CSI: An Investigative Multi-Agent Framework for Explainable Short Video Fake News Detection

Yuxin Wang* **Yang Yang*** **Huaiwen Zhang†**
Inner Mongolia University
yuxinwang@mail.imu.edu.cn, {yang.yang, huaiwen.zhang}@imu.edu.cn

Abstract

The proliferation of short video fake news threatens social stability. Current detection methods rely either on black-box Multimodal Small Language Models (MSLMs), which suffer from poor explainability and superficial understanding, or on specific prompt strategies for Multimodal Large Language Models (MLLMs) that underutilize their reasoning capabilities and knowledge. To address these challenges, we propose a novel multi-agent framework named CSI for short video fake news detection. CSI implements two key units: 1) Multimodal Forensics Unit (MFU), which performs synchronous multimodal deconstruction and external knowledge retrieval to collect comprehensive evidence. 2) Case Review Unit (CRU), which first employs collaborative discussion to facilitate viewpoint interaction to obtain the review result. Subsequently, the Adjudicator integrates evidence and the review result via multiple attention mechanisms to interact with the news, ensuring a robust verdict. Extensive experiments on two real-world datasets demonstrate that CSI provides rigorous explanations while achieving state-of-the-art performance. Our code is available at: https://github.com/VFCenter/CSI.

1 Introduction

Short video platforms like TikTok and Kuaishou dominate global social media but also facilitate misinformation propagation (Sun et al., 2020). Such short video fake news is generally defined as misleading multimodal content (video, audio, text) contradicting objective facts (Qi et al., 2023a). The multimodal nature and scene transitions in short videos significantly increase the complexity and ambiguity of both content and logic (Yu et al., 2025). Disinformation creators exploit these characteristics via editing, emotional manipulation, and tampering (Bu et al., 2024) to mislead the public.

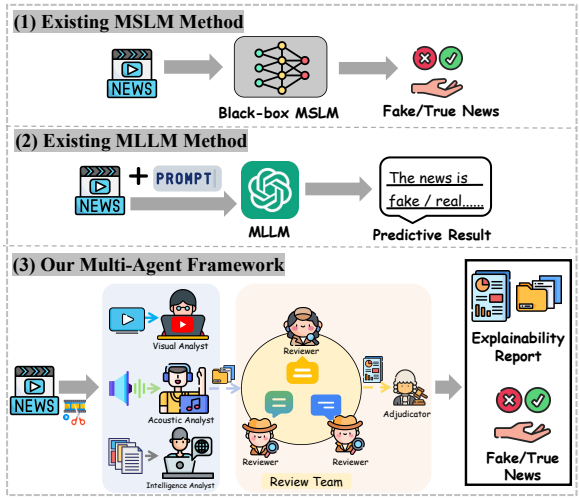


Figure 1: Illustrations of existing methods and our proposed multi-agent framework.

As illustrated in Figure 1, existing methods mainly rely on black-box MSLM for classification only (Bu et al., 2024; Qi et al., 2023a,b; Zhang et al., 2025a) or employ MLLM with specific prompts to directly generate results (Xu et al., 2025b; Tahmasebi et al., 2024). While task-specific MSLMs achieve outstanding performance in classification, their black-box nature and limited knowledge constrain deep logical reasoning and explainability, significantly undermining user trust. To overcome these limitations, recent works utilize MLLM with specific prompts to directly reason about short video news. However, when encountering cross-modal conflicting content, the standalone MLLMs tend to over-rely on unimodal or spurious cross-modal correlations, leading to hallucinations and sharp performance degradation (Bie et al., 2025; Zhang et al., 2025b). Moreover, their isolated nature precludes the integration of diverse viewpoints, resulting in premature and one-sided conclusions (Liu et al., 2025; Minaee et al., 2025). The gap between general pre-training and specific

* Equal contribution.
† Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 25508–25528

classification tasks means that MLLMs often underperform MLSMs when making decisions via comprehensive analysis (Zheng et al., 2025; Hu et al., 2024). This gap arises from the mismatch between MLLM training corpora and task demands, and their focus on multimodal understanding and generation rather than precise classification. Although fine-tuning MLLMs can enhance decision-making capabilities, the high costs of data, computation, and annotation limit their practicality.

To address these challenges, drawing inspiration from human collaborative investigative processes, we propose a novel multi-agent framework named CSI, **C**ollaborative **S**hort-video **I**nvestigation for fake news detection. First, within the Multimodal Forensics Unit (MFU), multiple independent, modality-specific agents synchronously deconstruct information from their respective modalities and aggregate all findings into a structured master casefile. This specialized division of labor effectively mitigates the interference of conflicting cross-modal content on the reasoning capabilities of MLLMs. Second, the Case Review Unit (CRU) deliberates on the compiled master casefile. By employing a collaborative discussion mechanism, the unit facilitates a thorough exchange of diverse viewpoints to derive the review result, thereby overcoming the perspective limitations inherent in standalone MLLMs. Finally, the Adjudicator employs multiple attention mechanisms to integrate the master casefile and the review result with the original news to reach a robust verdict. This adjudicator utilizes the insightful knowledge of MLLMs to efficiently guide the analysis of original news and enhance the accuracy of detection.

Our contributions are summarized as follows:

- We introduce CSI, a novel multi-agent framework for short video fake news detection that achieves rigorous explainability and precise detection.
- We propose a novel investigative reasoning architecture. Through multi-agent collaboration, it employs specialized multimodal digital forensics and structured discussion to fully leverage the reasoning capabilities of MLLMs.
- We design an efficient adjudicator that integrates the insightful knowledge with the news to make precise predictions, avoiding expensive fine-tuning of MLLMs and achieving state-of-the-art performance on two real-world datasets.

2 Related Work

2.1 Short Video Fake News Detection

Traditional short video fake news detection methods typically rely on MSLMs to classify news using multimodal information (Bu et al., 2024; Qi et al., 2023b; Qian et al., 2021). For instance, SVFEND (Qi et al., 2023a) enhances content representation by modeling cross-modal correlations and incorporating social context. Although these approaches achieve certain performance improvements, they are limited by insufficient external knowledge and reasoning capability, making it difficult to provide convincing explanations. Thus, human verification is still required to ensure accountability. With the rapid development of MLLMs, recent methods (Wang et al., 2025a) such as ExMRD (Hong et al., 2025) attempt to leverage an independent MLLM to reason over refined visual and textual inputs. However, these approaches do not fully utilize the reasoning potential. We construct a multi-agent system composed of multiple MLLMs, introducing modality-aware role specialization, structured deliberation, and a decision mechanism to enhance explainable short video fake news detection.

2.2 Multi-Agent Systems

Multi-Agent Systems (MAS) (Park et al., 2023) are developed to handle distributed complex tasks, enhancing problem-solving capabilities through a division of labor and collaboration among multiple agents. In the domain of fake news detection, TED (Liu et al., 2025) employs LLM-based agents to simulate the debate process and uses the attention mechanism to simulate the interaction between role embedding and debate to draw the final detection result. However, current MAS only use LLMs for unimodal fake news detection and have not yet explored the capability to detect more complex short video news. In contrast, we introduce a MAS that further enhances explainability and accuracy of short video fake news detection through specialized multimodal digital forensics, discussion mechanism, and effective Adjudicator.

3 Method

3.1 Problem Formulation

Let $\mathcal{D} = \{S_1, S_2, \ldots, S_N\}$ denote a collection of real-world short video news, where each sample S is represented as $S = (V, A, T)$, comprising the raw visual, auditory, and textual modalities, respec-

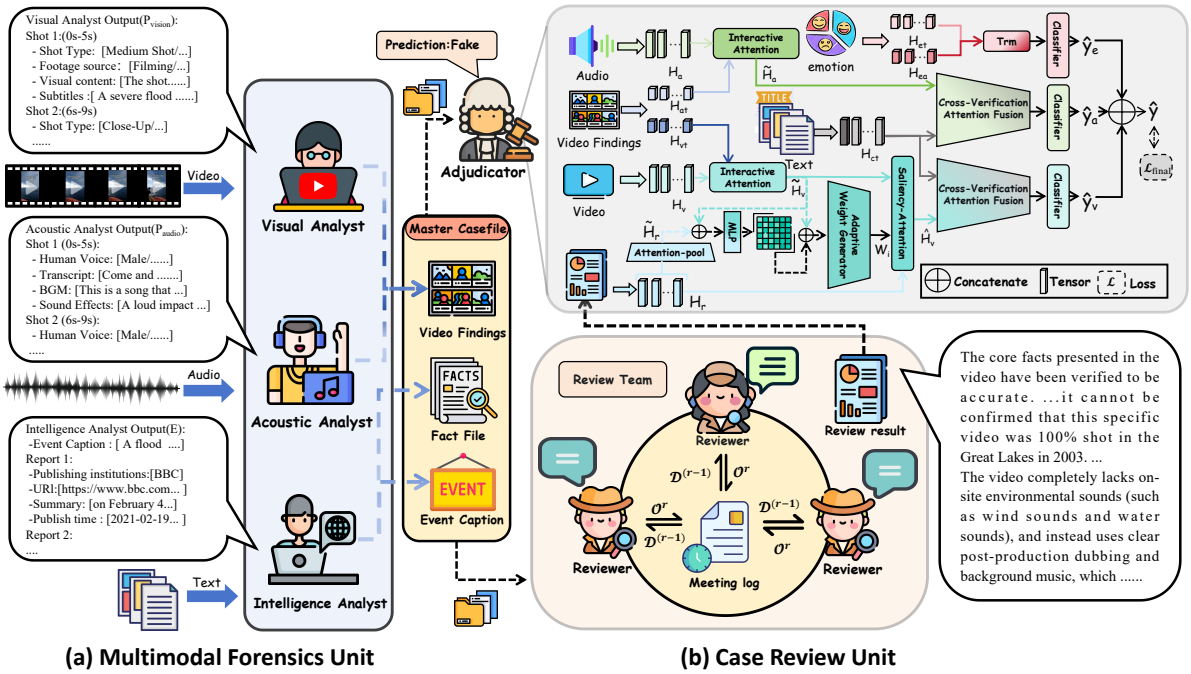


Figure 2: An overview of our framework, CSI, for explainable short video fake news detection.

tively. In this work, we reformulate short video fake news detection as a collaborative investigation task, aiming to synthesize comprehensive clues for robust decision-making while providing rigorous explanations. Within our CSI framework, MLLM-based agents are organized into two units. The first unit is tasked with analyzing the raw modalities (V, A, T) to compile a comprehensive master casefile $\mathcal{C}$. The second unit engages in deliberation based on $\mathcal{C}$ to generate a review result R. Finally, the predicted score $\hat{y}$ is determined by integrating the fine-grained insights from $\mathcal{C}$ and R with the original input S. This new paradigm shifts the objective from mere prediction to a more comprehensive and explainable detection process, simultaneously delivering accurate verdicts and detailed justifications. The detailed introduction to the algorithm can be found at §H.

