Pioneering Explainable Video Fact-Checking with a New Dataset and Multi-role Multimodal Model Approach

Kaipeng Niu[1*], **Danni Xu**[2*], **Bingjian Yang**[1], **Wenxuan Liu**[3], **Zheng Wang**[1†]

[1]National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, China
[2]National University of Singapore, Singapore
[3]Peking University, China
{kaipengniu, yangbingjian, wangzwhu}@whu.edu.cn, dannixu@u.nus.edu, liuwx66@pku.edu.cn

Abstract

Existing video fact-checking datasets often lack detailed evidence and explanations, compromising the reliability and interpretability of fact-checking methods. To address these gaps, we developed a novel dataset featuring comprehensive annotations for each news item, including veracity labels, the rationales behind these labels, and supporting evidence. This dataset significantly enhances models' ability to accurately identify and explain video content. We also present an explainable automatic framework **3MFact**, utilizing **M**ulti-role **M**ultimodal **M**odels for video **Fact**-checking. Our framework iteratively gathers and synthesizes online evidence to progressively determine the veracity label, generating three key outputs: veracity label, rationale, and supported evidence. We aim for this work to be a pioneering effort, providing robust support for the field of video fact-checking.

Code — https://github.com/MeiTaylor/TRUE-3MFact

1 Introduction

Misinformation has been a persistent issue since the rise of digital media, with large models exacerbating the problem through their advanced text generation capabilities, enabling the creation of highly persuasive, multimodal misinformation (Kasneci et al. 2023; Xu, Fan, and Kankanhalli 2023). Despite numerous detection techniques and various fact-checking tools, these measures often fall short. The impressive AI-generated content (AIGC) can make simplistic classification results seem trivial, and studies show that merely labeling content as misinformation has limited persuasive power on affected users (Sanderson, Farrell, and Ecker 2022). In contrast, providing correct answers alongside errors can significantly enhance the corrective impact (Mera, Rodríguez, and Marin-Garcia 2022; Mullet and Marsh 2016). To increase persuasiveness and retention, it is crucial to offer convincing rationale and robust evidence.

Early feature-based supervised models often struggle to fully capture the context of specific claims, rendering them less effective against unseen or complex misinformation. While fact-checking websites like Snopes and PolitiFact can verify suspicious claims using external evidence, they heavily depend on human labor, making them impractical for addressing the vast volume of AI-generated misinformation. Although zero-shot methods using large language models (LLMs) have been applied to fact-checking, they often focus on isolated text (Pan et al. 2023; Zhang and Gao 2023), limiting their effectiveness in multimodal scenarios. While recent multimodal models address text-image tasks (Tahmasebi, Müller-Budack, and Ewerth 2024; Liu et al. 2024a), fact-checking for video-based content remains largely unexplored. Moreover, many early approaches lack comprehensive methodological rigor and robust experimental validation, leading to unstable and uncertain performance.

Furthermore, existing fact-checking datasets that include evidence or rationales are limited and suffer from several drawbacks: 1) Most explainable fact-checking datasets focus primarily on text modality, with limited attention to the

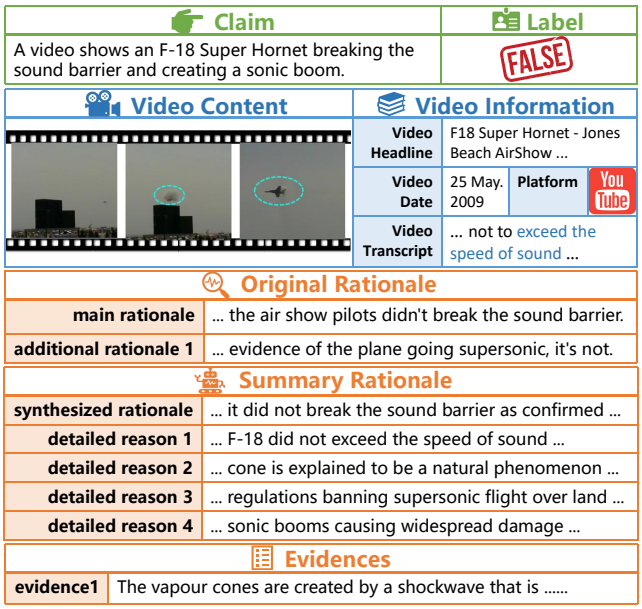


Figure 1: A sample in the proposed TRUE Dataset. It includes the claim, video, and video background information. Besides, three types of annotations are provided: 1) label, 2) evidences, and 3) original and summary rationales.

*These authors contributed equally.
†Corresponding author

Dataset	Modality T	I	A	V	Label Categories	Explanation Related ECR	LSR	Evidence	Time 2,000+	Size True / False
CHECKED	✓	✓	×	✓	✅, 🚫	×	×	×	19-20	1760 ✅ / 344 🚫
(Shang et al. 2021)	✓	×	✓	✓	✅, 🚫	×	×	×	21-21	665 ✅ / 226 🚫
Mocheg	✓	✓	×	×	✅, 🚫, 🔄	×	×	✓	-	5,144 ✅ / 5,855 🚫 / 4,602 🔄
FakeSV	✓	×	✓	✓	✅, 🚫, 🔨	×	×	✓	17-22	1827 ✅ / 1,827 🚫 / 1,884 🔨
VMH	✓	×	✓	✓	✅, 🚫	✓	×	×	14-16	341 ✅ / 1,906 🚫
TRUE (Ours)	✓	×	✓	✓	✅, 🚫 + 8 sub [1]	✓	✓	✓	16-24	1,097 ✅ / 1,828 🚫 [2]

Table 1: Comparison of the proposed TRUE with other datasets. ECR indicates the Expert-Crafted Rationale, LSR stands for the specific LLM-Summary-Rationale. [1,2]See Figure 2(a) for the sub_labels and their sizes. *Abbreviations*: T: text, A: audio, V: video, I: image. 🔄: Not Enough Information, ✅: True (Real), 🚫: False (Fake/Misleading), 🔨: Debunk

multimodal nature of online media, particularly video content (Yao et al. 2023). 2) Some datasets include machine-generated evidence and ratings, raising concerns about their reliability (Mishra et al. 2022; Abdelnabi, Hasan, and Fritz 2022). 3) Rationales and evidence are seldom well-organized within a single dataset, which undermines the robustness of the labels.

