From Detection to Understanding: Multi-Turn Reasoning for Video Misinformation Analysis

Zhi Zeng[1,2] **Jiaying Wu**[3] **Minnan Luo**[1,2*]
Di Zhang[1] **Yifei Yang**[1,2] **Xiangzheng Kong**[1,2] **Herun Wan**[1,2] **Zihan Ma**[1,2]
[1]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China
[2]Ministry of Education Key Laboratory of Intelligent Networks and Network Security, China
[3]National University of Singapore, Singapore

Abstract

Video misinformation detection is often approached as a binary veracity classification problem, overlooking the **complex reasoning** required to explain *how* and *why* content misleads. Existing benchmarks fail to capture the diversity of manipulation strategies, such as AI-generated edits and out-of-context manipulation, and do not evaluate whether models can provide process-level justifications for their judgments. We address these limitations with MISVIDEOQA, a multi-turn benchmark designed to assess comprehensive understanding and reasoning in video misinformation analysis. MISVIDEOQA covers 12 fine-grained video categories and evaluates models along six dimensions, progressing from perceptual attribution to intent and persuasion analysis. Recognizing that standard MLLMs struggle to sustain such structured, evidence-based deduction, we propose MISAGENT, a Delphi-inspired multi-agent framework in which specialized agents collaboratively integrate multimodal cues with external evidence. Experimental results show that state-of-the-art multimodal large language models perform poorly on MISVIDEOQA, while MISAGENT consistently improves reasoning accuracy and explanation quality. Together, our benchmark and framework establish a unified foundation for reliable, interpretable, and evidence-grounded video misinformation analysis.[1]

1 Introduction

Short-video platforms are a major channel for information consumption and a key vector for misinformation. Unlike text-only rumors, misleading videos combine visual, auditory, and textual signals to produce narratives that appear coherent yet are deceptive (Choi and Ko, 2021; Shang et al., 2021; Zong et al., 2024; Zeng et al., 2025b). These narratives arise from diverse strategies, including AI-generated media, multimodal manipulation, and the out-of-context reuse of authentic footage (Zeng et al., 2025a). Effective misinformation analysis therefore requires more than binary veracity judgments. It demands models that can explain *how* and *why* a video misleads by identifying manipulated signals, inconsistencies, and persuasive intent.

Despite the urgency of this challenge, current research remains constrained by fundamental limitations in both **benchmarking** and **modeling**. Existing benchmarks often oversimplify video misinformation analysis by focusing on narrow deception categories, such as deepfakes or caption inconsistencies (Gao et al., 2024; Batra et al., 2025), and by reducing evaluation to binary veracity classification (Qi et al., 2023a; Bu et al., 2024). This framing sidelines reasoning and prevents assessment of whether models can attribute deceptive cues or explain how manipulation leads to misleading interpretation (Xu et al., 2025a; Zeng et al., 2025a). As illustrated in Figure 1, comprehensive understanding requires tracing evidence from perceptual inconsistencies to underlying intent.

These benchmark limitations, in turn, shape how models are trained and evaluated. When supervision emphasizes label prediction rather than reasoning fidelity, models are neither incentivized nor evaluated on evidence grounding. Consequently, although recent work uses Multimodal Large Language Models (MLLMs) to generate natural language rationales for misinformation (Hong et al., 2025; Tran, 2025; Niu et al., 2025), such explanations are often fragile, hallucinated, and weakly grounded in verifiable external evidence. This weakness limits their applicability in real-world misinformation governance, where explanations must be transparent and auditable.

To address these gaps, we present two complementary contributions that jointly advance evaluation and reasoning for video misinformation anal-

*Corresponding Author
[1]Data and code are available at: https://github.com/zzeng1998/MisVideoQA.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 37009–37027
July 2-7, 2026

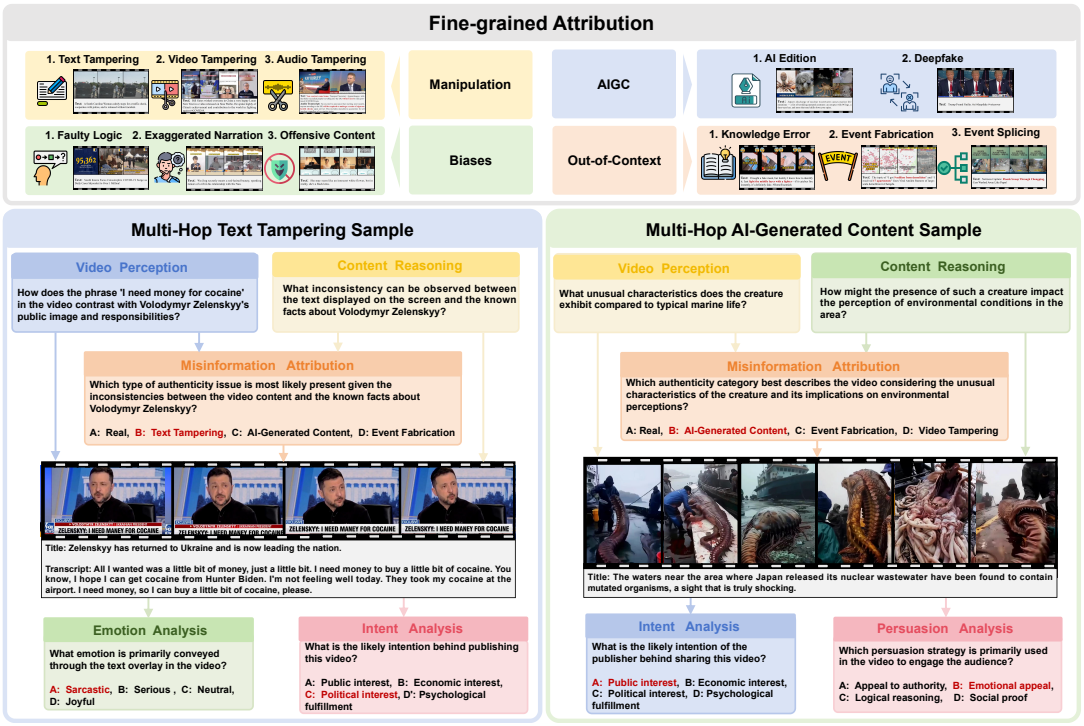


Figure 1: Two representative instances from MISVIDEOQA. Each example follows a multi-hop, multi-turn reasoning process that progresses from video perception and content reasoning to fine-grained misinformation attribution, and further supports post-detection analysis of emotion, intent, and persuasion.

ysis. We first introduce MISVIDEOQA, a multi-turn benchmark designed to assess process-level understanding and reasoning. MISVIDEOQA spans 12 diverse deception subcategories and evaluates models across six progressive dimensions, from pre-detection perception and attribution to post-detection analysis of emotion, persuasion, and intent. This enables systematic evaluation of not only what models detect, but how they justify their conclusions across diverse manipulation strategies. To tackle the modeling challenges exposed by MISVIDEOQA, we further propose MISAGENT, a multi-agent framework inspired by the Delphi framework's structured consensus-building process (Linstone et al., 1975; Xiong et al., 2025). Rather than relying on a single monolithic model, MISAGENT coordinates specialized agents for background verification, event-level examination, and evidence integration, enabling collaborative synthesis of multimodal cues with external evidence. Experiments show that while state-of-the-art MLLMs struggle with the progressive reasoning demands of MISVIDEOQA, MISAGENT consistently improves both accuracy and explanation quality, establishing a unified foundation for reliable and interpretable video misinformation analysis.

2 Related Work

Benchmarks for video misinformation detection and video question answering have advanced multimodal understanding, as summarized in Table 1.

2.1 Benchmarks

Video Misinformation Benchmarks. Early datasets collected misleading micro-videos and their textual context from social platforms. FVC (Papadopoulou et al., 2019) is one of the first large-scale benchmarks based on YouTube and Twitter, and later work expanded to additional platforms (Palod et al., 2019; Hou et al., 2019). Domain-specific datasets targeted Covid-19 on TikTok (Serrano et al., 2020; Shang et al., 2021), while FakeTT (Bu et al., 2024) broadened topics for cross-context evaluation. FakeSV (Qi et al., 2023a) further introduced a large-scale Chinese benchmark with multimodal signals. With the rise of MLLMs, newer benchmarks also consider AI generated or edited content (Xu et al., 2025a; Batra et al., 2025), yet they still underrepresent in the wild misinformation, especially out-of-context reuse of real footage (Luo et al., 2021). Overall, most datasets remain detection oriented and provide limited coverage of attribution and post-detection interpretation.

