Detecting and Explaining Fake News Short Videos with Multimodal Content and Real-World Evidence

Yifeng Luo[1,2] **Yupeng Li**[1,2] * **Ming Tang**[4] **Jianxiong Guo**[5] **Liang Lan**[1,3]
[1]Dept. of Interactive Media; [2]AI and Social Good Lab, AI Media Centre; [3]AI Media Centre
Hong Kong Baptist University
[4]Dept. of Computer Science and Engineering, Southern University of Science and Technology
[5]Institute of AI and Future Networks, Beijing Normal University

Abstract

Short-video platforms have become a primary news source for the public, which has also enabled the widespread dissemination of fake news videos. We study the task of fake news video detection and explanation (FNVDE). Existing methods face two critical limitations. First, commonly used frame selection strategies may omit veracity-relevant cues or provide insufficient temporal context for understanding news videos. Second, prior methods neglect either multimodal understanding or evidence retrieval. To address these limitations, we propose NVKE-CEI, a unified system that integrates a news video keyframes extraction method (NVKE) and an FNVDE framework leveraging both content and evidence information (CEI). NVKE selects keyframes based on chronological changes in combined visual and OCR-text similarity. CEI employs two specialized LLM-based fact checkers (content-based and evidence-based) whose outputs are fused by a lightweight judge model. Extensive experiments show that NVKE-CEI outperforms state-of-the-art baselines while generating high-quality content-grounded explanations.

1 Introduction

Short-video platforms have become a major channel for daily news consumption and for the dissemination of fake news videos (Wang et al., 2021), i.e., videos containing false or misleading information (Bu et al., 2023; Qi et al., 2023a; Bu et al., 2024; Luo et al., 2026). Recent studies have found that video-based misinformation tends to spread faster and is perceived as more trustworthy by users than the same misinformation when presented in text or image formats, posing significant risks to public opinion and social good (Vosoughi et al., 2018; Sundar et al., 2021; Li et al., 2024). To combat this, recent studies (Hong et al., 2025; Yan et al., 2025; Zhang et al., 2025) have utilized the pretrained knowledge and generative capabilities of Large Language Models (LLMs[1]) (Gong et al., 2023; OpenAI, 2023) to detect fake news videos.

In this paper, we consider the task of fake news video detection and explanation (FNVDE). Given a news video, FNVDE aims to predict its veracity label and generate a news content-grounded explanation, i.e., an explanation of why the news video is real or fake based on its news content (Cheng et al., 2021; Wang and Shu, 2023; Wang et al., 2024a; Khaliq et al., 2024; He et al., 2026b). This differs from post-hoc interpretation methods that highlight influential regions (Wu et al., 2024) and from supervised explanation-generation methods that require human-annotated explanations (Chen et al., 2026). Existing LLM-based methods either summarize news videos into text and leverage pretrained LLM knowledge (Hong et al., 2025; Yan et al., 2025), or fine-tune LLMs with annotated text-video pairs (Zhang et al., 2025). However, the hallucination problem (Ji et al., 2023; Perković et al., 2024) persists, as the LLM reasoning process is not grounded in real-world evidence, leading to incorrect veracity predictions (Wang et al., 2024d; Hu et al., 2024a). To mitigate hallucinations, recent methods ground LLM reasoning in real-world evidence by either training models for Deep Research (Tongyi DeepResearch Team, 2025) or prompting them to use retrieval-augmented generation (RAG) (Asai et al., 2024; Lewis et al., 2020; He et al., 2026a). These methods generate search queries by extracting keywords from news (Liu et al., 2025) or providing query examples to LLMs (Singal et al., 2024), then submit them to web search engines and integrate retrieved evidence to predict news veracity.

*Yifeng Luo conducted this work under the supervision of Yupeng Li and Liang Lan. Yupeng Li (ivanypli@gmail.com) is the corresponding author.

[1]Throughout this paper, we use "LLMs" as an umbrella term for large foundation models, including text-only, vision-language, and video-language models.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 6614–6633
October 24-29, 2026 ©2026 Association for Computational Linguistics

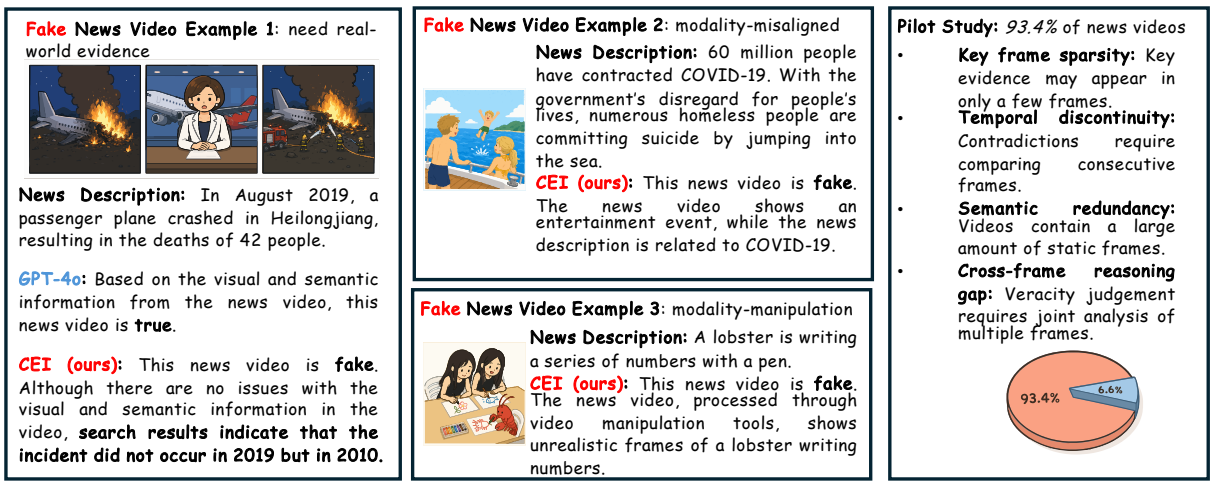


Figure 1: In fake news video example 1, the opening and concluding segments both depict the accident scene with minimal visual variation but convey different information: the opening introduces the accident, whereas the concluding segment shows rescue operations. Density-based keyframe extraction methods may miss these semantically distinct frames due to their visual similarity. Furthermore, verifying the veracity requires real-world evidence. Fake news video example 2 (resp. 3) illustrates the category of modality-misaligned (resp. modality-manipulated) fake news video. In the pilot study, we identify four frame-selection challenges in the FakeSV dataset.

Despite this progress, applying LLM-based methods to FNVDE presents several challenges. First, frame selection determines what visual information is available to an LLM for understanding a news video. Directly inputting all frames of a news video into the LLM context window is costly and often redundant. Existing LLM-based FNVDE methods rely on random frame sampling (Hong et al., 2025; Yan et al., 2025; Zhang et al., 2025), which may omit briefly occurring but veracity-relevant cues or fail to preserve sufficient temporal context, thereby limiting the LLM's ability to form a clear understanding of the news video. Existing LLM-based video understanding methods (Zhou et al., 2024; Jin et al., 2024) employ clustering-based visual compression, but these general strategies are not specifically designed to select news video frames according to joint changes in visual and OCR-text content over time. They may therefore underrepresent visually similar frames that convey different information at different time points. Our pilot study (see Appendix A) on the FakeSV dataset (Qi et al., 2023a) finds that 93.4% of the analyzed news videos exhibit at least one of four frame-selection challenges, highlighting the difficulty of preserving both semantically important cues and sufficient temporal context in a limited frame set (see Fig. 1). Second, both multimodal understanding and evidence retrieval are important. Effective FNVDE methods require both multimodal understanding to identify manipulation (e.g., modality misalignment or content fabrication) and real-world evidence retrieval to ground reasoning and avoid hallucinations (see Fig. 1). However, existing LLM-based methods either focus solely on multimodal content analysis (Hong et al., 2025; Yan et al., 2025; Zhang et al., 2025) or exclusively on external evidence retrieval (Wang et al., 2024a; Choi and Ferrara, 2024; Khaliq et al., 2024; Liu et al., 2025). This singular focus leads to limitations when handling diverse fake news types. Moreover, evidence-based methods employing RAG assume web-retrieved evidence is reliable, yet such evidence can be noisy and contain misleading information (Ge et al., 2025; Luo et al., 2024; He et al., 2026c), further complicating the FNVDE task.

To address these limitations, we propose NVKE-CEI, which contains NVKE, a news video keyframes extraction method, and CEI, a novel FNVDE framework that leverages both content and evidence information for veracity prediction and content-grounded explanation generation. CEI employs two specialized LLM-based fact checkers, namely a content-based LLM and an evidence-based LLM, whose complete reasoning outputs constitute the content-grounded explanation; a lightweight judge model integrates their outputs to produce the final prediction. We propose NVKE to address the first limitation by selecting keyframes according to the magnitude of chrono-

logical changes in visual-text similarity, with the goal of preserving semantic changes and temporal context in news videos. To address the second limitation, CEI utilizes LLM role-playing (Wang et al., 2024b; Shanahan et al., 2023) to handle diverse fake news video types via a decoupled architecture. We argue that a decoupled architecture is essential for FNVDE rather than using a single LLM. A single LLM may be affected by knowledge cutoffs (Cheng et al., 2024) and confirmation bias (Wan et al., 2025). When the same LLM analyzes content and retrieves evidence jointly, its initial assessment may influence how it interprets subsequently retrieved evidence (Wang et al., 2026). CEI therefore adopts a dual-perspective fact-checking process. The content-based LLM focuses on internal inconsistencies and provides content-driven predictions, while the evidence-based LLM focuses primarily on external verification. The evidence-based LLM, inspired by human fact-checking processes (Vlachos and Riedel, 2014; Brandtzaeg et al., 2018), employs reasoning and acting steps (Yao et al., 2023) to generate search queries, gather real-world evidence, and make evidence-grounded predictions. This decoupling keeps the two reasoning streams distinct before fusion, allowing the lightweight judge model to combine their signals through text feature attention and veracity score fusion to produce the final prediction. We address retrieval noise at two stages: the query regeneration module refines search queries before retrieval, while the text feature attention module allows the lightweight judge to adaptively weight retrieved evidence alongside the other textual features. The main contributions are summarized as follows:

- We propose NVKE to select keyframes for the LLM context window by tracking chronological changes in visual-text similarity, aiming to preserve content changes and temporal context while reducing redundancy.
- We combine multimodal content analysis and evidence retrieval through specialized LLM checkers, employing query regeneration in evidence-based LLM and attention mechanisms in a lightweight judge model to handle noisy web evidence effectively.
- Extensive experiments on FakeSV and FakeTT with thirteen baselines demonstrate that NVKE-CEI achieves state-of-the-art (SOTA) performance. Compared with the second-best results, NVKE-CEI achieves an increase of 0.0071 (resp. 0.0234) in Acc. and 0.0002 (resp. 0.0190) in Macro $\mathbf{F}_1$ score on FakeSV (resp. FakeTT)[2].