To address these challenges, we propose a zero-shot video fact-checking framework 3MFact (Multi-role Multimodal Models Fact-checking), that leverages both video and text information through Large Multimodal Models (LMMs) and LLMs. This framework analyzes both internal and external information sourced from online search, to automatically fact-check unseen and complex multimodal content without requiring additional human intervention. To support this effort and benefit the video fact-checking community, we have also developed a comprehensive dataset, **TRUE** (**T**ruthfulness and **R**ationale with **U**nderlying **E**vidence). As illustrated in Figure 1, this dataset includes rating labels, detailed reasons, evidence for each claim, featuring both original human rationales and LLM-summarized rationales sourced from reliable fact-checking websites. Extensive experiments using both traditional and novel reasoning-related metrics demonstrate that our framework produces more accurate results, supported by well-founded reasoning and robust evidence. Our contributions focus on three key areas:

- Exploratory Video Fact-Checking Dataset: We present the **first explainable video fact-checking dataset** designed to address general video-related claims. This dataset emphasizes the importance of explainable analysis supported by robust evidence and is the first to include clearly organized reasons (Human and LLM versions) directly linked to well-sourced evidence.
- Insightful Framework: We introduce a multi-role multimodal models framework 3MFact with a structured division of labor, where specific models are assigned tasks such as evidence retrieval, reasoning, and explanation. It addresses unseen multimodal misinformation by using text, video, and image data from both internal and external sources, ensuring transparent and thorough decision-making while reducing overlooking critical details.
- Innovative Standard: We establish novel metrics and a benchmark for multimodal fact-checking. Our metrics assess both accuracy and the quality of reasoning and evidence, setting a new standard for evaluating fact-checking systems in real-world scenarios.

2 Related Work

2.1 Video Fact-checking Datasets

Existing video datasets for fact-checking generally focus on truthfulness labels. For instance, the Checked dataset (Yang, Zhou, and Zafarani 2021) includes only truthfulness labels, while the FakeSV dataset (Qi et al. 2023a) offers a more comprehensive approach by incorporating social context, multimodal information, and debunking videos. The VMH dataset (Sung, Boyd-Graber, and Hassan 2023) provides some explanations but is primarily limited to misleading errors. The Mocheg dataset (Yao et al. 2023) pushes the boundaries by emphasizing multimodal fact-checking and text/image evidence collection. However, none of these datasets provide comprehensive explanations or supported evidence for general video fact-checking.

Our TRUE dataset addresses these gaps by integrating both human and LLM versions of rationales and evidence, sourced from credible fact-checking websites, targeting the video-post related claim, thereby enhancing interpretability, credibility, and comprehensiveness for video fact-checking evaluation. (see Table 1 for a detailed comparison).

2.2 Video Fact-checking Methods

Cross-modality learning improves fact-checking accuracy by integrating text, images, and videos. Models like SV-FEND (Qi et al. 2023a) and BMR (Ying et al. 2023) utilize feature fusion, while NEED (Qi et al. 2023b) leverages attention mechanisms. Traditional methods often rely on pre-trained language models like BERT and BART (Yao et al. 2023) for basic explanation generation. However, recent advancements in LLMs have significantly enhanced explanation generation and inference. For instance, QAcheck (Pan et al. 2023) and DAFND (Liu et al. 2024b) applies LLMs and exterior searching for multi-hop fact-checking, and HiSS (Zhang and Gao 2023) uses LLMs for claim decomposition and verification.

Despite these advancements, traditional methods often lack logical reasoning, failing to establish reliability, while LLM-based techniques face challenges with multimodal

content and exhibit unstable performance. Our 3MFact framework addresses these issues by integrating LMMs and LLMs for multi-modal analysis, incorporating credible online retrieval and multi-role analysis to enhance the robustness and effectiveness of video fact-checking.

3 Our TRUE Dataset

3.1 Dataset Construction

Data Collection. The dataset was sourced from Snopes[1], focusing on fact-checking articles containing videos, while excluding those with ambiguous labels like "unproven". We extracted the relevant video information for the videos in the articles from platforms like YouTube and TikTok. Following specific guidelines, we manually selected the claim-sourced video posts (target videos). Transcripts for these target videos were generated using the Deepgram API [2]. The raw dataset contains Snopes article texts, claim-related information, associated videos and video-related information, including dates, headlines, platforms, and transcripts(See the complete dataset fields on our Github).

Data Annotation. Our dataset takes a pioneering approach to enhancing the explainability and credibility of video fact-checking by being the first to specifically annotate rationales and corresponding evidence [3].We introduce two types of rationales: Expert-Crafted Rationale (ECR, also referred to as Original rationales) and LLM-Summary Rationale (LSR, also referred to as Summary rationales). ECR are **extracted directly** from Snopes articles by LLMs, preserving the original reasoning presented in the articles. Specifically, we extract both the main rationales for direct rating justification and additional supporting rationales. In contrast, LSR are generated by LLMs through **synthesization** of the article. Specifically, we transform the article into both concise yet comprehensive summaries and detailed reasons by decomposing the verification process and forming structured and traceable reasoning chains. We also collect Evidence that supports these rationales from the articles. Drawing from the Mocheg dataset's methodology (Yao et al. 2023), we extract textual evidence and external links from the `<blockquote>` tags in the source HTML of Snopes articles. Then, we employ LLMs to interpret the evidence and link it to the rationales. These annotations ensure that each claim is supported by well-defined rationales and evidence. The resulting dataset, as shown in Figure 1, showcases our detailed annotations.

