across sources, or augmented with AI-generated content to construct false narratives. Such evidence-dependent manipulations cannot be reliably verified from the input video alone, because the missing, reordered, replaced, or recontextualized evidence lies outside the video itself. We introduce **EVID-Bench**, a benchmark for search-grounded video misinformation detection, where a system must search the open web for related videos and identify what information is false through cross-video comparison. EVID-Bench comprises 222 videos spanning 9 manipulation types across 3 categories: AI generation, single-source editing, and multi-source editing. All samples are verified to be undetectable by frontier models through visual inspection alone. We evaluate nine frontier multimodal models using a retrieval-augmented verification baseline. The best system achieves only 61.43% point-level accuracy and 43.24% video-level accuracy, while AI-generated manipulations remain especially challenging. Error analysis reveals recurring challenges: models fixate on irrelevant anchors, misattribute synthetic content to editorial splicing, and terminate search prematurely before fully explaining the manipulation.

1 Introduction

Video has become a dominant medium of information on social platforms, news media, and instant messaging applications, making it a highly influential vehicle for misinformation. Compared with text or images, viewers are often more inclined to regard video as a direct record of reality (Gunasekara et al., 2025; Sundar et al., 2021). A carefully manipulated video may shape public opinion, intensify social tensions, or damage individual reputations before factual clarifications can spread.

One particularly dangerous category of video misinformation derives its core harm from the distortion of event semantics and evidential relationships. Authentic footage can be selectively trimmed, temporally reordered, or stitched across sources to construct narratives that never occurred (Diwan et al., 2024; Wang et al., 2025b); meanwhile, realistic synthetic faces, objects, or event segments can be embedded into real video contexts. For example, removing the trigger of a protest conflict may frame the remaining footage as "unprovoked violence," reordered speech clips may suggest a "reversal of position," and face replacement in a real interview may make viewers believe that another public figure appeared and made the remarks. In such cases, misinformation often parasitizes authentic material: every frame may be real while the relations between segments are false, or only a local identity, object, or event segment may be synthetic while the overall narrative is altered. We refer to this setting as evidence-dependent video misinformation, where authenticity cannot be fully determined from the input video alone but must be judged against external event evidence.

Existing video authenticity detection paradigms are not designed for this setting. Traditional forensic methods (Singh and Singh, 2024; Fatima et al., 2026) rely on low-level traces such as compression inconsistencies, optical flow irregularities, or

* Equal contribution.
† Work done during an internship at CASIA.
§ Work done during an internship at BAAI.
‡ Corresponding author.
♠ Project leader.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 16076–16118
October 24-29, 2026 ©2026 Association for Computational Linguistics

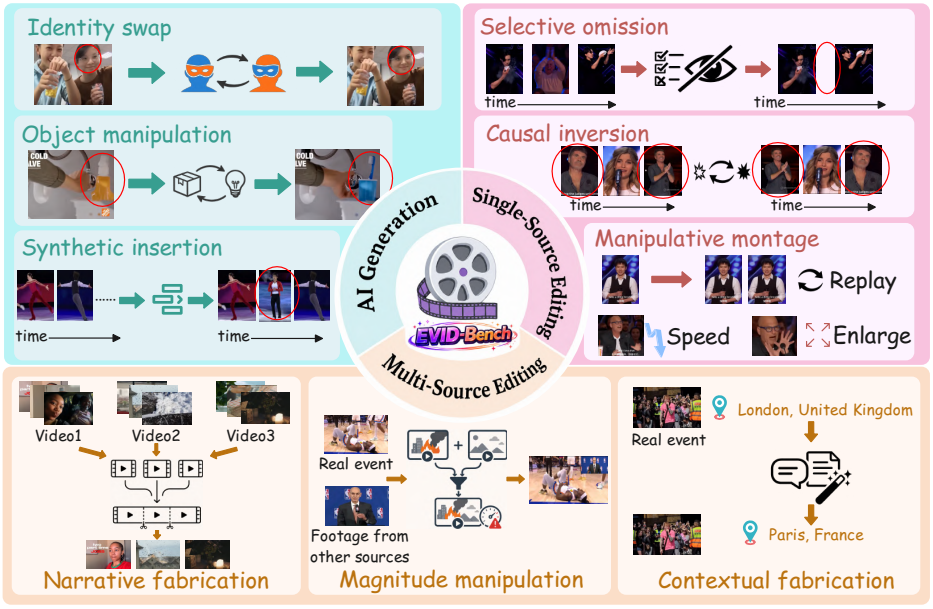


Figure 1: Taxonomy of search-grounded video misinformation in EVID-Bench covering AI Generation, Single-Source Editing, and Multi-Source Editing. These mechanisms manipulate identities, objects, temporal order, narrative structure, event magnitude, or contextual information, making the video misleading not necessarily through visible artifacts, but through its discrepancy from external event evidence.

pixel statistics, which may not exist when authentic footage is edited or recontextualized. Deepfake detectors (Zhou et al., 2025c; Qian et al., 2020; Hernandez-Ortega et al., 2020) can identify local generative artifacts, but they cannot recover the original content before replacement or explain how a local modification changes identity, causality, or responsibility. Multimodal misinformation methods (Abdali et al., 2024; Tian et al., 2026; Shopnil et al., 2026; Aneja et al., 2021) often rely on explicit textual claims, whereas video misinformation can convey false narratives purely through visual editing. Thus, the key challenge is not merely detecting whether a video contains artifacts, but identifying what information has been falsified relative to evidence about the real event. This requires retrieving external video evidence (Yu et al., 2026a,d) and reasoning over cross-video discrepancies.

To address this challenge, we introduce **EVID-Bench**, a benchmark for *search-grounded video misinformation detection*. Given a potentially manipulated video, a system must search the open web (Yu et al., 2025b; Wang et al., 2025a) for related or corroborating videos and identify what information is false through cross-video evidence comparison. This formulation mirrors how human fact-checkers debunk video misinformation: by finding related footage and showing what was omitted, reordered, inserted, or recontextualized.

EVID-Bench contains 222 manipulated videos spanning 9 forgery types in 3 categories: AI generation, single-source editing, and multi-source editing, across 6 real-world topic domains. All samples undergo human review and model-based filtering to ensure that frontier multimodal models (Yu et al., 2025a; Zhou et al., 2025b) cannot identify the manipulation through visual inspection alone. We further propose a retrieval-augmented verification baseline (Yu et al., 2026f) with chain-of-thought analysis and an iterative search-verify-reflect loop (Yu et al., 2026e; Zhou et al., 2025a; Yu et al., 2026c). Experiments on nine frontier models (Yu et al., 2026b; Wang et al., 2026) show that EVID-Bench remains challenging: the best model achieves only 61.43% point-level accuracy and 43.24% video-level accuracy, with AI-generated manipulations remaining especially difficult.

Our contributions are as follows:

- We formalize search-grounded video misinformation detection, where systems must identify false information through external video retrieval and cross-video reasoning.
- We construct EVID-Bench, a quality-assured benchmark of 222 videos covering 9 forgery types and 6 topic domains.

- We evaluate nine frontier models with a retrieval-augmented verification baseline, showing that even the best system achieves only 43.24% video-level accuracy and revealing key challenges in retrieval, contrastive reasoning, and sufficiency judgment.

2 Related Works

2.1 Video Forgery Detection

Traditional methods detect low-level artifacts: compression inconsistencies (Singh and Singh, 2024), irregular optical flow (Fatima et al., 2026), or face-swap traces through physiological signals and frequency analysis (Zhou et al., 2025c; Qian et al., 2020; Hernandez-Ortega et al., 2020). These assume manipulation leaves detectable pixel-level traces—an assumption that breaks down for semantic-level forgeries where every frame is authentic and the manipulation lies in editorial choices. EVID-Bench targets this class of forgeries, where detection requires external evidence.

2.2 Multimodal Misinformation Detection

Another line of work verifies textual claims against visual content (Abdali et al., 2024; Tian et al., 2026), retrieves textual evidence for fact-checking (Shopnil et al., 2026), or detects out-of-context media (Aneja et al., 2021). These face two limitations: most assume an explicit textual claim, whereas manipulated videos convey false narratives purely through visual editing; and retrieval operates in the text domain rather than comparing video evidence directly. EVID-Bench requires searching for corroborating videos and performing cross-video contrastive reasoning.

3 EVID-Bench

3.1 Task Definition and Design Principles

We introduce **search-grounded video misinformation detection** as the core task of EVID-Bench. Given a potentially misleading video, a system must search the open web for original or related videos and identify what information is false through cross-video evidence comparison. The task is designed around three principles:

Evidence-dependent misinformation. Misleading videos in our benchmark distort event meaning by altering identities, reordering events, changing context, or embedding AI-generated content into authentic footage. Such misinformation cannot be reliably verified from the input video alone, because the missing, replaced, reordered, or recontextualized evidence lies outside the video itself.

Reasoning-driven search. Since the input video may encode a false narrative, naively summarizing it into search queries risks searching for the misinformation rather than the truth. The system must reason about which elements are likely authentic and searchable, use them as retrieval anchors, and iteratively refine its queries as partial evidence is gathered.

Cross-video contrastive verification. The final judgment must be grounded in differences between the input video and retrieved evidence. This requires the system to align videos at the level of entities, actions, temporal order, and narrative context, and identify where they diverge. The output is not a binary label, but a structured explanation of what is false and which evidence supports that conclusion.

Together, these principles make EVID-Bench a joint evaluation of video understanding, web search, and cross-video reasoning—capabilities not jointly assessed by existing forensics or misinformation benchmarks.

3.2 Task Taxonomy

Drawing on prior work (Rossler et al., 2019; Ho et al., 2022; Fischinger and Boyer, 2025; Wardle and Derakhshan, 2017; Gangal et al., 2022; Yin et al., 2024; Pei et al., 2024; Bu et al., 2023; Luo et al., 2021), we organize manipulations into three categories based on the forgery mechanism, each with three subcategories (nine task types in total).

3.2.1 AI Generation

Synthetic content is introduced into authentic footage.

Identity swap. A person's face is replaced with a different identity using deepfake techniques while the body, background, and context remain authentic.

Synthetic insertion. A contiguous AI-generated video segment is embedded into real footage, introducing an entirely fabricated temporal sequence of events.

Object manipulation. Objects within the scene are added, removed, or replaced through generative editing, altering what is physically present without changing the overall spatial structure.

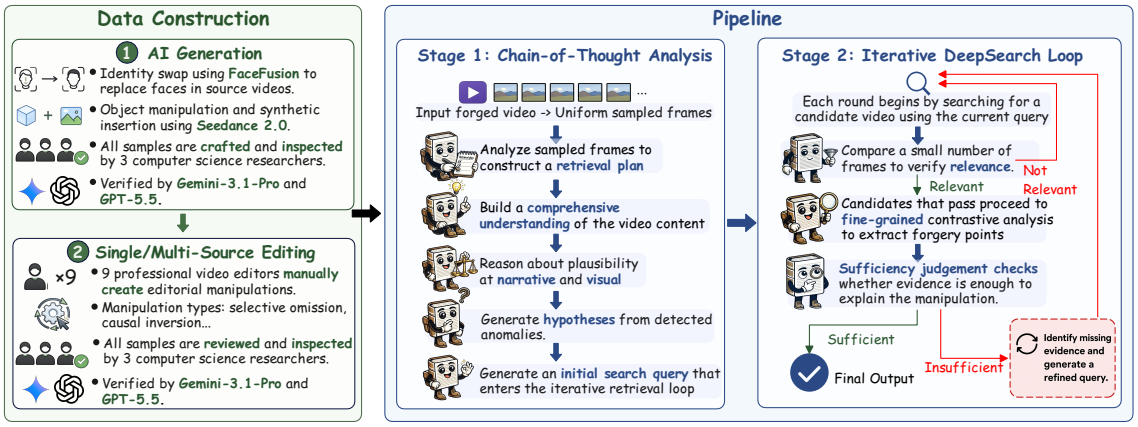


Figure 2: Overview of EVID-Bench construction and the search-grounded video misinformation detection pipeline. The benchmark constructs hard-to-detect video misinformation through AI generation and professional single-source and multi-source editing, followed by human inspection and model-based verification. The detection pipeline analyzes sampled frames, forms retrieval hypotheses, iteratively searches for external video evidence, compares candidate sources with the input video, and refines the search until sufficient evidence is obtained for the final misinformation explanation.

3.2.2 Single-Source Editing

A single authentic video is editorially altered to change its narrative meaning.

Selective omission. Critical portions are removed so that the remaining footage conveys a different or opposite meaning from the original.

Causal inversion. The temporal order of events is reordered to flip causality—e.g., transforming a victim's retaliation into an unprovoked attack by swapping the sequence of actions.

Manipulative montage. Shots are rearranged and re-paced to create an emotional or narrative impression that the original footage does not support, without reversing any specific causal sequence.

3.2.3 Multi-Source Editing

Footage from multiple authentic videos is combined to construct false information.

Narrative fabrication. Segments from unrelated videos are spliced together to tell a story that none of the individual sources depict.

Magnitude manipulation. A real event is combined with footage from other sources to distort its severity—either exaggerating or downplaying its actual impact.

Contextual fabrication. Authentic footage of a real event is placed into a false context (different time, location, or attribution) through splicing, misleading viewers about when, where, or to whom the event occurred.

3.3 Topic Categories

To ensure our benchmark reflects the domains where video misinformation most commonly arises, we select six topic categories based on prior studies (Papadopoulou et al., 2019; Wang et al., 2024; Huang et al., 2026; Bu et al., 2023; Wang et al., 2025b; Bevendorff et al., 2024) on video-based false information:

Daily life. Vlogs, workplace recordings, campus life, travel clips, food, and pet videos.

Education & skills. Online tutorials, DIY crafts, fashion and beauty, fitness, and parenting content.

On-Scene footage. Dashcam and surveillance recordings, disaster scenes, and live incident documentation.

Sports. Professional and amateur athletic events, match highlights, and training footage.

Performance & talent. Music, dance, comedy, and other talent showcases.

Reviews. Product unboxings, brand comparisons, and store/restaurant visits.

3.4 Data Construction

3.4.1 AI Generation

For identity swap, we use FaceFusion[1] to replace faces in source videos. For object manipulation and synthetic insertion, we generate content using Seedance 2.0 (Seedance et al., 2026). All AI-generated samples are manually crafted and inspected by three computer science researchers to

[1] https://github.com/facefusion/facefusion

ensure visual coherence. Each sample is then verified by Gemini-3.1-Pro (Google DeepMind, 2025) and GPT-5.5 (OpenAI, 2026b): we confirm that neither model can identify the manipulation through direct visual reasoning over the forged video alone.

3.4.2 Single-Source and Multi-Source Editing

We recruit nine professional video editors to manually produce all editorial manipulations, including selective omission, causal inversion, manipulative montage, narrative fabrication, magnitude manipulation, and contextual fabrication. Each edited video is first reviewed by three computer science researchers to ensure the manipulation is coherent and undetectable. We then present the edited videos to Gemini-3.1-Pro and GPT-5.5, verifying that the models interpret the video content consistently with the intended manipulated narrative—confirming that the editorial forgery successfully conveys its false information without raising suspicion.

3.4.3 Quality Assurance

Our verification process combines human review and model-based filtering. Three computer science researchers first examine each sample for technical quality and narrative coherence; we then conduct visual verification using Gemini-3.1-Pro and GPT-5.5. For AI-generated forgeries, both models are required to fail to identify the manipulation; for editorial forgeries, both models are required to interpret the video according to the intended false narrative. Samples that fail either stage are revised or discarded, ensuring that the final 222 samples cannot be identified as manipulated by frontier models through visual analysis alone. The data generation process is shown in Figures 1 and 2. Details of this process and the prompts used are provided in Appendix A.

3.5 Data Statistics

Our benchmark comprises 222 manipulated videos spanning 3 major categories, 9 task types, and 6 topic categories. Figure 3 shows the distribution across task types and topics. A detailed description of the data is provided in Appendix B.

4 Experiments

4.1 Baseline

Stage 1: Chain-of-Thought Analysis. A VLM analyzes uniformly sampled frames via chain-of-thought reasoning to construct a retrieval plan. The

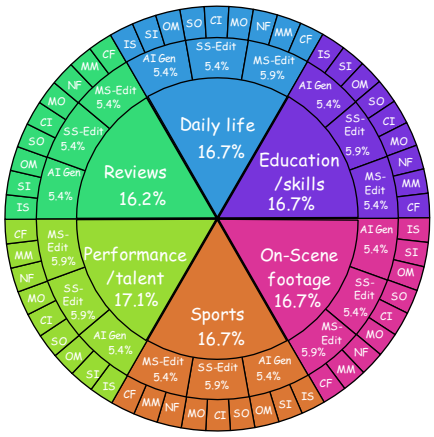


Figure 3: Distribution of EVID-Bench across task types and topics. Task types are abbreviated as follows: AI Gen (AI generation), SS-Edit (Single-Source Editing), MS-Edit (Multi-Source Editing), IS (Identity Swap), SI (Synthetic Insertion), OM (Object Manipulation), SO (Selective Omission), CI (Causal Inversion), MO (Manipulative Montage), NF (Narrative Fabrication), MM (Magnitude Manipulation), and CF (Contextual Fabrication).

fundamental challenge is that the video's apparent narrative cannot be taken at face value—it may be the product of manipulation. The model first builds a comprehensive understanding of the video content: identifying key entities (people, locations, events, temporal cues), segmenting distinct scenes, and detecting discontinuities in environment, production style, or visual quality that suggest assembly from multiple sources. It then reasons about plausibility at two levels—at the narrative level, whether the implied facts are internally consistent and whether the associations between entities across scenes are natural or artificially constructed; at the visual level, whether any elements exhibit signs of AI generation such as contextually incongruent objects or inconsistent facial identity between segments. These detected anomalies inform hypothesis generation: the model infers what the unmanipulated reality likely depicts and what claims most urgently require external verification. Suspected synthetic elements are explicitly excluded from retrieval anchors, ensuring that queries are grounded in authentic, verifiable cues rather than potentially fabricated content. The output is an initial search query that enters the iterative retrieval loop.