Datasets	Domain	Reasoning	#Video	Type	QAs	Annotation	Source
FVC	VMD	×	5,006	2	-	Human	YT/TW
(Hou et al. 2019)	VMD	×	250	2	-	Human	TT
(Li et al. 2022)	VMD	×	700	2	-	Human	BB
FakeSV	VMD	×	3,654	2	-	Human	DY/KS
FakeTT	VMD	×	1,991	2	-	Human	TT
AdsQA	Ads	×	1,544	5	7,859	Auto	Synthetic
MTMEUR	EUAR	√	1,451	6	5,101	Auto	Synthetic
MISVIDEOQA (Ours)	VMUAR	√	1,174	12	5,235	Human, Auto	WB/DY/KS YT/IS/BB

Table 1: Comparison of related benchmarks. **MISAGENT** supports multi-turn misinformation understanding and reasoning, with the broadest coverage of deception types and the most diverse real-world sources. Domains: VMD (Video Misinformation Detection), Ads (Advertisement), EUAR (Emotion Understanding and Reasoning), VMUAR (Video Misinformation Understanding and Reasoning). Annotation: Human (human annotation), Auto (automatic annotation). Source platforms: YT (YouTube), TW (Twitter), TT (TikTok), BB (Bilibili), WB (Weibo), DY (Douyin), IS (Instagram), KS (Kuaishou).

Video Question Answering Benchmarks. Video QA benchmarks evaluate spatiotemporal perception and event understanding by answering questions grounded in videos (Cao et al., 2025; Long et al., 2025; Hu et al., 2025). Many focus on human activities and everyday scenarios, such as NextQA (Xiao et al., 2021), ActivityQA (Yu et al., 2019), and VideoMME (Fu et al., 2025), with sources ranging from movies to web and social media videos (Long et al., 2025). Beyond general understanding, MTMEUR targets emotion understanding and reasoning (Hu et al., 2025).

Our Benchmark. We introduce MISVIDEOQA, a multi-turn benchmark for comprehensive evaluation of video misinformation understanding. MISVIDEOQA jointly assesses pre-detection reasoning, and post-detection analysis. This design enables process-level evaluation of whether models can not only detect misinformation, but also justify decisions and provide interpretable explanations.

2.2 Methods

Video Misinformation Detection. Early video misinformation detection relied on handcrafted cues from titles, captions, and user comments (Serrano et al., 2020; Hou et al., 2019). Further, later work learned multimodal representations and improved performance via cross-modal fusion and contextual modeling (Li et al., 2022; Choi and Ko, 2021; Shang et al., 2021; Zong et al., 2024; Wu et al., 2024). Recent studies further enhance robustness by incorporating external or structured knowledge. For example, SV-FEND (Qi et al., 2023a) integrates multimodal signals with Transformer-based architectures, while NEED (Qi et al., 2023b) leverages event and debunking knowledge with graph attention. Other methods exploit editing traces or reduce spurious correlations through debiasing strategies (Bu et al., 2024; Zeng et al., 2024).

MLLM-based Misinformation Detection. MLLM-based approaches leverage instruction following and broad world knowledge for multimodal claim analysis via pipeline frameworks (Zheng et al., 2025; Wang et al., 2024b; Qi et al., 2024; Liu et al., 2024; Li et al., 2025). Recent work introduces multi-view knowledge and role-based prompting to strengthen reasoning (Wang et al., 2025a; Tahmasebi et al., 2024). However, rationales may hallucinate or be weakly grounded, limiting reliability. To improve faithfulness, prior studies adopt structured reasoning and evidence-centric designs, such as CoT prompting (Hong et al., 2025), retrieval-augmented generation (Yue et al., 2024), reinforcement learning for verifiable reasoning (Zhang et al., 2025), and explicit external evidence integration (Niu et al., 2025; Zeng et al., 2026; Xu et al., 2025b).

Our Method. Existing methods often use MLLMs mainly as knowledge enhancers for detection, with limited support for collaborative, multi-step investigation that seeks, cross-checks, and consolidates evidence across modalities. To address fine-grained attribution and post-detection analysis, we propose MISAGENT, a Delphi-inspired multi-agent framework that retrieves and integrates real-world external evidence for more comprehensive and interpretable understanding and reasoning.

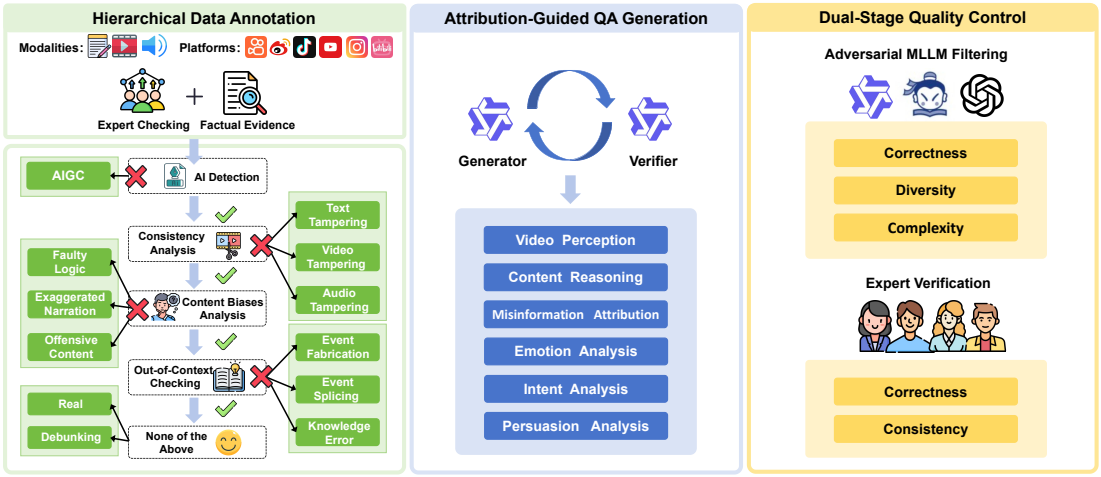


Figure 2: Overview of our MISVIDEOQA benchmark: data annotation, QA generation, and quality control.

3 MISVIDEOQA: Multi-Turn Reasoning for Video Misinformation

We introduce MISVIDEOQA, a multi-turn benchmark for evaluating video misinformation understanding and reasoning. Unlike prior datasets, MISVIDEOQA is grounded in verified evidence and spans multiple social platforms, enabling evaluation under realistic and diverse misinformation scenarios (see Figure 2).

3.1 Data Curation

To ensure factual reliability, we source videos exclusively from authoritative fact-checking repositories, namely *PolitiFact*[2] and the *China Internet Joint Rumor-Refuting Platform*[3]. Our initial collection contains 10,366 videos from 2017 to 2025 across six major platforms: Weibo, Douyin, Kuaishou, YouTube, Instagram, and Bilibili. We retain only videos containing objectively verifiable claims. To eliminate redundancy and prevent data leakage, we apply rigorous textual similarity filtering to remove near-duplicate videos while preserving topical diversity.

3.2 Hierarchical Data Annotation

In contrast to prior benchmarks that rely on binary labels or synthetic data (Qi et al., 2023a; Bu et al., 2024; Hu et al., 2025), MISVIDEOQA adopts an evidence-based annotation framework that explicitly categorizes the *mechanism* of deception. Each video is **(1)** first labeled as Real or Fake and is then **(2)** further classified into a fine-grained taxonomy.

As shown in Figure 2, we design a four-stage reasoning schema that progressively detects and characterizes misinformation in videos:

- **AI-Generated Content Detection:** Content synthesized or substantially edited by generative models, labeled as **(1)** AIGC.
- **Consistency Analysis:** Identification of cross-modal inconsistencies involving **(2)** text tampering, **(3)** video tampering, or **(4)** audio tampering.
- **Bias Analysis:** Detection of psychological manipulation, including **(5)** faulty logic, **(6)** exaggerated narration, and **(7)** offensive content.
- **Out-of-Context Detection:** Deceptive reuse of authentic footage, including **(8)** knowledge error, **(9)** event fabrication, and **(10)** event splicing.