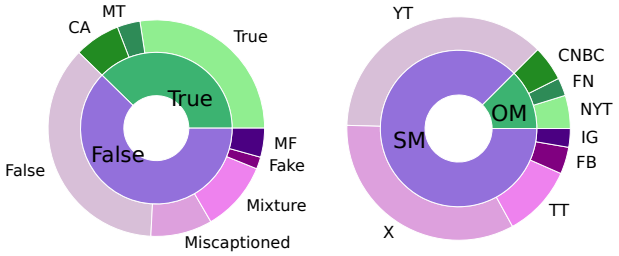
Quality Assessment. To evaluate our dataset quality, we randomly selected 27 representative samples across different time periods and sub-labels. Each sample underwent evaluation by three independent annotators from a pool of seven experts, following a systematic framework with standardized scoring criteria across three critical dimensions (Originality, Accuracy, and Comprehensiveness), as detailed in Table 2. The evaluation results show that LSR achieves significantly higher accuracy (92.3%) and comprehensiveness (98.4%) compared to ECR (70.9% and 33.7% respectively). This performance gap stems from their different construction approaches: ECR captures the natural progression of human reasoning by extracting representative segments where reasoning points are progressively developed, while LSR excels in providing structured, comprehensive summaries through systematic synthesis of the entire content.

	Originality[1]	Accuracy[2]	Comprehensiveness[3]
ECR	100%	70.90%	33.70%
LSR	-	92.30%	98.40%

[1] Originality: Unchanged from the original texts in the article.
[2] Accuracy: Consistent with the rationale's definition and semantically accurate relative to the article content.
[3] Comprehensiveness: Covering all aspects of the reasons in the article content.

Table 2: Dataset Assessment Results

Benefits of Dual Rationales. This dual annotation approach—comprising both ECR and LSR—leverages the strengths of each (reliability *vs.* systematicity) and provides diverse perspectives for evaluating explainable fact-checking results. It also facilitates comparisons between human and AI-generated fact-checking, making the dataset valuable for a wide range of research and applications.



(a) (b)

Figure 2: (a) Label Composition: False labels include False (1062), Miscaptioned (268), Mixture (305), Fake (52), and Mostly False (MF, 126). True labels include True (798), Mostly True (MT, 100), and Correct Attribution (CA, 199). (b) Social Media (SM): Platforms include Instagram (IG), TikTok (TT), Facebook (FB), X (formerly Twitter), and YouTube (YT). Official Media (OM): Sources include the New York Times (NYT), CNBC, and Fox News (FN).

3.2 Dataset Statistics

Claims in the TRUE dataset are from 2016 to 2024, with associated videos under 5 minutes in length. To preserve data authenticity, we have avoided any balancing operations, keeping the original distribution of true and false cases from controversial events. As shown in Figure 2(a), our dataset includes 1,097 true videos and 1,828 false videos, encompassing various types of misinformation. While some sub-labels appear less frequently, this distribution reflects the natural occurrence in real-world fact-checking sources (Snopes) and maintains ecological validity. The claim sources, as illustrated in Figure 2(b), include diverse official and social media platforms, enhancing the dataset's generalizability.

[1]Snopes refers to the website www.Snopes.com.
[2]Deepgram refers to the website deepgram.com.
[3]Rationale refers to the reasoning behind claims and ratings.

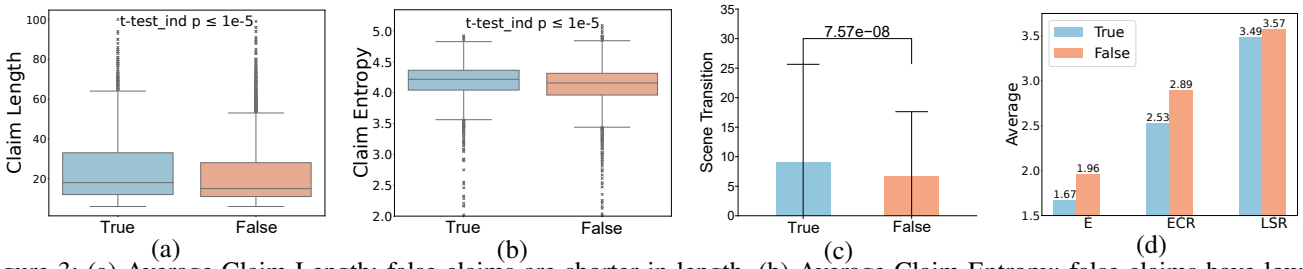


Figure 3: (a) Average Claim Length: false claims are shorter in length. (b) Average Claim Entropy: false claims have lower information richness. (c) Scene Transition Counts: Videos associated with false claims tend to have more uniform scenes. (d) Average Count of Fact-Checking Materials (E: Evidence, ECR: the Expert-Crafted Rationale, LSR: LLM-Summary-Rationale.) per Claim: False claims require more evidence and rationales for accurate judgment.