3.2 Multimodal Forensics Unit

The Multimodal Forensics Unit (MFU) comprises modality-specific analysts tasked with collecting comprehensive evidence simultaneously.

• **Preprocessing**: First, shot boundary detection is performed to facilitate the deconstruction of narrative structures and improve MLLMs' understanding of multi-shot videos (Han et al., 2025). We employ the shot boundary detection model TransNetV2 (Soucek and Lokoc, 2024) to segment V. It outputs the start frame f_s^i and end frame f_e^i for each detected shot i of the V. The segmentation result is expressed as $\mathcal{V}_f = \{(f_s^1, f_e^1), \ldots, (f_s^m, f_e^m)\}$, where m represents the total number of detected shots, and $\mathcal{A}_f$ represents the audio segmentation result aligned with $\mathcal{V}_f$.

• **Visual Analyst**: The Visual Analyst generates a structured visual analysis P_{vision} by deconstructing the visual content within each clip based on $\mathcal{V}_f$. The agent targets five key aspects, selected for their significance in visual analysis and manipulation mode (Wang, 2021; Qin et al., 2025): (1) shot type, which reveals narrative and emotional intent; (2) visual content, which serves as the basis for verifying factual and cross-modal consistency; (3) subtitles, which are the primary vehicle for informational distortion; (4) footage source, which verifies authenticity against repurposed or synthetic media; and (5) Technical authenticity, which uncovers subtle digital traces and blending artifacts.

• **Acoustic Analyst**: The Acoustic Analyst agent generates a structured auditory analysis P_{audio} by deconstructing the auditory information within each audio segment $\mathcal{A}_f$. The agent targets four key aspects, selected for their central role in auditory narrative (Barbosa and Dizon, 2020): (1) human voice, which helps identify the speaker; (2) voice transcription, which converts speech into text, enabling direct fact-checking; (3) background music (BGM), which creates a specific atmosphere and guides the audience's emotions; and (4) sound ef-

fects, which provide crucial contextual clues about the depicted events. Ultimately, P_{vision} and P_{audio} are compiled into the unified video findings P.

• **Intelligence Analyst**: The Intelligence Analyst agent employs a search engine $\mathcal{W}$ [1] to retrieve credible evidence. Since textual cues (e.g., location, date) provide the most explicit event grounding (Li et al., 2022), we prioritize the text modality for query formulation. The retrieval process operates in three stages: 1) First, the agent distills an event caption T_c from the noisy raw text T (Title and On-screen Text) as a query to focus on the core event. The T_c is submitted to $\mathcal{W}$ to obtain an initial candidate set $\mathcal{D}_{\text{raw}}$. 2) Next, the agent re-ranks the documents in $\mathcal{D}_{\text{raw}}$ based on a three-tier source authority hierarchy: *Official Institutions* (Tier 1) > *International Media* (Tier 2) > *Domestic Media* (Tier 3), and select the top-k most authoritative documents $\mathcal{D}_{\text{top-k}}$. 3) Finally, the agent filters out events in $\mathcal{D}_{top-k}$ that are irrelevant to T_c and summarizes the remainder into a structured credible evidence set E. Subsequently, E, T_c, and P are integrated into the master casefile $\mathcal{C}$.

3.3 Case Review Unit

The Case Review Unit (CRU) consists of the Review Team and the Adjudicator, which aims to conduct group discussions on the master casefile $\mathcal{C}$ to provide the review result R and ultimately the adjudicator gives the predicted score $\hat{y}$.

• **Review Team**: The discussion process is first orchestrated by the Review Team, composed of 3 reviewer agents, denoted as $\mathcal{A}_r = \{a_1, a_2, a_3\}$. Each agent is assigned a distinct role $Role_i$. They are respectively adept at (1) verification of facts and common sense (also acting as the team leader), (2) cross-modal consistency verification, and (3) verification of logic and editing rationality of video clips. The entire process follows a roundtable conference format. It is divided into three stages, denoted as $\Phi = \{\phi_1, \phi_2, \phi_3\}$. The outputs of agents in each round are recorded in a text-based meeting log M. At the end of each round, the M containing the statements of everyone from the previous round, is shared among all agents via the shared-context communication protocol (Hong et al., 2024).

In the **Initial Response stage** (ϕ_1), each agent independently and synchronously scrutinizes the $\mathcal{C}$, and guided by the task of its $Role_i$, generates an initial analysis $O_i^{(1)}$:

$$O_i^{(1)} = f_{init}(C, Role_i), \quad i = 1, \ldots, N \tag{1}$$

where $O_i^{(1)}$ is the initial output of a_i, f_{init} represents the initial analysis prompt function implemented by the MLLM for a_i.

In the **Discussion stage** (ϕ_2), the process through rounds $r = 1, \ldots, R_{\max}$. At each round r, each agent reviews the set of analyses $\mathcal{D}^{(r-1)} = \{O_1^{(r-1)}, \ldots, O_{i-1}^{(r-1)}, O_{i+1}^{(r-1)}, \ldots, O_N^{(r-1)}\}$ generated by other agents in the preceding round to refine its response:

$$O_i^{(r)} = f_{\text{dis}}(\mathcal{D}^{(r-1)}, Role_i), \quad i = 1, \ldots, N \tag{2}$$

where $O_i^{(r)}$ is the output of a_i at discussion round r, f_{dis} represents the discussion prompt function implemented by the MLLM for a_i. In the **Team Answer Generation stage** (ϕ_3), the team leader aggregates the complete meeting log M to generate a comprehensive review result R:

$$R = f_{\text{team}}(M, Role_L) \tag{3}$$

where f_{team} represents the team answer synthesis prompt function implemented by the MLLM for the team leader ($Role_L$).

Finally, R, C, and S are passed to the Adjudicator for the final prediction.

• **Adjudicator.** The Adjudicator composed of the MSLM derives the final prediction $\hat{y}$ starting from features: visual H_v, acoustic H_a, and textual features (news text H_t, visual findings H_{vt}, acoustic findings H_{at}, event caption H_c, review result H_r) in $\mathbb{R}^{k_t \times d_t}$, alongside textual and acoustic sentiment features $H_{et}, H_{ea} \in \mathbb{R}^{d_e}$.

In the **Unimodal Cue Guidance** stage, we employ an Interactive-Attention mechanism, implemented via Multi-Head Attention (MHA), to enhance ($H_m, m \in \{v, a\}$) using their corresponding findings H_{mt}:

$$\tilde{H}_m = \text{MHA}(H_m, H_{mt}, H_{mt}). \tag{4}$$

Subsequently, $\tilde{H}_v$ undergoes a Saliency-Attention mechanism using the H_r to highlight suspicious frames. We compute a global representation $\bar{H}_r$:

$$\bar{H}_r = \text{MHA}(\text{Mean}(H_r), H_r, H_r) \tag{5}$$

In order to derive frame-level relevance scores $s_i = \sigma(\text{MLP}_{\text{sim}}([\tilde{h}_{v,i}; \bar{H}_r]))$. Subsequently, these

[1] https://www.googleapis.com/customsearch/v1

scores generate adaptive weights w_i, scaled to $[w_{\min}, w_{\max}]$ and amplified by s_i:

$$w_i = \Big(w_{\min} + (w_{\max} - w_{\min}) \cdot \sigma(\text{MLP}_w([\tilde{h}_{v,i}; s_i]))\Big) \cdot (1 + s_i). \tag{6}$$

Finally, w modulates the attention scores (via $\text{softmax}(\frac{QK^T}{\sqrt{d_k}} \odot w)$) to produce cue features $\hat{H}_v$:

$$\hat{H}_v = \text{LayerNorm}(\tilde{H}_v + \text{MHA}_{\text{weighted}}). \tag{7}$$

In the **Cross-modal Mutual Verification** stage, we fuse the $\hat{H}_v$ and $\tilde{H}_a$ with the complete text $H_{ct} = [H_c; H_t]$ via Cross-Verification Attention, where $[\cdot;\cdot]$ denotes concatenation along the time dimension. For the visual branch:

$$\begin{aligned} \hat{H}_v^* &= \text{MHA}(\hat{H}_v, H_{ct}, H_{ct}), \\ H_{ctv}^* &= \text{MHA}(H_{ct}, \hat{H}_v, \hat{H}_v). \end{aligned} \tag{8}$$

A similar process applies to $\tilde{H}_a$ to yield $(\tilde{H}_a^*, H_{cta}^*)$. The resulting cross-verified feature pairs $(\hat{H}_v^*, H_{ctv}^*)$ and $(\tilde{H}_a^*, H_{cta}^*)$, along with the emotion feature pair (H_{et}, H_{ea}) are then fused. Each pair is first concatenated along the time dimension and then passed through a Transformer (Trm) layer followed by mean-pooling to produce representations f_v, f_a, f_e. They are respectively input into two-layer MLP classifiers to obtain predicted scores $\hat{y}_v$, $\hat{y}_a$, $\hat{y}_e$, which are summed for the final predicted score $\hat{y}$:

$$\hat{y} = (\hat{y}_v + \hat{y}_a + \hat{y}_e) \tag{9}$$

In the **Training** stage, the Adjudicator is trained using a standard cross-entropy loss between the predicted score $\hat{y}$ and the ground-truth label y:

$$\mathcal{L}_{final} = \text{CrossEntropy}(\hat{y}, y). \tag{10}$$

4 Experiments

4.1 Datasets and Metrics

To verify the effectiveness and generalization of CSI, we conduct experiments on two real-world short video datasets: FakeSV (Qi et al., 2023a) and FakeTT (Bu et al., 2024). Following the benchmark (Bu et al., 2024; Hong et al., 2025), we adopt the time split strategy to divide the training set, validation set and test set in proportions of 70%, 15% and 15% respectively to simulate the real-world scenarios on short video platforms. Meanwhile, we use Accuracy, F1-score, Precision and Recall to evaluate the overall performance of these methods. And the G-Eval method (Liu et al., 2023) based on LLM is adopted to evaluate the quality of the explanations in five key dimensions (informative, fluent, readable, persuasive, and sound). To ensure the highest level of rigor, we also employ the ILORA method (Zhou et al., 2021) to conduct a more rigorous manual assessment of the five key aspects (informative, accurate, readable, objective, and logical) and to reduce the potential biases that might exist among the language model evaluators. Appendix §A provides a detailed introduction.