Videos without evidence of misinformation are categorized as either **(11)** real videos or **(12)** debunking videos. Ambiguous samples, accounting for approximately 1.3% of the data, are excluded when annotator consensus cannot be reached. All labels are finalized through unanimous agreement by at least three expert annotators with graduate-level training in computer science or social sciences.[4]

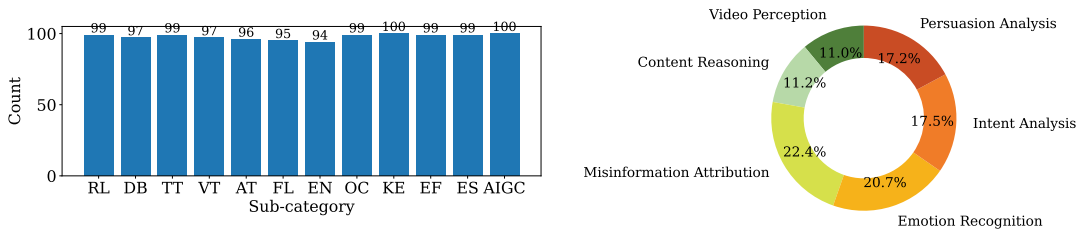
3.3 Attribution-Guided QA Generation

Manually constructing large-scale multi-turn reasoning chains is labor-intensive. To address this challenge, we design a semi-automated pipeline that generates high-quality QA pairs grounded in the hierarchical attribution taxonomy.

Sampling and Generation. From the 10,366 collected micro-videos, we curate a balanced subset with 100 samples per attribution type. We further

[2]https://www.politifact.com
[3]https://www.piyao.org.cn

[4]The detailed annotation protocol is provided in Appendix B.1.



(a) Distribution of video types in MISVIDEOQA. (b) Distribution of QA categories in MISVIDEOQA.

Figure 3: Data analysis of MISVIDEOQA. Types include Real (RL), Debunking (DB), Text Tampering (TT), Video Tampering (VT), Audio Tampering (AT), Faulty Logic (FL), Exaggerated Narration (EN), Offensive Content (OC), Knowledge Error (KE), Event Fabrication (EF), Event Splicing (ES), and AI-Generated Content (AIGC).

filter for title-level diversity to ensure broad coverage of real-world video content. For each video, we use Qwen-VL-Max (Bai et al., 2025b) to generate and verify a progressive six-step QA chain:

- **Pre-Detection Reasoning (QA1–QA3):** Evaluates MLLMs' understanding and reasoning abilities through **(1)** video perception, **(2)** content reasoning, and **(3)** misinformation attribution.
- **Post-Detection Analysis (QA4–QA6):** Assesses higher-level interpretation via **(4)** emotion recognition, **(5)** persuasion and **(6)** intent analysis.

Each question contains one correct option and three distractors. Distractors are adversarially constructed to be superficially plausible yet semantically distinct or contradictory, thereby increasing discriminative difficulty.[5]

3.4 Dual-Stage Quality Control

To ensure both reliability and appropriate difficulty, we apply a two-stage quality control process.

Adversarial MLLM Filtering. We first remove trivially solvable misinformation cases. Specifically, any QA pair correctly answered by all three representative models: GPT-4o (Hurst et al., 2024), InternVL-3 (Zhu et al., 2025), and Qwen2.5-VL (Bai et al., 2025b), is deemed insufficiently challenging and excluded. This adversarial filtering step ensures that the benchmark retains non-trivial questions that meaningfully probe reasoning capabilities. Figure 3(a) shows the resulting distribution, which preserves broad coverage across the deception types defined in Section 3.2.

Expert Verification. We further conduct expert verification to ensure logical soundness of the retained QA pairs. Annotators manually examine each multi-hop reasoning chain by decomposing complex questions into sub-questions and validating the progressive dependency from QA1 (video perception) to QA3 (misinformation attribution). Only QA pairs that pass both adversarial filtering and expert verification are retained, yielding a challenging and coherent evaluation set (Figure 3(b)).

[5]The detailed QA generation procedure is described in Appendix B.2.

4 MISAGENT

To address limitations of existing MLLMs in video misinformation analysis, we propose MISAGENT, a collaborative multi-agent framework for evidence-based reasoning as shown in Figure 4.

4.1 Motivation: The Delphi Framework

Delphi (Linstone et al., 1975) is a structured communication framework originally developed for systematic forecasting. It aggregates expert opinions through iterative rounds of anonymous inquiry, promoting consensus while preserving independent judgment. We adopt this principle as the design motivation for coordinating agents in our multi-agent reasoning framework.

Formally, let $\mathcal{P} = \{P_m\}_{m=1}^{M}$ denote a panel of specialized agents, let O represent a central organizer (the Reflection Agent), and let $f^{(0)} = \varnothing$ denote the initial shared state. At each iteration $i = 1, 2, \ldots$, agents independently generate opinions conditioned on the target input and the previous feedback state:

$$o_m^{(i)} = P_m\left(x, f^{(i-1)}\right), \quad f^{(i)} = O\left(\{o_m^{(i)}\}_{m=1}^{M}\right), \tag{1}$$

where x denotes the target micro-video content. The process continues until consensus is reached at round T. The final decision is obtained by aggregating agent opinions from the final round, typically

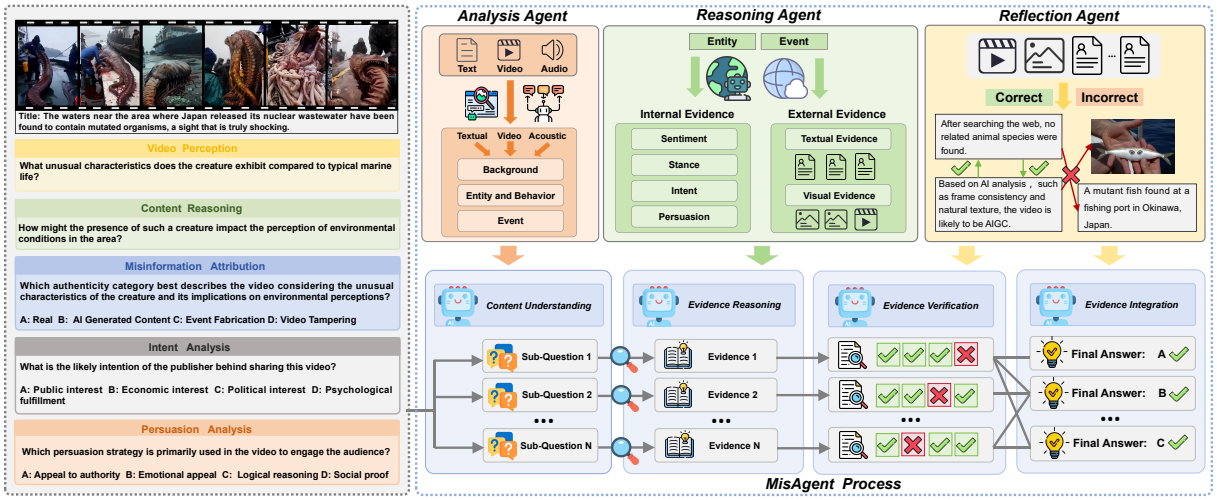


Figure 4: Overview of our proposed MISAGENT framework for multi-turn video misinformation analysis.

via confidence-weighted voting:

$$T = \min\left\{ i : \text{cons}\left(\{o_m^{(i)}\}_{m=1}^{M}\right) = 1 \right\},$$
$$o^* = \arg\max_{y \in \mathcal{Y}} \sum_{m=1}^{M} w_m^{(T)} \, \mathbb{I}\left[o_m^{(T)} = y\right], \tag{2}$$

where $\text{cons}(\cdot) = 1$ denotes a consensus function and $w_m^{(T)} \geq 0$ represents the reliability weight of agent m at round T.

4.2 Multimodal Content Analysis

Although MLLMs exhibit strong general capabilities (Wu et al., 2026; Jiang et al., 2025a), micro-videos often employ subtle deceptive strategies, such as multimodal manipulation (Chen et al., 2025a; Wang et al., 2025c) and implicit semantic twists (Wang et al., 2024a; Lang et al., 2025), that evade single-pass detection. To address this challenge, the **Content Analysis Agent** serves as the perceptual anchor by decomposing a video $x = \{x_t, x_v, x_a\}$, consisting of textual x_t, visual x_v, and acoustic x_a modalities, into granular observational streams.

The agent performs multi-view analysis across three levels: **(1)** *Background Description*, which captures setting and contextual cues; **(2)** *Entity and Behavior Identification*, which focuses on object interactions and gestures; and **(3)** *Event Narrative Analysis*, which examines temporal continuity and action logic. This structured decomposition produces a set of veracity-related rationales $\mathcal{R}$:

$$\mathcal{R} = \{r_j\}_{j=1}^{N}, \quad r_j = o_j(x), \tag{3}$$

where the function $o_j(\cdot)$ represents a single-view analysis expert, and each rationale r_j is produced through structured multi-turn interactions. By processing these perceptual signals, the agent provides the observational basis for subsequent reasoning. Prompt templates are provided in Appendix D.