3.3 Characteristics and Research Insights

Low Information Richness in Claims. To highlight the importance of incorporating new modalities and claim-related information, we conducted a preliminary analysis of the fact-checking dataset's conventional modalities—claims. Claims are typically considered key in fact-checking (Shu et al. 2017). We used entropy (Shannon 1948) to measure the information richness in claims, with lower entropy indicating higher redundancy. As shown in Figure 3(a) and Figure 3(b), false claims typically exhibit lower information richness and shorter lengths, encouraging the search for additional context or evidence for more accurate judgments.

Incomplete Video Context. With the rise of short video platforms, individual users can freely manipulate and upload videos, often resulting in incomplete footage, such as the removal of hypothetical earlier scenes. Previous work has highlighted the significant impact of editing traces in video-based fact-checking (Bu et al. 2023). On this dataset, we use PySceneDetect to identify scene transitions. As shown in Figure 3(c), false videos tend to have fewer scene transitions, suggesting a partial lack of temporal context.

These findings suggest that misinformation often lacks complete information, undermining the reliability of human and machine judgments. This underscores the importance of external information and emphasizes the role of rationales and evidence in clarifying misinformation for online consumers. We also count the average number of rationales and evidence in our dataset. As shown in Figure 3(d), both human and LLM-generated language require more information to clarify false claims compared to true ones.

4 Framework

We propose a multi-role collaboration framework, 3MFact, which systematically transfers data between roles to verify claims related to video posts. Inspired by the success of the Question-guided Multi-hop Structure in a text fact-checking system (Pan et al. 2023), our framework also incorporates a Question-guided process. Initially, the Video Descriptor converts the video into a textual description, which, along with the input claim and video information, is passed to the Claim Verifier. The Claim Verifier assesses the sufficiency of the existing available information. If sufficient, the data proceeds to the Reasoner to generate the final truthfulness, reasons, and evidence. If not, the Claim Verifier redirects the data to the Question Manager, initiating another cycle of inquiry and evidence gathering until adequate information is accumulated or a maximum number of cycles is reached.

4.1 Problem Definition

The problem of video fact-checking involves verifying a video-related claim $\boldsymbol{c}$. The model takes as input the claim $\boldsymbol{c}$, the video content $\boldsymbol{v}$, and the video background information $\boldsymbol{\mathcal{B}}$ (such as the video title, release time, etc.). The primary output is a veracity rating y. Additionally, explainable fact-checking generates rationale $\boldsymbol{r}$ supported by evidence $\boldsymbol{\mathcal{E}_r}$.

4.2 Video Descriptor

The Video Descriptor converts video content $\boldsymbol{v}$ into a textual description $\boldsymbol{t}$ for analysis by LLM and LMM following the following steps: 1) VideoLMM generates a video-based textual description $\boldsymbol{t}_{video}$ from $\boldsymbol{v}$. 2) ImageLMM produces a set of image descriptions $\boldsymbol{\mathcal{T}}_{image} = \{\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_n\}$ for the keyframes $\boldsymbol{\mathcal{F}} = \{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_n\}$ of the video. Each $\boldsymbol{t}_i$ represents the textual description of keyframe $\boldsymbol{f}_i$. The keyframes are extract by the Katna method. 3) The overall video content is synthesized by llm to produce the final textual description $\boldsymbol{t}$:

$$\boldsymbol{t} = \text{llm}(\boldsymbol{t}_{video}, \boldsymbol{\mathcal{T}}_{image}). \quad (1)$$

This process ensures that the textual descriptions $\boldsymbol{t}$ are detailed and accurate, taking into account both visual details and the overall semantics of the video.

4.3 Claim Verifier

The Claim Verifier assesses the sufficiency of existing available information for verifying the claim $\boldsymbol{c}$. This information includes the claim $\boldsymbol{c}$, the textual video description $\boldsymbol{t}$, the video background information $\boldsymbol{\mathcal{B}}$, and question-answer pairs with evidence $\boldsymbol{\mathcal{P}}$. This module helps the system establish a reliable judgment with high certainty, avoiding unnecessary reasoning. We use a Chain-of-Thought (CoT) strategy prompt to systematically guide the LLM in inference.

As the module's output, $\alpha \in \{0, 1\}$ represents the judgment of information sufficiency, $\delta \in [0, 1]$ indicates the confidence level, and $\boldsymbol{r}_{cv}$ provides reasoning for the judgment. The process continues with the Question Generator in two cases: 1) when $\alpha = 0$, indicating insufficient information, or 2) when $\alpha = 1$ but δ is below a certain threshold (default 0.93 for higher accuracy). Only when $\alpha = 1$ and δ exceeds the threshold does the process proceed to the Reasoner.

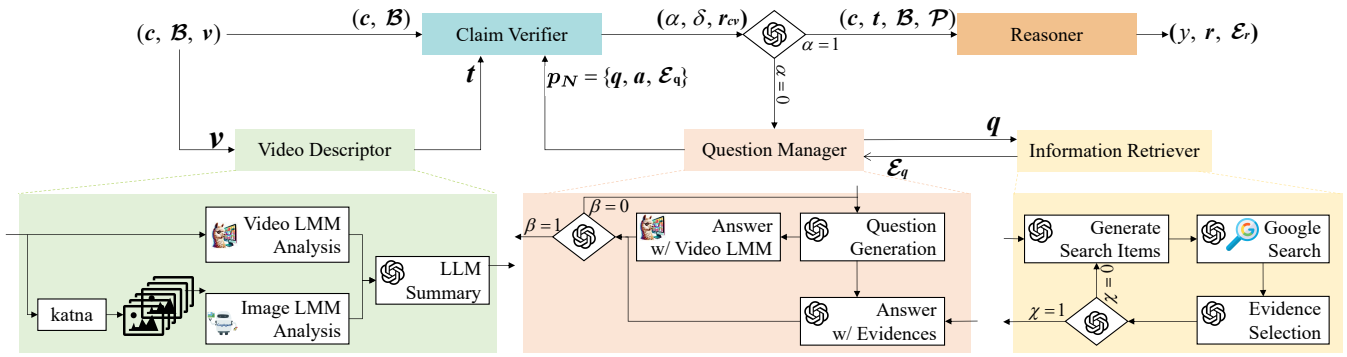


Figure 4: Overview of the proposed 3MFact framework, comprising five components: Video Descriptor (video-to-text conversion), Claim Verifier (assesses evidence sufficiency), Question Manager (generates questions and retrieves answers), Information Retriever (searches for evidence), and Reasoner (synthesizes judgment with rationale and evidence).

