Fine-Grained Fact-Checking for Short Videos: A Multi-Agent Report Generation Framework

Jingzhe Liu[1], Xiangyu Qiu[1, †], Wenze Ouyang[1, †], Yifan Wang[1, †], Jiayong Wen[1, †] and Guoyan Xu[1] (✉)

[1] College of Computer Science and Software Engineering, Hohai University, Nanjing, China
gy_xu@126.com

[†] These authors contributed equally to this work.

Abstract. With the rapid growth of short-video platforms, fake news presented in short-video format has become increasingly complex, multimodal, and difficult to verify. Existing text-only and image-text fact-checking methods typically assume a single explicit claim, while video-based fake news detection methods mainly provide coarse video-level labels. To address this gap, we introduce Fine-grained Fact-Checking Report Generation for Short Videos, a task that aims to identify specific misleading claims in suspicious short videos and generate evidence-based analyses. We construct a benchmark dataset of 255 short videos with 1,497 expert-annotated binary evaluation questions, enabling fine-grained assessment of claim identification and claim analysis. We further present Video Multi-agent Fine-grained Fact-checking (VMFF), a training-free, task-tailored multi-agent framework that organizes timestamped multimodal understanding, multi-claim task decomposition, and iterative Search-View evidence verification. Experiments show that VMFF outperforms adapted baselines in fine-grained claim identification and evidence-based claim analysis.

Keywords: Short Video Fact-checking, Agentic System, Multimodal Large Language Models.

1 Introduction

The rapid growth of short video platforms has greatly enriched the digital information ecosystem. However, the fragmented, multimodal, and highly narrative nature of short videos makes fake news more complex and deceptive. Given the massive volume of content, manual fact-checking is no longer sufficient, driving the need for automated fact-checking tailored to short videos.

The development of automated fact-checking has evolved from verifying text-only claims to multimodal image-text content. Recently, with the breakthroughs of large language models (LLMs) and multimodal large language models (MLLMs), many advanced fact-checking methods have emerged in the text and image-text domains. Through mechanisms like autonomous agents, multi-agent collaboration, and dynamic external knowledge retrieval [1–5], these methods can verify and explain individual claims effectively. Meanwhile, with the rise of short videos, researchers have started to focus on video fake news detection. Existing methods mostly target classification tasks

[6–8]. Some recent research [9, 10] leverages MLLMs to generate textual explanations for the video content, aiming to facilitate more accurate and interpretable classification.

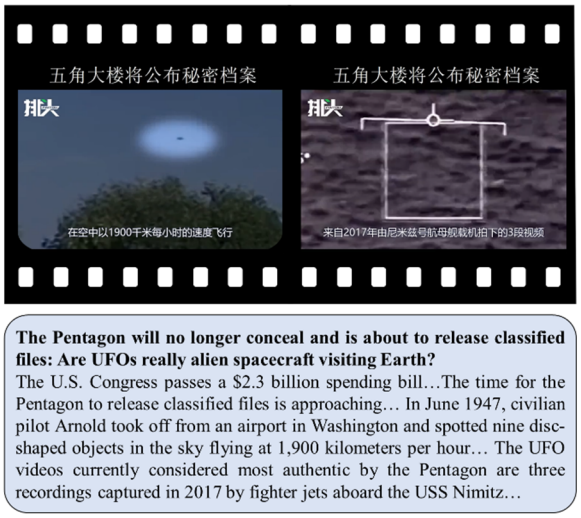

Fig. 1. Example of multi-claim multimodal disinformation. A single short video often contains multiple independent claims that require fine-grained identification and verification.

However, both text/image fact-checking and video fake news detection fall short when applied to short video fact-checking. First, fact-checking in text/image domains operates under a single-claim assumption, where a well-defined claim is explicitly provided. In contrast, as shown in Fig. 1, a single short video is highly information-dense, containing visual, textual, and audio modalities to present multiple claims simultaneously (e.g., "The US Congress passed a $2.3 billion spending bill in Dec 2020", "The Pentagon will release secret files", etc.). Existing text/image methods lack the capability to automatically decouple modalities and extract these multiple claims from a video. Second, video fake news detection yields coarse-grained true/false labels. Even when providing explanations, they merely summarize the video content rather than pinpointing exactly which specific claim is fake and why. Because they lack fine-grained external retrieval and reasoning for individual claims, they fail to provide evidence-based explanations. Furthermore, relying heavily on the parameterized knowledge of MLLMs makes these coarse explanations prone to hallucinations, severely preventing users from accurately understanding the underlying misleading mechanisms.