4.3 Multi-view Evidence Reasoning

Perception alone is insufficient for identifying complex misinformation. The **Reasoning Agent** contextualizes observations in $\mathcal{R}$ by synthesizing internal semantic cues with external verification.

Internal Evidence Synthesis. Leveraging the MLLM's inherent world knowledge (Wan et al., 2025; Liu et al.), this component extracts semantic evidence $\mathcal{I} = \{e_k\}_{k=1}^{K}$. It analyzes the content from the perspectives of emotion, stance, intent, and persuasion strategies (Wan et al., 2024; Yang et al., 2026; Tong et al., 2024; Lu et al., 2025; Tong et al., 2025), enabling detection of psychological manipulation patterns that do not require external information.

External Evidence Retrieval. To address out-of-context misinformation that exploits knowledge gaps in MLLMs (Xu et al., 2025a), the Reasoning Agent employs an adaptive retrieval strategy. Given the video title x_t, it constructs queries to retrieve verifiable facts from authoritative sources such as Wikipedia and Google News. The resulting external evidence corpus is defined as:

$$\mathcal{E} = \text{retrieval}(x_t, S_K), \tag{4}$$

where S_K denotes the top-K ranked results from trusted domains. This step anchors the reasoning

	Modality			Pre-detection Reasoning			Post-detection Analysis			Overall
	A	V	T	VP	CR	MA	EA	IA	PA	
MLLMs with Direct Prompt										
InternVL-3-8B (Direct)	×	√	√	92.03	90.26	58.01	50.00	43.39	45.21	58.95
LLaVA-OneVision-7B (Direct)	×	√	√	83.88	81.03	52.81	50.64	43.72	60.02	58.59
GLM-4.1V-9B (Direct)	×	√	√	63.26	68.55	41.48	62.17	60.44	55.23	56.87
Gemma3-12B (Direct)	×	√	√	88.56	32.31	51.45	59.76	50.60	45.66	53.98
InternVL-3.5-14B (Direct)	×	√	√	92.37	88.21	59.03	39.87	34.75	16.93	50.53
InternVL-3-38B (Direct)	×	√	√	92.72	90.60	60.05	81.03	71.48	72.72	75.59
Qwen3-VL-32B (Direct)	×	√	√	91.85	89.74	55.88	77.72	70.49	70.49	73.22
MLLMs with VoT-based Prompt										
VideoLLaMA2-7B (VoT)	√	√	√	72.62	73.50	45.40	65.47	75.30	61.02	63.61
GLM-4.1V-9B (VoT)	×	√	√	76.08	77.78	49.49	74.22	67.54	67.59	66.97
InternVL-3.5-14B (VoT)	×	√	√	92.89	87.69	56.64	**90.88**	69.95	48.66	72.17
Qwen3-VL-32B (VoT)	×	√	√	92.72	92.14	58.35	78.36	70.05	74.05	74.80
Closed-source MLLM										
GPT-4o-mini (Direct)	×	√	√	89.12	78.34	**60.41**	79.85	72.09	48.67	73.05
GPT-4o (Direct)	×	√	√	93.68	92.23	59.48	87.62	82.17	73.71	79.17
Our Proposed Approach										
GLM-4.1V-9B (MISAGENT)	×	√	√	83.86	83.39	46.00	79.28	71.96	74.24	68.41(11.54% ↑)
Qwen3-VL-32B (MISAGENT)	×	√	√	91.68	88.21	56.73	87.29	76.50	70.27	76.22 (3.00% ↑)
InternVL-3-38B (MISAGENT)	×	√	√	**94.21**	**92.75**	60.26	89.83	**83.48**	**77.19**	**80.75** (5.16% ↑)

Table 2: **Main results (%).** Performance of all models measured by Micro-Acc. QA types are abbreviated as follows: VP (Video Perception), CR (Content Reasoning), MA (Misinformation Attribution), EA (Emotion Analysis), IA (Intent Analysis), and PA (Persuasion Analysis). **Bold:** the best result. Underline: the second result.

process in verifiable evidence.

4.4 Evidence Reflection and Integration

The **Reflection Agent** acts as the Delphi organizer, filtering noise and guiding the system toward consensus. To ensure semantic alignment, it first filters the evidence to retain only contextually relevant information $\mathcal{F}$ that supports veracity assessment:

$$\mathcal{F} = \text{filter}(x_t, \mathcal{E}). \tag{5}$$

Here, $\mathcal{F}$ contains the filtered and context-aligned evidence that directly supports veracity assessment, allowing MISAGENT to adaptively integrate internal reasoning with external validation.

The agent then integrates content cues $\mathcal{R}$, internal evidence $\mathcal{I}$, and filtered external evidence $\mathcal{F}$ into a unified evidence set. Through I rounds of iterative feedback, as defined in Eq.6, the Reflection Agent synthesizes these complementary signals:

$$\mathcal{E}^{(I)} = \text{reflection}(\mathcal{R}, \mathcal{I}, \mathcal{F}). \tag{6}$$

In the final step, the agent aggregates $\mathcal{E}^{(I)}$ to produce the final prediction y^* and justification j^*, ensuring that the decision is accurate, auditable, and robust to hallucination.

5 Experiments

We conduct extensive experiments to address the following research questions:

- **RQ1** (§5.1): Does MISAGENT improve video misinformation understanding and reasoning?
- **RQ2** (§5.1): Is MISAGENT adaptable across MLLM backbones of varying scales?
- **RQ3** (§5.2): What is the contribution of each modality and agentic component to the framework's overall efficacy?
- **RQ4** (§5.3): How does MISAGENT reason about real-world cases?

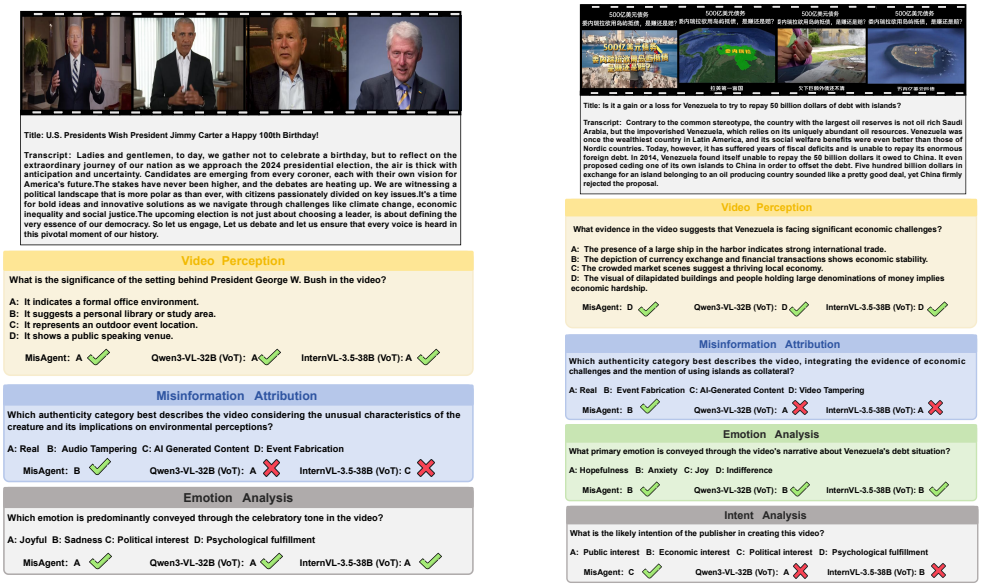
Details of the experimental setup, including model selection and implementation, are provided in Appendix C.

Model	InternVL3-38B	GLM-4.1V-9B
w/o Text	76.34	66.97
w/o Audio	76.62	67.09
w/o MCU	77.49	67.81
w/o MER	74.42	61.38
MisAgent	**80.75**	**68.41**

Table 3: Ablation results (%). Macro-level accuracy of different model variants on MISVIDEOQA.

5.1 Main Results (RQ1 & RQ2)

Overall Performance. Table 2 reports results on MISVIDEOQA. We make three key observations. **(1)** Even with Video-of-Thought (VoT) reasoning (Fei et al., 2024), small and medium MLLMs achieve at most ∼75% overall accuracy, underscoring the difficulty of video misinformation reasoning. **(2)** MISAGENT effectively narrows this gap. Across three representative MLLMs of varying sizes, GLM-4.1V-9B (Team et al., 2025b), Qwen3-VL-32B (Bai et al., 2025a), and InternVL-3-38B



(a) An audio tampering case.

(b) An event fabrication case.