2 NVKE-CEI

In this section, we begin by formally defining the FNVDE task. We propose NVKE to extract the news video keyframes by considering both semantic and temporal information. Building upon the NVKE, we introduce the CEI framework. The CEI framework consists of two specialized LLM fact checkers and one lightweight judge model. The overview of CEI is presented in Fig. 2.

2.1 Problem Setting

We denote a news video as $\mathcal{V} = \{\mathbf{F}, M\}$ with frames $\mathbf{F} = \{\mathbf{f}_1, \ldots, \mathbf{f}_n\}$ and text material M. Here, the text material M is the concatenation of the title, description, and audio transcript of the news video. Each news video $\mathcal{V}$ has a binary veracity label $y \in \mathcal{Y}$, where $\mathcal{Y} = \{\text{real}, \text{fake}\}$. The task of FNVDE is to predict the veracity label $\hat{y}$ and the news content-grounded explanation e.

2.2 NVKE: News Video Keyframes Extraction

Visual Text Embedding and Similarities. Given a news video $\mathcal{V}$ with frames $\mathbf{F} = \{\mathbf{f}_1, \ldots, \mathbf{f}_n\}$, we extract optical characters (OC) $\mathbf{O} = \{\mathbf{o_1}, \ldots, \mathbf{o_n}\}$ using the OCR tool, $\mathbf{O} = \text{Tool}_{\text{ocr}}(\mathbf{F})$. To capture both visual and semantic information, we use CLIP (Radford et al., 2021) encoders to embed them into a shared space: $\mathbf{H_F} = \text{CLIP}_{\text{visual}}(\mathbf{F})$ and $\mathbf{H_O} = \text{CLIP}_{\text{text}}(\mathbf{O})$, where $\mathbf{H_F} = \{\mathbf{h}_\mathbf{f}^1, \ldots, \mathbf{h}_\mathbf{f}^n\}$ and $\mathbf{H_O} = \{\mathbf{h}_\mathbf{o}^1, \ldots, \mathbf{h}_\mathbf{o}^n\}$ are the visual and semantic embeddings respectively. We compute the cosine similarity between consecutive frames and OCs, i.e., visual similarities $\mathbf{S_F} = \{s_\mathbf{F}^1, \ldots, s_\mathbf{F}^{n-1}\}$ and semantic similarities $\mathbf{S_O} = \{s_\mathbf{O}^1, \ldots, s_\mathbf{O}^{n-1}\}$ as $s_\mathbf{F}^i = \frac{\mathbf{h}_\mathbf{f}^i \cdot \mathbf{h}_\mathbf{f}^{i+1}}{||\mathbf{h}_\mathbf{f}^i|| ||\mathbf{h}_\mathbf{f}^{i+1}||}$ and $s_\mathbf{O}^i = \frac{\mathbf{h}_\mathbf{o}^i \cdot \mathbf{h}_\mathbf{o}^{i+1}}{||\mathbf{h}_\mathbf{o}^i|| ||\mathbf{h}_\mathbf{o}^{i+1}||}$ for $i \in \{1, 2, \ldots, n-1\}$. The overall visual text similarities $\mathbf{S} = \{s_1, \ldots, s_{n-1}\}$ are: $\mathbf{S} = \alpha \mathbf{S_F} + (1 - \alpha)\mathbf{S_O}$ with hyper-parameter $\alpha \in [0, 1]$.

Keyframes and OC Extraction. To capture temporal intensity of content variation, we compute the magnitude of similarity changes $\Delta S = \{\Delta s_1, \ldots, \Delta s_{n-2}\}$, where $\Delta s_i = |s_{i+1} - s_i|$

[2]Source code is available at https://github.com/TrustworthyComp/CEI.

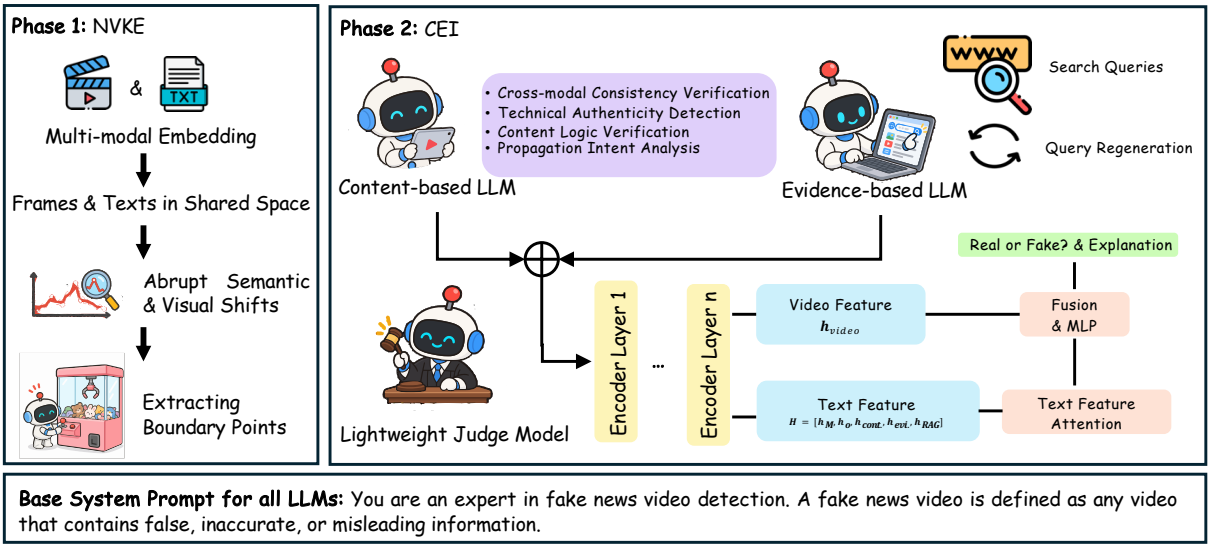


Figure 2: Overview of our proposed NVKE-CEI.

for $i \in \{1, 2, \ldots, n-2\}$. We sort all change points in descending order of Δs_i, where each Δs_i corresponds to three consecutive frames $\mathbf{f}_i$, $\mathbf{f}_{i+1}$, and $\mathbf{f}_{i+2}$ and their aligned OCR texts. Starting from the highest-ranked change point, we add its two boundary frames $\mathbf{f}_i$ and $\mathbf{f}_{i+2}$ and continue down the ranking until at least $2K$ distinct frames have been collected or no change points remain. The selected keyframes and OCs are denoted as $\mathbf{F}_{\text{key}} = \{\mathbf{f}_1, \ldots, \mathbf{f}_{2K}\}$ and $\mathbf{O}_{\text{key}} = \{\mathbf{o}_1, \ldots, \mathbf{o}_{2K}\}$.

2.3 CEI: FNVDE Framework with Content and Evidence Information

Base System Prompt and Structured Output. The two LLM checkers share a common base prompt that defines fake news videos and explains the task of FNVDE before assigning their specialized roles. The base system prompt is shown in Fig. 2. To extract well-defined components from the LLM fact-checkers' responses, we employ structured output[3]. For the content-based LLM, which operates in a single turn, the output includes a reasoning process $\mathcal{R}_{\text{cont.}}$ that justifies the predicted confidence score $\hat{\mathbf{y}}_{\text{cont.}}$. For the evidence-based LLM, which follows an iterative ReAct framework, the output structure depends on the dialogue stage. In intermediate turns, the LLM outputs a reasoning process $\mathcal{R}_{\text{evi.}}$ explaining its information needs and generates search queries $\mathbf{q}$ accordingly. In the final turn, after synthesizing retrieved evidence, it outputs $\mathcal{R}_{\text{evi.}}$ justifying the confidence score $\hat{\mathbf{y}}_{\text{evi.}}$. Both confidence scores represent the estimated probability of the fake class.

[3] https://openai.com/index/introducing-structured-outputs-in-the-api/

2.3.1 Content-based LLM

Content-based LLM Fake News Video Checker. The content-based LLM checker examines semantic and syntactic manipulations within news videos through comprehensive multimodal analysis. To enhance its analytical capability, we employ role-playing (Wang et al., 2024b; Shanahan et al., 2023) by instructing the LLM to act as *a seasoned fake news video detection expert with extensive expertise in media literacy, multimodal content analysis, and fact-checking methodologies*. Under this role, the LLM performs systematic verification across four core dimensions: (1) Cross-modal consistency verification analyzes potential contradictions between textual descriptions, visual content, and audio information; (2) Technical authenticity detection examines video frames for evidence of artificial modification, splicing, or fabrication; (3) Content logic verification assesses whether video content contains factual errors or logical fallacies; (4) Propagation intent analysis identifies inflammatory or manipulative expressions that may mislead the audience. Given a news video $\mathcal{V}$, we utilize structured output to simultaneously obtain the reasoning process $\mathcal{R}_{\text{cont.}}$ and confidence score $\hat{\mathbf{y}}_{\text{cont.}}$ from the content-based LLM $\text{LLM}_{\text{cont.}}(\cdot)$, $\hat{\mathbf{y}}_{\text{cont.}}, \mathcal{R}_{\text{cont.}} = \text{LLM}_{\text{cont.}}(\mathbf{F}_{\text{key}}, \mathbf{O}_{\text{key}}, M)$. The detailed prompt for the content-based LLM is presented in Appendix G.3.

2.3.2 Evidence-based LLM

While the content-based LLM addresses internal manipulations within news videos, many fake news videos contain factually incorrect claims that require real-world evidence to verify. To address this, we design an evidence-based LLM that grounds its reasoning in external evidence retrieved from the web. The checker operates through an iterative ReAct framework consisting of three integrated components: (1) the evidence-based LLM employs reasoning and acting steps to analyze video content, generate search queries, and use the retrieved evidence for veracity prediction (Yao et al., 2023); (2) the query regeneration module refines initial LLM-generated queries to improve evidence quality, addressing the problem that the retrieved evidence may contain noisy information at the retrieval stage; and (3) the RAG pipeline retrieves real-world evidence from web search engines.