Therefore, we propose the task of **Fine-grained Fact-Checking Report Generation for Short Videos**. Given a suspicious short video containing textual, visual, and audio modalities, the goal is to automatically generate a structured fact-checking report that identifies specific misleading or false claims and provides evidence-based analyses for each claim. This task is designed for videos that have already entered a suspicious-video verification pipeline, rather than for general open-world video classification. Accordingly, we focus on fine-grained claim identification and claim-level analysis within

suspicious videos, instead of determining whether arbitrary videos are entirely true or false.

The main contributions of this paper are as follows:

- We define a fine-grained report generation task for suspicious short videos. Unlike coarse video-level fake news detection, the task requires systems to identify multiple specific misleading or unsupported claims and explain them with evidence.
- We construct a benchmark dataset of 255 short videos and 1,497 expert-annotated binary questions. The evaluation framework measures two complementary dimensions, Claim Identification and Claim Analysis, so that report quality can be assessed at the level of concrete verification points.
- We propose **V**ideo **M**ulti-agent **F**ine-grained **F**act-checking (VMFF), a training-free framework tailored to this task. VMFF organizes short-video verification around three task-specific designs: timestamped multimodal representation, Questioner-Planner multi-claim decomposition, and iterative Search-View evidence checking for each sub-task.

2 Related Work

2.1 Short Video Fake News Datasets

With the growing interest in mitigating fake news on short video platforms, several public datasets have emerged in recent years. For example, FakeSV [6] is currently the largest Chinese dataset for short video fake news detection. FakeTT [7] is a large-scale English dataset, extending related research to cross-lingual and English platform scenarios. However, these existing datasets mainly target video-level classification tasks. They all lack the fine-grained annotations required by our study, making it difficult to support explainable fact-checking research for the complex narratives within short videos.

2.2 Short Video Fake News Detection

Early work on short video fake news detection mainly focuses on improving the performance of classification models. For example, SV-Fend [6] leverages multimodal news content and social context features. FakingRecipe [7] approaches from a content creation perspective to capture forgery features by analyzing material selection and editing behaviors. GLFE-SVFD [8] integrates global and local feature enhancement to effectively capture deep forgery features. Although these methods achieve high classification accuracy, they act as black-box models and lack interpretability.

To address this flaw, recent studies attempt to generate explanations for detection results. ExMRD [9] introduces a novel R^3CoT mechanism and uses MLLMs to provide chain-of-thought explanations. SAFE [10] generates explanatory text through a Self-Driven Divergent Reasoning (SDDR) paradigm, in the form of multi-turn question-answering between MLLMs regarding video authenticity. However, these two methods heavily rely on the parameterized knowledge of MLLMs without introducing external

knowledge. They easily produce hallucinations, making the generated explanations unreliable.

2.3 (M)LLM-based Fact-Checking

With the rapid development of LLMs and MLLMs, many studies propose LLM- or MLLM-based fact-checking frameworks to improve process transparency and evidence reliability. For example, HiSS [1] introduces a hierarchical prompting strategy. It guides LLMs to break claims into sub-claims and performs step-by-step verification by using search engines to obtain external information. DelphiAgent [3] uses LLM-driven multi-agents to simulate the Delphi method, achieving trustworthy text-only fact-checking. In the multimodal fact-checking domain, CP-FEND [2] proposes a Chain-of-Thought prompting-based method that guides LLMs to verify authenticity through reasoning and reflection. DEFAME [4] proposes to dynamically retrieve external evidence via MLLMs for image-text verification tasks. RAMA [5] further integrates retrieval augmentation and multi-agent collaboration mechanisms to improve image-text fake news detection.

Despite these advances, existing (M)LLM-based fact-checking methods remain limited for short-video fake news. Text-only and image-text methods typically assume an explicitly provided claim, whereas short videos require claim discovery from dynamic multimodal narratives. Existing video fake news detection and explanation methods mainly provide video-level judgments or global explanations. VMFF differs by constructing timestamped multimodal representations, decomposing videos into claim-oriented sub-tasks, and verifying each sub-task through iterative external evidence checking to produce structured claim-level analyses.