Figure 5: Qualitative examples illustrating how MISAGENT reasons about real-world video misinformation.

(Zhu et al., 2025), it consistently improves accuracy by 3.00% to 11.54% over direct prompting, demonstrating strong generalization and the ability to unlock reasoning capabilities in lighter models. **(3)** MISAGENT consistently outperforms all MLLM baselines, achieving a 1.58% absolute accuracy gain over strong GPT-4o (Hurst et al., 2024) when paired with InternVL-3-38B (Zhu et al., 2025).

Attribution and Reasoning Analysis. While MLLMs perform reasonably on basic video perception, they struggle with misinformation attribution, which requires identifying deceptive mechanisms. As shown in Table 2, adopting the VoT strategy yields marginal perception gains but degrades attribution accuracy, suggesting that unstructured reasoning can induce hallucinated causal links. In contrast, MISAGENT achieves higher accuracy across most fine-grained categories, indicating that multi-agent collaboration improves reasoning fidelity.

5.2 Ablation Study (RQ3)

Impact of Modalities. We evaluate variants of MISAGENT by removing textual or acoustic inputs. As shown in Table 3, eliminating any modality leads to a performance drop, confirming that MISVIDEOQA requires genuine cross-modal reasoning. These results also show that MISAGENT exploits interactions among multimodal signals to detect inconsistencies missed by unimodal approaches.

Impact of Agentic Components. We further examine the role of individual agents by comparing the full framework with two ablations: **w/o MCU**, which removes the Content Analysis Agent, and **w/o MER**, which removes multi-view evidence reasoning. Table 3 shows that the full framework consistently outperforms both variants. The larger degradation observed for **w/o MER** highlights that internal knowledge alone is insufficient and that dynamic retrieval with evidence integration is critical for robust misinformation reasoning.

5.3 Qualitative Analysis (RQ4)

We present two representative case studies: **(1)** audio tampering as multimodal manipulation and **(2)** event fabrication as out-of-context reuse. Figure 5(a) illustrates performance on multimodal manipulation. While baseline models fail to detect subtle acoustic inconsistencies, MISAGENT correctly attributes the deception to audio tampering through cross-modal comparison. In the out-of-context scenario shown in Figure 5(b), MISAGENT retrieves authoritative external evidence to verify the claim and identifies that authentic footage has been repurposed to support a fabricated narrative. These examples demonstrate that MISAGENT moves beyond binary detection to produce explainable, evidence-grounded justifications.[6]

[6]Additional error analysis and case studies are provided in Appendix E.

6 Conclusion

This work advocates for a paradigm shift in video misinformation analysis from binary veracity prediction to transparent, process-level reasoning. We introduce MISVIDEOQA, a multi-turn benchmark that evaluates misinformation understanding across progressive dimensions, from fine-grained attribution to intent and persuasion analysis. We also propose MISAGENT, a Delphi-inspired multi-agent framework that mitigates the hallucination and grounding failures of MLLMs by enforcing a structured, evidence-based consensus among specialized agents. Experiments show that while MISVIDEOQA is challenging for excellent models, MISAGENT consistently improves reasoning accuracy and explanation quality.

Limitations

While MISVIDEOQA provides broad coverage across six major social platforms, it focuses primarily on micro-videos and may not fully capture the diversity of long-form or text-centric misinformation in other digital ecosystems. Moreover, although the benchmark defines 12 video categories, including 10 deceptive types, and six QA dimensions, online manipulation strategies are adversarial and continuously evolving. As a result, the current taxonomy may not cover all emerging forms of misinformation, such as novel generative AI attacks or rhetorical strategies that exploit cultural context, irony, or implicit framing. These limitations point to the need for future expansion building on the foundation of MISVIDEOQA.

Ethics Statement

We only used AI assistants to polish the writing of the paper, and did not use them to generate any content or images. We strictly followed the data-use and scraping policies of all platforms involved in this study. All annotators received formal training and were familiar with relevant data privacy and security regulations. During annotation, only content related to public figures or public events was considered, and posts involving private individuals were excluded. The annotation experts included twelve individuals with bachelor's or master's degrees in computer science and social science. All annotators were paid wages that comply with national standards.

Our MISVIDEOQA dataset incorporates 292 video samples from FMNV (Wang et al., 2025c), and it is released under the Attribution NonCommercial ShareAlike 4.0 International license, CC BY NC SA 4.0. We will adopt this license to align with the licensing terms of several constituent datasets, thereby providing the same level of access. MISVIDEOQA contains offensive content type video potentially containing harmful text and videos. This is intended for use as an evaluation dataset to support safer content moderation and help improve the security and integrity of the online information ecosystem.

To ensure privacy protection, all identifiable user information, including usernames and IDs, was anonymized. We implemented safeguards throughout data processing and model training to prevent any leakage of personal data. All collected data are securely stored on protected servers with access restricted to authorized research personnel only. Lastly, the proposed video misinformation detection method is designed to contribute to the safety and stability of the internet environment and public opinion.

Acknowledgments

This work is supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM101), the National Natural Science Foundation of China (No. 62272374, No. 62192781), the Natural Science Foundation of Shaanxi Province (No.2024JC-JCQN-62), the State Key Laboratory of Communication Content Cognition under Grant No. A202502, the Key Research and Development Project in Shaanxi Province (No. 2023GXLH-024), and the Ministry of Education, Singapore, under its MOE AcRF TIER 3 Grant (MOE-MOET32022-0001). The China Scholarship Council also supports this research.

References

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, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo

Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Arnesh Batra, Jashn Khemani, Arush Gumber, Anushk Kumar, Arhan Jain, and Somil Gupta. 2025. Socialdf: Benchmark dataset and detection model for mitigating harmful deepfake content on social media platforms. In *Proceedings of the 4th ACM International Workshop on Multimedia AI against Disinformation*, pages 81–89.

Jinhe Bi, Danqi Yan, Yifan Wang, Wenke Huang, Haokun Chen, Guancheng Wan, Mang Ye, Xun Xiao, Hinrich Schuetze, Volker Tresp, and Yunpu Ma. 2025. Cot-kinetics: A theoretical modeling assessing lrm reasoning process.

Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. 2024. Fakingrecipe: Detecting fake news on short video platforms from the perspective of creative process. *arXiv preprint arXiv:2407.16670*.

Meng Cao, Pengfei Hu, Yingyao Wang, Jihao Gu, Haoran Tang, Haoze Zhao, Chen Wang, Jiahua Dong, Wangbo Yu, Ge Zhang, et al. 2025. Video simpleqa: Towards factuality evaluation in large video language models. *arXiv preprint arXiv:2503.18923*.

Lizhi Chen, Zhong Qian, Peifeng Li, and Qiaoming Zhu. 2025a. Multimodal fake news video explanation: Dataset, analysis and evaluation. *arXiv preprint arXiv:2501.08514*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

Hyewon Choi and Youngjoong Ko. 2021. Using topic modeling and adversarial neural networks for fake news video detection. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 2950–2954.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong Li Lee, and Wynne Hsu. 2024. Video-of-thought: step-by-step video reasoning from perception to cognition. In *Proceedings of the 41st International Conference on Machine Learning*, pages 13109–13125.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118.

Yifei Gao, Jiaqi Wang, Zhiyu Lin, and Jitao Sang. 2024. Aigcs confuse ai too: Investigating and explaining synthetic image-induced hallucinations in large vision-language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9010–9018.

Rongpei Hong, Jian Lang, Jin Xu, Zhangtao Cheng, Ting Zhong, and Fan Zhou. 2025. Following clues, approaching the truth: Explainable micro-video rumor detection via chain-of-thought reasoning. In *Proceedings of the ACM on Web Conference 2025*, pages 4684–4698.

Rui Hou, Verónica Pérez-Rosas, Stacy Loeb, and Rada Mihalcea. 2019. Towards automatic detection of misinformation in online medical videos. In *2019 International conference on multimodal interaction*, pages 235–243.

Jinpeng Hu, Hongchang Shi, Chongyuan Dai, Zhuo Li, Peipei Song, and Meng Wang. 2025. Beyond emotion recognition: A multi-turn multimodal emotion understanding and reasoning benchmark. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5814–5823.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Kailin Jiang, Hongbo Jiang, Ning Jiang, Zhi Gao, Jinhe Bi, Yuchen Ren, Bin Li, Yuntao Du, Lei Liu, and Qing Li. 2025a. Kore: Enhancing knowledge injection for large multimodal models via knowledge-oriented augmentations and constraints.