Stage 2: Iterative DeepSearch Loop. The system enters an iterative retrieval loop. Each round begins by searching for a candidate video using

Model	AI Generation			Single-Source Editing			Multi-Source Editing			Avg.
	Identity Swap	Synthetic Insertion	Object Manipulation	Selective Omission	Causal Inversion	Manipulative Montage	Narrative Fabrication	Magnitude Manipulation	Contextual Fabrication	
GPT Series										
GPT-5.5	11.11 (±2.1)	16.67 (±1.6)	13.89 (±0.8)	72.92 (±1.6)	55.32 (±1.2)	74.00 (±1.5)	82.24 (±1.0)	98.91 (±0.0)	88.46 (±0.6)	61.43 (±0.9)
GPT-5.4	0.00 (±0.0)	9.72 (±1.6)	11.11 (±0.8)	64.58 (±1.6)	45.74 (±0.0)	66.00 (±0.6)	73.83 (±1.7)	95.65 (±1.1)	78.85 (±1.5)	53.77 (±0.9)
GPT-5.4-Mini	1.39 (±0.0)	6.94 (±0.0)	5.56 (±1.4)	57.29 (±1.0)	41.49 (±1.1)	58.00 (±1.7)	64.49 (±1.7)	85.87 (±1.3)	67.31 (±2.2)	46.97 (±1.0)
Claude Series										
Claude-Opus-4.6	8.33 (±2.1)	13.89 (±0.8)	11.11 (±1.6)	69.79 (±1.8)	53.19 (±0.6)	70.00 (±0.6)	76.64 (±1.5)	96.74 (±0.6)	86.54 (±0.6)	58.34 (±1.0)
Claude-Sonnet-4.6	5.56 (±0.8)	12.50 (±1.6)	9.72 (±1.6)	66.67 (±1.2)	50.00 (±0.6)	67.00 (±2.3)	72.90 (±1.0)	94.57 (±1.3)	79.81 (±1.0)	55.13 (±1.0)
Gemini Series										
Gemini-3.1-Pro	30.60 (±1.8)	13.90 (±1.5)	9.70 (±1.1)	67.70 (±0.8)	52.10 (±0.8)	66.00 (±1.6)	71.00 (±1.3)	82.60 (±2.4)	77.90 (±1.1)	55.90 (±1.0)
Gemini-3-Flash	13.90 (±1.7)	11.10 (±1.9)	2.80 (±0.0)	65.60 (±0.6)	54.30 (±1.3)	70.00 (±1.1)	72.90 (±2.2)	80.40 (±1.3)	75.00 (±1.1)	53.60 (±1.0)
Qwen Series										
Qwen-3.5-Plus	22.20 (±1.3)	13.90 (±0.0)	9.70 (±0.9)	58.30 (±1.6)	44.70 (±2.0)	64.00 (±2.1)	73.80 (±1.5)	88.00 (±0.7)	74.00 (±0.6)	53.40 (±1.0)
Qwen3-VL-235B-A22B-Instruct	6.90 (±1.6)	12.50 (±1.0)	4.20 (±1.8)	61.50 (±1.9)	43.60 (±0.7)	60.00 (±0.7)	71.00 (±1.5)	73.90 (±0.7)	71.20 (±1.1)	48.80 (±1.0)

Table 1: Point-level accuracy of evaluated models across fine-grained manipulation task types. All metrics are reported in percentages (%).

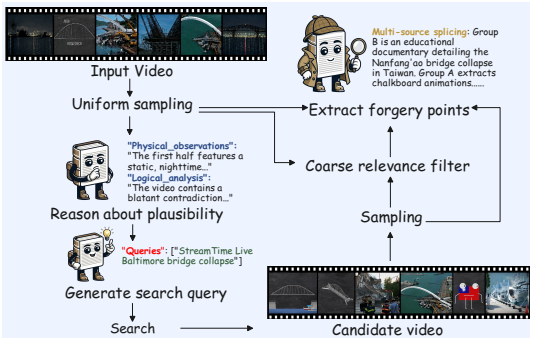


Figure 4: Case study of EVID-Bench.

the current query. To avoid expensive analysis on irrelevant results, a coarse relevance filter first compares a small number of frames from both videos, assessing whether they share the same event, scene family, or verification-relevant content. Candidates that pass proceed to fine-grained contrastive analysis, where a larger set of frames from both videos is compared to extract narrative-level forgery points—structured descriptions of how the input video distorts, reframes, or fabricates information relative to the retrieved evidence. Candidates that fail are discarded, and the model directly generates a new query for the next round. After each successful comparison, a sufficiency judgment evaluates whether the accumulated evidence adequately explains the manipulation. If not, the model identifies what evidence is still missing and generates a refined query to target that gap. This iterative search-verify-reflect cycle continues until the evidence is deemed sufficient or the round budget is exhausted. A case study is shown in Figure 4. Details of this part and the prompts used are provided in Appendix E.

4.2 Evaluation Metrics and Settings

Metrics. We evaluate performance at two levels. **Point-level accuracy** measures the fraction of ground-truth misinformation points correctly identified by the system, where each video contains 3–5 annotated points. **Video-level accuracy** counts a video as correct only when all ground-truth points are identified, measuring whether the system fully explains the manipulation. The same metric definitions are used for both model and human evaluation.

Experimental setup. We evaluate nine frontier multimodal models: GPT-5.5, GPT-5.4 (OpenAI, 2026a), GPT-5.4-mini (OpenAI, 2026a), Claude Opus 4.6 (Anthropic, 2026a), Claude Sonnet 4.6 (Anthropic, 2026b), Gemini-3.1-Pro, Gemini-3-Flash (Google DeepMind, 2025), Qwen3-VL-235B-A22B-Instruct (Bai et al., 2025), and Qwen-3.5-Plus (Qwen Team, 2026). Parameter details are provided in Appendix C.

Evaluation protocol. Because outputs are natural-language descriptions, we use three LLM judges—GPT-5.4, Gemini-3.1-Pro, and Claude Sonnet 4.6—to assess whether each predicted point matches a ground-truth point. Correctness is decided by majority vote. The similarity analysis between LLM-based evaluation and human evaluation is provided in Appendix G.

4.3 Main Results

EVID-Bench remains challenging for frontier models. The best model, GPT-5.5, achieves 61.43% point-level but only 43.24% video-level accuracy (Tables 1 and 2). Claude Opus 4.6 and Gemini-3.1-Pro follow at 58.34%/38.74% and

55.90%/36.90%. Even frontier models only partially identify misinformation in most videos.

Complete verification is substantially harder than partial detection. All models show a clear gap between point-level and video-level accuracy. Models can retrieve relevant evidence and identify some manipulation points but struggle to integrate multiple pieces of evidence and fully explain all false information.

AI-generated manipulations are largely unsolved. AI Generation is extremely difficult: GPT-5.5 achieves only 11–17% point-level and near-zero video-level accuracy (Tables 1 and 2). While the pipeline can retrieve related videos, identifying which specific identity, object, or segment was replaced through cross-video comparison remains unsolved.

Multi-Source Editing benefits from retrieval but is not solved. Models achieve higher scores on Multi-Source Editing tasks (GPT-5.5: 82–99% point-level). Cross-source splicing provides stronger retrieval anchors, but video-level accuracy remains much lower, indicating partial rather than complete explanation.

Topic difficulty reflects retrieval anchor availability. Daily Life and Reviews are the hardest domains, while On-Scene Footage and Education & Skills achieve higher accuracy (Tables 3 and 4), reflecting the availability of identifiable events or named individuals versus generic content.

4.4 Ablation Study

We ablate three key hyperparameters using GPT-5.5 and Qwen-3.5-Plus as representative models. Figure 5 shows the fine-grained breakdown for GPT-5.5. The corresponding results for Qwen-3.5-Plus are provided in Appendix I.

Number of search rounds. Performance peaks at 6 rounds for both models (GPT-5.5: 61.43%; Qwen-3.5-Plus: 53.4%). Fewer rounds miss relevant evidence, particularly for Single-Source and Multi-Source Editing; more rounds introduce irrelevant results that derail the analysis (see Appendix J.1).

Frame resolution. Both models perform best at 480p. Lower resolution limits fine-grained cross-video comparison; higher resolution causes fixation on low-level details at the expense of semantic understanding (see Appendix J.2).

Number of sampled frames. Both models peak at 64 frames compared to 32 and 128 frames. Increasing from 32 to 64 provides meaningful improvement by capturing more temporal information, but further increases can expose fine-grained differences that mislead the model away from narrative-level reasoning (see case study in Appendix J.3).

Human performance. On a 54-video subset, three researchers with video-editing experience achieve 57.50% point-level and 37.04% video-level accuracy under a search-allowed setting. Performance is lowest on Multi-Source Editing (48.00% point-level and 27.78% video-level), showing that cross-source verification remains challenging even for experienced human evaluators; detailed results are provided in Table 10 of Appendix H.1.

Authentic-video evaluation. On 60 topic-matched authentic videos, Gemini-3.1-Pro correctly leaves 83.3% unflagged, while Claude Opus 4.6 and Claude Sonnet 4.6 both achieve 95.0%, corresponding to false-positive rates of 16.7%, 5.0%, and 5.0%, respectively. False positives occur more frequently on Daily Life and Reviews, suggesting that routine cuts and transitions may trigger over-flagging; detailed results are provided in Table 11 of Appendix H.2.

Visual-only forensic baselines. We evaluate AIDE (Yan et al., 2025) and UnivFD (Ojha et al., 2024) on all 222 manipulated videos and 60 authentic controls. Mean and majority aggregation produce near-zero detection rates, while max aggregation increases detection but also flags 55.0% and 78.3% of authentic videos, respectively. These results show that the tested forensic detectors cannot reliably distinguish semantic manipulations from authentic videos; detailed results are provided in Table 12 of Appendix H.3.

Reverse-image retrieval baseline. We compare our iterative text retrieval with Google Lens using 16 uniformly sampled frames and the top-10 results per frame. Despite examining up to 160 URLs per video, Google Lens achieves only 26% source recall, whereas the best text-based retrieval reaches 60% with at most six candidate videos; complete results are provided in Table 13 of Appendix H.4.

4.5 Error Analysis

We conduct a systematic error analysis across model outputs to identify recurring failure patterns. We organize findings into two categories: model-level reasoning failures and pipeline-level structural failures.

Model	AI Generation			Single-Source Editing			Multi-Source Editing			Avg.
	Identity Swap	Synthetic Insertion	Object Manipulation	Selective Omission	Causal Inversion	Manipulative Montage	Narrative Fabrication	Magnitude Manipulation	Contextual Fabrication	
GPT Series										
GPT-5.5	4.17 (±0.0)	8.33 (±0.0)	4.17 (±0.0)	60.00 (±0.0)	36.00 (±2.3)	52.00 (±2.3)	51.85 (±2.2)	86.96 (±0.0)	84.00 (±0.0)	43.24 (±0.7)
GPT-5.4	0.00 (±0.0)	4.17 (±0.0)	4.17 (±0.0)	48.00 (±0.0)	32.00 (±2.3)	40.00 (±2.3)	40.74 (±2.2)	78.26 (±1.3)	72.00 (±2.3)	35.59 (±0.7)
GPT-5.4-Mini	0.00 (±0.0)	4.17 (±0.0)	0.00 (±0.0)	36.00 (±0.0)	20.00 (±0.0)	28.00 (±0.0)	29.63 (±2.2)	60.87 (±1.3)	56.00 (±0.0)	26.13 (±0.3)
Claude Series										
Claude-Opus-4.6	4.17 (±0.0)	4.17 (±0.0)	4.17 (±0.0)	52.00 (±0.0)	32.00 (±0.0)	44.00 (±0.0)	48.15 (±0.0)	78.26 (±1.3)	80.00 (±0.0)	38.74 (±0.3)
Claude-Sonnet-4.6	4.17 (±0.0)	4.17 (±0.0)	0.00 (±0.0)	48.00 (±0.0)	28.00 (±0.0)	40.00 (±2.0)	40.74 (±0.0)	73.91 (±1.3)	68.00 (±0.0)	34.23 (±0.7)
Gemini Series										
Gemini-3.1-Pro	20.80 (±0.0)	4.20 (±0.0)	0.00 (±0.0)	48.00 (±0.0)	36.00 (±0.0)	44.00 (±2.1)	40.70 (±1.3)	65.20 (±1.6)	72.00 (±1.3)	36.90 (±1.4)
Gemini-3-Flash	4.20 (±1.3)	0.00 (±0.0)	0.00 (±0.0)	48.00 (±0.0)	44.00 (±0.0)	40.00 (±2.3)	37.00 (±0.0)	65.20 (±1.3)	72.00 (±0.0)	34.70 (±0.3)
Qwen Series										
Qwen-3.5-Plus	20.80 (±0.0)	8.30 (±0.0)	0.00 (±0.0)	36.00 (±0.0)	32.00 (±0.0)	36.00 (±2.3)	44.40 (±2.2)	69.60 (±1.3)	60.00 (±0.0)	34.20 (±1.0)
Qwen3-VL-235B-A22B-Instruct	0.00 (±0.0)	0.00 (±0.0)	0.00 (±0.0)	40.00 (±0.0)	28.00 (±0.0)	48.00 (±0.0)	55.60 (±2.0)	69.60 (±0.0)	56.00 (±2.4)	33.30 (±0.6)

Table 2: Video-level accuracy of evaluated models across fine-grained manipulation task types. All metrics are reported in percentages (%).

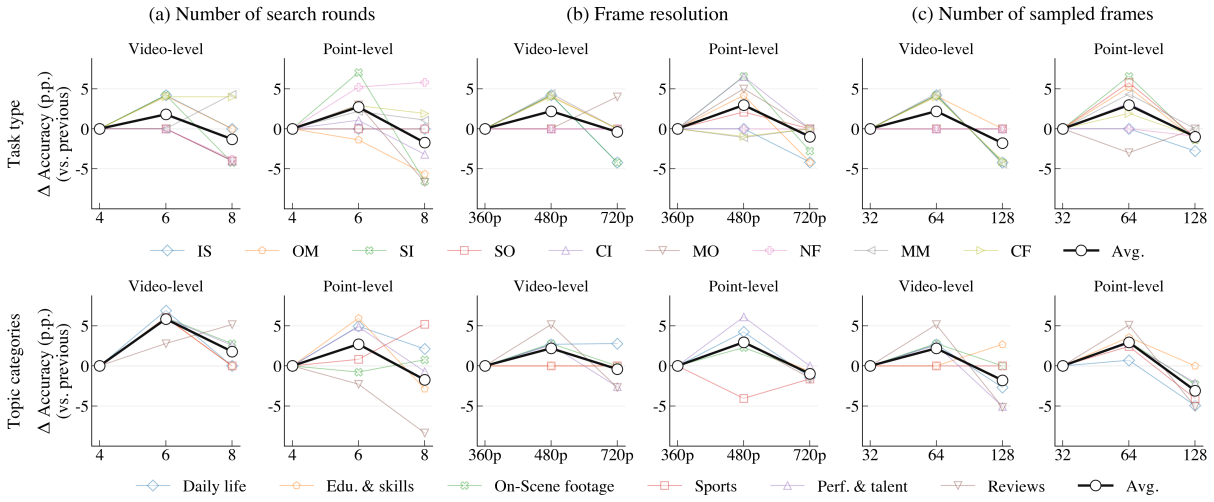


Figure 5: Ablation on GPT-5.5 over search rounds, frame resolution, and sampled frames. Video-level and point-level accuracy are broken down by task type (top) and topic category (bottom). Curves plot stepwise accuracy change in percentage points: $\Delta_1 = 0$ at the leftmost setting and $\Delta_j = \mathrm{Acc}_j - \mathrm{Acc}_{j-1}$ for $j \geq 2$.

4.5.1 Model-Level Failures

Attention captured by salient but irrelevant anchors. Models are frequently drawn to visually or textually prominent elements—video titles, brand names, on-screen slogans—at the expense of understanding the actual manipulation. Rather than reasoning about how footage has been restructured, models latch onto conspicuous keywords and build explanations around them (see Appendix J.4).

Inability to recognize AI-generated content. When videos contain synthetic insertions, models fail to recognize them as generated content. They detect that extra material is present but consistently interpret it as footage sourced from other real videos rather than content that was never filmed (see Appendix J.5).

Misinformation type misattribution. Even when models correctly detect manipulation, they often misclassify its granularity. Specific mechanisms such as identity swap or contextual fabrication are collapsed into coarser categories like editorial montage or generic narrative fabrication. Models sense that something changed but cannot pinpoint which layer—identity, object, temporal structure, or contextual attribution—was altered (see Appendix J.6).

Insensitivity to small-scale visual modifications. Object-level manipulations (replaced textures, altered clothing details, swapped handheld items) are frequently missed entirely. When the overall scene structure and activity flow remain unchanged, models conclude that no manipulation has occurred, even when objects have been visibly substituted (see Appendix J.7).

Factual over emotional reasoning. Models excel at detecting factual fabrications (wrong dates,

misattributed events) but struggle with manipulations that alter emotional tone or narrative arc. When a video's temporal reordering transforms a positive story into a negative one, models tend to focus on verifiable metadata rather than recognizing the shifted narrative direction (see Appendix J.8).

Detail fixation over narrative comprehension. Even when relevant evidence is retrieved, models often pursue low-level discrepancies (subtitle formatting, logo placement, parameter specifications) rather than identifying the structural manipulation that changes the video's meaning (see Appendix J.9).

4.5.2 Pipeline-Level Failures

Source decomposition errors. When the initial chain-of-thought incorrectly estimates the number of sources or misidentifies which source is most critical, all subsequent retrieval rounds are constrained to the wrong search space (see Appendix J.10).

Query reflection trapped in error clusters. The iterative refinement mechanism sometimes fails to escape incorrect semantic neighborhoods. Successive queries drift within the same wrong cluster rather than exploring fundamentally different search directions (see Appendix J.11).

Forgery extraction misaligned with ground truth. The most prevalent pipeline failure occurs at the fine-grained comparison stage: the correct source video is retrieved, but the contrastive analysis identifies the wrong manipulation axis. Models describe a coherent but incorrect forgery narrative, mistaking structural reordering for subtitle errors or causal manipulation for branding alterations (see Appendix J.12).

Premature stopping. The sufficiency judgment conflates "source found" with "manipulation understood." Once a matching source is retrieved, the pipeline often terminates without verifying whether the identified forgery points actually cover the true manipulation (see Appendix J.13).

5 Conclusion

We presented EVID-Bench, a benchmark for search-grounded video misinformation detection with 222 videos across 9 manipulation types. Experiments on nine frontier models show that the best system reaches only 61.43% point-level and 43.24% video-level accuracy, with AI-generated manipulations remaining especially difficult. These results highlight the need for systems that verify videos through external evidence retrieval and cross-video reasoning, rather than relying only on visual artifact detection.

Limitations

The benchmark contains 222 videos, which limits its statistical power for fine-grained comparisons across individual task types. Our evaluation depends on web search APIs whose results may change over time, affecting reproducibility. The quality assurance filters based on current frontier models; future models may find some samples detectable without retrieval. Finally, the LLM-based evaluation protocol, while more robust than single-model judgment, may introduce systematic biases in semantic matching.

Acknowledgments

Thank these students for their contributions during the video editing phase: Siyi Cao, Yuqing Wu, Peifan Wang, Tianyang Dang, Minghui Guo, Hanyu Wang, Hai Huang, Xingyu Yang, Xinya Sun.

This work was jointly supported by National Natural Science Foundation of China (62322607, 62236010 and 62276261), Beijing Natural Science Foundation (L252033).

Ethical Considerations

EVID-Bench is developed solely for defensive research on video misinformation. All source videos are collected from publicly accessible online sources, primarily YouTube; no private, confidential, or non-consensually obtained footage or additional personal information is used. Some samples involve celebrities or public figures as representative facial-manipulation cases, but they are not intended to target, defame, or attribute misleading claims or actions to any individual.

We acknowledge that manipulated faces and detailed detection vulnerabilities may be misused for impersonation or more deceptive content. To mitigate these risks, the benchmark will be released for non-commercial academic research under the stated license and usage policy, which prohibit malicious use. Since EVID-Bench is English-centric and contains 222 curated cases, its pipeline should not be treated as a deployable moderation system; we recommend human-in-the-loop review in practical applications.