3 Dataset

To support our proposed task of Fine-grained Fact-Checking Report Generation for Short Videos, we construct a benchmark dataset.

3.1 Data Collection

This dataset is derived from a subset of FakeSV [6], a widely used Chinese short video fake news detection dataset. Specifically, we select fake video topics from FakeSV and crawl related debunking articles from authoritative news platforms to serve as external factual references. After strict filtering, we retain 255 independent topics with complete evidence chains. For each topic, we manually select the most representative video based on multimodal richness and deceptive complexity. Ultimately, these steps resulted in 255 short videos.

3.2 Data Annotation

To enable objective and fine-grained evaluation, three professional fact-checkers were invited to design a scoring standard for each video based on its corresponding debunking news.

The standard consists of a series of binary questions to assess whether the model-generated report covers core verification points. These questions fall into two categories:

- Identification Questions: Evaluate whether the report accurately identifies a specific fake claim in the video.
- Analysis Questions: Evaluate whether the report provides specific logical analysis or external factual evidence for that claim.

Furthermore, experts assigned a difficulty level (1 to 3) to each question based on the human effort required for clue-finding, evidence-gathering, and reasoning (Level 1 is straightforward, while Level 3 requires deep reasoning). Table 1 presents typical examples of these binary questions.

Table 1. Example of Binary Assessment Question Annotations.

Questions	Difficulty	Type
Does it point out that "a car fell into the river between Chongqing and Lichuan recently" is a rumor?	1	Identification
Does it point out that "the incident happened between Chongqing and Lichuan" is a rumor?	2	Identification
Does it point out that the incident actually happened at Damu River, Houchang Town, Shuicheng County, Bijie City, Guizhou Province on August 30, 2017?	2	Analysis
Does it point out that the video and text were reported as early as 2017, and also circulated as fake news in August 2019, indicating it is an old rumor resurfaced?	3	Analysis

3.3 Data Features and Statistics

Despite its modest size, the dataset features fine-grained annotations and complex claim-level verification requirements. It is intended as a high-quality benchmark for evaluating fine-grained report generation, rather than a large-scale open-domain training corpus. Accordingly, we use it to examine whether methods can handle dense claim-level verification points, while leaving broader coverage of real-world short-video misinformation for future dataset expansion.

Data Scale. The dataset contains 255 short videos with an average duration of 50.9 seconds. Around these videos, experts constructed a total of 1,497 binary evaluation questions. On average, each video corresponds to about 5.9 questions.

Question Type Distribution. Out of the 1,497 evaluation questions, there are 859 identification questions (accounting for 57.4%) and 638 analysis questions (accounting for 42.6%). Fig. 2 shows the specific distribution of video durations and the number of binary questions.

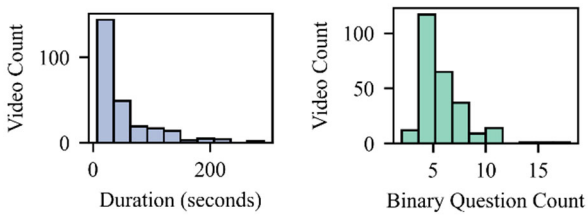

Fig. 2. Distribution of Video Duration and Binary Question Count.

Difficulty Level Distribution. As shown in Table 2, level-2 questions are the most common.

Table 2. Distribution of Binary Assessment Question Difficulty Levels.

Difficulty	Count	Percentage
1	504	33.7%
2	701	46.8%
3	292	19.5%

4 Methodology

We propose VMFF, a training-free multi-agent framework for fine-grained short-video fact-checking report generation. The framework integrates LLM, MLLM, and retrieval capabilities to address the core challenges of the proposed task: dense multimodal videos, multiple potential claims, and evidence-grounded report generation.

Specifically, VMFF consists of three modules, as illustrated in Fig. 3. **Video Understanding** converts complex multimodal inputs into timestamped textual information units. **Task Decomposition** uses Questioner-Planner collaboration to transform an ambiguous video-level verification problem into multiple claim-oriented sub-tasks. **Retrieval, Reasoning, and Generation** assigns each sub-task to a Fact-Checker agent that iteratively searches, views, and synthesizes external evidence before a Summarizer agent produces the final structured report.