Kailin Jiang, Ning Jiang, Yuntao Du, Yuchen Ren, Yuchen Li, Yifan Gao, Jinhe Bi, Yunpu Ma, Qingqing

Liu, Xianhao Wang, Yifan Jia, Hongbo Jiang, Yaocong Hu, Bin Li, and Lei Liu. 2025b. Mined: Probing and updating with multimodal time-sensitive knowledge for large multimodal models.

Jian Lang, Rongpei Hong, Jin Xu, Yili Li, Xovee Xu, and Fan Zhou. 2025. Biting off more than you can detect: Retrieval-augmented multimodal experts for short video hate detection. In *Proceedings of the ACM on Web Conference 2025*, pages 2763–2774.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Fanxiao Li, Jiaying Wu, Canyuan He, and Wei Zhou. 2025. CMIE: Combining MLLM insights with external evidence for explainable out-of-context misinformation detection. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9342–9354.

Xiaojun Li, Xvhao Xiao, Jia Li, Changhua Hu, Junping Yao, and Shaochen Li. 2022. A cnn-based misleading video detection model. *Scientific Reports*, 12(1):6092.

Harold A Linstone, Murray Turoff, et al. 1975. *The delphi method*, volume 1975. Addison-Wesley Reading, MA.

Xuannan Liu, Peipei Li, Huaibo Huang, Zekun Li, Xing Cui, Jiahao Liang, Lixiong Qin, Weihong Deng, and Zhaofeng He. 2024. Fka-owl: Advancing multimodal fake news detection through knowledge-augmented lvlms. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10154–10163.

Xuannan Liu, Zekun Li, Pei Pei Li, Huaibo Huang, Shuhan Xia, Xing Cui, Linzhi Huang, Weihong Deng, and Zhaofeng He. Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms. In *The Thirteenth International Conference on Learning Representations*.

Xinwei Long, Kai Tian, Peng Xu, Guoli Jia, Jingxuan Li, Sa Yang, Yihua Shao, Kaiyan Zhang, Che Jiang, Hao Xu, et al. 2025. Adsqa: Towards advertisement video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23396–23407.

Weihai Lu, Yu Tong, and Zhiqiu Ye. 2025. Dammfnd: Domain-aware multimodal multi-view fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 559–567.

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.

Kaipeng Niu, Danni Xu, Bingjian Yang, Wenxuan Liu, and Zheng Wang. 2025. Pioneering explainable video fact-checking with a new dataset and multi-role multimodal model approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28276–28283.

Priyank Palod, Ayush Patwari, Sudhanshu Bahety, Saurabh Bagchi, and Pawan Goyal. 2019. Misleading metadata detection on youtube. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part II 41*, pages 140–147. Springer.

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.

Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. 2023a. Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14444–14452.

Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. 2024. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13052–13062.

Peng Qi, Yuyang Zhao, Yufeng Shen, Wei Ji, Juan Cao, and Tat-Seng Chua. 2023b. Two heads are better than one: Improving fake news video detection by correlating with neighbors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11947–11959.

Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. 2020. Nlp-based feature extraction for the detection of covid-19 misinformation videos on youtube. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*.

Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2021. A multimodal misinformation detector for covid-19 short videos on tiktok. In *2021 IEEE international conference on big data (big data)*, pages 899–908. IEEE.

Sahar Tahmasebi, Eric Müller-Budack, and Ralph Ewerth. 2024. Multimodal misinformation detection using large vision-language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2189–2199.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025a. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

GLM-V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Wenkai Li, Wei Jia, Xin Lyu, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuxuan Zhang, Zhanxiao Du, Zhenyu Hou, Zhao Xue, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. 2025b. Glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning.

Yu Tong, Weihai Lu, Xiaoxi Cui, Yifan Mao, and Zhejun Zhao. 2025. Dapt: Domain-aware prompt-tuning for multimodal fake news detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 7902–7911.

Yu Tong, Weihai Lu, Zhe Zhao, Song Lai, and Tong Shi. 2024. Mmdfnd: Multi-modal multi-domain fake news detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1178–1186.

Khoa-Dang Tran. 2025. Explainable manipulated videos detection using multimodal large language models. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 725–728.

Herun Wan, Shangbin Feng, Zhaoxuan Tan, Heng Wang, Yulia Tsvetkov, and Minnan Luo. 2024. Dell: Generating reactions and explanations for llm-based misinformation detection. *arXiv preprint arXiv:2402.10426*.

Herun Wan, Jiaying Wu, Minnan Luo, Xiangzheng Kong, Zihan Ma, and Zhi Zeng. 2025. Difar: Enhancing multimodal misinformation detection with diverse, factual, and relevant rationales. *arXiv preprint arXiv:2508.10444*.

Bing Wang, Bingrui Zhao, Ximing Li, Changchun Li, Wanfu Gao, and Shengsheng Wang. 2025a. Collaboration and controversy among experts: Rumor early detection by tuning a comment generator. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 468–478.

Han Wang, Tan Rui Yang, Usman Naseem, and Roy Ka-Wei Lee. 2024a. Multihateclip: A multilingual benchmark dataset for hateful video detection on youtube and bilibili. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7493–7502.

Longzheng Wang, Xiaohan Xu, Lei Zhang, Jiarui Lu, Yongxiu Xu, Hongbo Xu, Minghao Tang, and Chuang Zhang. 2024b. Mmidr: Teaching large language model to interpret multimodal misinformation via knowledge distillation. *arXiv preprint arXiv:2403.14171*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. 2025b. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

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

Zhengjia Wang, Danding Wang, Qiang Sheng, Juan Cao, Siyuan Ma, and Haonan Cheng. 2025d. Exploring news intent and its application: A theory-driven approach. *Information Processing & Management*, 62(6):104229.

Jiaying Wu, Fanxiao Li, Zihang Fu, Min-Yen Kan, and Bryan Hooi. 2026. Seeing through deception: Uncovering misleading creator intent in multimodal news with vision-language models. In *The Fourteenth International Conference on Learning Representations*.

Kaixuan Wu, Yanghao Lin, Donglin Cao, and Dazhen Lin. 2024. Interpretable short video rumor detection based on modality tampering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9180–9189.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786.

Cheng Xiong, Gengfeng Zheng, Xiao Ma, Chunlin Li, and Jiangfeng Zeng. 2025. Delphiagent: A trustworthy multi-agent verification framework for automated fact verification. *Information Processing & Management*, 62(6):104241.

Qingzheng Xu, Heming Du, Szymon Łukasik, Tianqing Zhu, Sen Wang, and Xin Yu. 2025a. Mdam3: A misinformation detection and analysis framework for multitype multimodal media. In *Proceedings of the ACM on Web Conference 2025*, pages 5285–5296.

Wenyan Xu, Dawei Xiang, Tianqi Ding, and Weihai Lu. 2025b. Mmm-fact: A multimodal, multi-domain fact-checking dataset with multi-level retrieval difficulty. *arXiv preprint arXiv:2510.25120*.

Zhou Yang, Yucai Pang, Bin Yang, Haoyang Zhang, and Yunpeng Xiao. 2026. Dr-dgrnet: A model for intent-based disinformation recognition using dynamic graph representation learning. *IEEE Transactions on Computational Social Systems*.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134.

Zhenrui Yue, Huimin Zeng, Yimeng Lu, Lanyu Shang, Yang Zhang, and Dong Wang. 2024. Evidence-driven retrieval augmented response generation for online misinformation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5628–5643.

Zhi Zeng, Minnan Luo, Xiangzheng Kong, Huan Liu, Hao Guo, Hao Yang, Zihan Ma, and Xiang Zhao. 2024. Mitigating world biases: A multimodal multiview debiasing framework for fake news video detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6492–6500.

Zhi Zeng, Jiaying Wu, Minnan Luo, Xiangzheng Kong, Zihan Ma, Guang Dai, and Qinghua Zheng. 2025a. Understand, refine and summarize: Multiview knowledge progressive enhancement learning for fake news video detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9216–9225.

Zhi Zeng, Jiaying Wu, Minnan Luo, Herun Wan, Xiangzheng Kong, Zihan Ma, Guang Dai, and Qinghua Zheng. 2025b. Imol: Incomplete-modality-tolerant learning for multi-domain fake news video detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30921–30933.

Zhi Zeng, Yifei Yang, Jiaying Wu, Xulang Zhang, Xiangzheng Kong, Herun Wan, Zihan Ma, and Minnan Luo. 2026. From manipulation to mistrust: Explaining diverse micro-video misinformation for robust debunking in the wild. *arXiv preprint arXiv:2603.25423*.