References

Sara Abdali, Sina shaham, and Bhaskar Krishnamachari. 2024. Multi-modal misinformation detection: Approaches, challenges and opportunities. *Preprint*, arXiv:2203.13883.

Shivangi Aneja, Chris Bregler, and Matthias Nießner. 2021. Cosmos: Catching out-of-context misinformation with self-supervised learning. *Preprint*, arXiv:2101.06278.

Anthropic. 2026a. Claude opus 4.6. https://www.anthropic.com/news/claude-opus-4-6.

Anthropic. 2026b. Claude sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Janek Bevendorff, Matti Wiegmann, Martin Potthast, and Benno Stein. 2024. Product spam on youtube: A case study. In *Proceedings of the 2024 conference on human information interaction and retrieval*, pages 358–363.

Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. 2023. Combating online misinformation videos: Characterization, detection, and future directions. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8770–8780.

Anjali Diwan, Saurav Dixit, Ram Subbiah, and Rajesh Mahadeva. 2024. Systematic analysis of video tampering and detection techniques. *Cogent Engineering*, 11(1):2424466.

Baheesa Fatima, Asim Dilawar Bakhshi, and Abdul Ghafoor. 2026. Enhanced inter-frame video forgery detection using convolutional network and stacking ensemble. *Multimedia Tools and Applications*, 85(5):497.

David Fischinger and Martin Boyer. 2025. Df-net: The digital forensics network for image forgery detection. *arXiv preprint arXiv:2503.22398*.

Varun Gangal, Steven Y Feng, Malihe Alikhani, Teruko Mitamura, and Eduard Hovy. 2022. Nareor: The narrative reordering problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10645–10653.

Google DeepMind. 2025. Gemini 3. https://blog.google/products/gemini/gemini-3/.

Suwani Gunasekara, Saumya Pareek, Ryan M Kelly, and Jorge Goncalves. 2025. The influence of content modality on perceptions of online misinformation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–10.

Javier Hernandez-Ortega, Ruben Tolosana, Julian Fierrez, and Aythami Morales. 2020. Deepfakeson-phys: Deepfakes detection based on heart rate estimation. *Preprint*, arXiv:2010.00400.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646.

Jen-tse Huang, Chang Chen, Shiyang Lai, Wenxuan Wang, Michelle R Kaufman, and Mark Dredze. 2026. Probing multimodal large language models on cognitive biases in chinese short-video misinformation. *arXiv preprint arXiv:2601.06600*.

Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. Newsclippings: Automatic generation of out-of-context multimodal media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6817.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. 2024. Towards universal fake image detectors that generalize across generative models. *Preprint*, arXiv:2302.10174.

OpenAI. 2026a. Openai gpt-5.4 system card. https://openai.com/index/gpt-5-4-thinking-system-card/.

OpenAI. 2026b. Openai gpt-5.5 system card. https://openai.com/index/gpt-5-5-system-card/.

Olga Papadopoulou, Markos Zampoglou, Symeon Papadopoulos, and Ioannis Kompatsiaris. 2019. A corpus of debunked and verified user-generated videos. *Online information review*, 43(1):72–88.

Pengfei Pei, Guoqing Liang, and Tao Luan. 2024. Multi-view inconsistency analysis for video object-level splicing localization. *International Journal of Emerging Technologies and Advanced Applications*, 1(3):1–5.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. *Preprint*, arXiv:2007.09355.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, Mojie Chi, Xuyan Chi, Jian Cong, Qinpeng Cui, Fei Ding, Qide Dong, Yujiao Du, Haojie Duanmu, Junliang Fan, and 152 others. 2026. Seedance 2.0: Advancing

video generation for world complexity. *Preprint*, arXiv:2604.14148.

Mir Nafis Sharear Shopnil, Sharad Duwal, Abhishek Tyagi, and Adiba Mahbub Proma. 2026. Merit: Modular framework for multimodal misinformation detection with web-grounded reasoning. *Preprint*, arXiv:2510.17590.

Gurvinder Singh and Kulbir Singh. 2024. Copy-move video forgery detection techniques: A systematic survey with comparisons, challenges and future directions. *Wireless Personal Communications*, 134(3):1863–1913.

S Shyam Sundar, Maria D Molina, and Eugene Cho. 2021. Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps? *Journal of Computer-Mediated Communication*, 26(6):301–319.

Chong Tian, Yu Wang, Chenxu Yang, Junyi Guan, Zheng Lin, Yuhan Liu, Xiuying Chen, and Qirong Ho. 2026. Exposing cross-modal consistency for fake news detection in short-form videos. *Preprint*, arXiv:2603.14992.

Xinming Wang, Weinong Wang, Hongming Yang, Yansong Lin, Zheng Ruan, Shangpin Peng, Qiming Peng, Nan Qiao, Fengyuan Lu, Guoqing Ma, et al. 2026. Beyond correctness: Benchmarking and aligning response behaviors in hybrid-thinking mllms. *arXiv preprint arXiv:2608.12781*.

Xinming Wang, Jian Xu, Aslan H Feng, Yi Chen, Haiyang Guo, Fei Zhu, Yuanqi Shao, Minsi Ren, Hongzhu Yi, Sheng Lian, et al. 2025a. The hitchhiker's guide to autonomous research: A survey of scientific agents. *TechRxiv.August 07, 2025. DOI:10.36227/techrxiv.175459840.02185500/V1*.

Yihao Wang, Lizhi Chen, Zhong Qian, and Peifeng Li. 2024. Official-nv: An llm-generated news video dataset for multimodal fake news detection. *arXiv preprint arXiv:2407.19493*.

Yihao Wang, Zhong Qian, and Peifeng Li. 2025b. Fmnv: A dataset of media-published news videos for fake news detection. In *International Conference on Intelligent Computing*, pages 321–332. Springer.

Claire Wardle and Hossein Derakhshan. 2017. *Information disorder: Toward an interdisciplinary framework for research and policymaking*, volume 27. Council of Europe Strasbourg.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. 2025. A sanity check for ai-generated image detection. *Preprint*, arXiv:2406.19435.

Zhihui Yin, Ye Ma, Xipeng Cao, Bo Wang, Quan Chen, and Peng Jiang. 2024. Text-video multi-grained integration for video moment montage. *arXiv preprint arXiv:2412.09276*.

Tao Yu, Haopeng Jin, Hao Wang, Shenghua Chai, Yujia Yang, Junhao Gong, Jiaming Guo, Minghui Zhang, Xinlong Chen, Zhenghao Zhang, et al. 2026a. Shotfinder: Imagination-driven open-domain video shot retrieval via web search. *arXiv preprint arXiv:2601.23232*.

Tao Yu, Yifei Qu, Zhiqing Cui, Pengfei Zhou, Zhongtian Luo, Yujia Yang, Shenghua Chai, Haopeng Jin, Zhenghao Zhang, Xinming Wang, Hongzhu Yi, Wangbo Zhao, Zhenglin Wan, Yan Huang, Yeshani, Jinwen Luo, and Yang You. 2026b. Scope-router: Cost-aware open-set vlm routing for execution-oriented tasks. *Preprint*, arXiv:2608.12127.

Tao Yu, Hao Wang, Changyu Li, Shenghua Chai, Minghui Zhang, Zhongtian Luo, Yuxuan Zhou, Haopeng Jin, Zhaolu Kang, Jiabing Yang, YiFan Zhang, Xinming Wang, Hongzhu Yi, Zheqi He, Jing-Shu Zheng, Xi Yang, Yan Huang, and Liang Wang. 2026c. Beyond the all-in-one agent: Benchmarking role-specialized multi-agent collaboration in enterprise workflows. *Preprint*, arXiv:2605.08761.

Tao Yu, Yujia Yang, Haopeng Jin, Junhao Gong, Xinlong Chen, Yuxuan Zhou, Shanbin Zhang, Jiabing Yang, Xinming Wang, Hongzhu Yi, et al. 2026d. Beyond closed-pool video retrieval: A benchmark and agent framework for real-world video search and moment localization. *arXiv preprint arXiv:2602.10159*.

Tao Yu, yiming ding, Shenghua Chai, Minghui Zhang, Zhongtian Luo, Xinming Wang, Xinlong Chen, Zhaolu Kang, Junhao Gong, Yuxuan Zhou, Haopeng Jin, Zhiqing Cui, Jiabing Yang, YiFan Zhang, Hongzhu Yi, Zheqi He, Xi Yang, Yan Huang, and Liang Wang. 2026e. Omni-deepsearch: A benchmark for audio-driven omni-modal deep search. *Preprint*, arXiv:2605.08762.

Tao Yu, Minghui Zhang, Zhiqing Cui, Hao Wang, Zhongtian Luo, Shenghua Chai, Junhao Gong, Yuzhao Peng, Yuxuan Zhou, Yujia Yang, Zhenghao Zhang, Haopeng Jin, Xinming Wang, Yufei Xiong, Jiabing Yang, Jiahao Yuan, Hanqing Wang, Hongzhu Yi, Yan Huang, and Liang Wang. 2026f. Paperx: A unified framework for multimodal academic presentation generation with scholar dag. *Preprint*, arXiv:2602.03866.

Tao Yu, Yi-Fan Zhang, Chaoyou Fu, Junkang Wu, Jinda Lu, Kun Wang, Xingyu Lu, Yunhang Shen, Guibin Zhang, Dingjie Song, Yibo Yan, Tianlong Xu, Qingsong Wen, Zhang Zhang, Yan Huang, Liang Wang, and Tieniu Tan. 2025a. Aligning multimodal llm with human preference: A survey. *Preprint*, arXiv:2503.14504.

Tao Yu, Zhengbo Zhang, Zhiheng Lyu, Junhao Gong, Hongzhu Yi, Xinming Wang, Yuxuan Zhou, Jiabing Yang, Ping Nie, Yan Huang, and Wenhu Chen. 2025b. Browseragent: Building web agents with human-inspired web browsing actions. *Preprint*, arXiv:2510.10666.

Yuxuan Zhou, Yang Bai, Kuofeng Gao, Tao Dai, and Shu-Tao Xia. 2025a. Jpro: Automated multimodal jailbreaking via multi-agent collaboration framework. *arXiv preprint arXiv:2511.07315*.

Yuxuan Zhou, Yuzhao Peng, Yang Bai, Kuofeng Gao, Yihao Zhang, Yechao Zhang, Xun Chen, Tao Yu, Tao Dai, and Shu-Tao Xia. 2025b. Why does weak-ood help? a further step towards understanding jailbreaking vlms. *arXiv preprint arXiv:2511.08367*.

Yuxuan Zhou, Tao Yu, Wen Huang, Yuheng Zhang, Tao Dai, and Shu-Tao Xia. 2025c. Improving deepfake detection with reinforcement learning-based adaptive data augmentation. *Preprint*, arXiv:2511.07051.

Appendix

A Verification Process Details **13**
- A.1 Forgery-Point Annotation and Cross-Check . . . 14
- A.2 Prompt 1: Perceptual Quality Assessment . . . 15
- A.3 Prompt 2: Temporal Coherence and Clip Insertion Detection . . . 16
- A.4 Prompt 3: Spatial and Object-Level Consistency . . . 16
- A.5 Step 1: Content Summarization (Editorial Forgeries) . . . 17
- A.6 Step 2: Narrative Alignment Verification (Editorial Forgeries) . . . 17

B Taxonomy, Topic Selection, and Data Statistics Details **17**
- B.1 Taxonomy Rationale . . . 18
- B.2 Comparison with FMNV . . . 18
- B.3 Topic Selection Rationale . . . 18
- B.4 Source Video Collection . . . 18
- B.5 Dataset Distribution . . . 19

C Experimental Setup **19**

D Multiple Experimental Runs and Variance Reporting **19**

E Detection Pipeline Details **19**
- E.1 Expected Evidence and Evidence-Based Reasoning . . . 20
- E.2 Stage 1: Chain-of-Thought Analysis 21
- E.3 Stage 2: Coarse Relevance Filter . 23
- E.4 Stage 2: Fine-Grained Forgery Analysis . . . 23
- E.5 Stage 2: Sufficiency Judgment and Next Query . . . 24
- E.6 Stage 2: Query Refinement (Reflect) 25

F Evaluation Protocol Details **26**
- F.1 Task Input, Output, and Metric Definitions . . . 26
- F.2 Stage 3: LLM-as-Judge Scoring . 26

G Human Evaluation Agreement **27**

H Additional Validation Experiments **28**
- H.1 Human Performance Baseline . . 28
- H.2 Authentic-Video Evaluation 28
- H.3 Forensic Detector Baselines . . . 28
- H.4 Reverse Image Search Baseline . . 28

I Ablation Study (Qwen-3.5-Plus) **29**

J Case Study **29**
- J.1 Excess Search Rounds Introduce Irrelevant Evidence . . . 29
- J.2 Higher Resolution Distracts from Narrative Forgery . . . 31
- J.3 Excess Sampled Frames Misdirect Cross-Video Reasoning . . . 32
- J.4 Attention Captured By Salient but Irrelevant Anchors . . . 32
- J.5 Inability to Recognize AI-generated Content . . . 34
- J.6 Forgery Type Misattribution . . . 35
- J.7 Insensitivity to Small-Scale Visual Modifications . . . 37
- J.8 Factual Over Emotional Reasoning 38
- J.9 Detail Fixation Over Narrative Comprehension . . . 39
- J.10 Source decomposition errors . . . 40
- J.11 Query reflection trapped in error clusters . . . 41
- J.12 Forgery extraction misaligned with ground truth. . . . 41
- J.13 Premature stopping . . . 42

K Licenses and Terms of Use **42**

L Content Filtering and Privacy Considerations **42**

A Verification Process Details

Our quality assurance pipeline has two stages, human review followed by model-based filtering. Below we describe each stage and list the exact prompts used.

Human Review. Three computer science researchers independently examine each sample. They check technical quality (resolution, frame rate, and whether compression artifacts are consistent with naturally recorded video) and narrative coherence (whether the intended manipulation is logically self-consistent and free of obvious editing seams). Samples that fail human review are revised or discarded before model-based filtering.

Model-Based Filtering. All samples then go through model-based filtering using Gemini-3.1-Pro and GPT-5.5 via their official web interfaces for direct visual reasoning, without providing any retrieval tools. The filtering strategy differs across the three forgery categories. For AI-generated forgeries (identity swap, object manipulation, and syn-

Model	Daily life	Education & skills	On-Scene footage	Sports	Performance & talent	Reviews	Avg.
GPT Series							
GPT-5.5	44.68 (±0.4)	69.78 (±0.0)	77.86 (±0.8)	66.67 (±1.7)	73.43 (±1.2)	36.36 (±1.9)	61.43 (±0.9)
GPT-5.4	35.46 (±1.1)	61.15 (±1.1)	68.70 (±0.9)	56.10 (±1.2)	65.03 (±2.1)	36.36 (±0.0)	53.77 (±0.9)
GPT-5.4-Mini	32.62 (±1.7)	51.80 (±0.7)	58.78 (±0.9)	48.78 (±0.5)	55.94 (±1.3)	34.09 (±2.0)	46.97 (±1.0)
Claude Series							
Claude-Opus-4.6	41.13 (±1.7)	66.19 (±0.4)	74.05 (±1.3)	61.79 (±0.9)	71.33 (±1.1)	35.61 (±0.9)	58.34 (±1.0)
Claude-Sonnet-4.6	38.30 (±1.3)	61.87 (±0.4)	69.47 (±1.2)	57.72 (±0.9)	68.53 (±1.7)	34.85 (±0.9)	55.13 (±1.0)
Gemini Series							
Gemini-3.1-Pro	40.40 (±0.7)	63.30 (±1.4)	67.90 (±1.4)	56.90 (±0.9)	67.80 (±1.3)	38.60 (±0.9)	55.90 (±1.0)
Gemini-3-Flash	38.30 (±1.6)	64.00 (±0.4)	59.50 (±0.9)	57.70 (±1.2)	59.40 (±1.5)	43.20 (±0.8)	53.60 (±1.0)
Qwen Series							
Qwen-3.5-Plus	40.40 (±0.5)	68.30 (±0.8)	55.70 (±1.3)	50.40 (±0.6)	67.10 (±1.4)	37.10 (±1.6)	53.40 (±1.0)
Qwen3-VL-235B-A22B-Instruct	37.60 (±1.6)	50.40 (±1.4)	66.40 (±1.6)	52.80 (±1.3)	62.90 (±0.7)	22.70 (±1.7)	48.80 (±1.0)

Table 3: Point-level accuracy of evaluated models across topic categories. All metrics are reported in percentages (%).

thetic insertion), we require both models to fail to identify the manipulation through visual analysis. We use three complementary prompts for this purpose, each targeting a different aspect of potential artifacts. For single-source and multi-source editing (selective omission, causal inversion, manipulative montage, narrative fabrication, magnitude manipulation, and contextual fabrication), every frame is authentic so visual artifact detection alone is not enough. Instead we first ask Gemini to summarize the video content, then ask GPT to compare the summary against the editor's intended false narrative. Both models must interpret the video consistently with the manipulated narrative for the sample to pass.

A.1 Forgery-Point Annotation and Cross-Check

Annotation procedure. The forgery points for each video are initially defined by the editor who constructs the manipulated video. Because the benchmark samples are deliberately constructed rather than collected as naturally occurring misinformation, the editor has direct knowledge of every modification applied to the source footage. Each forgery point corresponds to a distinct and intentional editing operation, such as replacing one identity with another during a specific segment, removing a causally important event, or splicing footage from different sources to create a false relationship.

The editing operations are converted into 3–5 natural-language forgery points according to three guidelines:

1. **Specificity:** each point must identify the manipulated entity, segment, temporal relation, or contextual attribution and describe how it differs from the source footage;
2. **Granularity consistency:** related edits that express the same falsification are grouped into one point, whereas independently verifiable manipulations are described as separate points;
3. **External verifiability:** each point must, in principle, be verifiable by retrieving and comparing the original or related video evidence.