This design is tailored to short-video fact-checking in three ways. First, timestamped information units preserve temporal alignment among the title, OCR, keyframes, and transcript. Second, the Questioner-Planner mechanism prevents a dense video from being treated as a single vague claim. Third, the Search-View loop separates broad

retrieval from query-focused deep reading, reducing the risk of accepting fragmented snippets as final evidence.

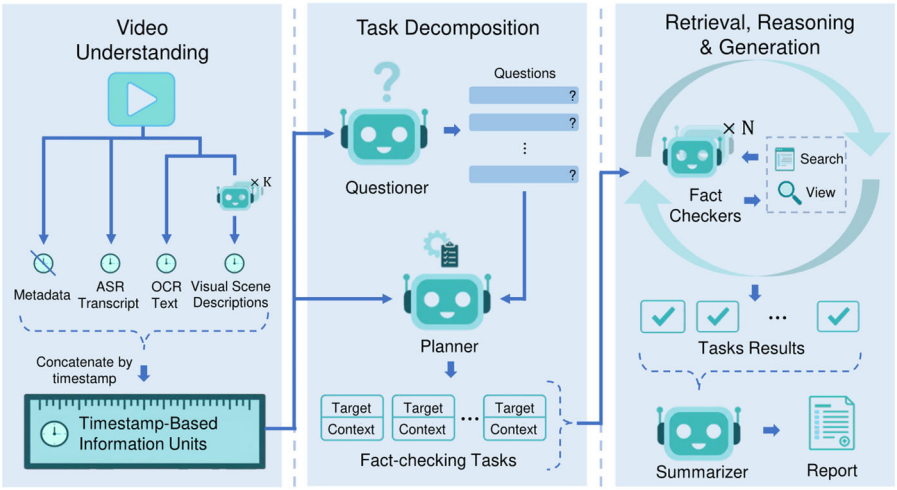

Fig. 3. Architecture of the proposed multi-agent framework for fine-grained short-video fact-checking report generation. The framework consists of three core modules: (1) Video Understanding, (2) Task Decomposition (Questioner & Planner), and (3) Retrieval, Reasoning & Generation (Fact-Checker and Summarizer Agents).

4.1 Video Understanding

Existing short-video fake news detection methods using MLLMs usually concatenate video titles, OCR text, and audio transcripts as textual input, while extracting keyframes as visual input. However, such approaches often ignore the temporal alignment between textual and visual inputs. They also rely on a single MLLM to generate a unified visual description for all keyframes, which may lose key visual details or introduce hallucinated descriptions.

To address this issue, we represent multimodal information as timestamped units to preserve the temporal structure of short videos. In addition, we use multiple MLLM calls to independently parse individual keyframes, enhancing the extraction of fine-grained visual information.

- **Text Modality** (T): We extract video metadata, including the video title and publication time, denoted as T_m.
- **Visual Modality** (V): We first extract a keyframe sequence $V_f = \{v_1, \ldots, v_k\}$ from the video and record the timestamp of each keyframe. Then, an MLLM generates a visual scene description V_d for each keyframe. We also extract OCR text from the video, denoted as V_t.
- **Audio Modality** (A): We extract the audio transcript and use Voice Activity Detection (VAD) to segment it into timestamped intervals, denoted as A_t. Since our task

focuses on verifiable semantic claims, the transcript serves as the primary audio-derived information source for subsequent claim decomposition and verification.

After these steps, we formalize all extracted content as timestamped information-unit triplets, $(timestamp, content, modality)$. Finally, we concatenate the information units of T_m, V_t, V_d, and A_t in chronological order. This constructs a temporally aligned textual representation of the video, $\mathcal{M}_{text}$.

4.2 Task Decomposition

Although we obtain the text representation $\mathcal{M}_{text}$, a short video usually contains multiple cross-modal claims. Given the effectiveness of task decomposition in solving complex reasoning problems [11], we introduce it into VMFF. The goal is to transform the macro and ambiguous video verification task into multiple sub-tasks with clear targets.

We consider that the parameterized knowledge of MLLMs is insufficient for direct fact-checking. However, MLLMs possess strong commonsense reasoning and logical error-correction capabilities. Therefore, we can use them to divergently question potential loopholes in the video content, indicating directions for subsequent verification.