4.4 Question Manager

The Question Manager formulates questions $\boldsymbol{q}$ when existing information is insufficient to verify the claim $\boldsymbol{c}$, subsequently deriving answers $\boldsymbol{a}$ and evidences $\boldsymbol{\mathcal{E}}_{\text{q}}$ via video content $\boldsymbol{v}$ or online retrieval, producing new QA&Evidence pairs $\boldsymbol{p_N}$ for further claim verification.Once a question $\boldsymbol{q}$ is generated, the Question Manager decides how to proceed:

- If $\boldsymbol{q}$ pertains to the video content, the Question Manager forwards the question $\boldsymbol{q}$ and video $\boldsymbol{v}$ to VideoLMM. VideoLMM processes $\boldsymbol{v}$ to generate both the answer and direct evidence from the video, resulting in a new QA&Evidence pair $\boldsymbol{p_N} = (\boldsymbol{q}, \text{videolmm}(\boldsymbol{v}))$.
- If $\boldsymbol{q}$ requires online retrieval, the Information Retriever module searches for relevant evidence based on the question $\boldsymbol{q}$ and selects evidences $\boldsymbol{\mathcal{E}}_{\text{q}}$. The Question Manager then uses $\boldsymbol{\mathcal{E}}_{\text{q}}$ to generate an answer $\boldsymbol{a}$, resulting in a new QA&Evidence pair $\boldsymbol{p_N} = (\boldsymbol{q}, \boldsymbol{a}, \boldsymbol{\mathcal{E}}_{\text{q}})$.

The Question Manager validates the usefulness of the generated QA&Evidence pair $\boldsymbol{p_N}$ by outputting $\beta \in \{0, 1\}$, enhancing the framework's accuracy and efficiency. If $\beta = 0$, the new $\boldsymbol{p_N}$ is not useful, so the process reiterates with a newly generated question $\boldsymbol{q}$. If $\beta = 1$, $\boldsymbol{p_N}$ is considered useful and passed to the Claim Verifier for further processing.

4.5 Information Retriever

The Information Retriever module extracts key search items from the raw question $\boldsymbol{q}$, facilitating the retrieval of diverse and credible evidence needed for a well-supported answer. Specifically, $\boldsymbol{q}$ is decomposed into key items $\boldsymbol{\mathcal{K}} = \{\boldsymbol{k}_1, \boldsymbol{k}_2, \ldots, \boldsymbol{k}_m\}$, where m defaults to 2 to balance workload and result quality. The module then conducts parallel online searches, retrieving up to 10 evidence links per item and subsequently evaluating the results based on website quality, recency, and relevance, scoring each item according to the following criteria:

- Website Quality Score: s_{wq} evaluates the reliability and quality of the website content.
- Newness Score: s_{new} assesses the recency of the evidence, favoring more recent information to the claim.
- Relevance Score: s_{rlv} measures how closely the content matches the search query, focusing on relevant sentences within 250 tokens of the identified key phrases extracted from the Google search snippet.

The overall score s for each piece of raw evidence is:

$$s = w_1 \cdot s_{\text{wq}} + w_2 \cdot s_{\text{new}} + w_3 \cdot s_{\text{rlv}}, \tag{2}$$

where $w_1 = 0.25$, $w_2 = 0.25$, and $w_3 = 0.5$. We prioritize relevance by assigning a higher weight to w_3 to ensure the retrieved content closely aligns with the query.

After scoring, the Information Retriever selects the top 3 pieces of evidence $\boldsymbol{\mathcal{E}}_q$ to balance precision and coverage within LLMs' context length limits. Before passing them to the Question Manager, the relevance and usefulness of $\boldsymbol{\mathcal{E}}_q$ are additionally validated by assessing whether the evidence can adequately addresses the query. The validation output is $\chi \in \{0, 1\}$, where $\chi = 1$ means the evidence is valid and can be passed on. If $\chi = 0$, the evidence is deemed invalid, triggering a new retrieval cycle until a valid $\boldsymbol{\mathcal{E}}_q$ appears.

4.6 Reasoner

The Reasoner serves as the final decision-making module, tasked with determining the truthfulness and providing explanations based on the existing available information. The information encompasses the claim $\boldsymbol{c}$, the textual video description $\boldsymbol{t}$, the video background information $\boldsymbol{\mathcal{B}}$, and the set of question-answer-evidence pairs $\boldsymbol{\mathcal{P}} = \{\boldsymbol{p_1}, \boldsymbol{p_2} \ldots \boldsymbol{p_N}\}$, where N is the number of effective QA&Evidence pairs. The Reasoner is activated after the Claim Verifier has validated the sufficiency of information or when the system reaches the maximum allowable iterations, ensuring a definitive and informed judgment is rendered.