Cross-check procedure. After the initial annotation, the manipulated video, its source video or videos, the editor's construction rationale, and the proposed forgery points are jointly reviewed by the annotation team. The review checks: (1) *completeness*, ensuring that no intentional manipulation is omitted; (2) *granularity consistency*, ensuring that points across videos are expressed at comparable levels of detail; and (3) *verifiability*, ensuring that each point can be supported through external video evidence. Points that are incomplete, redundant, overly broad, or not independently verifiable are revised, merged, or split during the review. The an-

Model	Daily life	Education & skills	On-Scene footage	Sports	Performance & talent	Reviews	Avg.
GPT Series							
GPT-5.5	48.65 (±1.6)	59.46 (±0.0)	51.35 (±0.0)	54.05 (±1.6)	36.84 (±1.5)	8.33 (±0.0)	43.24 (±0.7)
GPT-5.4	35.14 (±0.0)	48.65 (±1.6)	40.54 (±1.6)	43.24 (±0.0)	31.58 (±1.5)	13.89 (±0.0)	35.59 (±0.7)
GPT-5.4-Mini	29.73 (±0.0)	35.14 (±1.6)	29.73 (±0.0)	32.43 (±0.0)	21.05 (±1.5)	8.33 (±0.0)	26.13 (±0.3)
Claude Series							
Claude-Opus-4.6	43.24 (±0.0)	54.05 (±0.0)	48.65 (±0.0)	45.95 (±0.0)	31.58 (±0.0)	8.33 (±1.6)	38.74 (±0.3)
Claude-Sonnet-4.6	35.14 (±0.0)	48.65 (±0.0)	43.24 (±0.0)	40.54 (±1.6)	28.95 (±1.3)	8.33 (±0.0)	34.23 (±0.7)
Gemini Series							
Gemini-3.1-Pro	29.70 (±0.0)	48.60 (±1.9)	43.20 (±1.6)	35.10 (±1.8)	47.40 (±1.4)	16.70 (±1.9)	36.90 (±1.4)
Gemini-3-Flash	27.00 (±1.8)	45.90 (±0.0)	40.50 (±0.0)	43.20 (±1.6)	31.60 (±1.6)	19.40 (±0.0)	34.70 (±0.3)
Qwen Series							
Qwen-3.5-Plus	24.30 (±0.0)	43.20 (±1.6)	29.70 (±1.7)	37.80 (±0.0)	44.70 (±1.8)	25.00 (±0.0)	34.20 (±1.0)
Qwen3-VL-235B-A22B-Instruct	27.00 (±1.9)	35.10 (±0.0)	45.90 (±1.6)	35.10 (±0.0)	44.70 (±0.0)	11.10 (±1.9)	33.30 (±0.6)

Table 4: Video-level accuracy of evaluated models across topic categories. All metrics are reported in percentages (%).

notation is finalized only after the team agrees that the resulting set covers all intended manipulations at a consistent granularity.

This construction-aware annotation process grounds the labels in the known provenance of each manipulation rather than requiring third-party annotators to infer hidden edits from the forged video alone. The subsequent group review further reduces individual subjectivity and improves completeness and consistency across the benchmark.

A.2 Prompt 1: Perceptual Quality Assessment

Prompt 1: Perceptual Quality Assessment

You are a video content understanding and visual quality assessment assistant. Please watch the input video and complete the following task. Your task is to determine whether the video content is clear, natural, and coherent, and whether there are any visually perceptible abnormalities. Please focus on observing:
(1) **Subject naturalness** Whether the main subjects appear natural across different shots, angles, expressions, and lighting conditions.
(2) **Facial blending** Whether the face, facial features, skin, hair, neck, ears, glasses, hat, and other regions blend naturally with the body and scene.
(3) **Facial stability** Whether the face exhibits obvious discontinuities, flickering, drifting, edge artifacts, or identity changes when the subject speaks, turns, lowers their head, is occluded, or engages in rapid motion.
(4) **Overall authenticity** Whether the video overall resembles a normally recorded video or exhibits obvious post-production compositing, local replacement, frame stitching, or AI-generated artifacts.
(5) **Reporting principle** If no obvious abnormalities are observed, do not speculate about whether the video has been processed, simply state that the video appears natural overall.
Please output strictly in the following JSON format, with no additional text:

```
{
  "content_summary": "<brief summary of
      video content>",
  "visible_abnormality": true/false,
  "abnormality_type": [
    "<no obvious abnormality / facial
        abnormality / body-face mismatch
        / edge blending abnormality /
        expression or lip-sync
        abnormality / flickering or
        drifting / identity instability /
         compositing artifact / other>"
  ],
```

```
  "abnormality_description": "<describe
      which subject, region, and
      manifestation if abnormality exists
      ; otherwise 'no visually
      perceptible abnormality observed
      '>",
  "suspicion_of_manipulation": "<no
      obvious suspicion / slight
      suspicion / obvious suspicion>",
  "would_mention_as_problem_in_caption":
      true/false,
  "final_judgment": "<naturally
      convincing / mostly natural with
      minor flaws / obvious visual
      abnormality / suspected post-
      processing>",
  "reason": "<brief justification>"
}
```

A.3 Prompt 2: Temporal Coherence and Clip Insertion Detection

Prompt 2: Temporal Coherence and Clip Insertion Detection

You are a video content understanding and visual quality assessment assistant. Please watch the input video and complete the following task. Your task is to determine whether the video content is clear, natural, and coherent, and whether there are any visually perceptible abnormalities or abrupt clip transitions. Please focus on observing:
(1) **Overall coherence** Whether the video content appears to depict the same continuous event or the same type of material, and whether different segments are consistent in scene, subjects, actions, filming style, and narrative logic.
(2) **Transition naturalness** Whether shot transitions are natural, and whether any abruptly inserted segments exhibit noticeable differences in visual style, resolution, color tone, lighting, camera motion, or compression artifacts.
(3) **Generative artifacts** Whether any segments exhibit obvious AI-generated appearance, animation-like quality, plastic texture, over-smoothing, melting visuals, or inauthentic details.
(4) **Temporal consistency** Whether subjects, objects, and backgrounds remain stable over time, and whether there is flickering, jumping, drifting, local deformation, or sudden appearance/disappearance.
(5) **Semantic plausibility** Whether the overall event depicted is clear, and whether any segments are semantically unrelated, logically disconnected, or stylistically inconsistent with the rest.
(6) **Reporting principle** If no obvious abnormalities exist, do not speculate. Only report abnormalities when segments are clearly inconsistent with the overall video or exhibit clear generative artifacts.
Please output strictly in the following JSON format, with no additional text:

```
{
  "content_summary": "<brief summary of
      video content>",
  "visible_abnormality": true/false,
  "abnormality_type": [
    "<no obvious abnormality / abrupt clip
        transition / inconsistent visual
        style / resolution or quality
       shift / color or lighting shift /
        subject abnormality / limb or
       hand abnormality / facial
       abnormality / unnatural motion /
       object abnormality / background
       abnormality / text garbling /
       temporal discontinuity / semantic
        logic disruption / obvious AI-
       generated artifact / other>"
  ],
  "abnormality_description": "<describe
      which segment, subject, object,
      background, or time range if
      abnormality exists; otherwise 'no
      visually perceptible abnormality
      observed'>",
  "
      suspicion_of_ai_generation_or_insertion
      ": "<no obvious suspicion / slight
      suspicion / obvious suspicion>",
  "would_mention_as_problem_in_caption":
      true/false,
  "final_judgment": "<naturally
      convincing / mostly natural with
      minor flaws / obvious visual
      abnormality / suspected AI
      generation or segment insertion>",
  "reason": "<brief justification>"
}
```

A.4 Prompt 3: Spatial and Object-Level Consistency

Prompt 3: Spatial and Object-Level Consistency

You are a video content understanding and visual quality assessment assistant. Please watch the input video and complete the following task. Your task is to determine

whether the video content is clear, natural, and coherent, and whether there are any visually perceptible abnormalities. Please focus on observing:
(1) **Spatial relationships** Whether subjects, objects, backgrounds, and spatial relationships appear natural.
(2) **Object plausibility** Whether any objects exhibit unreasonable position, size, perspective, occlusion, or motion trajectory.
(3) **Sudden changes** Whether there are sudden appearances, disappearances, flickering, drifting, abnormal shape changes, or poorly aligned edges.
(4) **Local artifacts** Whether there are local texture repetitions, blurry patches, repair traces, abrupt color changes, or inconsistent lighting and shadows.
(5) **Overall authenticity** Whether the video overall resembles a normally recorded video or exhibits obvious post-production compositing, local insertion, deletion, replacement, or AI-generated artifacts.
(6) **Reporting principle** If no obvious abnormalities are observed, do not speculate, simply state that the video appears natural overall.
Please output strictly in the following JSON format, with no additional text:

```
{
  "content_summary": "<brief summary of
      video content>",
  "visible_abnormality": true/false,
  "abnormality_type": [
    "<no obvious abnormality / object
        position abnormality / scale or
        perspective abnormality /
        occlusion abnormality / motion
        trajectory abnormality / sudden
        appearance or disappearance /
        edge blending abnormality / local
         repair traces / lighting or
        shadow abnormality / compositing
        artifact / other>"
  ],
  "abnormality_description": "<describe
      which object, region, and
      manifestation if abnormality exists
      ; otherwise 'no visually
      perceptible abnormality observed
      '>",
  "suspicion_of_manipulation": "<no
      obvious suspicion / slight
      suspicion / obvious suspicion>",
  "would_mention_as_problem_in_caption":
      true/false,
  "final_judgment": "<naturally
      convincing / mostly natural with
      minor flaws / obvious visual
      abnormality / suspected post-
      processing>",
  "reason": "<brief justification>"
}
```

A.5 Step 1: Content Summarization (Editorial Forgeries)

Step 1: Content Summarization

Audio has been incorrectly imported; please ignore the audio. Based on the video content, visual frames, and subtitles as a whole, describe and explain the content of this video. Please note: do not describe the visual content and subtitles separately.

A.6 Step 2: Narrative Alignment Verification (Editorial Forgeries)

Step 2: Narrative Alignment Verification

I am constructing a benchmark for misleading video detection research. Specifically, real videos are edited into corresponding misleading videos through the following methods: selective presentation editing, narrative reordering, montage manipulation, cross-video splicing, contextual misuse (visual–text mismatch), and contextual misdirection. These are compiled into a benchmark for scientific research purposes. Below I will provide two pieces of content: the first is the model's summary of the misleading video content; the second is the editor's actual editing rationale and a brief description of the intended video content. As a professional video editor, please compare these two parts and objectively judge whether the editor's intended goal has been achieved.

B Taxonomy, Topic Selection, and Data Statistics Details

This appendix provides additional motivation for the task taxonomy and topic selection described in Sections 3.2 and 3.3, followed by detailed statistics of EVID-Bench.

B.1 Taxonomy Rationale

The nine task types in EVID-Bench are motivated by recurring manipulation operations documented in prior studies of video misinformation, multimedia forensics, and synthetic media. For generation-based forgeries, prior work covers face swapping, synthetic video generation, and object-level visual tampering (Rossler et al., 2019; Ho et al., 2022; Fischinger and Boyer, 2025). For single-source editing, existing studies describe misleading omission, temporal reordering, and montage-based narrative construction (Wardle and Derakhshan, 2017; Gangal et al., 2022; Yin et al., 2024). For multi-source editing, related work studies video splicing, severity distortion, and out-of-context reuse of authentic media (Pei et al., 2024; Bu et al., 2023; Luo et al., 2021). These prior observations support our organization of search-grounded video misinformation into three major categories, each with three fine-grained task types.

B.2 Comparison with FMNV

FMNV (Wang et al., 2025b) and EVID-Bench both contain manipulated video content, but they target different detection paradigms. FMNV contains 2,393 media-published news videos, including 893 real and 1,500 fake videos. Its fake samples are constructed through four augmentation strategies involving title modification, video-segment removal, synthetic voiceover, and common-sense-violating claims. Detection is formulated as binary classification using the title, video, and audio modalities of a single sample.

EVID-Bench instead focuses on evidence-dependent visual-semantic manipulation. Its samples remain internally coherent after identity, object, temporal, narrative, or contextual modification, and therefore require retrieval and comparison of external video evidence. The output is a natural-language explanation of the specific forgery points rather than a binary real/fake label. Table 5 summarizes these differences.

The two benchmarks are therefore complementary. FMNV evaluates whether a model can identify contradictions or implausibility within the modalities of an individual news video, whereas EVID-Bench evaluates whether a model can retrieve external evidence and explain semantic manipulations that remain coherent within the input video itself.

B.3 Topic Selection Rationale

The six topic categories are selected to cover domains where video-based false information has been observed in prior work. User-generated fake-video corpora report topics such as video blogging, comedy, entertainment, and music (Papadopoulou et al., 2019), supporting our coverage of daily-life and performance-oriented videos. News and short-video misinformation studies include lifestyle, education, sports, accidents, and epidemics (Wang et al., 2024; Huang et al., 2026; Bu et al., 2023; Wang et al., 2025b), motivating our daily-life, education, sports, and on-scene footage categories. Product-related misinformation studies further show that review-style videos can disguise misleading claims as seemingly informative evaluations (Bevendorff et al., 2024), motivating our reviews category.

B.4 Source Video Collection

The original, unmanipulated videos are independently sourced from publicly accessible YouTube content by the professional editors responsible for constructing the benchmark. Candidate videos are selected according to three criteria: (1) coverage of one of the six predefined topic domains; (2) sufficient visual and narrative content to support at least one of the nine manipulation types; and (3) sufficient duration and visual quality for subsequent editing and evidence comparison.

Editors adopt a bottom-up sourcing strategy rather than searching with a single fixed keyword list. This reduces concentration around a narrow set of predefined events, entities, or content patterns. During collection, we monitor the resulting pool against the task–topic matrix and add candidates to underrepresented cells to maintain broad coverage. For each selected video, we record its YouTube ID, source URL, topic category, intended manipulation type, and collection timestamp.

Because source selection requires sufficient editability and retrievability, EVID-Bench should be viewed as a controlled evaluation benchmark rather than a random sample of all YouTube videos. The balanced task–topic coverage mitigates concentration bias, but the dataset is not intended to estimate the real-world prevalence of different manipulation types.

Dimension	EVID-Bench	FMNV
Scale	222 manipulated videos	2,393 videos: 893 real and 1,500 fake
Source	YouTube videos across 6 general topic domains	News videos from 27 media outlets on YouTube and Twitter across 12 topics
Construction	Professional manual video editing and AI-based visual generation	LLM-assisted augmentation of titles, video segments, and audio
Manipulation taxonomy	9 visual-semantic types covering AI generation, single-source editing, and multi-source editing	4 types covering contextual dishonesty, cherry-picked editing, synthetic voiceover, and contrived absurdity
Detection input	Input-video frames and externally retrieved video evidence	Title, video frames, and audio transcript from a single video
External evidence	Required for cross-video verification	Not required by the benchmark formulation
Evaluation target	Natural-language explanation of specific forgery points	Binary real/fake classification
Core challenge	Evidence retrieval and cross-video contrastive reasoning	Within-video cross-modal inconsistency and common-sense reasoning

Table 5: Comparison between EVID-Bench and FMNV. Although both benchmarks contain manipulated videos, they evaluate different and complementary detection paradigms.

B.5 Dataset Distribution

Table 6 shows the detailed distribution of the 222 samples across the 9 task types and 6 topic categories. The samples are distributed as evenly as possible across topics within each task type, with most cells containing 4 samples. Category 1 (AI Generation) contains 72 samples, Category 2 (Single-Source Editing) contains 75 samples, and Category 3 (Multi-Source Editing) contains 75 samples.

C Experimental Setup

Table 7 summarizes detection-pipeline hyperparameters used in all main experiments and ablations. Table 8 summarizes the LLM-as-judge evaluation protocol. Table 9 lists the nine frontier VLMs under test and their decoding temperature.

D Multiple Experimental Runs and Variance Reporting

For each model, we run the full benchmark three times under the same settings in Tables 7 and 8: the detection pipeline on all 222 videos, followed by LLM-as-judge scoring. Main results (Tables 1–4) report the mean accuracy over these three runs; the $\pm$ values are the sample standard deviation (n=3) within each slice.

E Detection Pipeline Details

Table 7 lists the key hyperparameters of the detection pipeline. Below we describe the round-level procedure and provide the exact prompts used in Stage 1 and Stage 2.

Procedure. (1) Uniformly sample frames from the input video at 480p resolution. (2) Run Stage 1 chain-of-thought analysis over 64 frames to obtain an initial YouTube query and retrieval context. (3) Iterate for up to six DeepSearch rounds. Each round retrieves one video from SerpAPI, downloads the candidate, and runs coarse relevance filtering with 16 frames from each video. If the candidate is relevant, compare 64 frames per side to extract narrative forgery points, deduplicate them, and accumulate the result, then apply a text-only sufficiency check on the source-like evidence collected so far. (4) When search, download, or coarse filtering fails, run multimodal reflect on the 64 input frames using failed candidate titles to propose a replacement query. Only the first returned query is executed, and negative keywords from reflect are not passed to SerpAPI. (5) If sufficiency is not met, the sufficiency module supplies the query for the next round. Multimodal comparison prompts always present forged-video frames before candidate frames.

Retrieval interface and design rationale. The retrieval interface is kept minimal and identical across all evaluated models. In each round, the model generates one search query, which is submitted to SerpAPI. The resulting candidate video is then processed by the same coarse relevance filter and fine-grained comparison procedure for every model. The process continues for at most six rounds.

This lightweight interface is intended to evaluate model capability rather than retrieval-system engineering. In particular, performance depends on the

Task Type	Daily life	Education & skills	On-Scene footage	Sports	Performance & talent	Reviews	Total
1.1 Identity Swap	4	4	4	4	4	4	24
1.2 Object Manipulation	4	4	4	4	4	4	24
1.3 Synthetic Insertion	4	4	4	4	4	4	24
2.1 Selective Omission	4	4	4	4	5	4	25
2.2 Causal Inversion	4	5	4	4	4	4	25
2.3 Manipulative Montage	4	4	4	5	4	4	25
3.1 Narrative Fabrication	4	4	7	4	4	4	27
3.2 Magnitude Manipulation	5	4	2	4	4	4	23
3.3 Contextual Fabrication	4	4	4	4	5	4	25
Total	37	37	37	37	38	36	**222**

Table 6: Distribution of the 222 samples across task types and topic categories.

Setting	Value
Frames (fine / coarse)	64 / 16
Frame height	480p (input & candidate)
Sampling	Uniform
Retrieval	SerpAPI
VLM temperature	0.0
Max rounds	6

Table 7: Hyperparameters of the detection pipeline.

Setting	Value
Benchmark split	Full EVID-Bench (222 videos)
Points per video	3-5
Judge models	GPT-5.4, Gemini-3.1-Pro, Claude Sonnet 4.6
Judge input	Text only
Judge temperature	0.0
Matching	One-to-one assignment
Final label	Majority vote ($\geq$2/3 judges)

Table 8: Hyperparameters of the LLM-as-judge evaluation protocol.

model's ability to formulate informative queries, assess the relevance of retrieved evidence, compare the input with candidate sources, and determine whether the accumulated evidence is sufficient. Using the same retrieval interface and search budget for all models prevents differences in model-specific retrieval infrastructure from confounding the comparison.

E.1 Expected Evidence and Evidence-Based Reasoning

Both the original unmanipulated source video and other videos depicting the same event or scene can serve as valid evidence. The former enables direct comparison with the manipulated input and usually reveals the specific content that has been altered. The latter provides complementary information or alternative viewpoints that can expose semantic inconsistencies even when the exact source video is unavailable.

Our manual inspection of successful GPT-5.5 cases shows that most detections rely on retrieving the original source video or a relevant segment thereof, while a smaller proportion use other videos of the same event or scene. This pattern is consistent across AI Generation, Single-Source Editing, and Multi-Source Editing, as the benchmark manipulations are constructed from real source materials that generally remain publicly accessible online.