To this end, we introduce a dual-agent collaboration mechanism in this module, consisting of a Questioner and a Planner. First, based on the text representation and its parameterized knowledge, the Questioner agent is prompted to propose a divergent set of questions $\mathcal{Q} = \{q_1, \dots, q_k\}$ about the video content. Then, the Planner agent integrates and plans these questions to output n independent and specific sub-tasks $\mathcal{T}_{tasks} = \{t_1, \dots, t_n\}$. Each sub-task t_i contains a verification target $target_i$ and the extracted task-relevant context $context_i$ from the text representation $\mathcal{M}_{text}$.

4.3 Retrieval, Reasoning, and Generation

To ensure efficient and focused verification, this module assigns an independent Fact-Checker agent to each sub-task to perform retrieval and reasoning in parallel. We decouple information acquisition into Search and View. Search returns the top-10 web URLs and snippets for a query, while View reads a selected webpage with respect to the query and extracts targeted evidence. VMFF therefore does not directly treat search snippets as final proof; snippets provide retrieval breadth, and View provides evidence-oriented deep reading.

During execution, the Fact-Checker agent can autonomously use Search and View. The results of each call are appended to the agent context to guide the next action, allowing multi-step evidence gathering and cross-checking across sources. The loop continues until the agent judges the evidence sufficient or reaches the preset maximum step m. The agent then synthesizes all retrieved evidence and outputs the verification conclusion for its sub-task.

After all parallel sub-tasks are completed, a Summarizer agent aggregates the sub-conclusions and core evidence output by the n Fact-Checker agents. It eliminates potential redundant information and finally outputs a structured, fine-grained fact-checking report $\mathcal{R}$ for the short video.

5 Experiments

5.1 Experimental Setup

Baselines. Current research on video fake news detection mostly focuses on coarse-grained classification. Therefore, we select two advanced video fake news detection methods with explanation generation capabilities, together with two general MLLM-based generation approaches, as baselines:

- **$\mathbf{R^3}$CoT** [9]: A chain-of-thought method. It prompts MLLMs to refine video content, retrieve domain knowledge from parameterized knowledge, and generate explanations through logical reasoning.
- **SDDR** [10]: A Self-Driven Divergent Reasoning paradigm. It provides explanations for video content through multi-turn self-questioning and answering driven by MLLMs.
- **MLLM**: A single MLLM is directly used for end-to-end fact-checking report generation.
- **Retrieval-Augmented Generation** (RAG): It introduces a basic retrieval augmentation mechanism to the end-to-end MLLM. It prompts the MLLM to construct queries based on video information and concatenates the top-10 web snippets returned by the search engine into the prompt to assist generation.

To make the comparison transparent, we adapt each baseline to the proposed report-generation task while preserving its original method design. All methods receive the same available video information and are required to produce the same report-oriented output. The MLLM baseline is evaluated as a single-pass generation method without retrieval or decomposition. RAG uses query-based retrieval with top-10 search snippets, but does not include VMFF's task planning or View tool. R^3CoT and SDDR follow their original reasoning or self-questioning procedures, as these mechanisms are central to their designs. This setup enables a fair comparison under a unified input and output setting while retaining the methods' native differences in retrieval, decomposition, and interaction abilities.

Implementation Details. In our experiments, all baselines and VMFF use Qwen3.5-Plus [12] as the backbone, while Qwen3-Max [13] serves as the judge LLM. To assess generalization, we also employ Kimi-K2.5 [14], a strong open-source MLLM, as an alternative backbone for VMFF. For VMFF, given the complexity of short-video verification, we set the maximum number of verification tasks $n = 4$ and the maximum iteration loops $m = 5$. Prompts for all methods are minimally adapted to match the required report format, without altering their original reasoning, retrieval, or interaction mechanisms.

Evaluation Metrics. Since traditional text generation metrics cannot effectively evaluate factual correctness [15], we adopt an LLM-as-a-judge approach. Specifically, we prompt a judge LLM to answer the binary questions constructed in Section 3.2 based

on the generated reports. The proportions of identification and analysis questions answered as True are defined as the Identification Score and Analysis Score, respectively. These two metrics reflect the report's coverage of fake-claim identification and verification analysis.

To validate the reliability of the judge LLM, three experts independently labeled 432 binary evaluation questions sampled from 25 randomly selected videos, achieving Fleiss' kappa of 0.82. The LLM judge, Qwen3-Max, achieved a 96.5% agreement rate with the majority-voted human labels, with Cohen's kappa of 0.93, indicating strong reliability. As an additional robustness check, we asked a second LLM judge, DeepSeek-R1, to re-evaluate the same generated reports. The two LLM judges reached a 98.15% agreement rate, with Cohen's kappa of 0.96, suggesting that the automatic evaluation is stable across judge models.