Evidence-based LLM with ReAct Framework. To handle fake news videos containing factually incorrect claims, the evidence-based LLM checker verifies information through real-world evidence retrieval. Inspired by human fact-checking processes (Guo et al., 2022; Brandtzaeg et al., 2018), we employ the ReAct framework (Yao et al., 2023), where the LLM alternates between reasoning (understanding video content, formulating information needs, and assessing veracity) and acting (generating search queries to retrieve evidence). This iterative process continues until the LLM gathers sufficient evidence to produce a confidence score for its veracity prediction. Specifically, the evidence-based LLM operates through two types of reasoning steps. First, in the query generation reasoning step, the LLM analyzes the news video and outputs both its reasoning process $\mathcal{R}_{\text{evi.}}$ and search queries $\mathbf{q}$. Initially, this occurs by understanding the video content, $\mathbf{q}, \mathcal{R}_{\text{evi.}} = \text{LLM}_{\text{evi.}}(\mathbf{F}_{\text{key}}, \mathbf{O}_{\text{key}}, M)$. In subsequent iterations, if additional information is needed, the LLM generates new queries based on previously retrieved evidence $\mathcal{R}_{\text{RAG}}$, $\mathbf{q}, \mathcal{R}_{\text{evi.}} = \text{LLM}_{\text{evi.}}(\mathcal{R}_{\text{RAG}})$. Second, in the veracity prediction reasoning step, once the LLM determines that sufficient evidence has been gathered, it uses the retrieved information to output its confidence score $\hat{\mathbf{y}}_{\text{evi.}}$ and final reasoning process $\mathcal{R}_{\text{evi.}}$: $\hat{\mathbf{y}}_{\text{evi.}}, \mathcal{R}_{\text{evi.}} = \text{LLM}_{\text{evi.}}(\mathcal{R}_{\text{RAG}})$. We denote $\mathbf{q}^{\tau} = \{\mathbf{q}^1, \mathbf{q}'^1, \ldots, \mathbf{q}^{T-1}, \mathbf{q}'^{T-1}\}$ as the complete sequence of queries, $\mathcal{R}^{\tau}_{\text{evi.}} = \{\mathcal{R}^1_{\text{evi.}}, \ldots, \mathcal{R}^T_{\text{evi.}}\}$ as the complete sequence of reasoning process, and $\mathcal{R}^{\tau}_{\text{RAG}} = \{\mathcal{R}^1_{\text{RAG}}, \ldots, \mathcal{R}^{T-1}_{\text{RAG}}\}$ as the complete sequence of retrieved evidence. The complete output trajectory is denoted as $\tau = \{\mathbf{q}^1, \mathcal{R}^1_{\text{evi.}}, \mathbf{q}'^1, \mathcal{R}^1_{\text{RAG}}, \ldots, \mathcal{R}^{T-1}_{\text{RAG}}, \mathcal{R}^T_{\text{evi.}}, \hat{\mathbf{y}}_{\text{evi.}}\}$, where $t \in \{1, \ldots, T\}$ indexes an iterative step and T is the total number of iterative steps; query generation and evidence retrieval occur for $t \in \{1, \ldots, T-1\}$, while the final veracity prediction is produced at $t = T$. The detailed prompt for the evidence-based LLM is presented in Appendix G.2.

Query Regeneration. Initial LLM-generated queries may suffer from two critical issues that lead to low-quality evidence retrieval: they may contain redundant contextual information from the video description, and they may lack the specificity needed to retrieve precise, relevant evidence. To address this at the retrieval stage, we introduce a query regeneration module that refines each generated query $\mathbf{q}$ into a more focused version $\mathbf{q}'$: $\mathbf{q}' = \text{LLM}_{\text{query}}(\mathbf{q})$, where $\text{LLM}_{\text{query}}(\cdot)$ evaluates queries based on specificity and direct relevance to core factual claims in the video. By generating more targeted queries, this module improves the quality of retrieved evidence. The detailed prompt is presented in Appendix G.4.

RAG Pipeline. To retrieve real-world evidence, we implement a RAG pipeline using a web search engine as the retriever (Welz and Lanquillon, 2024; Borgeaud et al., 2022). Given a refined search query $\mathbf{q}'$, the RAG retriever $\text{RAG}(\cdot)$ returns the K_{RAG} most relevant webpage snippets: $\mathcal{R}_{\text{RAG}} = \text{RAG}(\mathbf{q}')$. These retrieved snippets provide real-world evidence that the evidence-based LLM uses to make grounded veracity predictions, avoiding hallucinations that may occur when relying solely on the LLM's pretrained knowledge.

2.4 Lightweight Judge Model

We employ a lightweight neural network as the final judge module instead of the meta-judge LLM for two primary reasons. First, while LLMs possess strong general reasoning capabilities, FNVDE often requires capturing subtle, dataset-specific distribution biases (e.g., specific linguistic patterns or platform-specific metadata correlations). A lightweight model trained via supervised learning can effectively capture these specific patterns, often outperforming a general-purpose LLM operating in a zero-shot or few-shot setting. Second,

the analyses generated by the content-based and evidence-based agents are typically lengthy and token-heavy. Concatenating these long contexts and feeding them back into a commercial LLM API (like GPT-4o) for a final verdict would incur prohibitive token costs, especially for large-scale datasets. In contrast, the lightweight judge processes the fused representations with negligible computational overhead and zero additional API costs, making the framework more practical for real-world deployment. The judge model consists of four key components: (1) *feature encoding* processes multiple text sources and video frames using a small model; (2) *text feature attention* employs self-attention to fuse heterogeneous text features while filtering noisy evidence; (3) *multimodal fusion* integrates text and video features through concatenation; (4) *veracity score fusion* incorporates LLM confidence scores as prior knowledge.

Feature Encoding. The judge model processes five text features (text material M, key OCs $\mathbf{O}_{\text{key}}$, content-based reasoning $\mathcal{R}_{\text{cont.}}$, evidence-based reasoning process $\mathcal{R}^{\tau}_{\text{evi.}}$, retrieved evidence $\mathcal{R}^{\tau}_{\text{RAG}}$) and one visual feature (keyframes $\mathbf{F}_{\text{key}}$). We employ CLIP encoders (Radford et al., 2021) and MLPs to project them into a shared hidden space via $\mathbf{h}_i = \text{MLP}_{\text{text}}(\text{CLIP}_{\text{text}}(\cdot)) + \mathbf{b}_i$, and $\mathbf{h}_{\text{video}} = \text{AvgPool}(\text{MLP}_{\text{visual}}(\text{CLIP}_{\text{visual}}(\mathbf{F}_{\text{key}})))$ respectively, where $i \in \{M, \mathbf{O}, \text{cont.}, \text{evi.}, \text{RAG}\}$ and $\mathbf{b}_i$ are learnable biases.

Text Feature Attention and Multimodal Fusion. To integrate heterogeneous text features while filtering noisy evidence, we apply self-attention over stacked text features $\mathbf{H} = [\mathbf{h}_M; \mathbf{h}_{\mathbf{O}}; \mathbf{h}_{\text{cont.}}; \mathbf{h}_{\text{evi.}}; \mathbf{h}_{\text{RAG}}] \in \mathbb{R}^{5\times d}$. The attended features are computed as: $\mathbf{H}_{\text{attended}} = \text{softmax}\left(\frac{\mathbf{HQ}(\mathbf{HK})^{\top}}{\sqrt{d}}\right)\mathbf{HV}$, where $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{d\times d}$ are learnable projection matrices and d is the hidden dimension. The fused text feature is obtained by $\mathbf{h}_{\text{text}} = \frac{1}{5}\sum_{i=1}^{5}\mathbf{H}_{\text{attended}}[i]$, which adaptively down-weights noisy retrieved evidence. We then concatenate text and video features as $\mathbf{h}_{\text{fused}} = [\mathbf{h}_{\text{text}}; \mathbf{h}_{\text{video}}] \in \mathbb{R}^{2d}$, preserving information from both modalities while allowing the classifier to learn their interactions. The judge logits are obtained through a two-layer MLP classifier: $\hat{\mathbf{y}}_{\text{judge}} = \text{MLP}(\mathbf{h}_{\text{fused}}) \in \mathbb{R}^2$, where the two entries correspond to the `real` and `fake` classes.

Veracity Score Fusion. To leverage the complementary strengths of both LLM checkers, we introduce a logit-level veracity score fusion mechanism that injects LLM confidence scores as prior knowledge into the model's predictions. Given the confidence scores $\hat{\mathbf{y}}_{\text{cont.}}$ and $\hat{\mathbf{y}}_{\text{evi.}}$ from the content-based and evidence-based LLMs respectively, we transform their fake-class probabilities into log-odds so that they can be combined additively with the judge logits. Specifically, we compute the logit of the fake news probability from each LLM: $\ell_{\text{cont.}} = \log\left(\frac{p_{\text{cont.}}}{1-p_{\text{cont.}}}\right), \ell_{\text{evi.}} = \log\left(\frac{p_{\text{evi.}}}{1-p_{\text{evi.}}}\right)$, where $p_{\text{cont.}} = \hat{\mathbf{y}}_{\text{cont.}}[\texttt{fake}]$ and $p_{\text{evi.}} = \hat{\mathbf{y}}_{\text{evi.}}[\texttt{fake}]$ are clamped to $[\epsilon, 1-\epsilon]$ with $\epsilon = 10^{-6}$. The combined prior is $\ell_{\text{prior}} = 0.5 \cdot \ell_{\text{cont.}} + 0.5 \cdot \ell_{\text{evi.}}$, which is injected into the model's fake class logit, $\hat{\mathbf{y}}_{\text{final}} = \hat{\mathbf{y}}_{\text{judge}} + [0, \gamma \cdot \ell_{\text{prior}}]$. Here, γ is a learnable scale parameter initialized to 1.0, enabling the model to balance its predictions with LLM priors automatically. The judge model is trained using standard cross-entropy loss with label smoothing ($\epsilon_{\text{smooth}} = 0.1$). The class corresponding to the larger entry in $\hat{\mathbf{y}}_{\text{final}}$ is returned as the final veracity label $\hat{y}$. The content-grounded explanation e consists of the content-based reasoning $\mathcal{R}_{\text{cont.}}$ and the complete sequence of evidence-based reasoning processes $\mathcal{R}^{\tau}_{\text{evi.}}$.

3 Experiment

In this section, we introduce the experiment settings, then evaluate NVKE-CEI in terms of detection performance and explainability. Additional experimental details and results are provided in Appendices B, C, D, E, and F.

3.1 Experiment Settings

Datasets and Evaluation Metrics. In our experiments, we consider two widely used datasets in the literature on the detection of fake news videos, FakeSV (Qi et al., 2023a) and FakeTT (Bu et al., 2024). To evaluate the performance of our proposed fake news video detection method, we employ two evaluation metrics: accuracy (Acc.) and Macro $\mathbf{F}_1$ score. More details of the datasets and evaluation metrics can be found in Appendix B.1.