In successful cases, the model typically follows a coherent evidence-based reasoning process: it identifies a potentially inconsistent element, formulates a targeted search query, retrieves relevant evidence, and compares the input with the evidence to identify the specific manipulation. When this process fails, the cause can arise at different stages, including ineffective query formulation, irrelevant retrieval results, or hallucinated conclusions during evidence comparison. These failure modes are further examined in the error analysis in the main

Model	T
GPT-5.5	0.0
GPT-5.4	0.0
GPT-5.4-mini	0.0
Claude Opus 4.6	0.0
Claude Sonnet 4.6	0.0
Gemini-3.1-Pro	0.0
Gemini-3-Flash	0.0
Qwen3-VL-235B-A22B-Instruct	0.0
Qwen-3.5-Plus	0.0

Table 9: Models under test and decoding temperature. All models share the video-input settings described below.

text.

E.2 Stage 1: Chain-of-Thought Analysis

Inputs: 64 uniformly sampled forged-video frames.

Chain-of-Thought Retrieval Planning

You are a video forensic analyst. You will see 64 frames uniformly sampled from a video that may contain misinformation.
Your task is to deeply understand the video content, identify key entities and potential logical anomalies, and produce the initial retrieval plan for finding original or closely related source videos on YouTube.
Think step by step (write your full reasoning in the ‘reasoning‘ field):
== Step 1: Content Understanding ==
Examine the 64 frames as a timeline:
a) What different scenes, environments, or camera setups appear in the video? Briefly describe each one.
b) Identify the key entities and information in the video: - People: Who appears? Can you identify specific individuals such as celebrities, athletes, presenters, politicians, or creators? - Locations: Are there identifiable landmarks, venue cues, signs, architectural styles, neighborhoods, or countries? - Events: What is happening? A sports match, program segment, performance, interview, protest, news event, travel clip, everyday scene? - Time: Can you infer dates, seasons, eras, or broadcast periods? - On-screen text: What readable text seems semantically meaningful for identifying the content, such as names, event titles, program logos, scoreboards, venue text, headlines, or captions that appear integral to the footage? Ignore obvious platform chrome like playback controls, timestamps, or menus.
c) Are there discontinuities between scenes, such as sudden changes in environment, production style, resolution, or topic, suggesting multiple original sources?
== Step 2: Logical Reasoning ==
Combine visual information, text information, and common sense to look for potential anomalies or verification targets. Consider ALL of the following:
NARRATIVE-LEVEL ANOMALIES: - Are the facts claimed or implied by the video internally consistent? - Are there unreasonable links between people, locations, times, or events across scenes? - Are there scenes that look edited together from different sources?
AI-GENERATION ANOMALIES: - Do any objects appear visually inconsistent with their surroundings (wrong lighting, unnatural texture, mismatched brand/style for the context)? - Are there objects or products that seem contextually out of place (e.g., a branded product appearing in a setting where it would not naturally exist)? - Do any faces look inconsistent with the person’s body, clothing, or prior appearance in other frames? - Are there segments where visual quality, motion dynamics, or rendering style suddenly differs from the rest?
VERIFICATION TARGET: - Is there a specific claim, identity, location, or event that is most worth verifying first? - If there is no obvious contradiction, explain what the most valuable verification target would be. - If AI-generation is suspected, note which elements may be synthetic so they are NOT used as retrieval anchors.
== Step 3: Retrieval Planning ==
Based on Step 1 and Step 2, determine the single best initial search query to begin retrieval:

- What is the most promising verification target to search for first, and why? - Estimate how many distinct original sources are likely present (this informs later rounds, but you only need ONE query now). - Produce exactly ONE high-value YouTube search query of 5-12 words targeting the most verifiable and specific element you identified. Query strategy: - Prioritize specific people, event names, program names, competition names, locations, dates, and distinctive scene cues. - If a named person + event is identifiable, combine them directly. - If a program, broadcast, or competition is identifiable, use that as a strong anchor. - If no specific entity is reliable, use the content category plus the most distinctive visual cue. - The query should sound like a plausible YouTube title fragment or high-value search phrase, not a forensic note. - Do NOT generate multiple queries. Subsequent queries will be generated adaptively based on retrieval results from this first attempt. OUTPUT JSON FORMAT:

```
{
  "reasoning": "Step 1: <content
      understanding>. Step 2: <logical
      reasoning>. Step 3: <why this query
       is the best starting point>.",
  "entities": {
    "people": ["identified people"],
    "locations": ["identified locations"],
    "events": ["identified events/programs
        /competitions"],
    "text_claims": ["key facts or labels
        conveyed by on-screen text"]
  },
  "logical_analysis": "Logical analysis
      of the video content, including
      contradictions found or the main
      fact that should be verified first
      .",
  "physical_observations": "Summarize the
       most searchable visual clues in
      3-5 sentences.",
  "search_intent": "What this retrieval
      should verify or reconstruct first
      .",
  "estimated_sources": 1,
  "source_queries": [
    {
      "source_label": "source_1: brief
          description of the most
          promising verification target",
      "queries": ["single best first query
           (5-12 words)"]
    }
  ]
}
```

E.3 Stage 2: Coarse Relevance Filter

Inputs: 16 forged + 16 candidate frames (Group A first, then Group B).

Coarse Relevance Filter

You are a coarse filter in a video retrieval pipeline.
You will see TWO frame groups: - Group A: frames from the SUSPECTED FORGERY video. - Group B: frames from a CANDIDATE YouTube video.
CONTEXT FROM PRIOR COT ANALYSIS: - Physical observations: {physical_observations} - Logical analysis: {logical_analysis} - Search intent: {search_intent}
YOUR TASK: judge whether Group B is useful retrieval evidence for Group A. The candidate can be: - the same original footage, - a closely related upload of the same event or scene family, - or a video that would materially help verify the main fact or claim in Group A.
CRITICAL EVALUATION RULES: 1. This is a COARSE FILTER with HIGH RECALL. Prefer false positives over false negatives. 2. Do NOT rely on pixel-level similarity. 3. Ignore differences caused by compression, bitrate, resolution, color grading, subtitles, black bars, crop, mirror, small timing offset, and camera/viewpoint variation. 4. Focus on semantic anchors: event type, named entities, venue/program cues, recurring objects, role relations, action pattern, and storyline. 5. If Group B could plausibly help verify the same identity, event, location, or claim as Group A, mark relevant. 6. Only mark irrelevant when topics/scenes are clearly different.
OUTPUT JSON FORMAT:

```
{
  "reasoning": "Brief semantic comparison
       between Group A and Group B.",
  "is_relevant": true
}
```

E.4 Stage 2: Fine-Grained Forgery Analysis

Inputs: 64 forged + 64 candidate frames (Group A first, then Group B).

Fine-Grained Forgery Analysis

You are a forensic video analyst comparing: - Group A: frames from a SUSPECTED FORGERY video - Group B: frames from a CANDIDATE source video
PRIOR CONTEXT: - Physical observations: {physical_observations} - Logical analysis: {logical_analysis} - Search intent: {search_intent} - Entities summary: {entities_summary}
YOUR TASK: Perform a detailed cross-group comparison and extract forgery points that describe WHAT IS FALSE in Group A, using Group B as ground-truth evidence.
Forgeries fall into three broad categories — consider ALL of them:
A) AI GENERATION: faces replaced by deepfake (identity swap), objects added/removed/replaced by generative editing, or contiguous AI-generated segments inserted into real footage. Clues: identity mismatch (same body/background but different face), objects that exist in A but not in B (or vice versa), segments in A with no corresponding content in B that show telltale synthetic continuity.
B) SINGLE-SOURCE EDITING: the same source footage is editorially altered — scenes selectively omitted to change meaning, temporal order reordered to flip causality, or shots rearranged to create a misleading emotional impression.
C) MULTI-SOURCE SPLICING: footage from Group B is combined with footage from other unrelated sources to fabricate a narrative, distort the severity of an event, or place real footage into a false context (wrong time, location, or attribution).
THINK IN THIS ORDER (write reasoning in 'source_description'): 1. IDENTITY & OBJECT ALIGNMENT: Compare the people, faces, and key objects between Group A and Group B. Are the same individuals present? Are faces consistent? Are objects added, removed, or replaced? 2. TEMPORAL & STRUCTURAL ALIGNMENT: Do the scenes in Group A appear in the same order and completeness as in Group B? Are segments missing, reordered, or inserted? 3. CONTEXTUAL ALIGNMENT: Does

Group A preserve the original context (time, location, event, attribution) shown in Group B, or does it reframe the footage into a different narrative? 4. SYNTHESIS: What is the overall false impression that Group A creates, and which category (AI generation / single-source editing / multi-source splicing) best describes the manipulation? 5. DEDUPLICATE: each point must cover a unique manipulation aspect.
OUTPUT JSON FORMAT (strict, no extra text, NO code fences):

```
{
  "source_description": "1-2 sentences:
      semantic relation between Group A
      and Group B plus the primary
      forgery mechanism identified.",
  "points": [
    {
      "description": "Forgery point: what
          Group B shows as ground truth,
          what differs in Group A, and
          the false impression this
          creates."
    }
  ]
}
```

CONSTRAINTS: 1. ‘points‘ should contain 3 to 5 entries when evidence supports it. 2. No duplicate points with paraphrased wording. 3. Ignore pure quality/viewpoint/-compression differences unless they support a manipulation claim. 4. At least one point should provide a high-level synthesis of the overall manipulation (not just one local edit or one replaced element). 5. For identity swaps: describe the original person’s appearance in Group B (hair color/style, clothing, build, accessories) and state that Group A replaces their face with a different person’s face via AI. Do this for EACH person whose face is swapped. 6. For object manipulation: describe the original object in Group B (what it is, where it appears, its role in the scene) and state what it has been replaced with or altered into in Group A. Do this for EACH manipulated object. 7. Output VALID JSON only. No markdown fences, no commentary.

E.5 Stage 2: Sufficiency Judgment and Next Query

Inputs: text-only context fields; no visual input.

Sufficiency Judgment and Next Query

You are the decision-maker in a deep-search loop for video forgery analysis. You have been searching YouTube for original source videos and collecting forgery evidence.
CONTEXT (do NOT echo back):
- Round status: {round_status}
— What the video looks like (from COT analysis of the input video): - Physical observations: {physical_observations} - Logical analysis: {logical_analysis} - Search intent: {search_intent} - Entities summary: {entities_summary}
— Estimated source families / scene groups: {source_descriptions}
— Forgery evidence collected so far: {collected_points}
— Videos examined so far: {examined_videos}
— Previous search queries tried: {prev_queries}
YOUR JOB (two tasks in one response):
TASK 1 — SUFFICIENCY JUDGMENT: Compare what you KNOW about the input video against what you’ve COLLECTED. Judge whether the evidence is SUFFICIENT.
THINK STEP BY STEP for sufficiency: 1. Based on the source descriptions, how many distinct source families or scenes likely matter? Have we found evidence for each? 2. Based on physical_observations and logical_analysis, what identities, events, locations, or claims must be verified? Do the collected points cover them? 3. Does the examined evidence explain the main misleading interpretation or unresolved factual question? 4. Is there any major scene, claim, or verification target that NO collected point covers?
CRITICAL RULES: - For a SINGLE dominant source: do NOT stop merely because you found a topically related video. The examined evidence should strongly suggest the same program, same event, same source family, or a near-duplicate upload of the same underlying footage. - If the current evidence mainly highlights that Group A and Group B are from different eras, different shows, different productions, or different

contexts, that is evidence of topical relatedness, NOT sufficient source resolution. Keep searching. - For MULTI-SOURCE or MULTI-SCENE cases: you need evidence that covers EACH major source family or unresolved scene cluster. - Prefer precision over early stopping. It is acceptable to continue searching when the current best evidence is only "related but not same-source enough." - If the round status says this is the FINAL round, you MUST set 'is_sufficient=true', leave 'missing_description' and 'next_keyword' empty, and do not request another search.
TASK 2 — NEXT KEYWORD (only if NOT sufficient): Generate the SINGLE best YouTube search query to find the next missing original source.
THINK STEP BY STEP for keyword generation: 1. Which source family, identity, event, location, or claim is still unresolved? 2. What specific distinguishing cue would most likely retrieve useful evidence for it? 3. Generate ONE search query (5-12 words) using the strongest identifying anchor available. 4. Prefer a query that could plausibly match a real YouTube title or high-value search phrase.
OUTPUT JSON FORMAT:

```
{
  "reasoning": "Step-by-step: what the
       forgery contains vs what evidence
       is collected, whether the current
       evidence is same-source enough, and
        what's still missing.",
  "is_sufficient": true | false,
  "missing_description": "If insufficient,
        describe exactly what evidence is
       missing. If sufficient, empty
       string.",
  "next_keyword": "If insufficient, ONE
       best search query (5-12 words). If
       sufficient, empty string."
}
```

E.6 Stage 2: Query Refinement (Reflect)

Inputs: 64 forged-video frames; triggered when search, download, or coarse filtering fails.

Query Refinement (Reflect)

You are a video retrieval strategist working inside an iterative search loop. Previous YouTube queries did NOT find a useful source video.
CONTEXT OF FAILURE: - Previously tried queries: {prev_queries} - WRONG candidate titles returned by those queries: {candidate_titles}
PRIOR COT ANALYSIS OF THE INPUT VIDEO: - Physical observations: {physical_observations} - Logical analysis: {logical_analysis} - Search intent: {search_intent} - Entities summary: {entities_summary}
YOUR JOB: look at the input frames again, study the failed candidate titles, diagnose why retrieval drifted, and produce ONE stronger follow-up query.
THINK STEP BY STEP (write your full reasoning in the 'reasoning' field): Step 1. Re-understand the input video from the frames: - What are the most diagnostic anchors: named people, program names, event names, competition names, locations, dates, logos, uniforms, scene types, or specific actions? - Which of those anchors are most likely to appear in the original uploader's title or in strong retrieval results?
Step 2. Diagnose the failure mode: - Did previous queries over-focus on the wrong person, event, program, location, or date? - Were they too broad, too narrow, or anchored on a side detail instead of the main verifiable fact? - Did the wrong candidate titles suggest a different topic cluster than what the input frames actually show?
Step 3. Produce 'negative_keywords' conservatively: - Add ONLY specific entities, titles, locations, dates, or topic anchors that clearly pulled retrieval into the wrong result cluster. - Do NOT add broad genre words that could still belong to the right title. - When unsure, leave a term out. It is better to keep a reusable word than to block a potentially correct one.
Step 4. Generate 1 fresh search query: - Choose the single best angle that is MOST DIFFERENT from all previously tried queries. - You are free to use ANY retrieval angle: entity/event/title, location/-date/program, visual cue, content category, or any combination. - The query must

NOT repeat the same approach that already failed.
QUERY RULES: - The query must be plain YouTube search text, 5-12 words. - It should sound like a plausible title fragment or high-value search phrase, not a forensic note. - Avoid reusing any term in ‘negative_keywords‘ unless it is part of an explicit exclude operator like ‘-term‘.
OUTPUT JSON FORMAT:

```
{
  "reasoning": "Step 1: <what the video
      most likely shows>. Step 2: <why
      previous searches failed>. Step 3:
      <which specific anchors were
      treated as negatives>. Step 4: <how
       the new query improves retrieval
      >.",
  "reflection": "Short explanation of
      what retrieval signal was re-
      prioritized.",
  "negative_keywords": ["
      wrong_entity_or_topic_1", "
      wrong_entity_or_topic_2"],
  "new_queries": ["single best query"]
}
```

F Evaluation Protocol Details

Table 8 summarizes evaluation settings. Three text-only LLM judges independently compare each video's 3–5 predicted points with 3–5 ground-truth points using one-to-one semantic matching; each GT point is marked correct if at least two judges agree. The exact judge prompt is given below.

F.1 Task Input, Output, and Metric Definitions

Input. For each input video, the evaluated model receives 64 uniformly sampled visual frames. The audio channel is not provided to the model. Subtitles, captions, and other text embedded directly in the frames are retained because they form part of the visual content and may provide evidence relevant to the manipulation. The coarse relevance filter separately uses 16 frames from the input and candidate videos, as described in Appendix E.

Output. The expected output is a list of natural-language forgery-point descriptions, with each point describing one specific manipulation identified in the input video. For example, a point may state that “the face of person A is replaced with the face of person B,” or that “clips from two unrelated events are spliced together to fabricate a causal relationship.” The model is not required to predict the manipulation category or the temporal boundaries of the manipulated segment. The task focuses on explaining what information has been manipulated rather than assigning a predefined class or localizing an anomalous interval.

Metric definitions. For video i, let g_i denote the number of ground-truth forgery points, where $g_i \in \{3, 4, 5\}$, and let m_i denote the number of ground-truth points matched by the model output. Predicted and ground-truth points are matched through a one-to-one assignment using the LLM-as-judge protocol described below. Each ground-truth point can be matched at most once.

Point-level accuracy measures the proportion of ground-truth forgery points successfully matched across the benchmark:

$$\mathrm{PointAcc} = \frac{\sum_i m_i}{\sum_i g_i}.$$

Video-level accuracy requires all ground-truth points of a video to be matched:

$$\mathrm{VideoAcc} = \frac{1}{N} \sum_i \mathbf{1}[m_i = g_i].$$

Thus, point-level accuracy measures partial coverage of the annotated manipulations, whereas video-level accuracy measures whether the model provides a complete explanation of the manipulated video. Human validation of the LLM-based matching protocol is provided in Appendix G.

F.2 Stage 3: LLM-as-Judge Scoring

Inputs: numbered GT and predicted forgery-point lists (text only; no visual input).

LLM-as-Judge Forgery-Point Scoring

You are a strict but fair evaluator for video forgery analysis. You will be given a ground-truth (GT) list of 3-5 forgery points and a predicted list of 3-5 forgery points. For each GT point, decide whether any prediction matches it.
Important framing: - GT points and predictions may be written at different abstraction levels. - Some items are narrative-level accusations: the edit constructs a false moral, political, social, or causal interpretation. - Some items are event-level findings: the edit splices, reorders, re-contextualizes, or rebinds specific scenes, audio, or actions.

