Perception, Understanding and Reasoning, A Multimodal Benchmark for Video Fake News Detection

Anonymous ACL submission

Abstract

The advent of multi-modal large language models (MLLMs) has greatly advanced research into applications for Video fake news detection (VFND) tasks. Traditional video-based FND benchmarks typically focus on the accuracy of the final decision, often failing to provide fine-grained assessments for the entire detection process, making the detection process a black box. Therefore, we introduce the MVFNDB (Multi-modal Video Fake News Detection Benchmark) based on the empirical analysis, which provides foundation for tasks definition. The benchmark comprises 10 tasks and is meticulously crafted to probe MLLMs' perception, understanding, and reasoning capacities during detection, featuring *9,730* human-annotated video-related questions based on a carefully constructed taxonomy ability of VFND. To validate the impact of combining multiple features on the final results, we design a novel framework named MVFND-CoT, which incorporates both creator-added content and original shooting footage reasoning. Building upon the benchmark, we conduct an in-depth analysis of the deeper factors influencing accuracy, including video processing strategies and the alignment between video features and model capabilities. We believe this benchmark will lay a solid foundation for future evaluations and advancements of MLLMs in the domain of video fake news detection. We have released the MVFNDB in the following anonymous link: https://anonymous.4open.science/r/video_fnd_paper-EC8DHGOPJOH/

1 Introduction

Large Language Models (LLMs) have achieved remarkable success in tackling a wide range of general AI tasks in recent years. However, certain tasks may require multi-modal capabilities to process visual and textual information and obtain better performance. To address these needs, MLLMs have been developed and widely adopted for more complex, multi-modal tasks. With the advancement of MLLMs, numerous benchmarks have been proposed to evaluate their capacities when processing multi-modal information. These assessments typically focus on general problem-solving abilities, such as video captioning, grounding, and object counting; However, due to expert knowledge requirements, task-specific scenarios, such as video-based FND, remain under explored. Firstly, the majority of previous video-based FND benchmarks were primarily designed for classification models. However, the mechanism of these benchmarks for VFND tasks differs fundamentally from that of MLLMs in processing VFND. Therefore, proposing a novel benchmark that aligns with the paradigm of MLLMs can evaluate the cross-modal understanding ability and knowledge generalization capability of MLLMs, and pave the way for further exploration of MLLMs in this field. Secondly, lack of an assessment of interpretable results. Existing benchmarks for video-based VFND focus exclusively on the distribution of classification results. However, the detection process is practically not a black-box, it requires evidence-based inference generation. The lack of evaluation of interpretable outcomes prevents us from tracing back process of detection and deprives us of direct access to more targeted optimization guidance. Thirdly, clues for VFND lie in multiple features of the video. Consequently, VFND is a process-and-result-oriented task. However,the majority of existing benchmarks predominantly focus on result-oriented evaluation; they lack the design of corresponding tasks to assess these specific capabilities of MLLMs and the extent to which these capabilities influence the final detection results. This limitation prevents us from optimizing FND performance based on in-depth analyses of the multimodal capabilities inherent in MLLMs. Consequently, there is a pressing need to establish a more comprehensive and targeted benchmark tailored to these specialized requirements.

Given the mentioned gap, the benchmark presents the following critical advantages:

1) Comprehensive evaluation of both open-source and proprietary MLLMs on MVFND. Specifically, we evaluated the performance of both open-source and proprietary MLLMs across 10 tasks on the new test dataset, which features a more balanced news category and label distribution. Meanwhile, the selected models include both the sota models in the current video processing field and the most widely used open-source models. To explore the impact of model scale on performance, we also incorporated 3 versions of the same-type model with different scales. Our work serves as a dedicated benchmark for MLLMs in the field of MVFND.

2) Process-oriented evaluation for MVFND. We propose 10 distinct tasks that collectively evaluate the perception, understanding, and reasoning capabilities of MLLMs for MVFND process and results. On one hand, these tasks not only include general video-related tasks but also incorporate specialized tasks that are , and they ensure that the evaluation covers the entire detection process. On the other hand, for the metrics, the alignment between task properties and metric formats enables both tasks with high accuracy requirements (e.g., counting, recognition) and tasks with high semantic matching requirements (e.g., reasoning) to be effectively evaluated.

3) Applicable evaluation tasks for MVFND. Each task in the benchmark is defined based on the differentiated features of real and fake video news—these features are identified through empirical analysis. For one thing, the source video data come from original content on social platforms. This helps reduce bias in these features. For another, these features match real-world situations, and some have been validated by authoritative studies in the news field. Thus, these tasks should theoretically have certain general applicability in the MVFND.