Baselines and Implementation Details. We evaluated the performance of our proposed fake news video detection framework with thirteen baselines, seven zero-shot LLM-based baselines, and six supervised training-based baselines. The seven zero-shot LLM-based baselines are **GPT-4o** (OpenAI, 2023), **DeepSeek-V3** (DeepSeek-AI, 2024),

Table 1: Performance compared with the baselines.

Method	FakeSV (Qi et al., 2023a)				FakeTT (Bu et al., 2024)			
	Acc.	Macro $\mathbf{F}_1$	Prec.	Rec.	Acc.	Macro $\mathbf{F}_1$	Prec.	Rec.
GPT-4o (OpenAI, 2023)	0.6518	0.6373	0.7527	0.6865	0.7090	0.6509	0.6633	0.6448
DeepSeek-V3 (DeepSeek-AI, 2024)	0.6974	0.6920	0.7500	0.7043	0.7224	0.6789	0.6845	0.5354
Qwen3-vl-plus (Wang et al., 2024c)	0.5657	0.5516	0.6300	0.6007	0.6738	0.5131	0.6044	0.5409
Claude-3-5-sonnet	0.5918	0.5833	0.6387	0.6177	0.6333	0.5990	0.5978	0.6011
CoRAG (Khaliq et al., 2024)	0.3326	0.3292	0.6764	0.2189	0.6720	0.6889	0.8204	0.5981
ReAct (Yao et al., 2023)	0.6084	0.6016	0.7625	0.5417	0.6519	0.6643	0.7682	0.5904
3MFact (Niu et al., 2025)	0.8122	0.8359	0.7970	**0.8788**	0.7661	0.7920	0.7339	**0.8601**
FANVM (Choi and Ko, 2021)	0.7952	0.7881	0.7982	0.7788	0.7180	0.7030	0.7023	0.7289
TikTec (Shang et al., 2021)	0.7343	0.7326	0.7324	0.7349	0.6622	0.6508	0.6584	0.6785
FactR1 (Zhang et al., 2025)	0.7662	0.7478	0.7885	0.7111	0.7444	0.7270	0.7925	0.6715
SVFEND (Qi et al., 2023a)	0.8086	0.8055	0.8018	0.8060	0.7714	0.7563	0.7512	0.7756
FakingRecipe (Bu et al., 2024)	0.8460	0.8434	<u>0.8585</u>	0.8435	0.7903	0.7770	0.7721	0.7980
ExMRD (Hong et al., 2025)	<u>0.8490</u>	<u>0.8512</u>	0.8286	<u>0.8771</u>	<u>0.8428</u>	<u>0.8315</u>	<u>0.8227</u>	0.8519
NVKE-CEI (ours)	**0.8561**	**0.8514**	**0.8615**	0.8471	**0.8662**	**0.8505**	**0.8473**	<u>0.8541</u>

Table 2: Ablation study of the proposed modules.

Dataset Variant	FakeSV (Qi et al., 2023a) Acc.	 Macro $\mathbf{F}_1$	FakeTT (Bu et al., 2024) Acc.	 Macro $\mathbf{F}_1$
w/o NVKE	0.8210	0.8198	0.8395	0.8138
w/o Content-based LLM	0.8266	0.8254	0.8227	0.7994
w/o Query Regeneration	0.8395	0.8327	0.8428	0.8300
w/o Evidence-based LLM	0.7989	0.7985	0.8027	0.7493
w/o Text Feature Attention	0.8081	0.8081	0.7960	0.7880
w/o Veracity Score Fusion	0.8229	0.8227	0.8194	0.8051
w/o Fine-tuning	0.7968	0.7985	0.7543	0.7253
NVKE-CEI (ours)	**0.8561**	**0.8514**	**0.8662**	**0.8505**

Qwen3-vl-plus (Wang et al., 2024c), **Claude-3-5-sonnet**[4], **CoRAG** (Khaliq et al., 2024), **ReAct** (Yao et al., 2023), and **3MFact** (Niu et al., 2025). The six supervised training-based baselines are **FANVM** (Choi and Ko, 2021), **TikTec** (Shang et al., 2021), **SVFEND** (Qi et al., 2023a), **FakingRecipe** (Bu et al., 2024), **FactR1** (Zhang et al., 2025), and **ExMRD** (Hong et al., 2025). The implementation details of the baselines can be found in Appendix B.2, and the implementation details of our proposed NVKE-CEI framework can be found in Appendix B.3.

3.2 Experiments Results

The detection performance of the baselines and our proposed NVKE-CEI framework is presented in Table 1. To comprehensively evaluate the individual module contribution of the NVKE-CEI, we conduct extensive ablation studies, the results of which are detailed in Table 2. We use bolded (resp. underlined) numbers to represent the best (resp. second best) experiment results. Moreover, our proposed NVKE-CEI achieves the optimal trade-off between time efficiency and performance compared to other baselines (see Appendix B.4).

Performance compared with baselines. The NVKE-CEI framework achieves the best Acc. and Macro $\mathbf{F}_1$ scores on both FakeSV (Qi et al., 2023a) and FakeTT (Bu et al., 2024) datasets. Compared with the second-best results, our proposed NVKE-CEI framework achieves an increase of 0.0071 (resp. 0.0234) in Acc. and 0.0002 (resp. 0.0190) Macro $\mathbf{F}_1$ score on FakeSV (resp. FakeTT) dataset. The superior performance of NVKE-CEI over all other baselines, particularly ExMRD (Hong et al., 2025) and FakingRecipe (Bu et al., 2024), demonstrates that our proposed NVKE-CEI framework has achieved SOTA performance. We also present NVKE comparison experiments, case studies, and error analysis in the Appendices D and F.

Ablation Studies. Results in Table 2 evaluate: (1) NVKE effectiveness, (2) LLM components (content-based LLM, query regeneration, evidence-based LLM), and (3) judge model components (text feature attention and veracity score fusion). First, NVKE improves performance. Without it, performance drops by 0.0351 (resp. 0.0267) in Acc. and 0.0316 (resp. 0.0367) in Macro $\mathbf{F}_1$ on FakeSV (resp. FakeTT). Additional NVKE comparisons and case studies are in Appendix D. Second, all NVKE-CEI modules are essential. Removing either content-based or evidence-based LLM reduces performance on both datasets. Among the three LLM-component ablations, removing the evidence-based LLM causes the largest degradation, 0.0572

[4] https://www.anthropic.com/news/claude-3-5-sonnet/

Table 3: Comparison of explanation quality for the NVKE-CEI and baseline methods capable of producing explanations on the FakeTT dataset.

	I	S	P	R	F
CoRAG (Khaliq et al., 2024)	4.27	4.20	4.33	4.68	4.47
ReAct (Yao et al., 2023)	4.45	4.81	4.72	4.69	4.82
ExMRD (Hong et al., 2025)	4.92	4.72	4.63	4.62	4.89
NVKE-CEI (ours)	4.96	4.90	4.87	4.70	4.94

(resp. 0.0635) in Acc. and 0.0529 (resp. 0.1012) in Macro $\mathbf{F}_1$ on FakeSV (resp. FakeTT), indicating that the evidence-based branch contributes substantially under this setup. Finally, judge model components are essential for integrating predictions from both LLM checkers. The text feature attention and veracity score fusion filter noisy evidence and combine complementary signals. Without fine-tuning the judge model, performance drops substantially.

3.3 Model Explainability

To assess the explainability of NVKE-CEI, we conduct the evaluation from three perspectives. (1) We compare the quality of explanations generated by NVKE-CEI against methods capable of producing explanations: (2) We evaluate the NVKE-CEI's explainability through a qualitative case study that demonstrates how a content-grounded explanation is generated (see Appendix E.2). (3) To better understand how the judge model integrates predictions from both fact-checker LLMs, we analyze the prediction distribution among the content-based LLM, evidence-based LLM, and judge model, to examine how their predictions relate to the final decision.

Quality of Explainability. Following previous studies (Hong et al., 2025; Wang et al., 2024a), we use G-Eval (Liu et al., 2023b), an LLM-based reference-free evaluation, to assess explanation quality across five metrics: (1) *Informativeness* (I): provides new information and context; (2) *Soundness* (S): valid and logically coherent; (3) *Persuasiveness* (P): convincing and well-supported; (4) *Readability* (R): proper grammar and structure; (5) *Fluency* (F): smooth flow with coherent ideas. Each metric uses a 5-point Likert scale (1=poorest, 5=best). Appendix E.1 presents the detailed settings of G-Eval. Results on FakeTT (Table 3) show NVKE-CEI achieves the highest scores across all metrics. NVKE-CEI scores 4.96 in I, outperforming ExMRD by 0.04. Notably, NVKE-CEI achieves 4.90 in S and 4.87 in P, surpassing ReAct by 0.09 and 0.15, respectively. These improvements stem from NVKE-CEI's dual-perspective fact-checking that grounds explanations in both multimodal content analysis and real-world evidence. While all methods maintain comparable R and F (> 4.6) (excluding CoRAG), NVKE-CEI's superior I, S, and P demonstrate its ability to generate comprehensive, logically coherent, and evidence-backed explanations.

Table 4: Prediction Distribution of the NVKE-CEI. R: Real, F: Fake.

Evid.	Cont.	Judge	Label	# for FakeSV	# for FakeTT
R	R	R	R	109	77
R	R	R	F	8	3
R	R	F	R	30	4
R	R	F	F	65	10
R	F	R	R	3	18
R	F	R	F	3	1
R	F	F	R	1	0
R	F	F	F	11	0
F	R	R	R	67	42
F	R	R	F	9	7
F	R	F	R	18	5
F	R	F	F	69	25
F	F	R	R	5	41
F	F	R	F	4	7
F	F	F	R	5	13
F	F	F	F	135	46

Prediction Distribution of the evidence-based LLM, content-based LLM, and judge model. To better understand how the predictions of the judge model are formed based on the content-based and evidence-based LLMs, we present the prediction distribution in Table 4. Through empirical analysis, we find that the judge model prefers to trust evidence-based LLM when strong evidence exists during conflicts, while defaulting to content-based LLM when evidence is insufficient or contradictory.

4 Related Work

This section reviews prior work on FNVDE and distinguishes it from our approach. Additional work on text-image fake news detection, and explainable fake news detection is reviewed in Appendix H.