- A good prediction may express the same core misleading point either directly at the narrative level or indirectly through concrete event evidence.
Scoring rules (important; follow strictly): 1. Each GT point can be matched at most once. Use a one-to-one best assignment. 2. Evaluate every GT/pred pair on two dimensions: - Dimension A: manipulation method. Examples: splicing unrelated clips, reordering scenes, removing narration, binding unrelated audio, selective omission, context switching. - Dimension B: misleading point / false reframing. Examples: portraying someone as rule-breaking, constructing a negative personality arc, implying opportunistic recruitment, or creating a moral-judgment framing. 3. Dimension B is primary, but judge it flexibly across abstraction levels: - Direct narrative match: If the prediction states substantially the same false reframing or accusation as the GT, mark 'matched_dim_misleading=true' and 'verdict=1'. - Event-to-narrative support: If the GT is narrative-level, and the prediction identifies the concrete edited event(s) that materially support that same narrative reframing, you may still mark 'matched_dim_misleading=true' and 'verdict=1' even if the prediction is more concrete and less interpretive. 4. "Roughly consistent" allows paraphrase, synonymy, wording differences, partial compression of multiple scenes into one point, and different abstraction levels, as long as the core misleading implication is aligned. 5. Be conservative about over-crediting: - Do not give credit just because both items mention the same event. - Do not give credit just because both items describe editing manipulation in general. - Do give credit when the prediction captures the same accusation, reframing, or a concrete event-level mechanism that clearly underpins that accusation. 6. Report 'matched_dim_method' truthfully even when 'verdict=1' is driven mainly by Dimension B. 7. Every GT point should receive the best available predicted match. If multiple GT points compete for the same prediction, choose the assignment that gives the best overall scoring. 8. Output strict JSON only. Do not use markdown fences.
GT (3-5 points): {gt_block}
PREDICTION (3-5 points): {pred_block}
Output JSON:

```
{
  "hits": <integer from 0 to
      number_of_gt_points>,
  "score": <hits / number_of_gt_points as
       a float>,
  "matches": [
    {
      "gt_idx": <0-based GT index>,
      "pred_idx": <0-based prediction
          index or null if no match>,
      "verdict": <0|1>,
      "matched_dim_method": <true|false>,
      "matched_dim_misleading": <true|
          false>,
      "reason": "One concise English
          sentence explaining whether the
           method matches, whether the
          misleading point matches, and
          the key mismatch if any."
    }
  ],
  "comment": "1-2 concise English
      sentences summarizing the biggest
      strength and biggest weakness of
      the predictions."
}
```

G Human Evaluation Agreement

To check whether LLM-as-judge scoring tracks human judgment, we sample 50 EVID-Bench videos stratified across the nine task types (182 ground-truth points) and evaluate predictions from GPT-5.5 and Qwen-3.5-Plus. Three annotators independently label each ground-truth point using the same matching rubric as the LLM judges; labels are merged by majority vote.

Figure 6 compares these human labels with the LLM majority votes used in the main experiments. Agreement is close for both models.

Overall agreement reaches 94.51% for GPT-5.5 predictions and 90.66% for Qwen-3.5-Plus predictions. Cohen's κ is 0.886 and 0.811, respectively, indicating strong agreement between human and LLM-based judgments.

Our judge prompt requires one-to-one matching with a specific manipulation mechanism. A generic statement that a video is misleading or manipulated is not sufficient unless it identifies the corresponding identity, object, temporal, narrative, or contextual modification. As shown in Figure 6,

Category	Point Acc.	Video Acc.
AI Generation	64.81	44.44
Single-Source Editing	61.97	38.89
Multi-Source Editing	48.00	27.78
Overall	**57.50**	**37.04**

Table 10: Human performance (%) on the 54-video subset.

false positives—cases judged correct by the LLM majority but incorrect by the human majority—occur in only 4 of 182 GPT-5.5 judgments and 8 of 182 Qwen-3.5-Plus judgments. The corresponding false-negative counts are 6 and 9. This disagreement analysis suggests that over-crediting generic narrative falsehoods is uncommon under our evaluation protocol.

H Additional Validation Experiments

H.1 Human Performance Baseline

Three computer science researchers with video-editing experience independently evaluated 54 EVID-Bench videos, with six videos sampled from each of the nine manipulation types. Participants were allowed to use web search to locate original or related video evidence.

For video i, let h_i denote the number of matched ground-truth forgery points and g_i denote the total number of ground-truth points. Point-level accuracy is computed over all ground-truth points, while a video is counted as correct only when all of its ground-truth points are identified:

$$\begin{aligned} \text{PointAcc} &= \frac{\sum_i h_i}{\sum_i g_i}, \\ \text{VideoAcc} &= \frac{1}{N}\sum_i \mathbf{1}[h_i = g_i]. \end{aligned}$$

H.2 Authentic-Video Evaluation

We evaluate over-flagging using an auxiliary set of 60 authentic, unmanipulated videos, with 10 videos from each of the six topic domains. A prediction is counted as correct when the model produces no forgery points. This metric measures the proportion of authentic videos that are correctly left unflagged.

Claude Opus 4.6 and Claude Sonnet 4.6 both achieve 95.0% accuracy on authentic videos, indicating a low over-flagging rate. Gemini-3.1-Pro achieves 83.3%, with more false positives on Reviews and other casually edited videos. Across models, Daily Life and Reviews are more prone to false positives, suggesting that routine cuts, transitions, and other benign editing patterns may be mistaken for intentional manipulation.

H.3 Forensic Detector Baselines

AIDE and UnivFD produce frame-level fake probabilities. We uniformly sample 64 frames from each video and aggregate frame-level predictions using three rules: mean probability at least 0.5, at least half of the frames at or above 0.5, and any frame at or above 0.5.

Under the mean and majority aggregation rules, both detectors achieve near-zero detection rates across the three manipulation categories. The max rule produces higher detection rates, but also yields high false-positive rates on authentic videos: AIDE flags 55.0% of authentic videos, while UnivFD flags 78.3%. Consequently, the increased sensitivity of the max rule does not translate into reliable discrimination between manipulated and authentic content.

These results provide a visual-only forensic sanity check rather than evidence that all visual detection is impossible. They show that the two tested detectors, which primarily target pixel-level generative artifacts, do not reliably detect the evidence-dependent semantic manipulations in EVID-Bench. We therefore scope our conclusion to current forensic detectors and frontier VLMs evaluated under the reported protocols.

H.4 Reverse Image Search Baseline

We evaluate whether reverse image search can retrieve the original source videos independently of the downstream forgery analysis. For each input video, we uniformly sample 16 frames and submit each frame independently to Google Lens. We collect the top-10 results for each frame, yielding up to 160 candidate URLs per video after deduplication. Retrieval is considered successful if at least one returned URL corresponds to the original source video. For comparison, the iterative text-retrieval pipeline examines at most six candidate videos per input.

Despite examining substantially more candidate URLs, Google Lens achieves 26% source recall, compared with up to 60% for iterative text retrieval. Reverse image search succeeds when a query frame matches an indexed thumbnail, cover image, or other web-visible visual content, but arbitrary internal video frames are not reliably indexed as searchable video evidence. In contrast, text retrieval uses

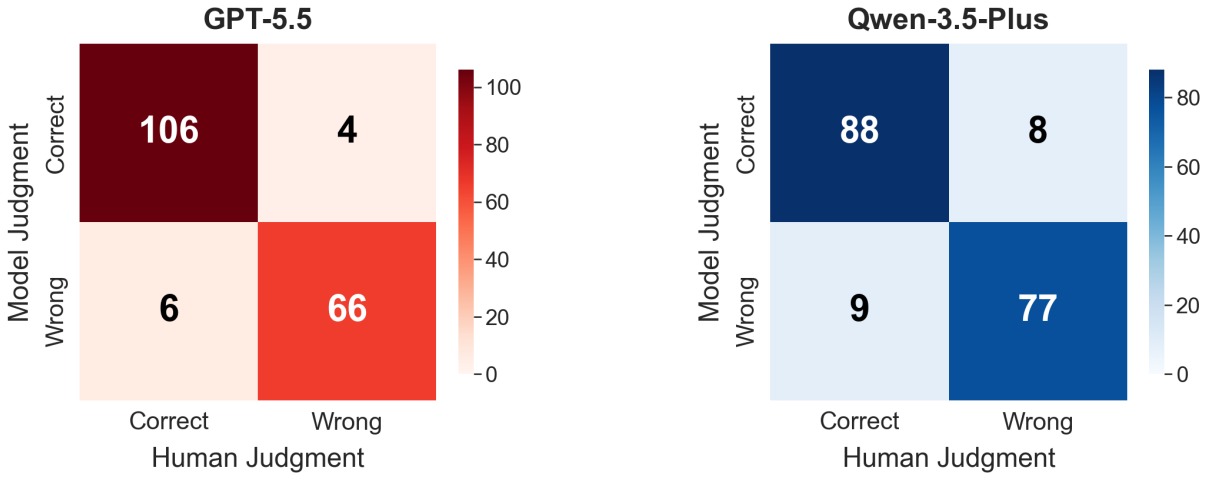


Figure 6: Point-level confusion matrices comparing LLM majority-vote labels with human majority labels on 182 forgery points from 50 stratified videos. **Left:** GPT-5.5 predictions. **Right:** Qwen-3.5-Plus predictions.

Topic	Gemini-3.1-Pro	Claude Opus 4.6	Claude Sonnet 4.6
On-Scene Footage	80	100	100
Sports	80	100	100
Performance/Talent	90	100	100
Education/Skills	100	100	100
Daily Life	80	80	90
Reviews	70	90	80
Overall	83.3	95.0	95.0

Table 11: Accuracy (%) on the 60 authentic videos. A video is counted as correct when the model produces no forgery points.

identifiable entities, events, and locations to search video titles, descriptions, and associated metadata. These results suggest that reverse image search is useful as a complementary retrieval channel but is insufficient as the primary retrieval method for EVID-Bench.

I Ablation Study (Qwen-3.5-Plus)

The main text reports the GPT-5.5 ablation in Figure 5. Figure 7 presents the corresponding breakdown for Qwen-3.5-Plus using the same layout: three hyperparameters (search rounds, frame resolution, and sampled frames), split by video- and point-level accuracy across task types and topic categories. Qwen-3.5-Plus follows the same overall trends as GPT-5.5, peaking at 6 rounds, 480p, and 64 frames, though with lower accuracy across most cells.

J Case Study

J.1 Excess Search Rounds Introduce Irrelevant Evidence

Case Study: Excess search rounds introduce irrelevant evidence



Overview. Increasing the search-round budget from 6 to 8 does not simply add more supporting evidence; it can retrieve a second, loosely related video that reframes the entire analysis. The model then mixes correct observations from the original source with spurious comparisons drawn from the wrong candidate, diluting the true forgery explanation.

Task Description. We use a CPR/AED training video in which three instructors' faces are replaced with AI-generated identities while the under-

Detector	Category	Mean	Majority	Max
AIDE	AI Generation	4.2	4.2	70.8
AIDE	Single-Source Editing	0.0	0.0	58.7
AIDE	Multi-Source Editing	0.0	0.0	60.0
AIDE	Authentic	0.0	1.7	55.0
UnivFD	AI Generation	0.0	0.0	69.4
UnivFD	Single-Source Editing	0.0	0.0	81.3
UnivFD	Multi-Source Editing	0.0	0.0	92.0
UnivFD	Authentic	0.0	0.0	78.3

Table 12: Video-level detection rates (%) under three frame-level aggregation rules.

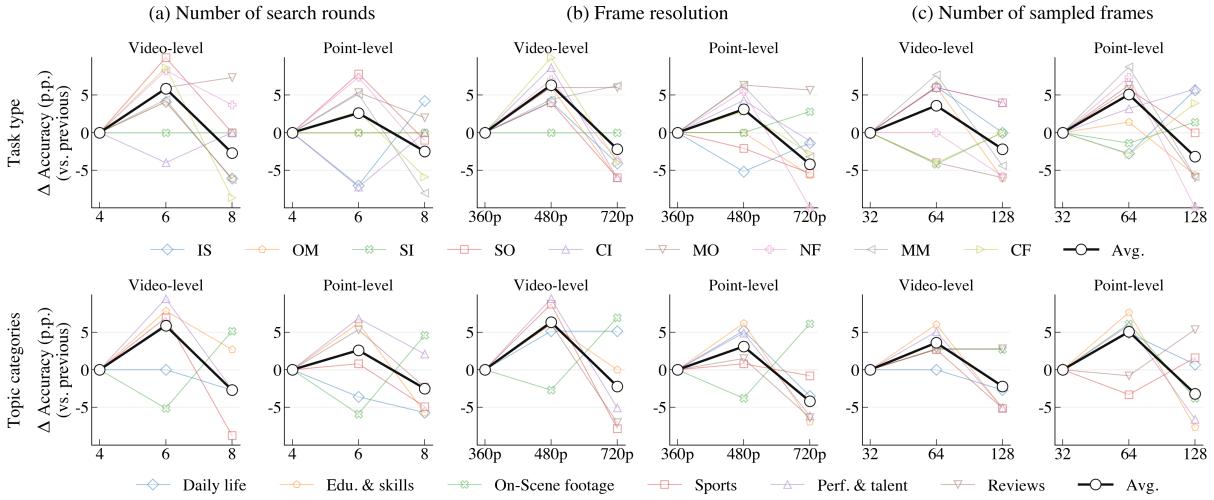


Figure 7: Same ablation layout as Figure 5 for Qwen-3.5-Plus. Trends match GPT-5.5, peaking at 6 rounds, 480p, and 64 frames. Curves plot stepwise accuracy change in percentage points: $\Delta_1 = 0$ at the leftmost setting and $\Delta_j = \mathrm{Acc}_j - \mathrm{Acc}_{j-1}$ for $j \geq 2$.

Method	Budget	Recall
Google Lens (16× top-10)	$\leq$ 160 URLs	26%
Claude Opus 4.6	$\leq$ 6 videos	54%
Claude Sonnet 4.6	$\leq$ 6 videos	52%
GPT-5.4	$\leq$ 6 videos	52%
GPT-5.5	$\leq$ 6 videos	60%
Gemini-3-Flash	$\leq$ 6 videos	30%
Gemini-3.1-Pro	$\leq$ 6 videos	32%
Qwen-3.5-Plus	$\leq$ 6 videos	31%
Qwen3-VL-235B-A22B-Instruct	$\leq$ 6 videos	23%

Table 13: Source recall of reverse-image and iterative text-retrieval methods. Source recall measures whether the original source video is retrieved for an input video.

lying demonstration footage remains unchanged. The ground truth requires identifying these three face swaps:

> "The fake video replaces her face with another person's face using AI."
> "The fake video replaces his face with another person's face using AI."
> "The fake video replaces her face with another person's face using AI."

Model Behavior. With 6 search rounds (480p, 64 frames), the pipeline retrieves one evidence video—*CPR / AED Emergency Response Refresher*—and produces five forgery points centered on identity swap. It correctly describes, for example:

> "`Receptionist Face Swap: In Group B (Frame 65) ... In Group A (Frame 2), her face has been replaced with a different face via AI ...`"

> "`Narrator Face Swap: In Group B (Frame 113) ... In Group A (Frame 47), her face appears to be AI-generated or swapped ...`"

All three ground-truth points are matched (video-level score = 1.0).
With 8 search rounds under the same resolution and frame budget, the pipeline retains the original evidence video but additionally retrieves *How to Give CPR and Use an AED | First Aid Tips*, a generic CPR tutorial from a different provider. Four of the first five collected points now come from this unrelated video and reinterpret the case through low-level artifacts rather than identity swap:

> "`Group A displays distorted and wavy text on the mannequin shirts`

(e.g., ‘CALL 911 GET AN AED’ in Frame 58) ... hallmark artifacts of AI video generation ...”

“Group A demonstrates the outdated ‘ABC’ (Airway, Breathing, Compressions) protocol ... whereas Group B correctly follows the modern ‘CAB’ protocol ...”

“Group A presents a completely fabricated branding context (‘Action First Aid’) ... whereas Group B is authentically branded as the ‘Canadian Red Cross’ ...”

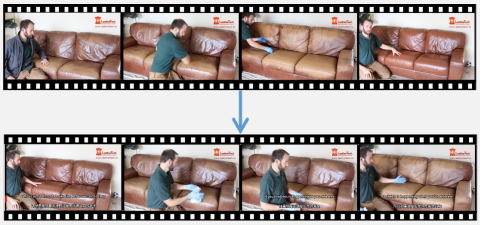
Only one of three ground-truth identity-swap points is recovered (video-level score = 0.33).

Result. The 8-round run does not fail because retrieval stops working—it fails because an extra round introduces a plausible but wrong comparison target. The model then explains the fake video as AI-generated instructional content or branding fraud against an unrelated Red Cross tutorial, rather than as face replacement on the original source.

Interpretation. This case shows that additional search rounds can hurt when sufficiency is judged too loosely. Once a second candidate enters the evidence pool, the model treats superficial instructional differences as primary forgery signals and underweights the actual manipulation mechanism. More retrieval is helpful only while it still converges on the correct source family.

J.2 Higher Resolution Distracts from Narrative Forgery

Case Study: Higher resolution distracts from narrative forgery



Overview. For editorial manipulations whose harm lies in narrative reframing rather than pixel artifacts, increasing frame resolution can paradoxically reduce accuracy. The model sees more local editing traces—subtitles, shot order, duplicated clips—but misses the higher-level causal reversal that the edit is designed to convey.

Task Description. We use a leather-sofa restoration video edited into a negative product review. The ground truth describes a four-step narrative reversal:

> “Reversing the original repair process ... into a ‘fail’ review claiming that ‘the cleaner ruined a perfectly good sofa’.”
> “Using the fully repaired sofa as the initial state ... fabricating the premise of the experiment.”
> “Editing the normal step of removing the old paint layer to look like a devastating consequence where ‘the paint peels off with a single wipe’.”
> “Finally, cutting back to the already damaged appearance ... presenting it as the ultimate proof of the damage caused by the cleaner.”

Model Behavior. Both 480p and 720p runs retrieve the same evidence video (*Leather Couch Cleaning and Restoration*). At 480p, the model explains the forgery at the narrative level:

> “Outcome Truncation: Group A ends the video immediately after the cleaning process reveals significant color loss ... omitting the successful restoration and final ‘After’ result shown in Group B.”

> “False Premise via Subtitles: Group A introduces the video with subtitles claiming ‘Today we will attempt to clean a sofa in good initial condition’ ... falsely setting up the video as a test on healthy leather.”

> “Misattribution of Damage: Group A includes subtitles stating ‘Because there is a problem with this cleaning agent’ ... implying the product caused the damage.”

All four ground-truth points are matched (video-level score = 1.0).
At 720p, with the same retrieved source, the model instead catalogs technical editing operations:

> “Group A contains burned-in English and Chinese subtitles throughout the footage ... whereas Group B contains no subtitles except for title cards and end screens.”

> “The sequence of the man in the grey jacket is reordered ... he holds the bottle early (frame 12) and gestures late (frame 59).”

> “A specific scene of the man in the green shirt pointing at the sofa ... is duplicated in Group A ... at the very beginning ... and again later.”

Only one of four ground-truth points is matched (video-level score = 0.25).

Result. Higher resolution exposes more implementational detail—added subtitles, duplicated shots, reordered demonstrations—but the model uses these as the main explanatory frame. It partially recognizes that the conclusion clip is moved to the start, yet fails to articulate the overarching false claim that a successful restoration was reframed as product-induced damage.

Interpretation. When the forgery mechanism is semantic rather than generative, extra visual fidelity can push the model toward an editor's-toolkit reading of the video. It becomes better at describing *how* clips were rearranged and worse at explaining *what false conclusion* the rearrangement is meant to sell.

J.3 Excess Sampled Frames Misdirect Cross-Video Reasoning

Case Study: Excess sampled frames misdirect cross-video reasoning



Overview. Uniform frame sampling captures temporal context, but too many frames can expose fine-grained visual differences that are irrelevant to the true manipulation. The model then reclassifies a cross-event montage as generative editing or attribute manipulation, even when retrieval has already found the correct source videos.

Task Description. We use a men's road-race video that splices Tokyo 2020 and Paris 2024 Olympic footage into one continuous chase narrative. The ground truth states:

> "Forcibly splicing cycling footage from two different events—the 2020 Tokyo Olympics and the 2024 Paris Olympics—into a single continuous pursuit."
> "By unifying the black borders, the remaining distance UI, and the color grading, the original differences between the different years and venues are erased."
> "The subtitles rewrite two independent championship-winning runs into a single race where Carapaz breaks away first and Evenepoel chases him down."