4) Further Exploration based on MVFNDB. For one thing, we proposed a framework——MVFND-COT, which derived from evaluation result and verify the effectiveness of combination of creator-added text and original shooting footage reasoning. For another, the process-based evaluation enables us to identify the impacts of video processing strategies and non-visual features (e.g., duration, category) on VFND. Furthermore, this evaluation leads to the first discovery of a performance bottleneck that arises when there is a mismatch between model capabilities and video features.

2 Related Work

2.1 Micro-Video Fake News Detection

Fake News Detection (FND) seeks to evaluate the veracity of news by analyzing a range of characteristics across multiple modalities (Wang et al., 2024b; D'ulizia et al., 2021; Qi et al., 2023; Papadopoulou et al., 2019), including text, images, audio, and video. Early research in this field (Cao et al., 2020; Bu et al., 2024) primarily concentrated on extracting discriminative features from textual data. For instance, (Wei et al., 2025) proposed a dual-granularity adversarial training approach that integrates document-level and entity-level features (Wu et al., 2024; Shen et al., 2025), enabling models to better adapt to unseen target domains.

Compared to textual formats, videos inherently convey richer information because they combine multiple data types and complex cross-modal interactions (Zong et al., 2024), (Zheng et al., 2022), (Wu et al., 2023). For example, (Qi et al., 2023) demonstrated that integrating video content, user comments, and textual metadata (such as titles) collected from platforms like DouYin and KuaiShou can significantly enhance the effectiveness of fake news verification.

2.2 Multi-modal Perception, Understanding and Reasoning over Video

The emergence of extensive multi-modal datasets (Yuan et al., 2025), alongside architectural innovations, has significantly propelled the development of multi-modal large language models (MLLMs). These advancements have enabled MLLMs to achieve competitive performance in core multi-modal tasks such as cross-modal perception (Tang et al., 2025a), object grounding (Wasim et al., 2024), and image captioning (Bucciarelli et al., 2024). As a result, there has been growing research interest in exploring the multi-modal reasoning capabilities of MLLMs, exemplified by approaches such as multi-modal Chain-of-Thought (CoT) (Wang et al., 2025b) and VisualPRM (Wang et al., 2025a).

Videos inherently possess dynamic temporal features—including motion patterns (Qian et al., 2024b; Zhao et al., 2023; Wang et al., 2024c; Maaz et al., 2023), audio rhythms, and color tran-

sitions—that go beyond what static text and images can offer, providing unique signatures crucial for analysis. For instance, Video-of-Thought (Fei et al., 2024) introduced spatial-temporal grounding-aware tuning, which effectively bridges perceptual insights with cognitive reasoning. Video-R1 (Feng et al., 2025) further advanced reasoning capabilities, achieving superior performance across diverse video benchmarks compared to previous methods Additionally, Momentor (Qian et al., 2024a) performed segment-level reasoning and localization while CaRDiff (Tang et al., 2025b) utilized a three-stage reasoning process to reach state-of-the-art results in video saliency prediction.

2.3 MLLM Benchmark over Video

As MLLMs continue to demonstrate increasingly robust multi-modal capabilities, a variety of benchmarking frameworks (Ning et al., 2023; Zhang et al., 2025; Caba Heilbron et al., 2015; Liu et al., 2024) have been developed to evaluate their performance in perception, understanding, and reasoning. For example, MMMU (Yue et al., 2024) provides a systematic framework for assessing visual perception and reasoning abilities across a broad range of disciplines, including the arts and sciences.

MMBench-Video (Fang et al., 2024) incorporates long-form videos to assess MLLMs' spatial-temporal understanding, while MLVU (Zhou et al., 2025) evaluates models on tasks such as topic reasoning, anomaly recognition, and summarization. However, most existing benchmarks focus on generic multi-modal capabilities (Han et al., 2025; Sun et al., 2025; Wang et al., 2024a), often overlooking specialized domain-specific tasks that demand nuanced and targeted reasoning. Fake news detection uniquely requires theme-dependent verification, involving analyses that range from low-level feature integrity (e.g., HSV anomalies) to high-level semantic coherence. Unlike general benchmarks, our proposed benchmark is designed to specifically evaluate MLLMs' discriminative perception, thematic understanding, and evidential reasoning skills, using veracity-relevant feature differentiators as the basis for assessment.