5.2 Experimental Results and Analysis

Overall Performance. As shown in Table 3, overall absolute scores are relatively low due to the challenging nature of this task, yet VMFF consistently outperforms all baselines. While zero-shot MLLM struggles with complex fact-checking, introducing basic retrieval (RAG) or reasoning (R^3CoT, SDDR) yields only moderate gains. In contrast, VMFF significantly improves both scores across two backbones. This gap demonstrates that VMFF's fine-grained video understanding, task decomposition, and iterative retrieval are crucial for accurately identifying fake claims and generating analysis. Moreover, the statistical significance of VMFF's superiority is supported by a Friedman test followed by a Nemenyi post-hoc test ($p < 0.05$, Fig. 4), highlighting its stable effectiveness.

Table 3. Performance Comparison of Methods on Identification and Analysis Questions

Method	Identification Score	Analysis Score
MLLM	0.444	0.200
RAG	0.491	0.240
R^3CoT	0.486	0.250
SDDR	0.483	0.200
$VMFF_{\text{Qwen3.5-Plus}}$	**0.590**	0.347
$VMFF_{\text{Kimi-K2.5}}$	0.586	**0.389**

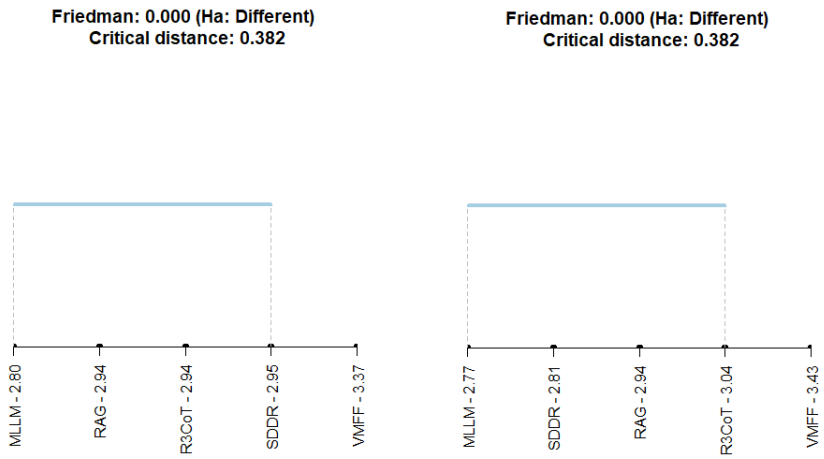

Fig. 4. Friedman test and Nemenyi post-hoc test results. Left: Identification scores. Right: Analysis scores. The results confirm that VMFF significantly outperforms all baselines.

Performance across Difficulty Levels. As shown in Fig. 5, we evaluate the results across expert-annotated difficulty levels. As expected, all models experience a performance drop on harder tasks; however, VMFF consistently outperforms all baselines across the board.

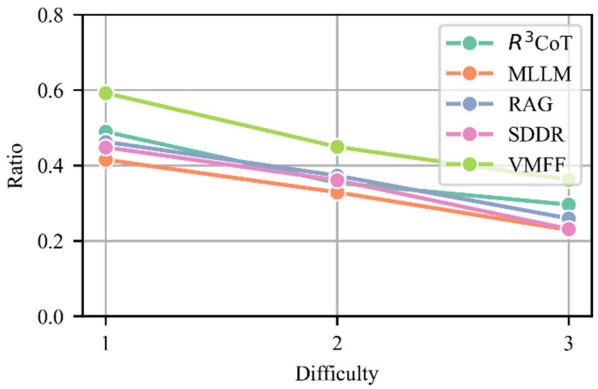

Fig. 5. Performance stratified by difficulty levels. VMFF consistently outperforms all baselines.

5.3 Ablation Study

Table 4. Results of the Ablation Study (All Experiments Use Qwen3.5-Plus)

Method	Identification Score	Analysis Score
VMFF	**0.590**	**0.347**
w/o Understanding	0.470	0.268
w/o Decomposition	0.471	0.300
w/o Retrieve	0.462	0.210
w/o View	0.344	0.256

To evaluate the specific contributions of core components in VMFF, we conduct an ablation study with four variants. As shown in Table 4, removing any module leads to a significant performance decline.