To guide the LLM in this crucial evaluation, we employ a meticulously designed prompt based on CoT strategy to enhance the reasoning capabilities, enabling it to integrate relevant information and obtain a well-substantiated decision with evidence cited for each rationale. The output of this module includes the binary truthfulness label $y \in \{0, 1\}$, where $y = 1$ indicates that the claim is true, and $y = 0$ indicates that it is false. Additionally, $\boldsymbol{r}$ provides the rationale

supporting the decision, and $\mathcal{E}_r$ comprises the evidence that substantiates each rationale.

5 Experiments

5.1 Experiment Setup

Datasets and Evaluation Metrics. Our experiments are conducted on the TRUE dataset. Traditional datasets either focus solely on text claims or provide incomplete reasons, making them unsuitable for our study. For experimental evaluation, we randomly selected 433 claims (15% of the complete dataset) as our test set, ensuring temporal coverage and maintaining the original True/False ratio for unbiased representation and computational efficiency. The previous quality assessment results indicate that LLM-Summary Rationale(LSR) is more accurate and comprehensive, making it the comparison rationale used in the main experiments.

For evaluating veracity accuracy, we use standard metrics: Accuracy (Acc.), Recall, Precision (Prec.), and F1-Score. For explanation evaluation, in addition to traditional metrics like BLEU (B.) and ROUGE (RG.), following (Kim et al. 2024), we refer to G-Eval (Liu et al. 2023)(G-E.) with GPT-4o-mini (OpenAI et al. 2024) to comprehensively evaluate the quality of the generated explanations on new designed metrics. Specifically, we follow G-Eval's methodology where LLM evaluates outputs through carefully designed prompts to generate a 5-point score for each metric, taking claims, fact-checking results, and ground truth as inputs. The complete set of evaluation metrics can be found in Table 4 [4].

Implementation Details. We utilize the GPT-4o-mini as the LLM and MiniCPM-V 2.6 (Yao et al. 2024) as both the VideoLMM and ImageLMM. During framework development, we optimized prompts and hyperparameters through iterative experiments on modules and the overall framework, accounting for cascading effects. The Video Descriptor module takes 7 keyframes per video as input, to balance the need to capture essential content without overloading the model. To balance accuracy and efficiency, the Claim Verifier, Question Manager, and Information Retriever each limit their iterations to three rounds.

5.2 Model Comparison Experiments

The Model Comparison Experiments evaluate models on detection accuracy (Table 3) and explanation quality (Table 4). We compared 3MFact with traditional methods (e.g., SV-FEND) and standalone VideoLMMs (VideoLLaVa (Lin et al. 2024) and MiniCPM-V 2.6). SV-FEND performed poorly in accuracy and lacked explainability, highlighting the challenges of the TRUE dataset and was excluded from Table 4. MiniCPM-V 2.6 outperformed VideoLLaVa in accuracy, but both suffered from imbalanced recall, precision, and weak explainability. In contrast, 3MFact excelled in accuracy, balance, and explainability, especially when integrated with strong components like MiniCPM-V 2.6.

Approach	Model	Acc.	Recall	Prec.	F1
Trad.	SV-FEND	62.80%	50.00%	31.40%	38.56%
LLM.	VideoLLaVa	44.11%	**95.40%**	41.50%	57.84%
	MiniCPM-V 2.6	62.73%	20.81%	60.00%	30.90%
	3MFact$_{(C+V)}$	74.83%	72.41%	67.38%	69.81%
	3MFact$_{(M+M)}$	**79.63%**	87.23%	**71.93%**	**73.85%**

Table 3: Comparison of Models for 2-Class Classification. **Trad.**: traditional methods; **LLM.**: LLM-based approaches. 3MFact$_{(C+V)}$ uses CogVLM (Wang et al. 2023) as ImageLMM and VideoLLaVa as VideoLMM, and 3MFact $_{(M+M)}$ uses MiniCPM-V 2.6 as both ImageLMM and VideoLMM.

Model	Trad.	Content Credibility Analysis (G-E.)								
	B.	RG.	R-O	LC	SoE	COM	CUR	CON	FAI	FH
VideoLLaVa	0.00	0.16	1.18	3.20	1.82	2.24	1.93	3.90	3.11	3.74
MiniCPM-V	0.00	0.16	1.43	3.69	2.08	2.67	2.12	4.03	3.55	4.16
3MFact$_{(C+V)}$	0.03	**0.26**	2.24	**4.48**	**3.98**	3.93	**3.06**	4.02	**4.32**	**4.56**
3MFact$_{(M+M)}$	**0.04**	0.25	**2.37**	4.47	3.96	**3.94**	3.01	**4.03**	4.28	4.55

Table 4: Evaluation of Models on Explanation Metrics. **R-O**: Reasons/Evidence Omission, **LC**: Logical Consistency, **SoE**: Strength of Evidence, **COM**: Comprehensiveness, **CUR**: Currency, **CON**: Conciseness, **FAI**: Faithfulness, **FH**: Fact Hallucination.

5.3 Ablation Study

The Ablation Study evaluates the impact of individual components within the 3MFact framework, identifying the key elements that significantly enhance model performance and explanation quality. The experimental results for detection accuracy (Table 5) and explanation quality (Table 6) demonstrate the contributions of these components.

Based on the ablation study results, we can broadly rank the impact of different modules or components on overall accuracy as follows: Information Retriever $>$ Validation (including Validation of $\mathcal{P}$ and Validation of $\mathcal{E}_\mathbf{q}$) $>$ Question Answering with VideoLMM $>$ Video Descriptor. This ranking reflects the relative importance of these modules in enhancing the framework's performance. Notably, due to the limitations in video descriptive capabilities, the Video Descriptor module shows limited impact on accuracy, despite advances in VideoLMMs(e.g., from VideoLLaVa to MiniCPM-V 2.6). As VideoLMM capabilities continue to improve, the Video Descriptor module is expected to bring more benefits to future video fact-checking field.