4.2 Baselines

To comprehensively validate the superiority of CSI, we select the 14 competitive baselines and divide them into four categories: (1) *Methods utilizing MLLM* for detection: Qwen2.5-VL (Bai et al., 2025), InternVL3 (Zhu et al., 2025), GPT-4o-mini (OpenAI et al., 2024), Gemini-2.5-Flash (Comanici et al., 2025), FakeSV-VLM (Wang et al., 2025b); (2) *Methods utilizing MSLM* for detection: TikTec (Shang et al., 2021), FANVM (Choi and Ko, 2021), HMCAN (Qian et al., 2021), SVFEND (Qi et al., 2023a), FakingRecipe (Bu et al., 2024); (3) *Methods utilizing MLLMs and MSLMs collaboration* for detection: CA-FVD (Wang et al., 2025a), MSAF-Net (Shen et al., 2025), ExMRD (Hong et al., 2025); (4) *Methods utilizing Multi-Agents*: TED (Liu et al., 2025). Appendix §B provides details about the baselines.

4.3 Performance Comparison

To comprehensively evaluate the performance of our method, we conduct comparisons with 14 competitive baselines. Extensive experiments are performed on two real-world datasets: FakeSV and FakeTT. To ensure the reliability of the results, each experiment is repeated five times, and the average performance is reported. The p-values obtained between CSI and the strongest baseline are all below 0.01, which confirms the statistical significance of CSI's improvement. More details on the experimental settings and implementation can be found in §C. Based on these results in Table 1, we can make the following observations:

First, CSI outperforms all existing methods across all evaluation metrics. Specifically, it improves Accuracy by 2.95% and F1-score by 3.12% on the FakeSV dataset, and by 4.35% and 4.38% on the FakeTT dataset, respectively. These results

Models	FakeSV				FakeTT			
	Accuracy	F1-score	Precision	Recall	Accuracy	F1-score	Precision	Recall
Qwen2.5-VL	64.21	60.79	64.55	61.52	53.18	52.95	56.77	57.35
InternVL3	65.13	61.07	66.46	62.12	52.51	52.23	64.26	62.21
GPT-4o-mini	68.08	68.05	69.88	69.49	61.54	61.20	64.41	65.89
Gemini-2.5-Flash	70.85	70.60	70.84	70.64	66.11	66.84	68.33	70.65
TikTec	73.43	73.26	73.23	73.54	66.22	65.08	65.84	67.87
FANVM	79.88	78.91	80.98	78.42	71.57	70.21	70.22	72.63
HMCAN	79.52	78.81	79.81	78.46	68.56	68.41	72.78	74.72
SVFEND	81.05	81.02	81.24	81.05	77.14	75.63	75.12	77.56
FakingRecipe	85.35	84.83	85.84	84.29	79.15	77.74	77.25	79.80
CA-FVD	85.79	85.28	86.57	84.78	81.61	80.26	79.50	82.17
MSAF-Net	87.45	86.97	88.50	86.40	83.61	81.45	81.54	81.38
ExMRD	86.90	86.52	87.31	86.13	84.28	83.13	82.27	85.19
TED	87.82	87.38	88.79	86.82	85.28	84.17	83.14	84.66
CSI (Ours)	**90.77**	**90.50**	**91.41**	**90.04**	**89.63**	**88.55**	**87.86**	**89.44**
p-values	$1.42e^{-3}$	$3.15e^{-3}$	$2.87e^{-3}$	$2.09e^{-2}$	$3.42e^{-3}$	$4.31e^{-3}$	$5.66e^{-3}$	$4.38e^{-3}$

Table 1: Performance comparison on FakeSV and FakeTT dataset. The best results are highlighted in **bold**.

Models	FakeSV		FakeTT	
	Acc.	F1	Acc.	F1
Qwen2.5-VL-CW	69.37	67.08	65.88	64.35
InternVL3-CW	71.91	71.67	69.90	68.87
GPT-4o-mini-CW	73.43	73.29	70.90	69.41
Gemini-2.5-Flash-CW	79.19	78.83	74.25	73.58
CSI (Ours)	**90.77**	**90.50**	**89.63**	**88.55**

Table 2: Performance comparison between CSI and MLLMs equipped with CoT and Web-RAG (CW). Best results are shown in **bold**.

Models	FakeSV		FakeTT	
	Acc.	F1	Acc.	F1
Qwen2.5-VL (Fine-Tuned)	85.79	85.24	84.62	83.65
InternVL3 (Fine-Tuned)	86.82	86.38	85.95	84.65
FakeSV-VLM	88.75	88.29	87.29	86.13
CSI (Ours)	**90.77**	**90.50**	**89.63**	**88.55**

Table 3: Performance comparison between CSI and fine-tuned MLLMs. Best results are shown in **bold**.

clearly demonstrate the outstanding capability of CSI in short video fake news detection.

Secondly, the performance of zero-shot MLLMs is significantly worse than that of MSLM. This indicates that generating results directly based on prompts is insufficient for detection.

Finally, by integrating the advantages of MLLM and MSLM through a collaborative approach, their combined performance is significantly superior to that of either MSLM or MLLM alone. This indicates that combining the two is effective in this task. And compared to the debate-based TED framework (With Web Search), our multi-agent framework achieves the state-of-the-art detection performance through specialized multimodal digital forensics, discussion and the adjudication mechanism.

To ensure experimental fairness, we equip MLLMs with Chain of Thought (CoT) (Wei et al., 2022) and Web-based Retrieval-Augmented Generation (Web-RAG) (Komeili et al., 2022), retrieving only information published prior to the video release. Table 2 shows that integrating CoT and Web-RAG significantly improves MLLM performance over zero-shot methods. This indicates that enhancing reasoning capabilities and external knowledge retrieval is highly effective for detection.

To comprehensively validate our approach, we fine-tune MLLMs with web-RAG on FakeTT and FakeSV datasets. Table 3 indicates that fine-tuning improves performance via domain adaptation. However, compared to our multi-agent framework, they suffer from a single reasoning perspective and high training costs, which hinders performance and scalability.

4.4 Ablation Study

Components Ablation Analysis. In Table 4, the absence of any agent in the MFU will lead to a decline in performance, which proves that each agent collects high-quality objective cues from external documents and multimodal content. And it further

Variant	FakeSV		FakeTT	
	Acc.	F1	Acc.	F1
CSI (Full Model)	**90.77**	**90.50**	**89.63**	**88.55**
w/o Visual Analyst	85.24	84.89	83.61	82.60
w/o Acoustic Analyst	86.53	86.24	84.28	82.51
w/o Intelligence Analyst	87.82	87.38	85.95	84.65
w/o Review Team	86.90	86.52	85.28	84.20
w/o MLLMs outputs	80.88	80.54	79.93	78.77
MLP-based Adjudicator	84.62	84.61	81.27	80.15
MLLM-based Adjudicator	85.98	85.65	83.61	82.60

Table 4: Ablation study of the components in CSI. We adopt Accuracy (Acc.), F1-score (F1) for evaluation.

Variant	FakeSV		FakeTT	
	Acc.	F1	Acc.	F1
CSI (Full Model)	**90.77**	**90.50**	**89.63**	**88.55**
CSI-UA	89.30	88.94	88.63	87.58
CSI-UR	88.38	88.04	86.96	85.53
CSI-Pipeline	86.35	85.98	85.95	84.65

Table 5: Ablation study of the architecture in CSI. We adopt Accuracy (Acc.), F1-score (F1) for evaluation.

proves its key contribution in subsequent discussions and decision-making. In CRU, the performance degradation caused by removing the Review Team indicates that multi-agent discussion effectively identifies potential disinformation. When all the outputs of MLLMs were removed, the significant performance drop of Adjudicator indicates that it can effectively utilize the insightful knowledge of MLLMs and improve accuracy. A simple MLP-based adjudicator applied to embeddings cannot effectively leverage comprehensive cues. Although an MLLM-based adjudicator has strong capabilities in information fusion and reasoning, its performance on the classification task is still inferior to that of the MSLM. These results further validate the rationality of the adjudicator design. Appendix §I provides an ablation study of Adjudicator.

Architecture Ablation Analysis. To further assess the impact of the architecture in CSI, we reshape it into three variants: (1) CSI-UA, where MFU is replaced by a Unified Analyst agent responsible for analyzing all modalities; (2) CSI-UR, where the review team is replaced by a Unified Reviewer agent handling master casefiles alone; (3) CSI-Pipeline, reshaping CSI into a linear architecture of Analyst, Reviewer and Adjudicator. Results in Table 5 confirm that performance improvement of CSI stems from the architecture.