- **Video Understanding:** Replacing the timestamped representation with an overall MLLM description and simply concatenated titles, OCR text, and audio transcripts (**w/o Understanding**) decreases both scores, demonstrating that fine-grained interpretation and temporal alignment are essential for resolving complex cross-modal interactions.
- **Task Decomposition:** Using a single fact-checking agent without the Questioner-Planner module (**w/o Decomposition**) causes a notable drop in the Identification Score. This shows that decoupling complex tasks prevents a single agent from suffering information overload in dense multimodal contexts.
- **Retrieval vs. Deep Reading:** Disabling all external tools (**w/o Retrieve**) severely degrades the Analysis Score, confirming the necessity of external knowledge. Interestingly, retaining Search but removing View (**w/o View**) causes the sharpest drop in the Identification Score, even underperforming **w/o Retrieve**. This suggests that fragmented search snippets can be actively misleading for complex claims. Search breadth must be coupled with query-focused webpage reading to extract usable evidence and reduce open-web retrieval noise.

5.4 Case Study

To intuitively evaluate model behaviors, we conduct a qualitative analysis comprising both a success and a failure case.

Success Case. We analyze a 2021 video falsely claiming a Suzhou woman contracted paragonimiasis from crayfish (Table 5). Baselines completely miss the fabricated context (I1, A2, A4) and hallucinate details (I2, A1, A3). In contrast, VMFF succeeds due to its multi-agent collaboration (Fig. 3). First, **Video Understanding** accurately identifies visual elements like the ultrasonic cleaner. Next, **Task Decomposition** formulates targeted sub-tasks (e.g., “Verify the 2017 Suzhou case”), guiding fact-checking agents in the **Retrieval, Reasoning, and Generation** module to locate the original 2016 news about raw crabs. Consequently, VMFF successfully exposes the tampered facts and strictly avoids logical hallucinations.

Table 5. Success Case: Manual Qualitative Evaluation of Specific Questions

Questions (Simplified)	MLLM	RAG	R^3CoT	SDDR	VMFF
I1: A woman in Suzhou recently contracted paragonimiasis from crayfish.	M	M	M	M	**T**
I2: Eating crayfish directly causes parasitic infections.	T	T	H	M	**T**
I3: Washing live worms out of crayfish proves they carry lung flukes.	T	T	T	T	**T**
A1: Eating fully cooked crayfish will not cause paragonimiasis.	T	T	H	M	**T**
A2: The woman actually ate raw crabs, and the case was not in Suzhou.	M	M	M	M	**T**
A3: The washed-out worms are not necessarily lung flukes from inside the crayfish.	H	H	T	H	**I**
A4: The video rehashes old news originally reported in 2016.	M	I	M	M	**T**

Labels: **T**: True (Correct) **M**: Missed (Ignores this point) **H**: Hallucinated (Fabricates details) **I**: Incomplete

Table 6. Failure Case: Error Analysis of VMFF

Questions (Simplified)	VMFF
I1: Two Tianjin fishing boats rammed each other and their crews engaged in a brawl.	I
I2: A crew member stood at the bow wielding a one-meter-long "scythe" provocatively.	T
A1: The "scythe" is a common auxiliary tool used at sea for cutting nets, not a weapon.	H
A2: They were actually helping to cut the entangled nets free.	M

Failure Case. To illustrate a specific failure mode, Table 6 evaluates a misleading video claiming two fishing boat crews engaged in a brawl with a scythe. In this instance, the low video resolution causes the Video Understanding module to misinterpret the scythe simply as a long stick. Misguided by this, Task Decomposition generates subtasks verifying the object's identity, completely missing the core issue of analyzing its actual usage. Furthermore, during Retrieval, Reasoning, and Generation, the fact-

checking agent retrieves numerous unverified reposts of this fake video but no useful debunking evidence. Unable to find contradictory information, VMFF misses the actual truth that they were cooperating to cut entangled nets (A2), and hallucinates the tool's intended use (A1). This case demonstrates how an initial visual misinterpretation, combined with a lack of authoritative retrieval evidence, can lead to downstream verification errors. It also reflects a general limitation of open-domain retrieval-based fact-checking: when the web contains repeated reposts but lacks authoritative debunking sources, evidence-grounded systems remain vulnerable.