3 Empirical Analysis

Our benchmark aims to comprehensively evaluate MLLMs' capabilities in both detection result and process of MVFND. Higher accuracy performance of the model can only be achieved when

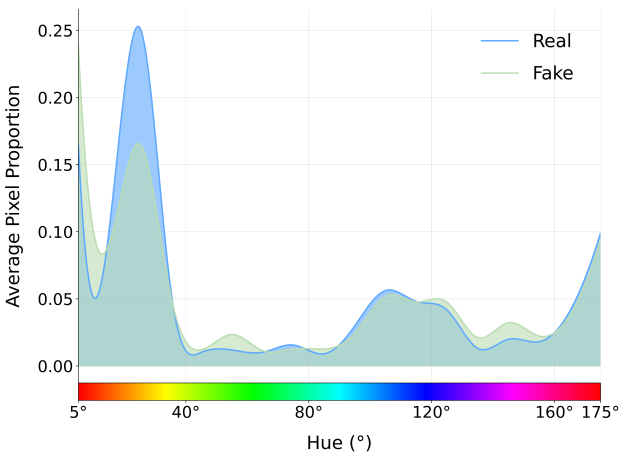


Figure 1: Hue distribution in creator-added content of fake/real news video

it accurately perceives the key features in the video during the recognition process. Therefore, we conduct an empirical analysis on a real-world dataset——FakeSV, to analyze and demonstrate the differences between real and fake news in terms of several multi-modal characteristics. Additionally, we provide plausible explanations for these phenomena from the perspectives of the purpose of news production(Firmstone, 2024) and the background of the producers. We analyses the creator-added content and original shooting footage respectively, as the former, with lower credibility, often embeds a higher degree of the author's creative intent (Guo et al., 2025). We only demonstrate some key findings from FakeSV here due to space limitations and attach remain results in the Appendix A.

3.1 Creator-added Content

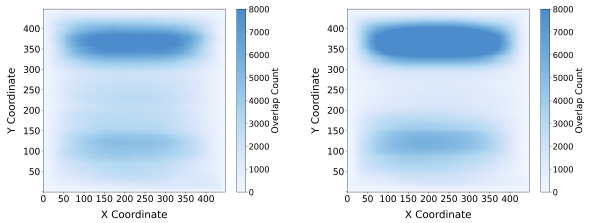
Creator-added text is typically used in video news to highlight key information, such as sites, date. We employee the PaddleOCR tools to extract text regions from video and process them with the pipeline in Figure 8.

3.1.1 Color Distribution

Finding. In terms of font color selection, fake news typically utilize colors with hue values within the range 0°-5°, whereas real news producers prefer to colors ranging 25°-30°.

Analysis. Given the prior work (Zhou et al., 2022) has indicated font color can affect the understanding ability of news readers and Hue acts as the primary identifier of a color, we conduct a census comparing the pixel-level distribution of hues in

real and fake video's creator-added text. As demonstrated in Figure 1, the results indicate a notable divergence in hue distribution between real and fake news videos. We attribute this differentiated distribution to the fake news producers tend to utilize high emotional color to influence audiences' cognition, however, real news creators employee formal color to guide users to focus on key information.



(a) Distribution of text region in fake news video

(b) Distribution of text region in real news video

Figure 2: Comparison of heatmaps

3.1.2 Spatial Distribution

Finding. In terms of text position arrangement, the text positions in fake news exhibit a higher degree of randomness, whereas those in real news show a more concentrated distribution.

Analysis. The placement selection of text shall not only ensure that the audience can effectively perceive the prompt information, but also minimize occlusion of the original shooting video frames—a balance that is critical to maintaining both the information delivery efficiency and the visual integrity of the news content. Our analysis focus on the text's spatial distribution on screen. Figure 2 highlights the distinctions between fake and real news videos. We ascribe this discrepancy to the following factors, 1) fake news producers often overlay text on video footage to obscure factual details in the original visuals. 2) fake news producers lack professional creation training about screen-text arrangement.

3.2 Original Shooting Footage

Typically, original footage in news videos conveys dynamic information about reported events. To investigate differences in original footage between fake and real news videos, we focus on analyzing consecutive video frames. These findings provide evidence supporting the tasks in our benchmark, designed to evaluate the ability of MLLMs to perceive dynamic information.