Regarding explainability, the 3MFact framework consistently provides strong and effective explanations across all configurations. Removing the VideoLMM component

[4]These metrics are aligned with theoretical and empirical studies in psychology and information science, particularly regarding factors relevant to evaluating information credibility (Metzger and Flanagin 2015). The detailed definitions of the metrics can be found in the supplementary material.

from the Question Manager unexpectedly improves some evidence-related metrics. This could because excluding VideoLMM forced the framework to rely solely on the Information Retriever to gather evidences for answering questions, which places more focus on evidence-based explanations. Conversely, eliminating the Information Retriever leads to less detailed explanations and higher conciseness scores. The results reflect the optimal performance of the complete 3MFact, where all modules work together to achieve a balanced and comprehensive explanation quality.

Model	Acc.	Recall	Prec.	F1
3MFact$_{(M+M)}$	**79.63%**	**87.23%**	**71.93%**	**78.85%**
w/o IR	69.98%	60.12%	63.03%	61.54%
w/o QA$_{LMM}$	76.18%	81.58%	66.31%	73.16%
w/o VD	79.42%	85.71%	68.67%	76.25%
w/o 2Val	72.09%	74.85%	64.00%	69.00%

Table 5: Ablation Study on Different Modules. IR: Information Retriever module, QA$_{LMM}$: Question Answering with VideoLMM, VD: Video Descriptor module, 2Val: Validation of $\mathcal{P}$ and Validation of $\mathcal{E}_q$

5.4 The Performance on Different Rationales

In this section we compare the 3MFact framework on both Expert-Crafted Rationale (ECR) and LLM-Summary-Rationale (LSR). The results of the 3MFact$_{(M+M)}$ framework are shown in Table 7. The results indicate that the 3MFact framework's performance on the original rationale is also fairly good, demonstrating its alignment with human explanations. The lower scores for the ECR compared to LSR may be due to some inaccuracies in the currently collected rationales (as shown in Table 2). Nevertheless, ECR can largely avoid fact-hallucination flaws while retaining the nuances of human reasoning. We believe that this part of the dataset can be further refined in the future, collaborating with LSR to provide a more reliable and comprehensive set of ground-truth rationales.

5.5 Case study

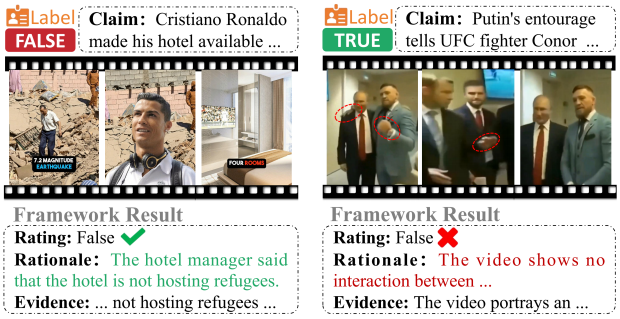
In addition to the quantitative analysis, we conducted a qualitative analysis with selected successful and unsuccessful cases to visually showcase the capabilities of our 3MFact framework. Figure 5(a) shows an example of successful detection. In this case, our framework successfully predicted the falsity of a claim by retrieving related news articles from the web that effectively refuted the claim. This demonstrates the framework's ability to utilize external evidence to challenge and verify claims. Conversely, as shown in Figure 5(b), a failed case is illustrated where an accurate claim was incorrectly classified by our framework. The error arose from inaccurate analysis of video details, leading to a miscued caption and a subsequent misjudgment. This highlights a current limitation in our framework's video detailed content analysis and the need for further improvements.

Model	Trad.		Content Credibility Analysis(G-E.)							
	B	RG	R-O	LC	SoE	COM	CUR	CON	FAI	FH
3MFact$_{(M+M)}$	**0.04**	0.25	**2.37**	**4.47**	3.96	3.94	3.01	4.03	**4.28**	**4.55**
w/o IR	0.03	0.24	1.70	4.22	2.72	3.28	2.16	**4.05**	4.03	4.15
w/o QA$_{LMM}$	0.03	0.25	2.30	4.46	**4.24**	**4.07**	**3.27**	4.00	4.26	**4.55**
w/o VD	0.03	**0.26**	2.25	4.41	4.01	4.01	3.15	4.01	4.27	4.44
w/o 2Val	**0.04**	0.25	2.17	4.41	3.92	3.91	3.11	4.02	4.27	4.45

Table 6: Ablation Study Evaluation on Explanation Metrics.

GT-R type	BLEU-4	ROUGE-L	R-O(G-E.)
ECR+LSR	0.024	0.242	2.157
ECR	0.021	0.194	2.166
LSR	**0.037**	**0.252**	**2.370**

Table 7: Comparison of Results on Different GT-R (Ground Truth Rationale) type. GT-R type: Ground Truth Rationale Type. ECR: the Expert-Crafted Rationale, LSR: the specific LLM-Summary-Rationale



(a) Successful detection example

(b) Failed detection due to video miscaption

Figure 5: Examples of successful and failed claim detection by the 3MFact framework.

6 Conclusion

We explored the first dataset TRUE for explainable video fact-checking, which emphasizes the essential of re-summarized rationale. It includes abundant multimodal information, providing an exploratory way for supporting explainable fact-checking research. The in-depth statistical analysis highlighted the necessity and practicality of TRUE. Based on this, we proposed an innovative multi-role structure 3MFact, tackling unseen misinformation among multimodals via diverse sources. We also established novel metrics to evaluate both accuracy and reasoning. Extensive experiments demonstrated the effectiveness of 3MFact in both misinformation detection and explanation. Nevertheless, significant room for improvement remains. LMMs, in particular, face challenges in fact-checking subtasks such as answering fact-related questions and generating video descriptions. Furthermore, the human-crafted rationales in our dataset require further refinement to enhance their completeness and accuracy.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62171325). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

Abdelnabi, S.; Hasan, R.; and Fritz, M. 2022. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *CVPR*, 14940–14949.