FNVDE. Detecting fake news in videos is more challenging than in text-image content because videos contain richer and more diverse modalities, making it difficult to understand and verify the information they convey (Qi et al., 2023a). Early fake news video detection studies (Shang et al., 2021; Li et al., 2022; Choi and Ko, 2021; Liu et al., 2023a)

primarily focused on feature extraction and integration across multiple modalities. Subsequent studies incorporate additional context, including social information (Qi et al., 2023a) and neighboring videos from the same event (Qi et al., 2023b), or model the creative process of fake news videos (Bu et al., 2024). This line of work broadens the available information beyond earlier content- or metadata-focused feature pipelines (Hou et al., 2019; Palod et al., 2019). With the pretrained knowledge and generative abilities of LLMs, recent studies (Hong et al., 2025; Zhang et al., 2025; Niu et al., 2025) have begun applying LLMs to the problem of FNVDE. These methods compress news videos into limited visual inputs, including sampled frames, extracted keyframes, or derived visual descriptions. Such representations may omit briefly occurring cues or temporal dependencies needed for understanding the video. Hong et al. (2025) propose ExMRD, which leverages LLM's pretrained knowledge to detect fake news videos, while Zhang et al. (2025) fine-tuning LLMs on video-text pairs for fake news video detection. However, these methods struggle when news videos require real-world evidence for verification, as they rely solely on the internal knowledge of LLMs.

5 Conclusion

We study the FNVDE task and identify two limitations in the existing literature. First, commonly used frame-sampling or clustering-based selection strategies may omit veracity-relevant cues or provide insufficient temporal context for news-video understanding. Second, prior methods neglect either multimodal understanding or evidence retrieval. To address these challenges, we propose NVKE-CEI, which combines a news video keyframes extraction method with an FNVDE framework that integrates content and evidence information. Experiments on two datasets with thirteen baselines show that the NVKE-CEI achieves SOTA detection performance while providing content-grounded explanations.

Limitations

We discuss four main limitations of our work. First, gradually evolving, veracity-relevant cues may not be fully captured by NVKE's similarity-based approach. NVKE selects frames around relatively large changes in combined visual and OCR-text similarity, which may miss cues that evolve gradually while maintaining high frame-to-frame similarity. Future versions will explore temporal modeling techniques and fine-grained manipulation detection methods to capture progressive content distortions. Second, as illustrated by the error cases in Appendix F, NVKE-CEI can fail when the content-based and evidence-based LLMs converge on the same incorrect conclusion. In these cases, their agreement may leave the judge model without a sufficiently strong contradictory signal to correct the error. Future work will explore incorporating external validation mechanisms or ensemble methods with diverse reasoning strategies to provide additional corrective signals in such cases. Third, NVKE-CEI relies on proprietary GPT-4o APIs and live web search, so its latency, cost, retrieved evidence, and outputs may vary with service versions and availability. Under our experimental setup, processing one batch of 15 samples takes 29.6 seconds, which may hinder practical deployment. Future work will explore adaptive routing to reduce unnecessary LLM calls and evaluations with open-source backbones. Finally, explanation quality is evaluated only on FakeTT using a single LLM evaluator and without human judgments. Therefore, the reported G-Eval scores may reflect evaluator-specific preferences and do not establish whether the generated explanations are faithful to the final judge prediction.

Acknowledgments

This work was supported by the Early Career Scheme (ECS) from the Research Grants Council of HKSAR (HKBU 22202423), the General Research Fund (GRF) from the Research Grants Council of HKSAR (HKBU 12203425), a grant from the Germany/Hong Kong Joint Research Scheme sponsored by the Research Grants Council of HKSAR and the German Academic Exchange Service of Germany (No. G-HKBU208/25), the Initiation Grant for Faculty Niche Research Areas 2023/24 (No. RC-FNRA-IG/23-24/COMM/01), Research Cluster Matching Scheme (No. RCMS/24-25/01) of Hong Kong Baptist University, Guangdong and Hong Kong Universities "1+1+1" Joint Research Collaboration Scheme (Project No. 2025A0505000001), National Natural Science Foundation of China (No. 62202402, and No. 61906161), and Startup Grant (Tier 1) for New Academics AY2020/21 of Hong Kong Baptist University.

References

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A survey on multimodal disinformation detection. In *Proc. of COLING*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *Proc. of ICLR*.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, and 9 others. 2022. Improving language models by retrieving from trillions of tokens. In *Proc. of ICML*.

Petter Bae Brandtzaeg, Asbjørn Følstad, and María Ángeles Chaparro Domínguez. 2018. How journalists and social media users perceive online fact-checking and verification services. *Journalism Practice*, 12(9):1109–1129.

Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. 2023. Combating online misinformation videos: Characterization, detection, and future directions. In *Proc. of ACM MM*.

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 *Proc. of ACM MM*.

Juan Cao, Peng Qi, Qiang Sheng, Tianyun Yang, Junbo Guo, and Jintao Li. 2020. Exploring the role of visual content in fake news detection. *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*, pages 141–161.

Canyu Chen and Kai Shu. 2024. Can llm-generated misinformation be detected? In *Proc. of ICLR*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proc. of AAAI*.

Lizhi Chen, Zhong Qian, Peifeng Li, and Qiaoming Zhu. 2026. Multimodal fake news video explanation: Dataset, model and evaluation. *Information Processing & Management*, 63(6):104769.

Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proc. of ACM WWW*.

Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2024. Dated data: Tracing knowledge cutoffs in large language models. In *Proc. of COLM*.

Lu Cheng, Kush R Varshney, and Huan Liu. 2021. Socially responsible ai algorithms: Issues, purposes, and challenges. *Journal of Artificial Intelligence Research*, 71:1137–1181.

Eun Cheol Choi and Emilio Ferrara. 2024. FACT-GPT: Fact-checking augmentation via claim matching with LLMs. In *Proc. of ACM WWW short papers*.

Hyewon Choi and Youngjoong Ko. 2021. Using topic modeling and adversarial neural networks for fake news video detection. In *Proc. of ACM CIKM*.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proc. of ACM SIGIR*.

DeepSeek-AI. 2024. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*.

Ziyu Ge, Yuhao Wu, Daniel Wai Kit Chin, Roy Ka-Wei Lee, and Rui Cao. 2025. Resolving conflicting evidence in automated fact-checking: a study on retrieval-augmented llms. In *Proc. of IJCAI*.

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*.

Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. 2003. Knn model-based approach in classification. In *Proc. of OTM*.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Haorui He, Xinwen Chen, Dacheng Wen, Reynold Cheng, Francis C. M. Lau, and Yupeng Li. 2026a. Novel claim or déjà vu? rethinking "contamination-free" dynamic evaluation for multimodal automated fact-checking. In *Proc. of ACM MM*.

Haorui He, Yupeng Li, Dacheng Wen, Yang Chen, Reynold Cheng, Donglong Chen, and Francis C. M. Lau. 2026b. Debating truth: Debate-driven claim verification with multiple large language model agents. In *Proc. of ACM WWW*.

Haorui He, Yupeng Li, Bin Benjamin Zhu, Dacheng Wen, Reynold Cheng, and Francis C. M. Lau. 2026c. Fact2fiction: Targeted poisoning attack to agentic fact-checking system. In *Proc. of AAAI*.

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 *Proc. of ACM WWW*.

Rui Hou, Verónica Pérez-Rosas, Stacy Loeb, and Rada Mihalcea. 2019. Towards automatic detection of misinformation in online medical videos. In *Proc. of ICMI*.

Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024a. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proc. of AAAI*.

Xuming Hu, Junzhe Chen, Xiaochuan Li, Yufei Guo, Lijie Wen, Philip S Yu, and Zhijiang Guo. 2024b. Do large language models know about facts? In *Proc. of ICLR*.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating llm hallucination via self reflection. In *Proc. of EMNLP Findings*.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proc. of IEEE CVPR*.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proc. of ACM MM*.

Mohammed Abdul Khaliq, Paul Yu-Chun Chang, Mingyang Ma, Bernhard Pflugfelder, and Filip Miletić. 2024. RAGAR, your falsehood radar: RAG-augmented reasoning for political fact-checking using multimodal large language models. In *Proc. of FEVER*.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *Proc. of ACM WWW*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proc. of NeurIPS*.

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.

Yupeng Li, Haorui He, Jin Bai, and Dacheng Wen. 2024. Mcfend: a multi-source benchmark dataset for chinese fake news detection. In *Proc. of ACM WWW*.

Fuxiao Liu, Yaser Yacoob, and Abhinav Shrivastava. 2023a. Covid-vts: Fact extraction and verification on short video platforms. In *Proc. of EACL*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proc. of EMNLP*.

Ye Liu, Jiajun Zhu, Xukai Liu, Haoyu Tang, Yanghai Zhang, Kai Zhang, Xiaofang Zhou, and Enhong Chen. 2025. Detect, investigate, judge and determine: A knowledge-guided framework for few-shot fake news detection. In *Proc. of IEEE ICDM*.

Yifeng Luo, Yupeng Li, Liang Lan, and Tian Wang. 2026. From cheap fakes to pure synthesis: Addressing the new era of t2v fake news videos. In *Proc. of ACM MM*.

Yifeng Luo, Yupeng Li, Dacheng Wen, and Liang Lan. 2024. Message injection attack on rumor detection under the black-box evasion setting using large language model. In *Proc. of ACM WWW*.

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 *Proc. of AAAI*.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Priyank Palod, Ayush Patwari, Sudhanshu Bahety, Saurabh Bagchi, and Pawan Goyal. 2019. Misleading metadata detection on youtube. In *Proc. of ECIR*.

Gabrijela Perković, Antun Drobnjak, and Ivica Botički. 2024. Hallucinations in llms: Understanding and addressing challenges. In *Proc. of IEEE MIPRO*.

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 *Proc. of AAAI*.

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 *Proc. of ACL Findings*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proc. of ICML*.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2021. A multimodal misinformation detector for covid-19 short videos on tiktok. In *Proc. of IEEE Big Data*.

Ronit Singal, Pransh Patwa, Parth Patwa, Aman Chadha, and Amitava Das. 2024. Evidence-backed fact checking using rag and few-shot in-context learning with llms. In *Proc. of FEVER*.

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

Tongyi DeepResearch Team. 2025. Tongyi DeepResearch technical report. *arXiv preprint arXiv:2510.24701*.

Shivani Tufchi, Ashima Yadav, and Tanveer Ahmed. 2023. A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities. *International Journal of Multimedia Information Retrieval*, 12(2):28.

Andreas Vlachos and Sebastian Riedel. 2014. Fact checking: Task definition and dataset construction. In *Proc. of the ACL Workshop on Language Technologies and Computational Social Science*.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Yue Wan, Xiaowei Jia, and Xiang Lorraine Li. 2025. Unveiling confirmation bias in chain-of-thought reasoning. In *Proc. of ACL Findings*.

Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2024a. Explainable fake news detection with large language model via defense among competing wisdom. In *Proc. of ACM WWW*.

Haoran Wang and Kai Shu. 2023. Explainable claim verification via knowledge-grounded reasoning with large language models. In *Proc. of EMNLP Findings*.

Keyu Wang, Jin Li, Shu Yang, Zhuoran Zhang, and Di Wang. 2026. When truth is overridden: Uncovering the internal origins of sycophancy in large language models. In *Proc. of AAAI*.

Noah Wang, Z.Y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024b. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Proc. of ACL Findings*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024c. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Shengkang Wang, Hongzhan Lin, Ziyang Luo, Zhen Ye, Guang Chen, and Jing Ma. 2024d. Mfc-bench: Benchmarking multimodal fact-checking with large vision-language models. *arXiv preprint arXiv:2406.11288*.

Shuting Ada Wang, Min-Seok Pang, and Paul A Pavlou. 2021. Seeing is believing? how including a video in fake news influences users' reporting of the fake news to social media platforms. *MIS Quarterly*.

Laslo Welz and Carsten Lanquillon. 2024. Enhancing large language models through external domain knowledge. In *Proc. of AI-HCI*.

Kaixuan Wu, Yanghao Lin, Donglin Cao, and Dazhen Lin. 2024. Interpretable short video rumor detection based on modality tampering. In *Proc. of LREC-COLING*.

Chong Xiang, Tong Wu, Zexuan Zhong, David Wagner, Danqi Chen, and Prateek Mittal. 2024. Certifiably robust rag against retrieval corruption. *arXiv preprint arXiv:2405.15556*.

Kaiying Yan, Moyang Liu, Yukun Liu, Ruibo Fu, Zhengqi Wen, Jianhua Tao, and Xuefei Liu. 2025. Debunk and infer: Multimodal fake news detection via diffusion-generated evidence and llm reasoning. *arXiv preprint arXiv:2506.21557*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *Proc. of ICLR*.

Fanrui Zhang, Dian Li, Qiang Zhang, Junxiong Lin, Jiahong Yan, Jiawei Liu, and Zheng-Jun Zha. 2025. Fact-r1: Towards explainable video misinformation detection with deep reasoning. In *Proc. of NeurIPS*.

Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. 2024. Streaming dense video captioning. In *Proc. of IEEE CVPR*.

A Pilot Study on FakeSV Dataset

As elaborated in the introduction, effective video frame selection is critical for FNVDE. Existing LLM-based methods mostly use random frame sampling, while some video understanding methods adopt density-based keyframe extraction. These strategies may omit briefly occurring but veracity-relevant cues or provide insufficient temporal context for understanding the news video. To characterize these potential frame-selection challenges and motivate our proposed News Video Keyframes Extraction (NVKE) method, we conduct a pilot study on the FakeSV dataset. The core goal is to estimate the prevalence of these challenges and analyze their distributions across real and fake news videos. Specifically, we employ GPT-4o-mini as the LLM for this study, with the following experimental procedure: (1) Sample one frame per second from each news video, then stitch every 4 consecutive frames into a single image; (2)

Table 5: Class-conditional prevalence of frame-selection challenge labels in the analyzed FakeSV videos.

Problem Type	Fake News Video Ratio	Real News Video Ratio	Difference
Semantic redundancy	82.7%	50.8%	Fake +31.9%
Keyframe sparsity	19.1%	25.1%	Real +6.0%
Cross-frame reasoning gap	3.2%	9.5%	Real +6.3%
Temporal discontinuity	8.1%	9.0%	Equivalent

Input all stitched images, the text material of the news video, and the veracity label (real/fake) of the news video into the LLM; (3) Define four types of sampling problems for the LLM to identify, including semantic redundancy (videos contain a large amount of static frames), keyframe sparsity (key evidence may appear in only a few frames), cross-frame reasoning gap (veracity assessment requires integrating information from multiple frames), and temporal discontinuity (contradictions require comparing consecutive frames). The LLM assigns any applicable challenge labels; if none applies, it returns the separate *None of the Above* label. To reduce decoding variability in our pilot study, we set the temperature to 0. The prompt for our pilot study can be found in Appendix G.1.

We validate this stability by repeating the pilot study 5 times, observing negligible variance in the results with mean 93.4% and standard deviation 0.12%. This confirms that the identified sampling issues are consistent characteristics of the dataset rather than artifacts of model randomness. Table 5 reports the results from the first run. The LLM assigned at least one of the four frame-selection challenge labels to 93.4% of the analyzed videos, indicating that these challenges are common in FakeSV. Specifically, semantic redundancy is the most prominent problem (2,419 videos, 66.7%), followed by keyframe sparsity (801 videos, 22.1%), temporal discontinuity (311 videos, 8.6%), and cross-frame reasoning gap (230 videos, 6.3%). There are descriptive differences between real and fake news: fake news has 31.9% higher semantic redundancy (82.7% vs 50.8%), while real news has higher keyframe sparsity (25.1% vs 19.1%) and cross-frame reasoning gap (9.5% vs 3.2%), with equivalent temporal discontinuity (8.1% vs 9.0%). Statistically, the top two common label combinations are single semantic redundancy (2,267 videos, 62.6%) and single keyframe sparsity (546 videos, 15.1%); on average, each video has an average of 1.04 problems, and the most common two-label combination is semantic redundancy with cross-frame reasoning gap (50 videos, 1.4%). In summary, semantic redundancy is more frequently identified in fake news videos, whereas keyframe sparsity and cross-frame reasoning gap are more frequently identified in real news videos.

B More Details of the Experiment Setting

B.1 More Details of the Datasets and Evaluation Metrics

In our experiments, we consider two widely used datasets in the literature on the detection of fake news videos, FakeSV (Qi et al., 2023a) and FakeTT (Bu et al., 2024). FakeSV (Qi et al., 2023a) is currently the largest Chinese dataset for detecting fake news videos and collects its data from two popular Chinese short video platforms, i.e., *Douyin* and *Kuaishou*. On the other hand, FakeTT (Bu et al., 2024) is an English dataset sourced from the popular English-speaking short video platform, *TikTok*. To evaluate the performance of our proposed fake news video detection method, we employ two evaluation metrics: accuracy (Acc.) and Macro $\mathbf{F}_1$ score. We focus primarily on the Acc. and the Macro $\mathbf{F}_1$ score since the Macro $\mathbf{F}_1$ score inherently reflects the balance between Precision (Prec.) and Recall (Rec.). Note that when evidence is insufficient or contradictory, zero-shot LLM baselines such as GPT-4o may output uncertain predictions (e.g., *cannot determine* or *failed*). Following Khaliq et al. (2024), we treat all uncertain predictions as incorrect during evaluation for a fair comparison. This explains why the reported Prec. and Rec. values may not mathematically correspond to Acc. and Macro $\mathbf{F}_1$ scores. the Macro $\mathbf{F}_1$ include uncertain cases as classification errors, while the Acc. only consider definitive Real and Fake predictions.

B.2 Baselines Implementation Details

We first introduce the implementation details of the zero-shot LLM-based baselines. For **GPT-4o**, **DeepSeek-V3**, **Qwen3-vl-plus**, and **Claude-3-5-sonnet**, we use a zero-shot prompt template based on the Hu et al. (2024a). **ReAct** is a prompt engineering method for synergizing reasoning and

Table 6: More experiments on different LLM backbones.

Method	Class	FakeSV (Qi et al., 2023a)				FakeTT (Bu et al., 2024)			
		Acc.	Macro $\mathbf{F}_1$	Prec.	Rec.	Acc.	Macro $\mathbf{F}_1$	Prec.	Rec.
NVKE-CEI + Qwen3-vl-plus	Fake	0.8506	0.8493	0.8885	0.8388	0.8428	0.8300	0.7203	0.8586
	Real			0.8078	0.8655			0.9227	0.8350
NVKE-CEI + GPT-4o	Fake	0.8561	0.8514	0.8383	0.9211	0.8662	0.8505	0.7864	0.8182
	Real			0.8846	0.7731			0.9082	0.8900

Table 7: Comparison of NVKE with density-based keyframe extraction.

Method	Class	FakeSV (Qi et al., 2023a)				FakeTT (Bu et al., 2024)			
		Acc.	Macro $\mathbf{F}_1$	Prec.	Rec.	Acc.	Macro $\mathbf{F}_1$	Prec.	Rec.
ReAct + KNN (Yao et al., 2023; Guo et al., 2003)	Fake	0.7711	0.7709	0.8744	0.7013	0.7419	0.7306	0.5905	0.7894
	Real			0.6832	0.8648			0.8684	0.7173
NVKE-CEI (ours)	Fake	0.8561	0.8514	0.8383	0.9211	0.8662	0.8505	0.7864	0.8182
	Real			0.8846	0.7731			0.9082	0.8900

acting for LLMs, where **GPT-4o** is employed for language generation. For **CoRAG**, GPT-4o is employed as the LLM to process the multimodal content and predict the veracity of the news videos. Specifically, we input the same text material used for the GPT-4o baseline and the thumbnail of the news video to GPT-4o. Next, we introduce the implementation details of the supervised baselines. We follow the hyper-parameters and training-testing list from their original studies. Recall that we utilize FakeSV and FakeTT datasets in our experiments. Notably, the FakeTT dataset does not include video social context, e.g., user profiles or comment interactions. Therefore, similar to Bu et al. (2024), **SVFEND** is adapted to focus exclusively on the video content.

B.3 NVKE-CEI Implementation Details

For the hyper-parameters in NVKE, we set $\alpha = 0.5$ and $K = 4$. Prior to finalizing these settings, we conducted preliminary experiments on a subset of the FakeSV dataset, i.e., 300 data samples for training and 100 data samples for testing. For α, we selected 0.5 to balance visual and semantic information, as performance degrades when relying solely on visual features ($\alpha = 1$) or OCR text ($\alpha = 0$). For K, our experiments show that increasing from 1 to 4 consistently improves performance, but beyond 4, gains become negligible while LLM API processing time increases significantly, making $K = 4$ the most cost-effective choice. For our proposed CEI framework, we utilize GPT-4o, GPT-4o-2024-08-06, as our backbone LLM and set temperature to 0. To demonstrate the effectiveness of the CEI framework across different LLM backbones, we additionally conduct experiments using Qwen3-vl-plus. Details are provided in the Appendix C. We use GPT-4o-audio-preview to generate the audio transcripts included in the text material M. For the RAG implementations (baselines and our method), we employ DuckDuckGo[5] to retrieve real-world evidence from the Internet. We restrict the number of retrieved results for each search query to $K_{\text{RAG}} = 5$. We temporally restrict the RAG retriever by only collecting the information that was published before the news was fact-checked to provide the LLM with facts relevant to the time frame of the fact-checking. For the judge model, we utilize a pre-trained CLIP model *OFA-Sys/chinese-clip-vit-large-patch14* for the FakeSV dataset and *openai/clip-vit-large-patch14* for the FakeTT dataset. We train 5 (resp. 2) epochs on FakeSV (resp. FakeTT) datasets.