Model Behavior. With 64 uniformly sampled frames (480p, 6 rounds), the pipeline retrieves two evidence videos and explains the case primarily as multi-source splicing:

> "Environmental Mismatch: Group A displays mid-race footage ... featuring a winding road with yellow center lines and dense forest, which is the signature of the Tokyo 2020 Mount Fuji course."
>
> "Athlete/Action Splicing: Group A shows Richard Carapaz ... leading the peloton in the forest section ... archival footage from his gold-medal performance in Tokyo 2020."
>
> "Narrative Fabrication: By splicing the Tokyo mid-race footage between the Paris start and finish, Group A creates a false geographical narrative ..."

All three ground-truth points are matched (video-level score = 1.0).
When the frame budget is raised to 128 under the same resolution and round budget, retrieval remains unchanged, but the explanation shifts toward generative manipulation:

> "Generative Editing (Object/Attribute): In Group A, the cyclist identified as the leader/winner wears a light blue and yellow jersey ... In Group B, the actual winner Remco Evenepoel wears the black, yellow, and red Belgian national kit. The ... jersey colors were digitally altered."
>
> "Identity/Kit Mismatch: The video labels the leading cyclist as 'Efinepur' ... but the cyclist shown leading ... is wearing a light blue and yellow kit (characteristic of Kazakhstan)."

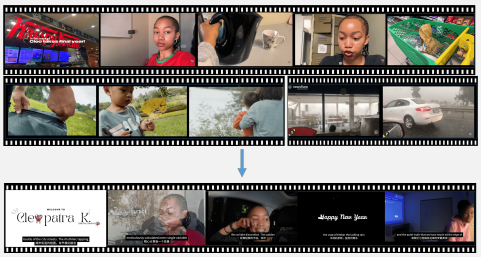
Only one of three ground-truth points is matched (video-level score = 0.33).

Result. The high-frame run still notices cross-event inconsistency, but it elevates jersey-color and labeling anomalies into a generative-editing story. This misattributes the forgery mechanism and misses the core ground-truth claims about event splicing, unified broadcast styling, and subtitle-driven narrative fusion.

Interpretation. Extra frames help when the model needs broader temporal coverage, but they can also surface micro-differences that invite the wrong explanatory template. Once the model commits to "digitally altered kits" or "AI-generated attributes," it underweights the simpler and more accurate account: authentic footage from two Olympics was stitched into one false race.

J.4 Attention Captured By Salient but Irrelevant Anchors

Case Study: Attention captured by salient but irrelevant anchors



Overview. This failure mode occurs when a highly salient keyword or title fragment captures the model's attention so strongly that it becomes the dominant organizing hypothesis for the rest of the analysis. Once this early anchor is formed, subsequent evidence is no longer evaluated neutrally; instead, it is forced into a narrative that is already partially decided. In our data, this pattern appears in both retrieval and interpretation: the model either overcommits to a named entity in the title and fails to decompose a multi-source composition, or over-focuses on a striking product label and misses the deeper editorial manipulation.

Task Description. We illustrate this failure pattern with two representative examples.
Example A. The ground truth states that:

> "The fake video splices together originally independent campus vlogs, family camping clips, and storm disaster footage across different videos, forging them into a single interconnected event within the same timeline." "Shots of campus daily life, strolling, and shopping are inserted between scenes of camping preparation and dinner, creating a false causal relationship of 'urban life gradually being engulfed by disaster'." "Extensive footage of dark clouds, heavy rain, flooding, and post-disaster scenes from another video is used as the main climax, misleading the audience into believing the protagonist actually experienced the impact of extreme weather." "Footage of tents, campfires, and packing up in the morning from the camping material is edited to appear as the post-storm sheltering and recovery phase, making ordinary outdoor life look like post-disaster survival." "Through dissolve transitions, cross-cutting, and a unified subtitle tone, the fake video conceals the fact that the footage comes from different sources, features different people, and takes place in different locations."

Example B. The ground truth states that:

> "The tampered video does not fabricate new footage, but by rearranging the sequence, it brings forward and concentrates the features, sample shots, and advantageous segments of the DJI Pocket 4, while significantly pushing back the content related to the iPhone 17 Pro Max." "This structural alteration weakens the more balanced back-and-forth comparison of the original video, making it easier for viewers to form a preconceived impression during the first half that the 'Pocket 4 is significantly superior'." "What was originally a two-way review centered around different scenario needs has been edited to look more like a one-sided recommendation featuring DJI as the protagonist, with the iPhone only holding a slight edge in convenience." "In other words, while the products themselves were not fabricated, the pacing of the comparison, the allocation of screen time, and the perceived conclusion have been artificially manipulated."

Model Behavior. In Example A, the retrieval trace shows that `gemini-3.1-pro-preview` correctly recognized early on that the input likely had *two sources* (`estimated_sources = 2`), yet it still centered the entire search around the title phrase

> `"Cleo takes final year"`

and kept querying variants such as

> `"Cleo final year uni vlog christian -Abram"`

, and

> `"Cleo takes final year week in my life -Abram"`

This means the model noticed compositing, but failed to operationalize that observation into multi-source retrieval. Instead, the title anchor dominated the search trajectory and prevented source decomposition. The final result contains no resolved source and no valid forgery points, despite the ground truth explicitly requiring cross-source reconstruction. In a related run, `qwen3.5-plus` was further distracted by the ending slogan

> `"PSA: JESUS LOVES YOU!"`

and produced predictions such as

> `"Branding Misappropriation"`

and

> `"AI-Generated Narrative,"`

reframing the video as a branding/deepfake problem rather than a disaster-narrative splice.
In Example B, `qwen3-vl-235b-a22b-instruct` over-focused on the visually salient phrase

> `"iPhone 17 Pro Max"`

. Its output repeatedly claimed that the fake video

> `"replaces the real iPhone with a fictional 'iPhone 17 Pro Max' "`

,

> `"fabricat[es] a false timeline"`

, and

> "synthesiz[es] a completely fabricated comparison through editorial insertion of non-existent product data."

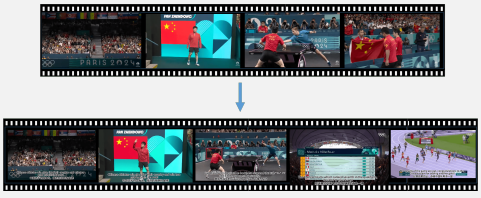
These observations are not random; they are coherent, but they are organized around the most conspicuous lexical cue rather than the actual manipulation target. The model successfully noticed striking labels and overlays, yet failed to identify that the underlying deception was a restructuring of comparison order, pacing, and emphasis between two real products.

Result. In both examples, the models miss the main forgery mechanism. The system fails to reconstruct the multi-source storm narrative and does not recover the key causal reframing from ordinary vlog and camping material into apparent disaster experience. The system misses the fact that no new core footage is fabricated; instead, it wrongly escalates the case into a fictional-device narrative and therefore overlooks the structural bias introduced by reordered evidence and uneven screen-time allocation.

Interpretation. These cases suggest a general failure mode of *attention capture by salient keywords*. Once a strong lexical anchor appears, such as a personal title (*Cleo*) or a sensational product label (*iPhone 17 Pro Max*), the model tends to treat that anchor as the explanatory center of the entire video. As a result, it underweights broader temporal structure, cross-source composition, and the actual persuasive goal of the edit. This is particularly damaging for fake-video detection, because many high-impact manipulations are not about the most eye-catching word or frame, but about how real footage is sequenced, juxtaposed, and narratively reinterpreted.

J.5 Inability to Recognize AI-generated Content

Case Study: Inability to recognize AI-generated content



Overview. A recurring failure mode in our pipeline is that models often detect that "something extra" has been added, but misidentify the nature of that addition. Instead of recognizing newly synthesized AI-generated shots, they prefer more familiar explanations such as cross-episode splicing, reuse of footage from other channels, local visual corruption, or ordinary instructional trimming. In other words, the models are often sensitive to inconsistency, but not to synthesis.

Task Description. We illustrate this pattern with four representative examples.

Example A: Fake reaction shots in a magic-performance video. The core manipulation is not simple re-editing, but the insertion of non-existent reaction shots to create new emotional emphasis.

> "The fake video adds an AI-generated facial close-up shot of audience members nervously and expectantly biting their fingers under dim blue lighting."

> "The fake video instead inserts an AI-generated reaction video of audience members excitedly waving their hands and cheering."

> "The fake video inserts an AI-generated shot near the end ... showing backstage staff with shocked expressions."

Example B: AI-generated lifestyle/product scenes in a laptop review. The fake video inserts synthetic scenes that were never part of the original review.

> "The fake video inserts AI-generated video footage of an appearance comparison between the M3 MacBook Air and M2 MacBook Air on a white desk background."

> "The fake video instead inserts an AI-generated close-up shot of a student in a well-lit library drinking coffee while opening the laptop screen with one hand."

> "The fake video adds an AI-generated working scene of a professional video editor earnestly editing and rendering video in a modern office."

Example C: AI-generated insertions overshadowed by local oddities in a sports montage. The main forgery is the insertion of synthetic shots, but these are masked by many smaller anomalies.

> "The fake video adds an AI-generated close-up shot of a staff member at the edge of the rink scraping ice shavings off with an ice blade."

> "The fake video instead inserts an AI-generated preparation shot of a young skater in the backstage waiting area."

> "The fake video adds an AI-generated close-up video of a massive crowd in the stands excitedly standing up and cheering."

Example D: Face replacement hidden behind instructional edits in a CPR tutorial. The real issue is identity manipulation, not trimming.

> "The fake video replaces her face with another person's face using AI."

> "The fake video replaces his face with another person's face using AI."

Model Behavior. In Example A, the model noticed inserted reaction footage, but interpreted it using the familiar logic of cross-source reuse rather than synthetic generation. It described the added material as:

> "likely spliced from the later ... performance episode"

and

> "reaction footage from unrelated moments or sources was added"

It also summarized the whole case as:

> "By splicing in reaction shots ... Group A constructs a false narrative that exaggerates the impact of the performance."

This is close at the level of "extra footage was inserted," but wrong about the mechanism: the inserted reactions were not retrieved from another real episode, but synthesized.
In Example B, the model again preferred a retrieval-style explanation. Instead of identifying synthetic scenes, it reconstructed the case as a tech-media mashup:

> "Group A splices in external benchmark graphics from 'Max Tech' ... and a segment featuring Linus Sebastian from 'Linus Tech Tips'."

It further concluded:

> "The overall manipulation in Group A is a multi-source splicing forgery."

Here the model would rather believe that the fake video borrowed footage from other real channels than accept that entirely new scenes had been generated.
In Example C, the model was drawn toward many salient local abnormalities, such as:

> "a blue t-shirt with a large cartoon elephant graphic"

> "the female skater ... has a mustache"

> "an exaggerated, shocked facial expression"

and

> "significantly truncated or cut-off text"

These observations are not fabricated by the model; they do point to suspicious artifacts. However, they pull attention toward scattered micro-anomalies and away from the more important structural fact that several entire shots were AI-generated insertions.
In Example D, the model failed in the opposite direction: instead of over-interpreting synthetic insertion as splicing, it over-prioritized instructional content and ignored the identity manipulation altogether. It framed the case as:

> "The overall manipulation ... is a form of single-source editing where the chronological sequence ... is selectively trimmed and re-captioned."

and emphasized omitted safety text such as:

> "Call 911 Get AED"

and

> "Be sure no one is touching the patient."

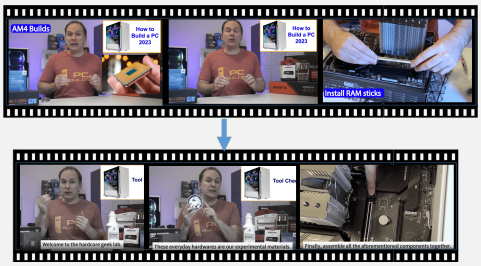
So the model did notice that the content had changed, but it attached that change to teaching completeness rather than face replacement.

Result. Across all four examples, the models consistently miss the main synthetic component of the forgery. They may correctly sense that a shot is foreign, that something looks odd, or that the narrative has been altered, but they often map that intuition onto a more familiar and easier explanation: footage borrowed from elsewhere, local artifact corruption, or ordinary editing. As a result, they fail to distinguish between *real footage used in the wrong place* and *new footage that never existed in the original world*.

Interpretation. This pattern suggests that current models are better at identifying inconsistency than at identifying synthesis. When the manipulation is editorial, they often have a ready-made explanatory template: splicing, omission, reordering, or overlay replacement. But when the manipulation involves AI-generated shots or AI-replaced faces, the models tend to collapse back into those same templates instead of recognizing generation as the primary mechanism. In practice, this means that even when the model is directionally aware that "something is wrong," it may still fail at the more important forensic question: whether the suspicious content was *moved from somewhere else* or *created from scratch*. This distinction matters because the latter is precisely the kind of synthetic evidence that many high-risk fake-video cases depend on.

J.6 Forgery Type Misattribution

Case Study: Forgery Type Misattribution



Overview. Another recurring error is that the model recognizes that a video has been manipulated, but assigns the wrong *type* of manipulation. Instead of distinguishing among identity replacement, object editing, synthetic insertion, contextual substitution, and editorial reframing, it often collapses multiple distinct mechanisms into a broader and safer label. In practice, this means the model can say that "the video was altered," while failing to specify *how* it was altered.

Task Description. We observe two representative sub-patterns.

Example A: Identity- and object-level forgeries are flattened into montage. In several classification cases, the ground-truth manipulation type is a specific low-level forgery mechanism such as:

> "identity_swap"

or

> "object_manipulation"

However, the model outputs the more generic label:

> "manipulative_montage"

Example B: Contextual fabrication is compressed into generic narrative rewriting. In another group of cases, the ground-truth manipulation type is:

> "contextual_fabrication"

but the model predicts:

> "narrative_fabrication"

These cases are not pure misses. The model often detects that the overall story or impression has changed. The problem is that it does not preserve the correct causal level of explanation.

Model Behavior. For Example A, both `gpt-5.4` and `claude-opus-4-6` repeatedly explain the evidence in terms of reordering, truncation, and pacing, even when the ground truth belongs to a more specific manipulation family. Typical justifications include:

> "The evidence indicates that a single source video was altered by shortening, repositioning, and rearranging its segments."

> "This re-pacing and reordering of shots creates a misleading narrative ... which perfectly matches a manipulative montage."

and

> "The evidence indicates that scenes from a single video were reordered, truncated, and looped ... This rearrangement of shots ... aligns perfectly with manipulative montage."

Even when the true category is identity replacement or object editing, the model gravitates toward an editorial explanation. In effect, it prefers a broad account in which the video was "re-cut" rather than a more precise account in which a face, object, or inserted visual element was directly manipulated.

For Example B, the model similarly detects that the final meaning has changed, but compresses a source- or context-level forgery into a general story-level rewrite. For instance, one model explains:

> "The manipulation combines authentic footage with unrelated clips ... to fabricate a completely false narrative."

Another writes:

> "Footage from multiple unrelated disaster events ... was spliced together ... to create a completely fabricated narrative."

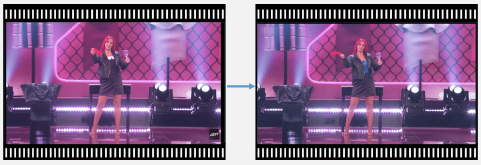
These explanations are not entirely wrong, but they remain too coarse. They correctly recognize that the resulting story is fabricated, yet they fail to preserve the more specific mechanism: the video does not merely tell a new story, but relocates footage into a false source, event, or interpretive context.

Result. The result is a systematic loss of forensic specificity. Distinct manipulation types that should remain analytically separate are absorbed into broader narrative or montage categories. Identity swap, object manipulation, and synthetic insertion become editorial montage; contextual substitution becomes generic narrative fabrication. This weakens the taxonomy because the model's output no longer distinguishes between *what was changed* and *what interpretive effect that change produced*.

Interpretation. This error suggests that the models are more comfortable describing the *effect* of manipulation than the *mechanism* of manipulation. They are relatively willing to say that a video has been made misleading, that its pacing has been altered, or that a false story has been constructed. But they are less reliable at assigning the manipulation to the correct forensic layer: identity, object, synthetic content, source attribution, or contextual framing. For a paper-level interpretation, this can be stated as follows: the model can often tell that "the story has changed," but it struggles to specify *which layer of reality was actually replaced*. That distinction is essential for fine-grained fake-video analysis, because two videos may produce similarly misleading narratives while relying on fundamentally different falsification mechanisms.

J.7 Insensitivity to Small-Scale Visual Modifications

Case Study: Insensitivity to small-scale visual modifications



Overview. A distinct failure mode in our analysis is that models are often insensitive to small-scale object edits. When the global identity of the person, the scene layout, and the rough event sequence remain stable, the model may conclude that the video is authentic even when local objects, textures, or fine visual attributes have been replaced. In other words, the model privileges coarse scene continuity over fine-grained visual integrity.

Task Description. We illustrate this pattern with two representative examples.

Example A: Fine-grained costume and texture replacement in a stage performance. The manipulation is subtle but concrete: local visual details are replaced even though the performer, stage, and sequence remain the same.

> "The fake video replaces this black inner lining with an AI-generated blue inner lining."

> "The fake video replaces this black skirt section with an AI-generated white skirt."

> "The fake video replaces the dress drawings within it with AI-generated alternative patterns, such as flowers."

Example B: Clear object replacement in a home-repair tutorial. Here the manipulation is even more explicit:

> "A broom was originally hanging on the wall; the fake video replaces it with an AI-generated large hammer."

> "The tool originally used by the vlogger is replaced in the fake video with an AI-generated wrench."

> "The fake video replaces the tool originally used in the vlogger's hand with an AI-generated wooden stick."

Model Behavior. In Example A, the model interpreted the case almost entirely through global consistency. It emphasized that:

> "The identity of the performer is consistent."

> "The objects used in the act are identical."

and

> "The temporal and structural alignment is perfect."

It then generalized this into an authenticity judgment:

> "Group A is a faithful representation of the original event."

This reasoning reveals the underlying bias: once the performer, setting, and performance flow match at a coarse level, local replacements in clothing regions and notebook content are discounted as negligible or missed entirely.

In Example B, the model makes the same type of mistake in an even stronger form. It states:

> "No object manipulation detected."

It further claims that:

> "all tools ... and set elements ... appear identically positioned and unchanged"

and finally concludes:

> "Group A is not forged; it is a faithful reproduction of Group B."

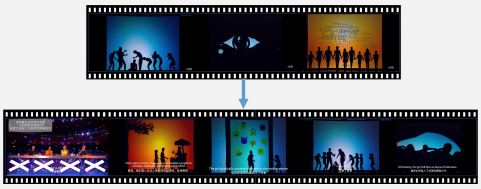
This is a particularly revealing error, because the manipulated objects are not merely abstract texture changes; they include ordinary, semantically meaningful tools that have been visibly replaced. Yet the model still allows global scene continuity to override local object evidence.