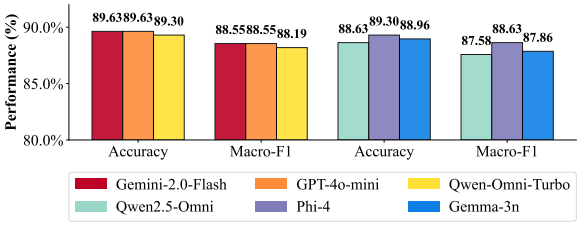
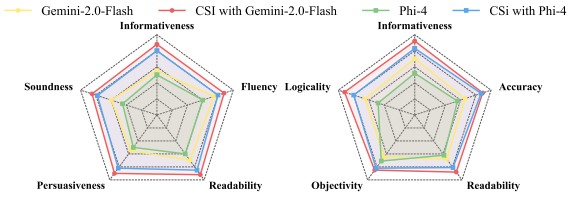


Figure 3: Performance of CSI with different MLLM backbones on the FakeTT dataset.



(a) LLM-based Evaluation **(b) Human Evaluation**

Figure 4: Comparison of explanation quality with and without the proposed CSI on the FakeTT dataset, employs a 5-point Likert scale to assess the quality.

4.5 Performance on Different Backbones

To evaluate the generalizability of CSI, we conduct experiments on the FakeTT dataset by replacing the backbone MLLMs with different closed-source models Gemini-2.0-Flash (Pichai, 2024), Qwen-Omni-Turbo (Xu et al., 2025a), GPT-4o-mini (OpenAI et al., 2024) and Open-Source Models (Qwen2.5-Omni (Xu et al., 2025a), Phi-4 (Abdin et al., 2024), Gemma-3n (Team et al., 2025)). As shown in Figure 3, CSI consistently improves detection performance across all backbone MLLMs. This confirms that our framework can achieve high-precision detection by leveraging the strengths of diverse backbones, even maintaining excellent performance with slightly weaker open-source models. Furthermore, these results underscore the broad adaptability of our framework, indicating that its efficacy is not dependent on any specific MLLMs.

4.6 Evaluations on Explanation

Quality of Explainability. As illustrated in Figure 4, CSI significantly enhances the quality of explanations across all evaluated dimensions. These results demonstrate that: (1) The MFU improves informativeness, soundness, and objectivity by systematically aggregating comprehensive multimodal clues and official reports. (2) The structured evidence collection and review process yields highly organized outputs, thereby improving readability, fluency, and logicality. (3) Through multi-agent

Title: Nature's Fury Unleashed: Jaw-Dropping Waterspout Fact.

On-screen text: WEATHER FACTS: In 2003 the Great Lakes saw the largest waterspout outbreak in recorded ....

Ground label: Fake

Case-ExMRD

Refining: The largest waterspout outbreak in Great Lakes ...This video presents a fact about **the largest waterspout outbreak in the Great Lakes**, which reportedly occurred in 2003.According to the video's text and transcript, the event lasted seven days and produced over 65 water spouts...

Retrieving: Waterspouts.... the largest recorded waterspout outbreak in the Great Lakes occurred between August 27 and September 3, 2003.

Reasoning: The video correctly describes a known event: the 2003 Great Lakes waterspout outbreak. **Both the visual and textual elements support this claim** ... about the largest waterspout outbreak in recorded history is **backed by verified historical data**, and the visuals match the description of severe weather conditions typical during such an event.

Prediction: Real ✗

Case-CSI (Ours)

Video Findings: shoot1 (0-9s) Vision: Shot Type: Medium and long shots. Footage source: appears to be footage shot in real life. Visual content: .... multiple waterspouts can be clearly seen... ; Audio: Human Voice: Male voice. BGM: **a suspenseful piano piece**. Sound Effects: **no ambient sounds or effects...**""

Fact File: noaa.gov: From 09/27 to 10/03/2003, the Great Lakes region experienced the largest and longest-lasting waterspout outbreak on record...

Review result : The core facts presented in the video have been verified to be accurate. ...**it cannot be confirmed that this specific video was 100% shot in the Great Lakes in 2003.** ...The video completely **lacks on-site environmental sounds** (such as wind sounds and water sounds), and instead uses clear post-production dubbing and background music, which makes its authenticity unverifiable.

Prediction: Fake ✓

Figure 5: Case study of a detected short video fake news on the FakeTT dataset.

collaborative investigation, a verifiable reasoning chain is formed, which substantially bolsters both persuasiveness and accuracy. More detailed explainability evaluation settings can be found at §G.

Qualitative Analysis on Explainability. To qualitatively evaluate the explanation quality of CSI, we present a case study from the FakeTT dataset, focusing on comparing how CSI and ExMRD explain and detect. As shown in Figure 5, this video actually depicts multiple waterspouts that formed above the ocean in 2003. Although the event in the video is verified by official reports, it is not a live video recorded in 2003. In ExMRD, reasoning solely based on modal semantics and logical flaws leads to missed key details and limited perspectives. In contrast, CSI detects suspicious cues by scrutinizing footage source, sound effects, and employing multi-angle discussions. This case demonstrates how CSI verifies authenticity through deep analysis and rigorous reasoning. Appendix §J provides more detailed case analyses.

4.7 Efficiency Analysis

To further assess the efficiency of CSI, we compare it against multiple advanced MLLMs. These models employ CoT and Web-RAG based on the search engine $\mathcal{W}$. We evaluate a total of 5,615 videos across the FakeTT and FakeSV datasets. We record the average reasoning time and token usage per sample, and respectively compared the accuracy on the FakeSV(Acc-SV) and FakeTT(Acc-TT).

Models	Time(s)	Token(K)	Acc-SV	Acc-TT
Gemini-2.5Pro-CW	32.7	11.1	83.03	81.93
Gemini-2.5Flash-CW	24.1	9.5	82.05	80.27
GPT-4o-CW	24.5	10.1	81.73	80.94
GPT-4o-mini-CW	21.3	8.9	80.77	78.26
CSI	**16.2**	**8.7**	**90.77**	**89.63**

Table 6: Efficiency comparison between CSI and MLLMs (CW). Best results are shown in **bold**.

As shown in Table 6, our proposed CSI framework outperforms the other explainability methods in terms of efficiency and performance. By leveraging a highly parallel architecture, structured output requirements, and streamlined communication protocols, CSI minimizes computational overhead in terms of both latency and token usage. This indicates that the framework not only delivers superior performance but also possesses high scalability and feasibility for practical deployment. Appendix §F provides detailed efficiency analysis.

5 Conclusion

This work introduces CSI, a novel multi-agent framework reframing explainable short video news detection into a collaborative investigation of digital forensics, discussion, and robust adjudication. Extensive experiments on two real-world datasets validate its superior accuracy and explanation qual-

ity. By leveraging MLLMs via transparent reasoning and traceable analysis reports, it fosters a healthier online ecosystem. In future work, we plan to further exploit the reliability of the proposed framework.

Limitations

Although our framework demonstrates excellent performance in detecting and explaining false news in short videos, there are still several limitations. In practical applications, there may be no relevant information about newly released news on the Internet, thus making it impossible to collect valid evidence. In future work, we plan to explore the detection capabilities of the framework in situations where there is no relevant credible evidence.

Ethics Statement

This study involved testing human subjects to evaluate the quality and reliability of the CSI explanation. To ensure the protection of participants and their ethical treatment, we followed the following principles: 1) All participants were informed of the nature of the study and their roles in it. Participation was completely voluntary, and participants had the right to withdraw at any time without any consequences. 2) A written informed consent form was obtained from all participants. The consent form detailed the purpose of the study, the procedures involved, potential risks, and the measures taken to protect the participants' data. 3) All data collected during the study was anonymized. Personal identification information was deleted to ensure confidentiality, and the data was properly stored to prevent unauthorized access. 4) The study posed minimal risk to the participants. The tasks performed were similar to daily activities, and no sensitive personal information was required or recorded. 5) All participants received detailed training on the evaluation criteria and were paid at the local average hourly wage to ensure a fair labor system.

The study shows that evaluating content related to misinformation may have negative effects. To protect our human evaluators, we established three guidelines: 1) Ensure they acknowledge exposure to potentially misleading content; 2) Limit the number of evaluations per day and encourage a more relaxed workload; 3) Suggest they stop working if they feel overly stressed.

The purpose of this work is to prevent the spread of false information in short videos and ensure that people do not come into contact with untrue information. However, we are also aware that malicious users may create false information through reverse engineering and based on the guidelines set by CSI. Such behavior is strictly prohibited and condemned. Finally, we strictly adhered to the intended use and license terms of all the tools used in this work. These datasets are publicly available and are only used for academic research purposes, which is consistent with the conditions of their initial release. Similarly, the use of all pre-trained models and commercial APIs also complies with their respective licenses and service terms.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62576179, 62576246, 62506178, 62532004, 62276257, and in part by the National Natural Science Foundation of Inner Mongolia under Grant 2025JQ012.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Álvaro Barbosa and Kristine Dizon. 2020. The film sound analysis framework: A conceptual tool to interpret the cinematic experience. *Journal of Science and Technology of the Arts*, 12:81–96.

Fuqing Bie, Shiyu Huang, Xijia Tao, Zhiqin Fang, Leyi Pan, Junzhe Chen, Min Ren, Liuyu Xiang, and Zhaofeng He. 2025. Omniplay: Benchmarking omni-modal models on omni-modal game playing. *Preprint*, arXiv:2508.04361.

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. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 1351–1360.

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*, page 2950–2954.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Mingfei Han, Linjie Yang, Xiaojun Chang, Lina Yao, and Heng Wang. 2025. Shot2story: A new benchmark for comprehensive understanding of multi-shot videos. In *ICLR 2025, Singapore, April 24-28, 2025*.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. 2017. Cnn architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135.

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*, page 4684–4698.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. Metagpt: Meta programming for A multi-agent collaborative framework. In *ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22105–22113.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR 2022, Virtual Event, April 25-29, 2022*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-augmented dialogue generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8460–8478.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability. Technical report, University of Pennsylvania, Annenberg School for Communication. Postprint version.

Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. 2022. Clip-event: Connecting text and images with event structures. In *CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16399–16408.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Yuhan Liu, Yuxuan Liu, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2025. *The Truth Becomes Clearer Through Debate!* multi-agent systems with large language models unmask fake news. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*, pages 504–514.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2025. Large language models: A survey. *Preprint*, arXiv:2402.06196.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

Sundar Pichai. 2024. Introducing Gemini 2.0: our new AI model for the agentic era.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):14444–14452.

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.

Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 153–162.

Lixiong Qin, Yang Zhang, Mei Wang, Jiani Hu, Weihong Deng, and Weiran Xu. 2025. Fake-in-facext: Towards fine-grained explainable deepfake analysis. *Preprint*, arXiv:2510.20531.

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.

Jing Shen, Yanjia Wang, Shengze Wang, Yuping Zhang, and Haibo Liu. 2025. Multi-modal similarity guided adaptive fusion network for short video fake news detection. In *Proceedings of the 2025 International Conference on Multimedia Retrieval, ICMR 2025, Chicago, IL, USA, 30 June 2025 - 3 July 2025*, pages 1145–1153.

Tomás Soucek and Jakub Lokoc. 2024. Transnet v2: An effective deep network architecture for fast shot transition detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, page 11218–11221.

Li Sun, Haoqi Zhang, Songyang Zhang, and Jiebo Luo. 2020. Content-based analysis of the cultural differences between tiktok and douyin. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 4779–4786.

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*, page 2189–2199.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Junxi Wang, Jize Liu, Na Zhang, and Yaxiong Wang. 2025a. Consistency-aware fake videos detection on short video platforms. In *ICIC 2025, Ningbo, China, July 26-29, 2025, Proceedings, Part XVIII*, pages 200–210.

JunXi Wang, Yaxiong Wang, Lechao Cheng, and Zhun Zhong. 2025b. FakeSV-VLM: Taming VLM for detecting fake short-video news via progressive mixture-of-experts adapter. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4782–4798.

Yiting Wang. 2021. Multimodal analysis: Researching short-form videos and the theatrical practices.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025a. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

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

Hai Yu, Chong Deng, Qinglin Zhang, Jiaqing Liu, Qian Chen, and Wen Wang. 2025. Multimodal fusion and coherence modeling for video topic segmentation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17577–17593.

Liyuan Zhang, Yang Yajing, Yan Yang, Yong Liu, Zhongyan Gui, Ruofan Li, and Hao Fei. 2025a. MFSVFND: multimodal fusion network for detecting fake news on short video platforms. In *ICMR 2025, Chicago, IL, USA, 30 June 2025 - 3 July 2025*, pages 2123–2127.

Zongmeng Zhang, Wengang Zhou, Jie Zhao, and Houqiang Li. 2025b. Robust multimodal large language models against modality conflict. In *Proceedings of the 42nd International Conference on Machine Learning*.

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*, page 5364–5375.

Jianlong Zhou, Amir Gandomi, Fang Chen, and Andreas Holzinger. 2021. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10:593.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

A Datasets

The detailed statistics of the two datasets are shown in Table 7, where the average number of shots (Avg Shots) is obtained from the detection results of the TransNetv2 model (Soucek and Lokoc, 2024).

• **FakeSV** (Qi et al., 2023a), designed for a Chinese-speaking context. It collects news videos from popular Chinese short video platforms Kuaishou and Douyin (the Chinese equivalent of TikTok). Each sample in FakeSV includes the video, title, on-screen text, user comments, relevant metadata, and the publisher's profile.

• **FakeTT** (Bu et al., 2024), designed for an English-speaking context. It collects news videos from the global short video platform TikTok. Each sample in FakeTT includes the video, title, and corresponding metadata.

B Baselines

The 14 competitive baselines in this study are presented as follows:

- Qwen2.5-VL (Bai et al., 2025) introduces dynamic resolution processing and absolute time encoding, enabling it to process images of varying sizes and videos of extended durations with second-level event localization. We use it to detect short video news through Zero-Shot, CoT and Web-RAG.
- InternVL3 (Zhu et al., 2025) introduces a native multimodal pre-training paradigm that unifies visual and linguistic learning from scratch. This model offers performance comparable to GPT-4o while maintaining strong pure language capabilities. We use it to detect short video news through Zero-Shot, CoT and Web-RAG.
- GPT-4o-mini (OpenAI et al., 2024) is a new generation of natively omni-modal model developed by OpenAI, which seamlessly integrates and processes visual, audio, and text inputs through a single unified architecture. We use it to detect short video news through Zero-Shot, CoT and Web-RAG.
- Gemini-2.5-Flash (Comanici et al., 2025) is a lightweight MLLM developed by Google, optimized for speed and efficiency. We use it to detect short video news through Zero-Shot, CoT and Web-RAG.
- TikTec (Shang et al., 2021) aligns visual and speech signals via co-attention mechanisms tailored to short-form video content to identify misleading or manipulated clips.
- FANVM (Choi and Ko, 2021) models multimodal cues and topic-agnostic features to detect rumors in micro-videos by combining cross-modal interactions with robustness-promoting training.
- HMCAN (Qian et al., 2021) constructs hierarchical multimodal context with attentive fusion of text and visual features to capture multi-level semantics for rumor detection.
- SVFEND (Qi et al., 2023a) focuses on the joint modeling of multimodal information content and social context to improve detection performance.
- FakingRecipe (Bu et al., 2024) models the creative production process by analyzing material selection and spatio-temporal editing patterns to detect fabricated short videos.
- CA-FVD (Wang et al., 2025a) models content–affect consistency across modalities and

Dataset	Avg shots	Total	Fake	Real	Duration (s)	Language	Time Range	Platform
FakeSV	7	3624	1810	1814	39.88	Chinese	2017/10-2022/02	Douyin, Kuaishou
FakeTT	5	1991	1172	819	47.69	English	2019/05-2024/03	TikTok

Table 7: Statistics of two datasets.

uses this cross-check to detect manipulated or misleading short-video claims.

- MSAF-Net (Shen et al., 2025) performs multi-scale, self-adaptive fusion of multimodal features so as to aggregate complementary signals for robust misinformation detection.
- FakeSV-VLM (Wang et al., 2025b) fine-tunes large Vision Language Models (VLMs) using specialized adapters to detect fake short-video news via progressive reasoning and cross-modal alignment. To prevent the leakage of data and prior knowledge, we do not use the event fields of the dataset for fine-tuning.
- ExMRD (Hong et al., 2025) guides MLLMs with a pipeline (Refine–Retrieve–Reason) to produce evidence-aware, explainable micro-video rumor assessments.
- TED (Liu et al., 2025) employs a rigorous debate framework where the agent simulates human-like discourse analysis and the agent synthesizes global perspectives. By leveraging attention mechanisms to model the interactions between role embeddings and arguments, the system delivers a final verdict for fake news detection.

C Implementation Details

C.1 Experimental Setups

In our experiments, we follow the benchmarks (Bu et al., 2024; Hong et al., 2025; Wang et al., 2025a; Shen et al., 2025) and adopt only the videos, titles, and on-screen text from the original content, excluding comments, events, keywords and user profiles to ensure a fair comparison. During the fine-tuning process of the MLLM baseline model, we strictly limited the input content to titles, screen texts and visual frames, and removed the event fields in the FakeSV-VLM (Wang et al., 2025b) to ensure the fairness of the experiment. In the benchmark dataset, we observed that the event fields of short video samples belonging to the same event were completely identical. And the training set and test set had approximately 75.5% overlap in the event fields. This exclusion measure is crucial as it aims to prevent data leakage. All experiments are implemented in PyTorch and run on 8 NVIDIA GeForce RTX 5090 GPUs.

C.2 Implementation of MLLMs

For the MLLM baselines, we employ *Gemini-2.5-Flash* [2], *GPT-4o-mini-2024-07-1*[3], *Qwen2.5-VL-7B*, and *InternVL3-8B*. To ensure the reproducibility of the experiments, the temperature for all MLLMs is set to 0.0 without any sampling mechanism.

To ensure a fair comparison, we apply Chain-of-Thought (CoT) (Wei et al., 2022) to MLLM baselines, aligning their token budgets with CSI. Furthermore, we employ MLLM baselines with the Web-based Retrieval-Augmented Generation (Web-RAG) (Komeili et al., 2022) based on the search engine $\mathcal{W}$ and the number of the selected documents is K = 3, strictly filtering the retrieved documents to ensure their publication dates predate the video release via API and prompt constraints, thereby preventing data leakage.

We apply LoRA (Hu et al., 2022) to the fine-tuned MLLM baselines with a rank of 8, alpha of 32, and bfloat16 precision. These models were trained for 5 epochs with a batch size of 4.

To validate the generalization capability of CSI, we first evaluate three closed-source MLLMs: Gemini-*2.0-Flash*, *GPT-4o-mini-2024-07-1*, and *Qwen Omni Turbo*[4]. Furthermore, we include three advanced open-source MLLMs with parameter sizes below 10B: *Qwen2.5-Omni-7B*, *Phi-4-multimodal-instruct*, and *Gemma-3n-E4B-it*, which are particularly suitable for resource-constrained scenarios.

In the CSI framework, all MLLM backbones of the agents are uniformly set to Gemini-2.0-Flash (Pichai, 2024) in the form of calling APIs, including the Visual Analyst agent, Acoustic Analyst agent and Intelligence Analyst agent in MFU, as well as the Reviewer agents in the Review Team.

[2] https://aistudio.google.com/
[3] https://openai.com/
[4] https://www.aliyun.com/

You are a professional visual analysis expert responsible for deconstructing video content from a visual perspective.
Given the shot-level information {shot_information} and video{visual_frame }, analyze each shot across four aspects:
1.Shot Type (e.g., wide, close-up)
2.Visual Content (description of visual content)
3.Subtitles (on-screen text, if any)
4.Footage Source (e.g., Generated, Filming)
Format your output as a structured list, where each shot corresponds to one record containing fields: ShotID, ShotType, VisualContent, Subtitles, and FootageSource. Please output in a JSON format: {"shot_id": ...,"analysis": ...}.