Fanrui Zhang, Dian Li, Qiang Zhang, Junxiong Lin, Jiahong Yan, Jiawei Liu, Zheng-Jun Zha, et al. 2025. Fact-r1: Towards explainable video misinformation detection with deep reasoning. *arXiv preprint arXiv:2505.16836*.

Xiaofan Zheng, Zinan Zeng, Heng Wang, Yuyang Bai, Yuhan Liu, and Minnan Luo. 2025. From predictions to analyses: Rationale-augmented fake news detection with large vision-language models. In *Proceedings of the ACM on Web Conference 2025*, pages 5364–5375.

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

Linlin Zong, Jiahui Zhou, Wenmin Lin, Xinyue Liu, Xianchao Zhang, and Bo Xu. 2024. Unveiling opinion evolution via prompting and diffusion for short video fake news detection. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10817–10826.

A Appendix

B MisVideoQA Pipeline Details

B.1 Data Annotation

Unlike prior benchmarks that rely solely on binary veracity labels (Qi et al., 2023a; Bu et al., 2024) or use synthetic content (Xu et al., 2025a; Liu et al.), MisVideoQA adopts a fine-grained, multi-dimensional annotation framework grounded in factual evidence (Figure 2). In addition to binary Real/Fake labels, each sample receives a *fine-grained attribution label* describing the mechanism of deception.

Our annotation follows a four-stage reasoning process, with 12 subtypes capturing distinct deceptive strategies:

- **Stage 1: AI-Generated Content (AIGC).** Identify whether the micro-video is (1) **AIGC**, such as content synthesized or heavily edited using generative AI tools to simulate real-world events or evoke emotional reactions.
- **Stage 2: Multimodal Manipulation.** Detect inconsistencies or falsifications across modalities: (2) **Text Tampering (TT)** modifies captions, titles, or on-screen text to misrepresent the visual or factual content. (3) **Video Tampering (VT)** alters or splices video segments to visually distort the original narrative. (4) **Audio Tampering (AT)** manipulates voiceovers, background sounds, or overlays to fabricate claims or emotional cues.
- **Stage 3: Cognitive Biases.** Capture psychological or rhetorical strategies used to influence perception: (5) **Faulty Logic (FL)** introduces misleading causal or correlational reasoning, such as false analogies or post hoc conclusions. (6) **Exaggerated Narration (EN)** employs overstated or sensational language to heighten engagement. (7) **Offensive Content (OC)** leverages implicit hate speech or personal attacks to evoke moral outrage.
- **Stage 4: Out-of-Context Manipulation.** Identify cases where authentic material or partial truths are used deceptively: (8) **Knowledge Error (KE)** misinterprets legitimate information, often framed with pseudo-scientific or misleading narratives. (9) **Event Fabrication (EF)** invents events without factual basis, often supported by fabricated visuals or commentary. (10) **Event Splicing (ES)** combines unrelated real-world clips or scenes to construct a false narrative.
- **Stage 5: Non-Misinformation.** We divide rest of non-Misinformation into two types: (11) **Real (RL)** Videos that accurately reflects factual information and follows correct logical reasoning. (12) **Debunking (DB)** videos debunk misinformation or conducts fact checking of false claims.

Representative examples with corresponding debunking evidence are provided in Figure 1. Approximately 1.3% of micro-videos that could not be confidently categorized were excluded. Each sample was independently annotated by at least three experts, and final labels were determined through unanimous consensus. The annotation experts included twelve individuals with academic or master's degrees in computer science and social science.

B.2 QA Generation Details

For QA1–QA3, we construct a chain of misinformation-attribution reasoning questions. Specifically, QA1 and QA2 are generated via attribution-guided prompting conditioned on the annotated misinformation attribution in QA3, and are designed to be specific and concrete, covering factual dimensions such as *when*, *where*, *who*, *what*, and *how*. QA3 targets diverse forms of deception aligned with different categories of video misinformation attribution, as detailed in Appendix B.1.

For QA4–QA6, we define a set of major societal influence aspects to facilitate higher-level reasoning over misinformation, including emotion recognition (Hu et al., 2025), creator intent (Wang et al., 2025d), and persuasion strategies (Wan et al., 2024). These aspects are designed to capture the underlying communicative and manipulative intents of misinformation content. The selected aspects are summarized as follows:

Emotion Categories
Disgust, Curiosity, Sadness, Skepticism
Happy, Hope, Surprise, Relief
Excited, Anxiety, Angry, Fear
Sarcastic, Serious, Joyful, Neutral

Model	Model Card
GPT-4o (Hurst et al., 2024)	`gpt-4o-2024-08-06`
GPT-4o-mini (Hurst et al., 2024)	`gpt-4o-mini-2024-07-18`
InternVL3-8B (Zhu et al., 2025)	`InternVL3-8b-Instruct`
LLaVA-OneVision-7B (Li et al., 2024)	`llava-onevision-qwen2-7b-ov`
GLM-4.1V (Team et al., 2025b)	`GLM-4.1V-9B-Thinking`
Gemma3-12B (Team et al., 2025a)	`gemma-3-12b-it`
InternVL-3.5-14B (Wang et al., 2025b)	`InternVL3_5-14B`
InternVL3-38B (Zhu et al., 2025)	`InternVL3-38B`
Qwen3-VL-32B (Bai et al., 2025a)	`Qwen3-VL-32B-Instruct`
VideoLLaMA2(Cheng et al., 2024)	`VideoLLaMA2-7B-Base`

Table 4: Model cards for all MLLMs evaluated throughout our work.

Intent Categories
Public interest,
Economic interest,
Political interest,
Psychological fulfillment

Persuasion Categories
Political Polarization,
Appeal to Authority,
Social Proof,
Storytelling,
Bandwagon Effect,
Rhetorical Questioning,
Logical Reasoning,
Narrative Engagement,
Emotional Appeal

C Experimental Settings

Baselines. To evaluate both detection performance and explainability across diverse sources and misinformation types, we benchmark representative MLLMs. These include InternVL-2.5 (Chen et al., 2025b), Qwen2.5-VL (Bai et al., 2025b), Qwen3-VL (Bai et al., 2025a), GLM-4.1V (Team et al., 2025b), VideoLLaMA2 (Cheng et al., 2024), Gemma3 (Team et al., 2025a), InternVL3 (Zhu et al., 2025), LLaVA-OneVision (Li et al., 2024), GPT-4o (Hurst et al., 2024) and o1. In addition, we implement enhanced inference variants based on the Video-of-Thought (VoT) paradigm (Fei et al., 2024), which augments temporal reasoning for video-based misinformation understanding and reasoning. With MISVIDEOQA, we systematically evaluate 10 representative MLLMs spanning different families, sizes, and access levels. Detailed model names and versions are listed in Table 4.

Evaluation Metrics. In the era of MLLMs, video misinformation understanding and reasoning requires assessing understanding and reasoning abilities. We report Accuracy (Acc) for overall QAs, following prior work (Cao et al., 2025; Hu et al., 2025) (see Table 2). To further evaluate performance across different misinformation categories, we also report Micro-Accuracy for each type of QAs (Table 2).

Implementation Details. To enable the fair evaluation, we set the sampling hyperparameter of the off-the-shelf MLLMs, "do_sample = False" or "Temperature = 0", to guarantee consistency in the prediction outputs. For each micro-video, we uniformly sample 16 frames and uses Qwen2-Audio (Chu et al., 2024) to extract audio captioning for audio inputs. For our MISAGENT, we utilize Qwen3-VL (Bai et al., 2025a), InternVL-3 (Zhu et al., 2025) and GLM-4.1V (Team et al., 2025b) as the MLLM backbone, to support comprehensive multimodal understanding. In the reasoning agent, we set "Top-K = 3". All experiments are conducted on eight NVIDIA A100 GPUs, each equipped with 80 GB of memory.

D Prompts for MLLMs

MLLMs possess broad world knowledge and demonstrate strong generalization across diverse multimodal tasks (Bi et al., 2025; Jiang et al., 2025b). To evaluate their effectiveness in video misinformation understanding and reasoning, we employ carefully designed prompt templates. The specific prompts used for all baseline models are detailed below.