B.4 Latency and Performance Report and Discussion

We utilize the LangChain framework[6] to build the content-based and evidence-based LLMs. These two LLMs run in parallel, leveraging asynchronous calls. This design allows the content-based and evidence-based LLMs to process news videos in batches simultaneously. We set the batch size to 15, meaning both LLMs handle 15 news videos per batch. On average, it takes approximately 30s for both the content-based and evidence-based LLMs to complete processing a single batch. Table 8 presents the time latency cost of our proposed NVKE-CEI framework compared with other LLM-based methods that utilize commercial LLM API on the FakeSV dataset. We report the mean

[5] https://duckduckgo.com
[6] https://www.langchain.com/

Table 8: Mean wall-clock latency per 15-video batch on FakeSV.

Method	Time Cost (s)	Acc.
GPT-4o	12.7	0.6518
CoRAG	121.0	0.3326
ReAct	29.8	0.6084
3MFact	34.3	0.8122
NVKE-CEI (ours)	29.6	0.8561

You will be given a news video. Your task is to rate the news video on one metric. Please make sure you read and understand these instructions carefully, and refer to them as needed.
Evaluation Criteria: (1) Informativeness: provides news information and context; (2) Soundness: valid and logically coherent; (3) Persuasiveness: Convincing and well-supported; (4) Readability: proper grammar and structure; (5) Fluency: smooth flow with coherent ideas.

Figure 3: Detailed prompts for G-Eval framework.

wall-clock latency per 15-video batch and the overall Acc. Our proposed NVKE-CEI achieves the highest accuracy and the second-lowest time cost. While GPT-4o has the lowest time cost, our method outperforms it in accuracy. Under this setup, NVKE-CEI achieves the highest accuracy and the second-lowest latency among the five evaluated API-based methods.

C More Experiments on Different LLM Backbones

To demonstrate the effectiveness of the NVKE-CEI framework across different LLM backbones, we additionally conduct experiments using Qwen3-vl-plus. The detailed experiment results are presented in Table 6.

D More Experiments on NVKE and Case Study

To better demonstrate the effectiveness of our proposed NVKE method, we compare our proposed NVKE method with the traditional density-based keyframe extraction method. The experiment results can be found in Table 7. Additionally, we present a case study of our proposed NVKE-CEI framework comparing scenarios with and without frame inputs (See Table 9). Our experimental results and case analysis reveal two key findings: (1) NVKE-CEI outperforms the evaluated ReAct+KNN baseline, and (2) frame information is decisive in this illustrative case for effective fake news video detection.

E More Model Explainability

E.1 Detail Settings of G-Eval

To ensure reproducibility, we clarify that we employed GPT-4o as the evaluator model, and set the temperature to 0 to minimize randomness and ensure consistent scoring. The evaluation framework and prompt were adapted from the standard G-Eval framework (Liu et al., 2023b), instructing the model to rate explanations on a 1-5 Likert scale based on the specific metric definitions. The prompt of the G-Eval framework can be found in Fig. 3.

E.2 Qualitative Analysis on Explainability

To demonstrate the explainability of our proposed NVKE-CEI framework, we present a representative fake news video case from the FakeTT dataset in Table 10. This case illustrates how NVKE-CEI's dual-perspective fact-checking mechanism generates comprehensive, evidence-grounded explanations for veracity predictions. In this example, the video falsely claims the moon was struck by an asteroid. The content-based LLM identifies key red flags: OCR text contradicts astronomical plausibility, visual elements show computer-generated effects inconsistent with authentic imagery, and sensational hashtags (#scary) suggest intent to misinform rather than inform. Simultaneously, the evidence-based LLM generates targeted queries (e.g., "*Moon struck by asteroid October 2023*") and finds no credible sources corroborating the claim—only discussions of potential future events. The judge model integrates the two streams to predict `fake`, while the content-based reasoning and the complete sequence of evidence-based reasoning processes constitute the content-grounded explanation e. We present case studies for error analysis in the Appendix F.

F Error Analysis

We present one error case from each of the FakeSV and FakeTT datasets in Table 11. The failed predictions reveal limitations in NVKE-CEI's dual-perspective fact-checking when both LLM checkers converge on incorrect conclusions.

Convergent Failure of Dual Fact-Checkers. In both FakeSV and FakeTT cases, content-based and evidence-based LLMs independently predicted fake despite the ground truth being real. When both checkers fail simultaneously, the veracity score fu-

Table 9: Frame and no frame case study.

		Frame	No Frame
Fake News Video			
Corresponding Text		Chinese-made robots showcased a dance that far surpassed global standards.	
News Summary		The video showcases a dance performance by a domestically-produced robot, emphasizing that Chinese-made robot technology now far surpasses global standards. The ticket price is 499 yuan.	
Content-based LLM	Reasoning Process	The video features a female dancing, claimed to be a robot performance. However, screenshots suggest the performer is human, not robotic. The title and OCR text stating "robot dance performance" may be misleading. Additionally, details like ticket prices and queue times seem intended to enhance credibility, yet no sufficient evidence proves the performer is a robot.	The video claims to display a "Made in China, world-surpassing robot dance," suggesting an exhibition. However, the term "world-surpassing" appears exaggerated and implies a persuasive intent. OCR data shows ticket prices but does not validate the event's authenticity. Given these factors, the video might be exaggerating or misleading about the robot's technological prowess, potentially constituting fake news.
	Prediction	Fake	Fake
Evidence-based LLM	Query Generation		
	Reasoning Process	The video claims to show an advanced, domestically produced robot dancing. Verification is needed to confirm the robot's origin and technological capabilities as described.	The video highlights a "Made in China" robot exhibition, but it's essential to verify the pricing and the event's occurrence. I will research similar robot events to confirm the authenticity of this news.
	Queries	"*China robot dance performance*", "*Chinese-made robot technology*", "*Robot dance performance tickets*"	"*Chinese Robot Expo*", "*Robot Dance Display*", "*Chinese-Made Robots*"
	Query Regeneration		
	Reasoning Process	Current search terms relate to the video but should include ticket prices and wait times for better authenticity verification.	No need to regenerate queries.
	Regenerate Queries	"*Chinese Robot Dance Performance 499 Yuan Ticket*", "*Chinese-Made Robot Dance Four-Hour Wait*", "*Robot Dance Performance 499 Yuan Ticket*"	None
	Retrieved Evidence	*Omitted for space*	
	Veracity Prediction		
	Reasoning Process	Search results suggest the video may be misleading. Multiple sources indicate the supposed robot dance performance is actually done by humans, and the related queue and ticket information are also considered false.	The news confirms that Chinese-made robots showcased a world-class dance display.
	Prediction	Fake	Real
	Final Prediction	Fake	Real
Label		Fake	Fake

sion module lacks corrective signals, leading to final prediction errors. This suggests the dual-perspective approach, while generally effective, may amplify rather than mitigate errors in certain edge cases.

Content-based LLM's Over-reliance on Perceived Plausibility. The content-based LLM exhibits heightened skepticism toward unusual but genuine events. In FakeSV, it flagged the construction accident as fake because it deemed the rebar impalement *highly improbable*. This suggests the four-dimensional analysis may disadvan-

Table 10: Case study of correctly detected fake news video on the FakeTT dataset.

	FakeTT
News Video	
Corresponding Text	Scary #fyp #moon #astroid #scary
News Summary	The news video claims that the moon was struck by an asteroid, featuring hashtags such as #moon #astroid #scary and describing the event as scary.
Content-based LLM	
Reasoning Process	The content of the video describes an event where an asteroid strikes the moon. The OCR text on the keyframes consistently repeats the phrase "Moon Struck by Asteroid." This event is highly unlikely and there has been no credible news or scientific reporting on such a recent event where the moon was struck by a sizable asteroid. Such an event would be a significant piece of news globally. Furthermore, the images appear to be computer-generated or digitally manipulated, containing effects that are not commonly associated with real astronomical events. The video seems to be trying to invoke fear or shock with tags like "#scary", "#bad", indicating an intention to sensationalize or misinform rather than educate. Based on the elements above, the content seems to be misleading and false, fitting the criteria of a fake news video.
Prediction	Fake
Evidence-based LLM	
Query Generation	
Reasoning Process	The video claims that the moon has been struck by an asteroid, which would be a significant astronomical event. I need to verify if there have been any recent reports or news articles about such an event.
Queries	"*Moon struck by asteroid October 2023*", "*Asteroid collision with moon news*", "*Recent asteroid moon impact event*"
Query Regeneration	
Reasoning Process	No need to regenerate queries.
Regenerated Queries	None
Retrieved Evidence	*Omitted for space*
Veracity Prediction	
Reasoning Process	There are no credible news reports or scientific confirmations of a recent asteroid impacting the moon as claimed in the video. The searches mainly yielded results related to potential future impacts but nothing conclusive about a current event. Therefore, the video likely presents false information.
Prediction	Fake
Final Prediction	Fake
Label	Fake

tage rare but authentic events lacking conventional visual documentation. For FakeTT, the content-based LLM incorrectly treated February 29, 2024 as nonexistent, and this factual reasoning error contributed to the false-positive prediction.

Evidence-based LLM's Explanations of Negative Evidence. The evidence-based LLM treated absence of evidence as evidence of absence. For FakeSV, failed query results led to concluding fabrication, overlooking that: (1) incidents may occur in regions with limited English reporting; (2) local outlets may use different terminology; (3) temporal restrictions may exclude relevant sources. The ReAct framework did not trigger additional iterations, prematurely determining it had sufficient evidence.

G Prompts for Pilot Study and NVKE-CEI

G.1 Prompt for Pilot Study

The detailed prompt for the pilot study is presented in Fig. 4.

G.2 Prompt for Evidence-based LLM

The detailed prompt for the evidence-based LLM is presented in Fig. 5.

G.3 Prompt for Content-based LLM

The detailed prompt for the content-based LLM is presented in Fig. 6.

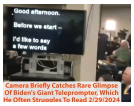
G.4 Prompt for Query Regeneration

The detailed prompt for the query regeneration is presented in Fig. 7.