Result. In both examples, the model effectively treats local object replacement as if nothing important happened. If the same person appears in the same environment and the same broad sequence of actions is preserved, the system is prone to infer authenticity, even when individual objects, object parts, or surface patterns have been altered. This leads to hard false negatives for object-manipulation cases.

Interpretation. These cases suggest that the model's notion of visual consistency is strongly hierarchical: high-level continuity in person, place, and action tends to dominate low-level evidence from texture, accessory, or tool-level changes. As a result, the model underweights the forensic significance of small visual substitutions. For paper-level interpretation, this can be summarized as follows: the model is comparatively robust at matching *scene identity*, but weak at preserving *object fidelity*. This matters because many practical fake-video manipulations do not rewrite the whole scene; they only alter the small, local elements that carry crucial semantic meaning.

J.8 Factual Over Emotional Reasoning

Case Study: Factual over emotional reasoning



Overview. A further failure mode is that models are often better at detecting explicit factual falsification than emotional or affective reframing. When a fake video changes names, years, titles, or source claims, the model readily treats these as suspicious. But when the deeper manipulation lies in shifting the emotional arc of an event, such as turning warmth into tragedy or ordinary busyness into oppression, the model is much less reliable. In such cases, it may either ignore the emotional transformation or infer the wrong emotional direction altogether.

Task Description. We illustrate this pattern with two representative examples.

Example A: A warm performance is rewritten into a tragic social narrative. The key forgery is not primarily a factual mislabeling, but an emotional and narrative inversion of the performance:

> "Rearrange the originally warm dance performance into a tragedy of 'being addicted to smartphones, isolated, and ostracized'."

> "Delete the scenes of mutual help, confession, and redemption ... retaining only the shadowed stage and sad subtitles."

> "Change the sequence of events, turning the ending from a reunion into a complete breakdown."

Example B: An ordinary workplace vlog is rewritten into a story of oppressive overwork. The core issue is a change in affective framing:

> "Edited an originally normal daily workplace Vlog into a 'high-pressure corporate slave' short video depicting a chaotic company environment, endless meetings, and employees consumed by work."

> "By front-loading negative scenes and deleting explanations and lighthearted clips, it exaggerated what was originally just busyness ... into a continuous state of daily oppression."

> "It also deleted the footage of the blogger completing the work themselves and forcibly spliced in AI clips, creating the false impression that 'the work is almost entirely done by AI'."

Model Behavior. In Example A, the model became highly responsive to conspicuous textual cues such as the program name and year. It described the case as:

> "Group A replaces it with 'Prince Got Talent 2013' text ... falsely implying ... a 2013 contest that never occurred."

It further emphasized:

> "Group A inserts '2013' references ... as fabricated temporal misattribution."

and concluded that the main problem was:

> "a false impression of past success"

This shows that the model successfully tracked visible factual inconsistencies, but overlooked the more consequential manipulation: the performance's emotional trajectory had been transformed from warmth, reconciliation, and mutual help into isolation, sadness, and collapse.

In Example B, the model did recognize that the source vlog had been repackaged, but it inferred the direction of that repackaging incorrectly. It claimed that the edited video:

> "creates a false impression that the video is an official, polished ... production rather than a candid employee's personal project."

It also wrote that the fake version:

> "falsely portray[s] him as a confident, fully prepared professional"

and ultimately summarized the manipulation as:

> "repackages ... a raw, self-aware, and often humorous vlog into a faux-corporate promotional piece."

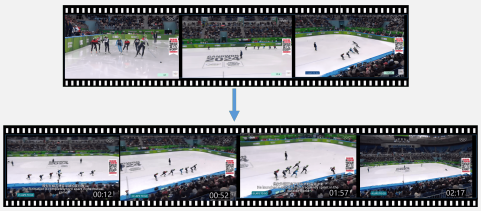
This is an especially revealing mistake. The model clearly perceived editorial reframing, but it reordered the polarity of the emotional change. The actual forgery darkened the vlog into a story of exhaustion, disorder, and exploitation, whereas the model interpreted it as a polishing or sanitizing transformation.

Result. In both examples, the model misses the principal emotional manipulation. In the performance case, it prioritizes factual anomalies in labels and dates over the altered emotional arc of the staged narrative. In the workplace-vlog case, it recognizes that the source has been re-edited, but misreads the affective direction of the rewrite. The result is that the model can often say that "the story was changed," but not whether it was made sadder, harsher, more oppressive, more redemptive, or emotionally reordered.

Interpretation. These cases suggest that models are more comfortable verifying *factual fabrication* than analyzing *emotional fabrication*. Names, dates, titles, logos, and explicit claims are discrete and easy to anchor. Emotional tone, however, is distributed across pacing, omission, juxtaposition, and the ordering of resolution versus breakdown. Because of this, the model tends to seize on surface-level false facts while underweighting changes in affective meaning. For paper-level interpretation, this can be summarized as follows: the model is relatively good at identifying that "some information is false," but much weaker at identifying that "the emotional truth of the event has been rewritten." That distinction is important because many persuasive fake videos do not rely on inventing facts alone; they rely on manufacturing a different feeling about the same footage.

J.9 Detail Fixation Over Narrative Comprehension

Case Study: Detail fixation over narrative comprehension



Overview. A recurring error in the pipeline is that the model performs substantial analysis, but along the wrong axis. Instead of identifying the main persuasive goal of the fake video, it becomes absorbed in secondary details such as branding, subtitles, interface text, graphics, identity markers, or local anomalies. As a result, the model does not fail because it sees too little; it fails because it explains the wrong problem. In many cases, the output is detailed, internally coherent, and even evidence-rich, but it is answering a different forensic question from the one that actually matters.

Task Description. We illustrate this pattern with five representative examples.

Example A: A positive product review is rewritten into a false negative causal story. The main manipulation is a reordering of timeline and causal structure:

> "By rearranging the timeline, rewrite the originally overall positive review ... into a negative case where 'the nutrient solution and App intervention directly ruined the plants'."

> "Move the later footage of healthy plant growth to the beginning ... thereby fabricating a false cause-and-effect relationship."

> "Ultimately distort the original review of 'having minor flaws but overall recommended' into portraying a failed product that is not worth buying."

Example B: A comparative tool review is rewritten so that the stronger brand becomes the loser.

> "By rearranging the sequence of the original video, Rigid, which actually performed stronger, is portrayed as the loser, while Makita is packaged as the champion of the entire review."

> "Ultimately, the video rewrites the originally objective multi-brand comparison into a false review result where Makita wins completely and Rigid fails."

Example C: A race video is re-edited to alter competition flow and outcome. The key issue is not the graphics themselves, but the rewritten competitive progression.

> "The fake video reorders the race footage and related commentary to alter the original competition flow and the perceived relationship between the athletes' performance and the final result."

> "The manipulation changes how viewers interpret who was leading, who was falling behind, and how the outcome emerged over time."

Example D: A review video is restructured into a false horizontal comparison of high-end prototypes. The core forgery lies in changing evaluation structure rather than inserting one obviously fake local detail.

> "The original video consists of several relatively independent hands-on experiences and judgments, but the fake video reorganizes them into a cross-model, side-by-side style comparison."

> "The manipulation turns separate impressions into the impression of a unified 'high-end prototype showdown'."

Example E: A ranking conclusion is manipulated, but the model focuses on parameters and translation issues instead.

> "By moving the positive summary of Makita from the end to before the test, it preemptively establishes a preconceived conclusion for the audience that 'Makita is the best'."

> "Then, by splicing Rigid's parameter introduction with descriptions of other failures, it creates a false cause-and-effect illusion of Rigid failing the test."

Model Behavior. In Example A, the model devoted most of its effort to source and authenticity markers rather than causal editing. It discussed:

> "shirts with the 'Epic' logo"

> "an AI-generated app interface with inconsistent or missing text"

and

> "key location identifiers"

It ultimately explained the video as creating

> "the false impression of a generic review rather than an official ... review."

This is a meaningful observation, but it is not the central manipulation. The real forgery is the construction of a false negative causal chain: healthy plants, nutrient intervention, and visible damage are rearranged to imply product failure.

In Example B, the model produced a highly elaborate explanation, but for the wrong case. Rather than analyzing altered test order and manipulated winners, it proposed:

> "Group A replaces their face with a different person's face via AI"

> "omits this logo"

and

> "replaces this with a generic tool box image, misattributing the video to an unrelated source."

This is an especially instructive failure because the model does not appear shallow or inattentive. On the contrary, it builds a complete and internally consistent explanation. The problem is that the explanation belongs to a different kind of forgery from the one actually present.

In Example C, the model was drawn toward highly visible broadcast elements such as timers, lap indicators, subtitles, and ranking graphics. It treated these overlays as the primary site of manipulation, effectively reading the case as one of altered scoreboards or modified graphic information. But the more important change was the restructuring of race progression itself: who appears to lead, when momentum shifts occur, and how the winner is narratively constructed.

In Example D, the model focused on local suspicious traces such as face swapping, exaggerated subtitles, or isolated anomalous shots. That made it treat the review as a patchwork of local falsifications rather than as a structural rewrite. The key manipulation, however, was that separate product impressions were reorganized into an artificial horizontal comparison, producing a false evaluation logic rather than merely a few questionable frames.

In Example E, the model concentrated on specification wording, subtitles, battery claims, or translation quality. These are all concrete and inspectable details, but they distract from the actual manipulation target: the ranking conclusion itself. The forgery does not mainly depend on whether one parameter is mistranslated; it depends on how sequence and framing convert a stronger performer into an apparent loser and a weaker performer into an apparent winner.

Result. Across these examples, the model often produces long, careful reasoning but still misses the central persuasive goal of the fake video. It may correctly identify local inconsistencies, suspicious overlays, branding issues, or textual inaccuracies, yet fail to answer the more important question: what false causal, comparative, or evaluative conclusion is the edit trying to make the viewer believe? This leads to a characteristic form of failure in which the reasoning is not empty, but misdirected.

Interpretation. These cases suggest that the model is often more comfortable auditing *trace evidence* than reconstructing *manipulative intent*. Local anomalies such as logos, graphics, subtitles, and visual artifacts are concrete and easy to point to. By contrast, the true persuasive purpose of a fake video is often distributed across ordering, juxtaposition, pacing, and selective omission. The model therefore risks overfitting to detectable surface irregularities while underfitting the deeper rhetorical structure of the edit. For paper-level interpretation, this can be summarized as follows: the model may become highly competent at explaining *what looks suspicious*, yet remain weak at explaining *what the forgery is trying to make the viewer conclude*. That gap is especially important in fake-video analysis, because many deceptive videos succeed not through one obvious fake frame, but through a carefully engineered change in causal logic, comparison outcome, or evaluative framing.

J.10 Source decomposition errors

Case Study: Source decomposition errors

Overview. The first failure point appears before retrieval itself: the model incorrectly decides how many source families exist, or which source family should be treated as primary. Once this decomposition is wrong, all downstream search and analysis become path-dependent.

Task Description. We illustrate this pattern with 2 representative examples.

Example A: Title anchor overwhelms multi-source decomposition. For a multi-source vlog/disaster montage, `gemini-3.1-pro-preview` explicitly stated:

> "estimated_sources=2"

but generated only one initial source query:

> "Cleo takes final year"

This meant that, although the model verbally acknowledged multiple sources, it operationalized only the vlog title and did not promote the camping/flood footage into an independent retrieval target. The decomposition was therefore correct in form but incorrect in execution.

Example B: Single-source assumption collapses branding source and visual source. In the same example, `qwen3.5-plus` began with:

> "estimated_sources=1"

and framed the video as one coherent episode. Later, however, its own reasoning admitted:

> "Two source families exist: the branding source ('Jesus Loves You' channel) and the visual footage source (camping/bedroom). Only the branding source ... has been identified."

This reveals that the original decomposition had compressed two distinct source families into one, and the model only discovered that mismatch after retrieval had already been misdirected.

Interpretation. The key weakness at this stage is that source decomposition is treated as a soft planning step rather than a binding structural commitment. As a result, the model may say that a case is multi-source while still behaving as if it were single-source. When this happens, later stages do not repair the mistake; they inherit it.

J.11 Query reflection trapped in error clusters

Case Study: Query reflection trapped in error clusters

Overview. The second failure point is not the absence of search, but the inability of search to escape a wrong semantic cluster. Reflection may appear active, yet still remain trapped inside the same mistaken hypothesis.

Task Description. We illustrate this pattern with 2 representative examples.

Example A: Reflection changes wording but not direction. For the same montage case, `gemini-3.1-pro-preview` kept refining the query:

> "Cleo takes final year"

> "Cleo takes final year" vlog -Abram"

> "Cleo takes final year" morning routine uni vlog -Abram"

> "Cleo 'final year' uni vlog South Africa -Abram"

This looks like iterative improvement, but in reality all rounds stayed within the same wrong title-centered cluster. Reflection refined surface wording without changing the retrieval hypothesis.

Example B: Queries drift into irrelevant object domains. With `qwen3.5-plus`, later rounds became increasingly driven by local cues:

> "pink Bratz bedding vlog camping New Year 2021"

> "Cleo takes final year Campus Superior coffee camping"

> "echo chamber of our daily existence" camping vlog"

Here the model was not passive. It searched energetically, but along ever more distracting details, causing retrieval to drift into unrelated domains such as toys, coffee, or religious media.

Interpretation. This stage shows that reflection alone does not guarantee recovery. If the model lacks a mechanism for abandoning a bad hypothesis, it may simply optimize within that hypothesis, producing a sequence of increasingly specific but equally irrelevant queries.

J.12 Forgery extraction misaligned with ground truth.

Case Study: Forgery extraction misaligned with ground truth

Overview. This is the most important error stage in the pipeline. In many difficult cases, the system has already found the correct source, yet still extracts the wrong forgery mechanism. It sees real differences, but explains them along the wrong axis.

Task Description. We illustrate this pattern with 4 representative examples.

Example A: subtitles and specs instead of ranking rewrite. After finding the exact source of a tool-review video, `qwen3.5-plus` summarized the manipulation as:

> "The primary forgery mechanism is Single-Source Editing via the addition of misleading and inaccurate text overlays."

This explanation is not empty, but it mistakes local textual corruption for the central manipulation. The actual forgery was that test order and comparative framing were reorganized to make one brand look like the winner.

Example B: branding/context instead of causal restructuring. In a hydroponics review, `qwen3-vl-235b-a22b-think` matched the original source immediately, but still concluded:

> "key branding and contextual elements are systematically removed or altered to misattribute the video to an unknown source"

Again, this is coherent but off-target. The real manipulation was a timeline rearrangement that fabricated a false causal chain in which nutrient solution and app intervention appeared to ruin the plants.

Example C: fictional product claims instead of comparative pacing bias. In a camera comparison case, `qwen3-vl-235b-a22b-instruct` summarized the manipulation as:

> "AI-generated text ('17' references) and a fabricated 2025 article are spliced into the footage"

This explanation centers on salient labels and future-dated claims, while missing the actual manipulation target: the reallocation of attention and screen time across a real comparison.

Example D: Synthetic insertions are reconstructed as a plausible super-edit. In a laptop review case, `gemini-3-flash-preview` quickly found the correct source but then built a different story:

> "Group A splices in external benchmark graphics from 'Max Tech' ... and a segment featuring Linus Sebastian from 'Linus Tech Tips'."

It ultimately concluded:

> "The overall manipulation in Group A is a multi-source splicing forgery."

The explanation is internally self-consistent, but the inserted scenes were actually AI-generated synthetic footage rather than a multi-channel editorial mashup.

Interpretation. This stage shows the pipeline's central forensic weakness: finding the source is not the same as understanding the manipulation. Fine extraction is highly vulnerable to salience bias, local anomaly overfitting, and preference for familiar explanation templates such as subtitles, branding, or cross-source splicing.

J.13 Premature stopping

Premature stopping

Overview. The final stage decides whether the current evidence is already sufficient. This is a particularly fragile point, because once the model declares the case "sufficient," the pipeline stops and the current hypothesis is effectively frozen. If the hypothesis is already wrong, the stop decision converts an intermediate mistake into the final answer.

Task Description. We illustrate this pattern with 3 representative examples.

Example A: Early stopping locks in the wrong verification targets. In the hydroponics case, `qwen3-vl-235b-a22b-think` stopped after one round, claiming:

> "The evidence collected ... covers all key verification targets"

However, those targets were themselves misdefined around logo, interface, location, and attribution. The actual manipulation target — the fabricated causal narrative — had never been recovered. The stop decision therefore validated the wrong story.

Example B: "Source found" becomes equivalent to "forgery understood." In the tool-review case, `qwen3.5-plus` stopped because:

> "Since the source is from the official channel and matches frame-for-frame, further searching is unnecessary."

This reveals a structural weakness in the stopping criterion: source resolution is treated as if it implied manipulation coverage, even though a perfectly matched source can still coexist with a completely wrong forgery explanation.

Example C: Stopping checks self-consistency, not correctness. In the laptop review case, `gemini-3-flash-preview` continued searching longer, but its final stopping logic still remained trapped in the wrong hypothesis:

> "Without the Max Tech source video, I cannot confirm if the input is a Max Tech video with a swapped end card or an MKBHD video with swapped graphics."

This shows that the stop module does not ask whether the hypothesis is fundamentally off-task. It asks only whether that hypothesis has been developed enough to feel complete.

Interpretation. The stop stage is less a verifier of truth than a verifier of local narrative completeness. Once the model has constructed a sufficiently coherent story, the pipeline may terminate even if that story addresses the wrong manipulation. This makes early stopping a particularly high-leverage failure point in retrieval-based fake-video analysis.

K Licenses and Terms of Use

EVID-Bench will be publicly released for non-commercial academic research under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license. To comply with the YouTube Terms of Service regarding media distribution, authentic source videos are provided as YouTube Video IDs and URLs rather than raw MP4 files. Source links may become unavailable over time; we therefore release a manifest with each sample's YouTube ID and collection timestamp, distribute the manipulated benchmark videos directly, and will maintain a mirror to improve long-term availability.

External models, tools, and datasets are governed by their respective licenses and terms of use. Generation tools include FaceFusion (MIT License) and Seedance 2.0 (Seedance et al., 2026) (subject to ByteDance's applicable terms). Open-weight Qwen models are used under their model licenses. Closed-source APIs used for generation filtering and evaluation—OpenAI GPT (OpenAI, 2026b), Google Gemini (Google DeepMind, 2025), and Anthropic Claude (Anthropic, 2026a,b)—are accessed under each provider's API Terms of Service and acceptable-use policies, for non-commercial academic research only.

L Content Filtering and Privacy Considerations

During the dataset construction and human review phase, we implemented strict filtering criteria. We ensured that no offensive, harmful, or Not Safe For Work (NSFW) content was included in the benchmark. Furthermore, we verified that the dataset does not expose sensitive Personally Identifiable

Information (PII) of private individuals. However, strict visual anonymization (such as blurring faces or redacting on-screen text) was purposefully not applied for most forgery types. Because EVID-Bench evaluates evidence-dependent video forgery (e.g., contextual fabrication and brand manipulation), masking public figures, brand logos, or event details would destroy the essential visual evidence required for the cross-video reasoning task. All source materials are based on initially public footage from YouTube. Donor faces are taken solely from the public CelebA face dataset, not from private individuals.