Prompt of the MLLMs(Direct)

Text Prompt: You are an expert system for video misinformation analysis. You are given a video represented by multiple frames and a description. Answer the following multiple choice questions. You MUST return the result in EXACTLY the following JSON format. Do not include any extra text.

```
{
  "answers": {
    "Q1": "A",
    "Q2": "B"
  }
}
```

News Text: {news title}
Video: {video frames}
Audio: {audio transcription}

Prompt of the MLLMs(VoT)

Text Prompt 1 (Background Analysis): You are an expert system for video misinformation analysis. Perform the following steps strictly in order. Step 1 Background Analysis Describe the scene, environment, objects, filming conditions, and visual context. Only state observable facts.
Text Prompt 2 (Entity and Behavior Analysis): Based on the analyses above, identify people or entities and analyze gestures, lip motion, eye gaze, and physical consistency. Point out any suspicious or unnatural patterns.
Text Prompt 3 (Event Analysis): Based on the above analyses, describe the event in the micro-video and analyze temporal continuity, action coherence, and possible editing or manipulation artifacts.
Text Prompt 4 (Answer Verification): Given the video frames and the accompanying title, now you need to verify the previous answer. You are given a video represented by multiple frames and a description. Answer the following multiple-choice questions. You MUST return the result in EXACTLY the following JSON format. Do not include any extra text.

```
{
  "answers": {
    "Q1": "A",
    "Q2": "B"
  }
}
```

News Text: {news title and content}
Video: {video frames}
Audio: {audio transcription}

Prompt of the MISAGENT

Text Prompt 1 (Content Analyst Agent): You are an expert investigator. Analyze the provided frames strictly following this **Structure of Thought**. Do not rush. Think step by step. Describe the scene, lighting, and objects. Is it high-quality footage or blurry/grainy? Analyze human behavior, facial expressions, and body language. Any signs of AI generation (hands, eyes)? describe the event in the video and analyze temporal continuity, action coherence, and possible editing or manipulation artifacts.

Text Prompt 2 (Reasoning Agent): You are a professional micro-video reasoning expert, please analyze: What emotion is being projected? (e.g., Panic, Anger, Happy) Is the video objective, or does it push a specific agenda/bias? Is the goal to inform, mislead, or provoke? Does it use fear inducement or emotional appeals? Based on the above, given specific keywords and core claims, retrieve external evidence or event details 🔍 that support or refute the understanding and reasoning of the content of the micro-video.

Text Prompt 3 (Reflection Agent):

- Consistency Check: Does the External Evidence confirm the Visual Analysis? If evidence is 'No results', rely on the Entity & Behavior and Intent analysis from reasoning agent
- Evidence Reflection: Evaluate whether the above logic is correct. If it is not correct, revise and improve it.
- Answer Generation: The answer should incorporate the synthesized logical framework above where appropriate.

Answer the following multiple-choice questions. Return ONLY a valid JSON object:

```
{
  "answers": {
    "Q1": "A",
    "Q2": "B"
  }
}
```

News Text: {news title and content}
Video: {video frames}
Audio: {audio transcription}

E Error Analysis

Although MISAGENT consistently improves video misinformation understanding and reasoning, it still fails on cases where correct perceptual understanding does not translate into correct misinformation attribution and pragmatic interpretation (Figure 6). In the solar flare example, models correctly answer the perception question about the physical impact of solar activity, yet MISAGENT misclassifies the video as *Real* rather than a *Knowledge Error*, likely because the claim implicitly links a solar flare to immediate symptoms such as dizziness, headaches, and insomnia, which requires specialized domain verification. Similarly, MISAGENT predicts *Curiosity* as the primary emotion, while the narrative framing is closer to fear or anxiety, indicating a gap in modeling affective cues and rhetorical framing beyond factual content.

These errors suggest that stronger domain aware grounding is needed, especially for health or science related claims, and that attribution and emotion oriented reasoning should be explicitly tied to verified evidence rather than inferred solely from surface semantics. Future work could explore adaptive retrieval that prioritizes domain specific sources, tighter evidence claim alignment for subjective or causal statements, and continual updating mechanisms that improve both veracity attribution and persuasion aware interpretation in rapidly evolving information environments.

F More Cases Analysis

Figure 7 shows two representative real world examples that further illustrate MISAGENT strengths and weaknesses across the progressive reasoning dimensions of MISVIDEOQA. In the debunking case (Figure 7(a), the video refutes the rumor that spraying perfume and smoking in a vehicle will immediately cause an explosion, and provides a plausible explanation that the accident was related to gas cylinders. MISAGENT correctly performs content reasoning and attributes the video to *Debunking*, while its post detection outputs are also consistent with the corrective framing, including recognizing *Fear* as the dominant emotion, inferring *Public interest* as the intent, and selecting *Logical reasoning* as the persuasion strategy. This case suggests that when the video contains explicit corrective cues and the evidence is easy to align with concrete events, MISAGENT can deliver coherent attribution and interpretation.

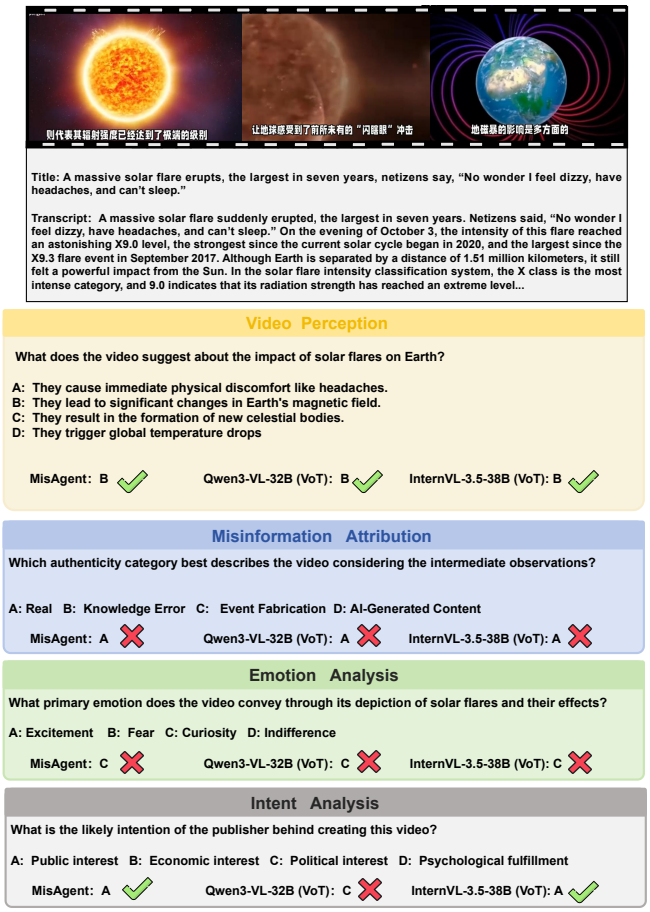
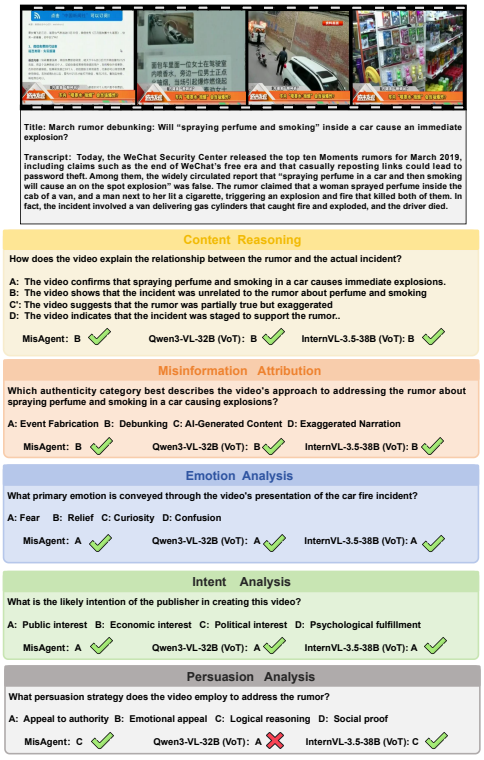


Figure 6: An error case. A real-world example highlighting the limitation of MISAGENT in handling domain-specific understanding and reasoning.

In the knowledge error case (Figure 7(b)), the video discusses the use of copper sulfate in fish farming and frames it as a potential safety risk, where the core issue lies in public confusion about its legitimate usage, dosage constraints, and possible misuse. MISAGENT correctly identifies the added substance at the perception level and attributes the video to *Knowledge Error*, and its post detection judgments, such as *Anxiety* as the primary emotion and *Public interest* as the publisher intent, are aligned with the video's warning oriented narrative. By contrast, direct MLLM baselines tend to misclassify the case as *Event Fabrication* and are less reliable in intent inference, suggesting that domain grounded clarification and evidence aligned attribution are crucial for distinguishing knowledge misconceptions from fabricated events.



(a) A debunking case.

(b) A knowledge error case.

Figure 7: More cases analysis. Two real-world examples highlighting the reasoning ability of MISAGENT in handling debunking and knowledge error videos.