6 Conclusion

We introduced the task of fine-grained fact-checking report generation for short videos, focusing on suspicious videos that require claim-level verification rather than arbitrary open-world video classification. This task addresses the limitations of prior approaches that struggle to handle multiple claims and provide evidence-grounded claim-level analysis. To support evaluation, we curated a high-quality benchmark dataset with dense expert annotations. We also presented VMFF, a training-free multi-agent framework that simulates professional fact-checking workflows through timestamped multimodal understanding, claim-oriented task decomposition, and iterative Search-View evidence verification. Experimental results show that VMFF outperforms adapted baselines in identifying misleading claims and generating evidence-based fact-checking analyses. The results also suggest that fine-grained short-video fact-checking remains challenging, especially under open-web retrieval and multi-claim verification settings. Future work will expand the dataset in scale, diversity, and coverage, and will explore source credibility assessment, cost-aware retrieval, and more effective multi-agent coordination for practical deployment.

Acknowledgements. This work was funded by Hohai University Undergraduate Innovation and Entrepreneurship Training Program Funding (Grant No. S202510294055).

References

1. Zhang X, Gao W (2023) Towards LLM-based Fact Verification on News Claims with a Hierarchical Step-by-Step Prompting Method. In: Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Long Papers, IJCNLP-AACL 2023. Bali, Indonesia, pp 996–1011
2. Xu Y, Ge J, Lyu G, et al (2024) Multimodal Fake News Detection Based on Chain-of-Thought Prompting Large Language Models. In: Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics. Kuching, Malaysia, pp 559–566
3. Xiong C, Zheng G, Ma X, et al (2025) DelphiAgent: A trustworthy multi-agent verification framework for automated fact verification. Information Processing & Management 62:104241. https://doi.org/10.1016/j.ipm.2025.104241

4. Braun T, Rothermel M, Rohrbach M, Rohrbach A (2025) DEFAME: Dynamic Evidence-based FAct-checking with Multimodal Experts. In: Proceedings of Machine Learning Research. Vancouver, BC, Canada, pp 5383–5417
5. Yang S, Yu Z, Ying Z, et al (2025) RAMA: Retrieval-Augmented Multi-Agent Framework for Misinformation Detection in Multimodal Fact-Checking. arXiv:2507.09174
6. Qi P, Bu Y, Cao J, et al (2023) FakeSV: A Multimodal Benchmark with Rich Social Context for Fake News Detection on Short Video Platforms. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence, AAAI 2023. Washington, DC, United States, pp 14444–14452
7. Bu Y, Sheng Q, Cao J, et al (2024) FakingRecipe: Detecting Fake News on Short Video Platforms from the Perspective of Creative Process. In: MM 2024 - Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne, VIC, Australia, pp 1351–1360
8. Wang H, Yang Y, Zhong Y (2025) Global and Local Feature Enhancement for Short Video Fake News Detection. In: Lecture Notes in Computer Science. Ningbo, China, pp 435–446
9. Hong R, Lang J, Xu J, et al (2025) Following Clues, Approaching the Truth: Explainable Micro-Video Rumor Detection via Chain-of-Thought Reasoning. In: WWW 2025 - Proceedings of the ACM Web Conference. Sydney, NSW, Australia, pp 4684–4698
10. Liu F, Li Y, Lang J, et al (2026) Enhancing fake news video detection with self-driven question–answer from LMMs. Information Processing and Management 63(2): 104432
11. Khot T, Trivedi H, Finlayson M, Fu Y, Richardson K, Clark P, Sabharwal A (2023) Decomposed Prompting: A Modular Approach for Solving Complex Tasks. In: The Eleventh International Conference on Learning Representations, ICLR 2023. Kigali, Rwanda.
12. Qwen Team (2026) Qwen3.5: Towards Native Multimodal Agents. Qwen Blog.
13. Qwen Team (2025) Qwen3-Max: Just Scale it. Qwen Blog.
14. Kimi Team, Bai T, Bai Y, et al (2026) Kimi K2.5: Visual Agentic Intelligence. arXiv:2602.02276
15. Maynez J, Narayan S, Bohnet B, McDonald R (2020) On Faithfulness and Factuality in Abstractive Summarization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 1906–1919