Figure 6: The prompt of Visual Analyst agent.

You are a professional auditory analysis expert responsible for deconstructing audio content from an acoustic perspective.
Based on the audio segment information {segment_information} and audio {audio}, analyze each segment from four aspects:
1.Human Voice (e.g., male, female, child)
2.Transcript (speech-to-text content)
3.Background Music (BGM) (its type and emotional tone)
4.Sound Effects (ambient or specific noises).

Format your output as a structured list, where each shot corresponds to one record containing fields: ShotID, HumanVoice, Transcript, BGM, and SoundEffects. Please output in a JSON format: {"shot_id": ...,"analysis": ...}.

Figure 7: The prompt of Acoustic Analyst agent.

Visual Analyst: The role of the Visual Analyst is to objectively analyze the visual content of the short video news. To ensure reproducibility and reliable content, we set the temperature of the Visual Analyst agent to 0.0 without any sampling mechanism. The instruction prompt is designed as shown in Figure 6.

Acoustic Analyst: The role of the Acoustic Analyst is to objectively analyze the audio content of the short video news. To ensure reproducibility and reliable content, we set the temperature of the Acoustic Analyst agent to 0.0 without any sampling mechanism. The instruction prompt is designed as shown in Figure 7.

Intelligence Analyst: To ensure the authenticity and credibility of the official reports retrieved by MLLMs, Intelligence Analyst agents use Google's search engine API as an external tool and check whether the retrieved information meets the requirements. The parameter temperature is set to 0.0, because we need the most reliable content to suppress the creativity of the generation. The instruction prompt is designed as shown in Figure 8.

The first Reviewer: The role of the first Reviewer is to verify facts and common sense, while also acting as the team leader to summarize the review result. To stimulate diverse reasoning paths and facilitate viewpoint interaction, we set the temperature of the agent to 1.0. The instruction prompt is designed as shown in Figure 9.

You are a professional intelligence analysis expert responsible for analyzing and verifying official reports related to events.
First, extract a neutral and objective event caption from the Title {Title} and On-screen Text {on-screen_text} of the short video as the query of the search engine.
Then retrieve official information using the search engine, ranking results by credibility: Tier 1 (official or institutional sources), Tier 2 (international media or fact-checking organizations), and Tier 3 (domestic mainstream media).
Select the top three documents and filter out the reports that are not related to the event caption (and only reports published earlier than {Publish_time} can be selected).Summarize them with the following fields: Source, URL, Summary, Publisher and PublishTime.

Format your output as a structured list, where each report corresponds to one record containing fields: Event id, Source, URL, Summary, Publisher and PublishTime. Please output in a JSON format: {"event_caption":.. "event_id":...,"summary": ...}.

Figure 8: The prompt of Intelligence Analyst agent.

1. Initial Response stage:
Role: You are a misinformation detection expert specializing in factual analysis and reasoning.
Task: Conduct an independent review of all modal analyses and official reports regarding the video in the master casefile {C}. Verify and confirm the factual claims of short video news and verify them with official reports in casefile (assess the credibility of each source to decide whether to adopt them and prevent the introduction of noise). If there are factual contradictions or common sense errors in the reasoning, please give your opinion. Please limit your response to within 100 words.

2. Discussion stage: The following are replies from other reviewers {Peer_Responses}. Objectively examine your responses based on these opinions. If there is a need for modification, please update your opinion. Please limit your response to within 100 words.

3.Team Answer Generation stage:
Role: As the team leader, objectively analyze everyone's opinion.
Task: View the meeting log of the discussion {M}. Briefly summarize everyone's opinion and provide a summary opinion without assessing veracity. Please limit your response to within 100 words.

Figure 9: The prompt of the first Reviewer agent.

The second Reviewer: The role of the second Reviewer is to specialize in cross-modal consistency verification, detecting contradictions among text, visual, and audio. To encourage the agent to scrutinize the case file from multi-perspective dimensions, we set the temperature of the agent to 1.0. The instruction prompt is designed as shown in Figure 10.

The third Reviewer: The role of the third Reviewer is to verify the logic and editing rationality of the clips. To ensure the agent effectively uncovers subtle logical loopholes through discussion, we set the temperature of the agent to 1.0. The instruction prompt is designed as shown in Figure 11.

Adjudicator: For shot segmentation, we use the TransNet-V2 model (Soucek and Lokoc, 2024) to detect start and end frames of video shots. For

Initial Response Stage:
Role: You are a misinformation detection expert specializing in cross-modal consistency reasoning.

Task: Conduct an independent review of all modal analyses and official reports regarding the video in the master casefile {C}. Carefully analyze the text, visual and audio modal clues in this short video news, and reason whether there are any logical inconsistencies among these modalities. If there are potential inconsistencies, please provide your opinion.Please limit your response to within 100 words.

Discussion Stage : The following are replies from other reviewers {Peer_Responses}. Objectively examine your opinions based on these reviewers. If there is a need for modification, please update your opinion. Please limit your response to within 100 words.

Figure 10: The prompt of the second Reviewer agent.

Initial Response Stage:
Role: You are a misinformation detection expert specializing in logical and editing consistency between video clips.

Task: Conduct an independent review of all modal analyses and official reports regarding the video in the master casefile {C}. Carefully analyze whether there is any malicious editing or taking things out of context in the logic and editing between different shots in this short video news. If there is any suspicious information that misleads the public, please give your opinion. Please limit your response to within 100 words.

Discussion Stage: The following are replies from other reviewers {Peer_Responses}. Objectively examine your responses based on these reviewers . If there is a need for modification, please update your opinion. Please limit your response to within 100 words.

Figure 11: The prompt of the third Reviewer agent.

feature extraction, we employ a pre-trained BERT model (Devlin et al., 2019) to extract textual features $H_t \in \mathbb{R}^{k_t \times d_t}$ from the T. We also extract the visual analysis features $H_{vt} \in \mathbb{R}^{k_t \times d_t}$ from P_{vision}, the auditory analysis features $H_{at} \in \mathbb{R}^{k_t \times d_t}$ from P_{audio}, the event caption features $H_c \in \mathbb{R}^{k_t \times d_t}$ from the T_c, and the review result features $H_r \in \mathbb{R}^{k_r \times d_t}$ from R. We employ a pre-trained Vision Transformer (ViT) (Dosovitskiy et al., 2021) to extract visual features $H_v \in \mathbb{R}^{k_v \times d_v}$ from the set of visual frames $\mathcal{V}_f$. We use a pre-trained VGGish model (Hershey et al., 2017) to extract audio features $H_a \in \mathbb{R}^{k_a \times d_a}$ from A. We employ fine-tuned XLM-RoBERTa (Conneau et al., 2020) and HuBERT models (Hsu et al., 2021) to extract sentimental features from T and A, respectively represented as $H_{et} \in \mathbb{R}^{d_e}$ and $H_{ea} \in \mathbb{R}^{d_e}$.

Audio features are extracted with `harritaylor/torchvggish`. For FakeTT datasets in English, we adopt `openai/clip-vit-base-patch32` to extract textual and visual features, and fine-tune `facebook/xlm-roberta-base` and `facebook/hubert-base-ls960` for extracting sentimental features from text and audio respectively. For FakeSV datasets in Chinese, we instead use `OFA-Sys/chinese-clip-vit-base-patch16` to extract textual and visual features. Sentimental features are obtained via `bhavikardeshna/xlm-roberta-base-chinese` (text) and `Gelel/chinese-hubert-base` (audio).

Regarding the training hyperparameters, the Adjudicator is trained for 20 epochs with the batch size of 128 using the Adam optimizer (Kingma and Ba, 2015). To optimize convergence, we set the learning rate to 1×10^{-4} for FakeTT and 1×10^{-5} for FakeSV.

D Stability Analysis

In order to fully verify the robustness of our system, we further developed two additional prompt versions for all agents: (1) Simplified version (CSI-simple), examples and supplementary details removed from task descriptions; (2) Complex version (CSI-complex), detailed elaboration for each task. Each agent prompt in our method strictly follows a standard role–task structured format, without specially crafted wording. The results in Figure 8 clearly demonstrate that our system is not sensitive to prompt variations, ensuring the stability of the framework.

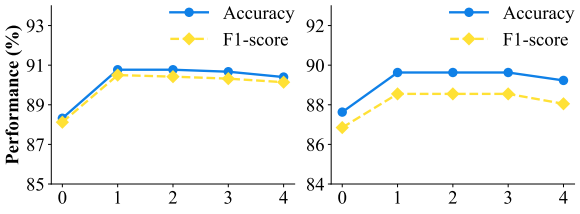
Variant	FakeSV		FakeTT	
	Acc.	F1	Acc.	F1
CSI (Full Model)	**90.77**	**90.50**	**89.63**	**88.55**
CSI (simple)	90.59	90.35	89.96	88.6
CSI (complex)	90.77	90.50	89.63	88.55

Table 8: Stability Analysis of our method.

E Hyperparameter setting

E.1 Discussion Setting

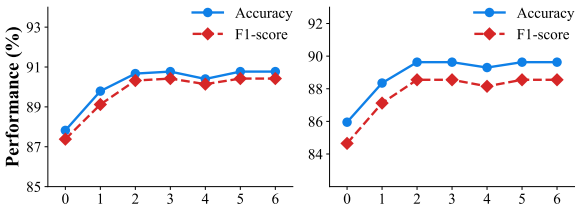
We analyze the optimal number of discussion rounds in Discussion stage(ϕ_2), with results in Figure 12. On both datasets, the performance of CSI improves significantly following the discussion process, with accuracy and F1-score reaching their peaks between the first and third rounds. Subsequent rounds yield only minor fluctuations without significant gains. This indicates one round is sufficient for agents to exchange key information and converge. Additional rounds introduce computational overhead for negligible benefit. Therefore, we set the number of discussion rounds to 1.



(a) Rounds on FakeSV. **(b) Rounds on FakeTT.**

Figure 12: Effect of the number of discussion rounds on model performance on the FakeSV and FakeTT datasets.