Bu, Y.; Sheng, Q.; Cao, J.; Qi, P.; Wang, D.; and Li, J. 2023. Combating Online Misinformation Videos: Characterization, Detection, and Future Directions. In *ACM Multimedia*, 8770–8780.

Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences*, 103: 102274.

Kim, K.; Lee, S.; Huang, K.-H.; Chan, H. P.; Li, M.; and Ji, H. 2024. Can LLMs Produce Faithful Explanations For Fact-checking? Towards Faithful Explainable Fact-Checking via Multi-Agent Debate. *arXiv preprint arXiv:2402.07401*.

Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2024. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *EMNLP*, 5971–5984.

Liu, X.; Li, Z.; Li, P.; Xia, S.; Cui, X.; Huang, L.; Huang, H.; Deng, W.; and He, Z. 2024a. MMFakeBench: A Mixed-Source Multimodal Misinformation Detection Benchmark for LVLMs. *arXiv preprint arXiv:2406.08772*.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In Bouamor, H.; Pino, J.; and Bali, K., eds., *EMNLP*, 2511–2522.

Liu, Y.; Zhu, J.; Zhang, K.; Tang, H.; Zhang, Y.; Liu, X.; Liu, Q.; and Chen, E. 2024b. Detect, Investigate, Judge and Determine: A Novel LLM-based Framework for Few-shot Fake News Detection. *arXiv preprint arXiv:2407.08952*.

Mera, Y.; Rodríguez, G.; and Marin-Garcia, E. 2022. Unraveling the benefits of experiencing errors during learning: Definition, modulating factors, and explanatory theories. *Psychonomic bulletin & review*, 29(3): 753–765.

Metzger, M. J.; and Flanagin, A. J. 2015. Psychological approaches to credibility assessment online. *The handbook of the psychology of communication technology*, 445–466.

Mishra, S.; Suryavardan, S.; Bhaskar, A.; Chopra, P.; Reganti, A. N.; Patwa, P.; Das, A.; Chakraborty, T.; Sheth, A. P.; Ekbal, A.; et al. 2022. FACTIFY: A Multi-Modal Fact Verification Dataset. In *DE-FACTIFY@ AAAI*.

Mullet, H. G.; and Marsh, E. J. 2016. Correcting false memories: Errors must be noticed and replaced. *Memory & Cognition*, 44: 403–412.

OpenAI et al. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Pan, L.; Lu, X.; Kan, M.-Y.; and Nakov, P. 2023. QACheck: A Demonstration System for Question-Guided Multi-Hop Fact-Checking. In Feng, Y.; and Lefever, E., eds., *EMNLP*, 264–273.

Qi, P.; Bu, Y.; Cao, J.; Ji, W.; Shui, R.; Xiao, J.; Wang, D.; and Chua, T.-S. 2023a. FakeSV: a multimodal benchmark with rich social context for fake news detection on short video platforms. In *AAAI*.

Qi, P.; Zhao, Y.; Shen, Y.; Ji, W.; Cao, J.; and Chua, T.-S. 2023b. Two Heads Are Better Than One: Improving Fake News Video Detection by Correlating with Neighbors. In *ACL*, 11947–11959.

Sanderson, J. A.; Farrell, S.; and Ecker, U. K. 2022. Examining the role of information integration in the continued influence effect using an event segmentation approach. *PloS one*, 17(7): e0271566.

Shang, L.; Kou, Z.; Zhang, Y.; and Wang, D. 2021. A Multimodal Misinformation Detector for COVID-19 Short Videos on TikTok. In *Big Data*, 899–908.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423.

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD*, 19(1): 22–36.

Sung, Y. Y.; Boyd-Graber, J.; and Hassan, N. 2023. Not all Fake News is Written: A Dataset and Analysis of Misleading Video Headlines. In Bouamor, H.; Pino, J.; and Bali, K., eds., *EMNLP*, 16241–16258.

Tahmasebi, S.; Müller-Budack, E.; and Ewerth, R. 2024. Multimodal Misinformation Detection using Large Vision-Language Models. In *CIKM*, 2189–2199.

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; Xu, J.; Xu, B.; Li, J.; Dong, Y.; Ding, M.; and Tang, J. 2023. CogVLM: Visual Expert for Pretrained Language Models. arXiv:2311.03079.

Xu, D.; Fan, S.; and Kankanhalli, M. 2023. Combating misinformation in the era of generative AI models. In *ACM Multimedia*, 9291–9298.

Yang, C.; Zhou, X.; and Zafarani, R. 2021. CHECKED: Chinese COVID-19 Fake News Dataset. *SNAM*.

Yao, B. M.; Shah, A.; Sun, L.; Cho, J.-H.; and Huang, L. 2023. End-to-End Multimodal Fact-Checking and Explanation Generation: A Challenging Dataset and Models. In *SIGIR*, 2733–2743.

Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800*.

Ying, Q.; Hu, X.; Zhou, Y.; Qian, Z.; Zeng, D.; and Ge, S. 2023. Bootstrapping multi-view representations for fake news detection. In *AAAI*, 5384–5392.

Zhang, X.; and Gao, W. 2023. Towards LLM-based Fact Verification on News Claims with a Hierarchical Step-by-Step Prompting Method. In *IJCNLP*, 996–1011.