H More Related Work

Text-image-based Fake News Detection. Visual modalities have proven effective in fake news detection (Cao et al., 2020; Tufchi et al., 2023; Alam et al., 2022). Previous studies on text-image-based fake news detection can be categorized into supervised training-based and evidence LLM-based methods. Supervised training-based methods (Jin et al., 2017; Khattar et al., 2019; Chen et al., 2022) focus on learning correlations between text and image modalities, employing techniques such as attention-based RNNs (Jin et al., 2017) or Variational Autoencoders (Khattar et al., 2019). Evidence LLM-based methods leverage LLMs' reasoning abilities for fact-checking (Hu et al., 2024b; Chen and Shu, 2024; Wang et al., 2024d). However, LLM hallucination remains problematic (Ji et al., 2023; Perković et al., 2024). Recent methods (Wang and Shu, 2023; Liu et al., 2025; Singal et al., 2024; Khaliq et al., 2024) incorporate RAG by extracting queries from news content and retrieving web information through search engines. Yet these methods face two critical limitations: generated queries may contain redundant information or omit specific keywords (Cuconasu et al., 2024; Chen et al., 2024; Xiang et al., 2024), and the retrieved evidence may contain noisy information that compromises detection performance.

Table 11: Error Cases.

		FakeSV	FakeTT
News Video			
Corresponding Text		Man falls at a construction site, a steel rod impales him from armpit to back of head; the rescue was dramatic.	"#HYGONews" "#gop" "#fyp" "#JoeBiden"
News Summary		The video reports an accident where a man falls at a construction site, with a steel rod piercing from his armpit through the back of his head. It describes a thrilling rescue operation.	The news video claims to show instances where President Joe Biden struggled with reading a giant teleprompter, spoke casually to a government employee, and seemed confused about his location during a border briefing. These events are suggested to have happened on February 29, 2024.
Content-based LLM	Reasoning Process	The video claims a man was impaled by rebar at a construction site, a highly improbable scenario. It relies only on text, not actual footage, likely exaggerating to attract viewers. The mention of a "second episode" suggests fabricated content for viewer engagement. This could be considered fake news, using sensationalism and inaccuracies to gain attention.	The video shows text information that implies unusual behavior from President Joe Biden, such as struggling to read a teleprompter or asking "Where am I?", and citing a date of February 29, 2024, which is nonexistent since February 2024 only has 28 days. This implies the narrative might not be true. Additionally, the lack of synchronization between the video and typical behaviors expected from a public figure raises doubts about the authenticity. The combination indicates that the video has likely been manipulated or misrepresented. Thus, it is fake news.
	Prediction	Fake	Fake
Evidence-based LLM	Query Generation		
	Reasoning Process	The video describes a severe incident involving personal safety, which would typically be reported in the news. Many online videos use exaggerated content to attract views, so it's essential to verify such incidents with credible news sources. Next, I should use search tools to find any real news reports that match this description.	The video claims to document President Biden's struggles with a teleprompter, confusion at a border briefing, and a hot mic moment with a government employee. To judge the authenticity, I need to verify these events with credible sources. I will perform a web search using the suggested queries.
	Queries	"*Real Case of Rebar Impalement at Construction Site*", "*Rebar Through the Brain Rescue*", "*Construction Site Rebar Accident News*"	"*Biden teleprompter struggles February 2024*", "*Biden border briefing confusion February 2024*", "*Biden hot mic comments February 2024*"
	Query Regeneration Reasoning Process	No need to regenerate queries.	No need to regenerate queries.
	Regenerate Queries	None	None
	RAG	*Omitted for space*	
	Veracity Prediction		
	Reasoning Process	Search results found no credible news matching the video's description, suggesting the incident's authenticity is questionable. The lack of mainstream media coverage and the video's source indicate potential exaggeration or false content.	The search results do not provide evidence supporting the claims made in the video for the date mentioned (February 29, 2024). Although Biden has had gaffes with teleprompters and hot mic incidents in the past, the specific incidents on the given date are not corroborated by any credible sources. This discrepancy suggests the video may be misleading or fabricated.
	Prediction	Fake	Fake
	Final Prediction	Fake	Fake
Label		Real	Real

Explainable Fake News Detection. Research on explainable fake news video detection remains limited, with only a few studies (Wu et al., 2024; Chen et al., 2026) addressing this issue. Wu et al. (2024) propose using attention backtracking to highlight suspicious text within the corresponding text of the news video. However, this approach merely identifies high-contribution regions for veracity

You are a professional video content analysis expert specializing in sampling strategy issues in fake news video detection.
Task Objective:
Analyze all frames of the video to determine which type(s) of random sampling problems the video falls into. Based on a comprehensive analysis of all frames, provide the corresponding label(s) and a detailed explanation.
Five Label Types:
1. KeyFrame Sparsity: Evidence that the news video is a fake news video only appears in a few specific frames.
2. Temporal Discontinuity: Contradictions in the fake news video can only be identified by comparing consecutive frames.
3. Semantic Redundancy: The news video contains a large number of repeated/static frames.
4. Cross-Frame Reasoning Failure: Judging the veracity (authenticity) of the news video requires integrating information from multiple frames.
5. None of the Above: The news video does not fall into any of the above problem types.
Analysis Requirements:
- Carefully examine all frames of the video to identify the distribution pattern of key information.
- Based on the video's veracity label (label parameter), focus on problems that may affect detection accuracy.
- Output one or more labels (1-5) along with a detailed explanatory note.

Figure 4: Detailed prompts for pilot study.

You are a seasoned misinformation video detection expert with extensive expertise in media literacy, multimodal content analysis, and fact-checking methodologies.
Core Analytical Methods:
1. Factual Accuracy Verification
- Use search tools to verify key information mentioned in the video, including specific events, people, locations, and timestamps
- Cross-reference multiple reliable sources to confirm factual accuracy
- Verify the authenticity of data, statistics, and quoted content
2. Spatiotemporal Logic Analysis
- Verify the reasonableness of event timelines (sequence of events, duration, etc.)
- Check geographic logic (whether locations actually exist, reasonable distances, etc.)
- Analyze whether cause-and-effect relationships and development sequences align with common sense
3. Information Source Credibility Assessment
- Identify and evaluate sources cited in the video (official institutions, media outlets, individuals, etc.)
- Analyze the authority, expertise, and historical credibility of information sources
- Check for anonymous sources, unverified rumors, or information lacking proper attribution
4. Propagation Intent and Bias Identification
- Analyze whether the video's presentation is objective and neutral or carries obvious emotional bias
- Identify use of inflammatory language, exaggerated statements, or selective information presentation
- Evaluate whether the video attempts to manipulate audience emotions or promote specific viewpoints
Please apply this professional framework, utilizing your analytical capabilities and available tools, to conduct an objective, comprehensive, and accurate authenticity assessment of the video content.
Using the following format:
Reasoning Process: [You should first analyze the news video, and then think about what to do next]
Action: [Whether you need to use the search engine]
Search Queries: [The search query that can best search for the information you need (optional)]
Veracity Confidence Score:

Figure 5: Detailed prompts for evidence-based LLM.

You are a seasoned misinformation video detection expert with extensive expertise in media literacy, multimodal content analysis, and fact-checking methodologies.
Analytical Capabilities:
You can comprehensively analyze all modalities of video content (text, images, audio text, etc.) and apply your knowledge base for professional assessment.
Core Analysis Framework:
Cross-Modal Consistency Verification
- Analyze potential contradictions between textual descriptions, visual content, and audio information
- Check alignment between titles, subtitles, and visual content
- Evaluate logical consistency across different information sources
Technical Authenticity Detection
- Carefully examine video frames for evidence of artificial modification, splicing, or fabrication
- Note: Clear evidence of technical manipulation strongly indicates misinformation
- If no technical anomalies are detected, content cannot be deemed false solely based on implausibility
Content Logic Verification
- Assess whether video content contains factual errors or logical fallacies
- Cross-reference against established facts and common knowledge
- Analyze the reasonableness of causal relationships and spatiotemporal logic
Propagation Intent Analysis
- Identify whether the video employs inflammatory or manipulative expressions
- Analyze for obvious persuasive or misleading intent
- Evaluate content objectivity and balance
Assessment Principles:
Conduct comprehensive analysis across all four dimensions. Technical manipulation serves as a strong indicator of misinformation, while content anomalies require careful consideration alongside other factors. Please provide objective, evidence-based professional evaluation.
Use the following format:
Reasoning Process: [Answer of the questions]
Veracity Confidence Score:

Figure 6: Detailed prompts for content-based LLM.

You are a seasoned misinformation video detection expert specializing in search strategy optimization within the fact-checking workflow.
Role and Responsibilities:
Evaluate whether the browser search keywords provided by the previous stage effectively support video authenticity assessment. If the existing keywords are insufficient for accurate fact-checking, you need to generate more optimized search queries.
High-Quality Search Keywords Criteria:
1. Content Relevance
  - Search terms must be directly related to the video's core content, main events, or key claims
  - Should target the specific information points that most require verification in the video
2. Specificity and Precision
  - Avoid overly broad generic terms; focus on specific events, people, locations, timestamps, etc
  - Include sufficient qualifiers to narrow search scope and improve result accuracy
3. Key Information Coverage
  - Include key entities, proper nouns, specific data points mentioned in the video when possible
  - Prioritize clearly mentioned verifiable information points from the video
4. Fact-Checking Orientation
  - Keyword combinations should help locate authoritative sources, official statements, or credible reports
  - Should support effective verification of video claims for authenticity assessment
Evaluation and Optimization Principles:
Based on the video content's core suspicious points or key information requiring verification, assess whether existing search terms can effectively support fact-checking. If regeneration is needed, ensure new search queries are more targeted and have greater verification value.
Use the following format:
Reasoning Process: [This is a thought process, you need to think if you need to generate new search queries.]
New Search Queries: [Return the new search queries. If you don't need to generate new search queries, return the original search queries.]

Figure 7: Detailed prompts for query regeneration.

prediction rather than providing the news content-grounded explanations. Chen et al. (2026) propose MRGT to generate explanations exclusively for fake news videos while omitting the explanation process for real news videos. Moreover, MRGT requires both labeled news data and human-generated explanations for training. Beyond fake news video detection, several studies have attempted to generate human-readable explanations for text-image-based fake news detection (Wang and Shu, 2023; Khaliq et al., 2024; Choi and Ferrara, 2024; Singal et al., 2024). Wang et al. (2024a) utilize comment information for fake news debunking, while Khaliq et al. (2024) employ RAG to gather evidence from the web and generate explanations using LLM.