E.2 Retrieval Setting



(a) Number K on FakeSV **(b) Number K on FakeTT**

Figure 13: Impact of the number of retrieved documents on model performance on the FakeSV and FakeTT datasets.

We analyze the effect of the retrieved document count K on detection performance, as shown in Figure 13. The performance improves rapidly as K increases from 0 to 3, demonstrating the necessity of incorporating external knowledge to mitigate hallucinations and ground the model's reasoning. Notably, the performance gains saturate when $K \geq 3$. This implies that the top-3 documents already cover the majority of the distinct information required for verification. Increasing K further primarily introduces redundant information rather than new evidence. To maintain computational efficiency without compromising accuracy, we adopt $K = 3$ for all subsequent experiments.

F Detailed Efficiency Analysis

To fully demonstrate the practicality and credibility of this high efficiency, we conduct a detailed time analysis. This framework achieves significant time savings in MFU through high concurrency, where visual, audio, and forensic agents operate synchronously for multimodal deconstruction and evidence collection. This parallel execution effectively compresses the MFU stage within a time window of approximately 7.2 seconds, mainly depending on the time of visual analysis.

After the main case file is constructed, the CRU initiates a structured discussion process. Although these stages require an inherent chronological order to ensure reasoning depth, the shared-context blackboard communication protocol minimizes inter-agent latency, compressing the total duration to 8.9 seconds. Ultimately, the lightweight adjudicator conducts rapid feature integration and makes the final judgment. With the highly parallelized and streamlined design in the framework and detailed time reports, the efficiency advantages reported in Table 6 are strongly demonstrated, not only maintaining higher performance but also ensuring lower costs.

G Explainability Evaluation Details

LLM-based Evaluation: We adopt the LLM-based reference-free evaluation method G-Eval (Liu et al., 2023) to evaluate the quality of the interpretation text generated by our framework. The LLM employed for evaluation utilizes the GPT-4o to serve as an independent third-party evaluator, thereby ensuring fairness and preventing evaluation bias associated with the Gemini and Phi models. Specifically, we adopt the following criteria:

- **Informativeness:** Evaluates whether the explanation provides new information, such as background or additional context.
- **Soundness:** Evaluates whether the explanation appears valid and logically sound.
- **Persuasiveness:** Evaluates whether the explanation presents a convincing argument.
- **Readability:** Evaluates whether the explanation adheres to correct grammatical and structural rules.
- **Fluency:** Evaluates whether the explanation is smooth and fluent, exhibiting coherent and interconnected thoughts.

For each criterion, a 5-point Likert scale is adopted, where 1 represents the worst quality and 5 represents the best.

Human Evaluation: We implement a stricter human subjective evaluation method, ILORA (Zhou et al., 2021), to assess explanation quality. A 5-point Likert scale is utilized to evaluate the overall quality of generated explanations across five key dimensions:

- **Informativeness:** Evaluates whether the explanation provides new information, such as background or additional context.
- **Logicality:** Evaluates whether the explanation follows a sound reasoning process and if a strong causal relationship exists between the explanation and the result.
- **Objectivity:** Evaluates whether the explanation is objective and free from excessive subjective emotions.
- **Readability:** Evaluates whether the explanation adheres to correct grammatical and structural rules, ensuring coherence and ease of understanding.
- **Accuracy:** Evaluates whether the generated explanation aligns with ground-truth labels and accurately reflects the results.

We employed a 5-point Likert scale to assess the quality of explanations, where 1 represents the lowest quality and 5 represents the highest, enabling a detailed and nuanced evaluation. We recruited 10 evaluators, all holding at least a Bachelor's degree. Guided by the first author, the evaluators received detailed training on the evaluation criteria and were compensated at the local average hourly rate. We randomly selected 60 samples from the FakeTT dataset, comprising 30 fake and 30 real news videos. To ensure reliability, each sample was evaluated by at least two independent evaluators. Furthermore, model names were anonymized and the presentation order was randomized to mitigate potential bias. Finally, we calculated Krippendorff's Alpha (Krippendorff, 2011) to measure inter-evaluator agreement. The resulting coefficient of 0.806 confirmed the consistency and reliability of our human evaluation results.

H Algorithm

The detailed structure of our CSI is presented in Algorithm 1. In the algorithm, f_{vision} represents the visual analysis prompt function implemented by the MLLM for the Visual Analyst, f_{audio} represents the acoustic analysis prompt function implemented by the MLLM for the Acoustic Analyst, f_{int} represents the intelligence analysis prompt function implemented by the MLLM for the Intelligence Analyst. And the rest of the notations are consistent with those used in the main text.

Algorithm 1 CSI Framework

Input: Short video news item $S = \{V, A, T\}$
Output: $\hat{y}$, R, $\mathcal{C}$

MFU:
for each agent in parallel **do**
  Visual Analyst: $P_{\text{vision}} \leftarrow f_{\text{vision}}(V)$
  Acoustic Analyst: $P_{\text{audio}} \leftarrow f_{\text{audio}}(A)$
  Intelligence Analyst: $(T_c, E) \leftarrow f_{\text{int}}(T)$
  video findings $P \leftarrow \{P_{\text{vision}}, P_{\text{audio}}\}$
  master casefile $\mathcal{C} \leftarrow \{P, T_c, E\}$
end for

CRU:
Initialize Review Team:
Assign agents $\mathcal{A}_r = \{a_1, \ldots, a_N\}$
Assign roles $\{Role_1, \ldots, Role_N\}$
Review Team:
for each $a_i \in \mathcal{A}_r$ **do**
  $O_i^{(1)} \leftarrow f_{\text{init}}(\mathcal{C}, Role_i)$ ▷ Initial response
end for
for round $r = 1$ to R_{max} **do** ▷ Discussion
  $\mathcal{D}^{(r-1)} \leftarrow \{O_1^{(r-1)}, \ldots, O_N^{(r-1)}\}$
  for each agent i **in parallel do**
    $O_i^{(r)} \leftarrow f_{\text{dis}}(\mathcal{D}^{(r-1)}, Role_i)$
  end for
end for
Meeting log $M \leftarrow \{\mathcal{D}^{(1)}, ..., \mathcal{D}^{(R_{max})}\}$
Review result $R \leftarrow f_{\text{team}}(M)$ ▷ Team Answer
Adjudicator:
$\hat{y} \leftarrow$ Adjudicator$(\mathcal{C}, R, S)$
return $\hat{y}, R, \mathcal{C}$

I Ablation Study of Adjudicator

Variant	FakeSV		FakeTT	
	Acc.	F1	Acc.	F1
CSI (Full Model)	**90.77**	**90.50**	**89.63**	**88.55**
w/o Unimodal Cue Guidance	82.84	82.51	81.61	79.62
w/o Cross-modal Mutual Verification	83.79	83.39	82.61	81.30

Table 9: Ablation study of the Adjudicator. We adopt Accuracy (Acc.), F1-score (F1) for evaluation.

To evaluate the Adjudicator's contribution, we conducted an ablation study on its two key components: *Unimodal Cue Guidance and Cross-modal Mutual Verification* in Table 9. Removing Unimodal Cue Guidance caused a sharp performance drop on both datasets. This highlights the Adjudicator's effectiveness in integrating modality-specific features

from the MFU with review results from the CRU. Similarly, removing Cross-modal Mutual Verification significantly reduced performance. This confirms the necessity of cross-verifying multimodal information to uncover subtle inconsistencies. This mechanism enables the Adjudicator to capture critical cross-modal conflicts.

J Other Case Studies

J.1 Successful Cases

To comprehensively demonstrate the effectiveness and explainability of CSI in handling complex real-world disinformation, we present a detailed analysis of two successful cases shown in Figure 14 and Figure 15. These cases aim to highlight how our framework achieves rigorous explanations and more accurate detections.

Figure 14 displays a piece of fake news involving a political figure from the FakeTT dataset. In this case, the MFU accurately identified the video as a genuine close-up shot and extracted the precise date and location. Through retrieval, it precisely obtained the official reports from the White House (a Tier 1 authoritative source), the content of which clearly contradicts the on-screen text in the video. Simultaneously, it carefully examined the audio, confirming the absence of specific sound effects that would support the false claim. The CRU cross-referenced official reports with the short video and combined with the absence of audio cues, precisely identified that the video's claims lacked a factual basis. This deep multi-perspective discussion mechanism renders the final verdict not only accurate but also sufficiently explainable.

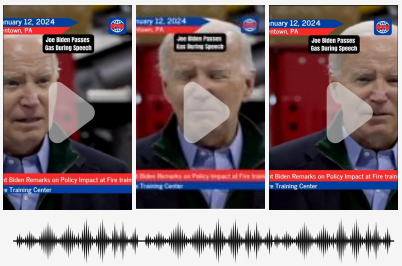
Figure 15 presents a scientific falsehood from the FakeSV dataset claiming that coarse salt and saliva can break car windows. In this case, the framework accurately described the demonstration process and keenly observed the lack of close-up shots showing the key substances of coarse salt and saliva; it then precisely retrieved a previously published debunking report (from MyGoPen), accurately revealing that the substance actually used in the video was ceramic spark plug fragments, rather than the salt and saliva claimed by the video. The Review Team comprehensively compared the accurate visual observations collected by the MFU with external reliable factual information, thereby effectively revealing the cross-modal inconsistency and logical fallacies present in the video. This in-depth multi-angle discussion ensures that the framework can reach highly accurate and persuasive judgments even when facing subtle manipulations.

J.2 Failure Cases

As illustrated in Figure 16, we present a failure case from the FakeTT dataset to examine the limitations of the CSI framework. This case features a video creator expressing intense dissatisfaction with Canadian government policies, where the content is ground-truth real. However, the Review Team misclassified this "highly agitated genuine voter" as a "malicious propagator attempting to incite emotional instability." Specifically, the agents identified a logical construct where the speaker juxtaposed refugee support against the neglect of indigenous communities. By deeply scrutinizing the underlying logic and intent, the framework interpreted the combination of "selective presentation of facts" and "intense emotional instigation" as characteristic signatures of disinformation campaigns designed to sow social discord.

This demonstrates that although the framework effectively showcases the ability for deep reasoning, its conservative ethical framework has led to excessive defense when distinguishing between extreme partisan opinions and malicious false news. The model prioritizes **the warning of social stability**, indicating that future work needs to improve this ethical coordination in order to handle such "grey area" content more carefully.

Case 1



Video News

Text

Title: Joe Biden Passes Gas #funny #trump2024 #fart #biden #trump #donaldtrump
On-screen Text: January 12, 2024, Allentown, PA" , "Joe Biden Passes Gas During Speech", "President Biden Remarks on Policy Impact at Fire Training Center".
Publish time: 2024.1.21

Video findings:
Shot 1: Shot Type: It's a close-up shot. Visual Content : We see a close-up of President Joe Biden as he's talking. He's wearing a blue shirt and a dark jacket with green trim around the collar. The background is a little blurry, but it looks like he's inside, possibly at a training facility. We can see some red equipment behind him. Subtitles: There are a few lines of text: In the top left corner: "January 12, 2024, Allentown, PA" (This tells us the date and location of the video).In the middle: "Joe Biden Passes Gas During Speech" (This is a text overlay highlighting the video's claim) At the bottom: "President Biden Remarks on Policy Impact at Fire Training Center" (This gives us more context about the event). Footage Source : It was most likely filmed, not computer-generated, as it shows a real-life event. HumanVoice: Male (primarily Joe Biden speaking). Transcript: "And I said, you know, I think we underestimate the impact that..." (The transcript captures Biden's speech during the segment, but it's an incomplete sentence). BGM: Absent (There appears to be no background music in this segment). SoundEffects: Possible subtle ambient noise (The recording might contain faint background sounds typical of an indoor setting, but no prominent or specific sound effects are distinctly audible).

Fact File:
Objective Event Title: On January 12, 2024, President Biden farted while delivering a speech at the fire training center.
Document 1 (Tier 1 - Official Source):
Source: The White House
URL:https://bidenwhitehouse.archives.gov/briefing-room/speeches-remarks/2024/01/12/
Summary: This is a transcript of President Biden's remarks after touring the Allentown Fire Training Academy. He discusses the impact of the COVID crisis and highlights the importance of supporting small businesses.
Publisher: The White House
PublishTime: 2024-01-12
Document 2 (Tier 3 - Domestic Mainstream Media):
Source: LehighValleyNews.com
URL:https://www.lehighvalleynews.com/photos-president-biden-visits-allentown-fire-training-academy.....
Summary: This article provides photos of President Biden's visit to the Allentown Fire Training Academy. It mentions that he also stopped at local stores to promote small businesses and his economic record.
Publisher: LehighValleyNews.com
PublishTime: 2024-01-12
.....

Review result: The team unanimously believes that the short video may contain incorrect information. The main claim that President Biden "farted" during his speech is not supported by reliable news or official reports. Furthermore, visual and audio analyses indicated that there were no obvious farting sound effects or related visual information about people.

Ground Truth: Fake **Prediction: Fake**

Figure 14: A successful case of short video fake news detection on the FakeTT dataset.

Case 2

Video News

Text

Title: Too Scary! Are Car Windows This Fragile Now? Everyone, Be Careful Not to Leave Valuable Items Inside Your Car!
On-screen Text: “Beware of New Window-Smashing Theft Just a handful of coarse salt and some spit can break the window Criminals demonstrated this on the spot Advice for friends and family driving: Do not leave valuable or expensive items inside your car
Publish time: 2021.10. 20

Video findings:
Shot 1: Shot Type: Medium shot .
Visual Content: A group of men, including one in a plaid shirt (possibly the individual being escorted), are walking past a silver car. Several other men are in the background, some observing.Subtitles: New type of window breaking theft. A handful of coarse salt, a handful of saliva, can break a window. FootageSource: Filming. HumanVoice: Multiple male voices. Transcript: The background noise is noisy. Someone is speaking Indonesian, making it difficult to transcribe accurately. BGM: None. SoundEffects: Speech, general crowd noise.

Shot 2: Shot Type: Medium shot. Visual Content: The group of men is now positioned around the silver car. The man in the grey shirt, identified as the criminal, is being guided towards the car window, appearing to prepare for a demonstration. Other men are watching and some are recording. Subtitles: Criminal demonstrates on the spot. FootageSource: Filming. HumanVoice: Multiple male voices. Transcript: The background noise is noisy. Someone is speaking Indonesian, making it difficult to transcribe accurately. BGM: None. SoundEffects: Speech, general crowd noise.

Shot 3: Shot Type: Medium shot. Visual Content: The man in the grey shirt proceeds with the demonstration, placing a substance on the car window and then applying force, which causes the window to shatter into pieces. Onlookers, including several men holding cameras or phones, react to the breaking glass. The man is then held by others, while shards of glass are visible on the window frame. Subtitles: Advise friends and relatives driving out not to leave valuable items in the car. FootageSource: Filming. HumanVoice: Multiple male voices. Transcript: The background noise is noisy. Someone is speaking Indonesian, making it difficult to transcribe accurately. BGM: None. SoundEffects: Speech, general crowd noise, car window shattering, camera clicks/shutter sounds.

Fact File:
Objective Event Title: Criminal demonstrates car window breaking method using salt and saliva.
Document 1 (Tier 2 - fact-checking organization)
Source: MyGoPen
URL:https://www.mygopen.com/2019/09/Coarse-salt-Car-theft.html
Summary: ..... Fact-checking website MyGoPen refuted that this was a scam video from Indonesia in 2017. It clarified that the substance in the suspect's hand was actually ceramic fragments of spark plugs, which contained high-hardness aluminum oxide, and the saliva might have only been used to bond the fragments and prevent them from scattering, rather than as an effective agent to break the glass.
Publisher: MyGoPen. PublishTime: 2019-09-09

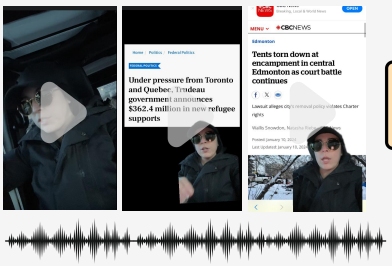
Review Result:
The team believes that the claim of "coarse salt and saliva breaking the window" explicitly put forward in the first shot of the video is completely contrary to the external fact-check (MyGoPen), which clearly pointed out that this was a fraud from 2017 and actually utilized ceramic fragments of spark plugs (alumina). This visually "certain substance" (rather than clearly visible salt and saliva) was applied to the window, causing it to break. Although the editing is persuasive, it deliberately misleads by taking advantage of "cross-modal inconsistency" : the video gives the visual content a false cause through subtitle text.... There is no close-up shot between the second and third shots, deliberately avoiding the unknown substance in the hand.....

Ground Truth: Fake **Prediction: Fake**

Figure 15: A successful case of short video fake news detection on the FakeSV dataset.

Case 3

Video News



Text

Title: Its time for this Liberal government to go. Justin Trudeau and his party have completely lost touch and I don't say that lightly. #cdnpolitics #cdnpoli #canadianpolitics.
On-screen Text: FEDERAL POLITICS. Under pressure from Toronto and Quebec, Trudeau government announces $362.4 million in new refugee supports.Marc Miller, @MarcMillerVM.....
Publish time: 2024.1.31

Video findings:
Shot 1: Shot Type: close-up shot. Visual Content : A person wearing a black baseball cap, gold-rimmed sunglasses, and a black winter coat is sitting inside a car. The interior of the car is visible, with the roof appearing to be either white or light grey. The person is speaking directly to the camera. A building can be seen through the car window. Subtitles: None. Footage Source : Filming. HumanVoice: Female. Transcript: I am so unbelievably sick and tired of this liberal government that it's not even funny. It's been three days since these politicians have returned to work after a six week break from the House of Commons, not sitting. And every day, every day you open up the news and it's like, what are these people thinking?" BGM: None. SoundEffects: Car interior sounds.
Shot 2: Shot Type: Medium Shot. Visual Content : The view switches to a digital news article on a phone screen, overlaid onto the previous video of the person in the car. The headline reads: 'Under pressure from Toronto and Quebec, Trudeau government announces $362.4 million in new refugee supports.Subtitles: FEDERAL POLITICS. Under pressure from Toronto and Quebec, Trudeau government announces $362.4 million in new refugee supports
.Footage Source : Filming. HumanVoice: Female. Transcript: Like, I'm sorry, what? $362 million announced today
.....

Fact File:
Objective Event Title: On January 31, 2024, the Canadian federal government announced an additional allocation of 362.4 million Canadian dollars to the Temporary Housing Assistance Program (IHAP) to assist in the resettlement of asylum seekers.
Document 1 (Tier 1 - Official Source):
Source: Government of Canada
URL:https://www.canada.ca/en/immigration-refugees-citizenship/news/2024/01/new-funding-to-support-housing-for-asylum-claimants.html
Summary: The Minister of Immigration, Refugees and Citizenship of Canada (IRCC), Marc Miller, announced an additional allocation of 362.4 million Canadian dollars for the Temporary Housing Assistance Program (IHAP) in the 2023-2024 fiscal year. This funding aims to support provinces and municipalities in addressing the temporary housing pressure caused by the surge in asylum seekers through cost-sharing, and to prevent refugees from falling into homelessness.
Publisher: Immigration, Refugees and Citizenship Canada
Time: 2024-01-31
Document 2 :
.....

Review result: This video expresses the speaker's strong dissatisfaction with the policies of the Canadian government. The potential problem with this video ....that the speaker criticized the government's zero investment in drug addiction in indigenous communities needs further verification. By selectively presenting news clips, the video may tend to create a negative atmosphere while ignoring other possible positive information. The content of the video largely depends on the speaker's subjective interpretation of the event. Viewers should maintain a critical mindset and conduct independent investigations to obtain more comprehensive information.

Ground Truth: Real **Prediction: Fake**

Figure 16: A failure case of short video fake news detection on the FakeTT dataset.