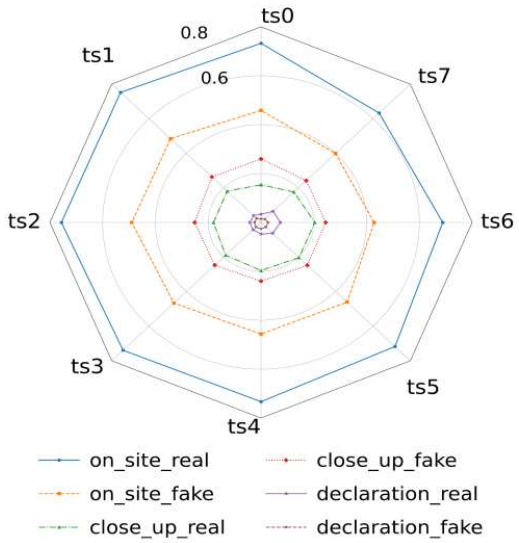


Figure 3: Distribution of key footage in fake/real news video

3.2.1 Key footage Distribution

Finding In terms of on-site shooting footage, real news contains a relatively higher proportion of such content across all time segments. Regarding close-up shooting footage of characters, real news tends to arrange such shots in the tail section of the video, whereas fake news prefer to place them at the beginning. As for official declarations, real news incorporates more these footage in the tail part of the video.

Analysis Key frames are closely tied to the news theme and hold significant evidentiary value for veracity detection. To examine the temporal distribution of key footage, we analyz the proportion of three type of key footage——on-site shooting, close-up shooting and official declarations across eight temporally sequential video segments. Figure 3 illustrates how the proportion of key footage in real and fake news videos varys over time.

4 The MVFNDB Benchmark

4.1 Overview of MVFNDB

We propose MVFNDB, which builds upon the video clips of the open-source FakeSV dataset (Qi et al., 2023). MVFNDB is a process and result oriented benchmark consisting 10 tasks evaluating MLLMs' perception, understanding and reasoning capacities during FND. The test set comprises 10.5k questions, which cover 3 distinct formats: single-choice, multiple-choice, and open-ended generation. The metrics include both exact

match and semantic match. The benchmark is distinguished by the following features:

Extensive Evaluation Datasets. Videos in FakeSV are collected from real-world platforms, so this dataset enables our evaluation to align as closely as possible with the prevalent distribution of fake news in terms of content styles, editing characteristics and misinformation creation patterns.

Diversified Evaluation Tasks. MVFNDB provides a diverse array of evaluation tasks, which are closely aligned with perception, understanding and reasoning capacities of MLLMs, such as temporal grounding, objects counting, video captioning. All tasks are tailored for MVFND process, that is to say, each task is associated with the final detection result. Specifically, the better a model performs on a single individual task, the higher the accuracy of its FND should be. The tasks encompass three categories: single-choice, multi-choice, and free-form generation. For each task category, its format is strictly tailored to align with the inherent attributes and requirements of the task type. This design enables the benchmark to assess MLLMs' both video contents perception and reasoning semantic expression capacities.

Comprehensive Process and Result Evaluation. To evaluate the end-to-end capabilities of MLLMs in FND, the MVFNDB uses the accuracy of binary classification results as a key metric. At the same time, it assesses the process of generating these results. This assessment includes evaluation of the final interpretable answer and 14 intermediate capabilities related to the detection process. The latter focuses on 3 core aspects of MLLMs in individual task: 1) Whether models can accurately identify important multi-modal features in videos based on task goals; 2) Whether they can correctly understand the main content and theme of the news; 3) Whether they can use appropriate general knowledge and video relevant characteristic to support reasoning process.

4.2 Data curation

The videos annotated in our benchmark are sourced from the FakeSV dataset, which is curated by extracting videos directly from two prevalent social media platforms——Douyin and Kuaishou. Using this dataset gives our benchmark several advantages: 1) The original labels have been cross-validated against official rumor-refutation platforms, which ensures the factual accuracy and reliability of our annotations. 2) The platform-native nature of the FakeSV dataset make the videos retain the authentic characteristics of real-world social media content style, which endows our benchmark captures generalized features reflective of actual fake news creation and achieves close alignment with the practical realities of FND. To comprehensively evaluate MLLMs' performance in MVFND, we constructed a novel test split dataset with a balanced label and category distribution. We classify news videos into six distinct categories, whose detailed descriptions are provided in Appendix B.1 and 18. This dataset enables a thorough assessment of MLLMs' ability to handle various content categories. The test split comprises 10 tasks, each containing 753 pairs.

For data annotation, as depicted in Figure 4, to reduce bias and hallucination, we utilize different types of MLLMs (Qwen2.5vl-72b and Gemini2.5-flash) with strong video process capabilities to generate and refine visual tasks. Then, we use a LLM(GPT-o1) with excellent reasoning capabilities to evaluate and filter the reasoning outputs of MLLMs. The GPT-o1 model evaluates these reasoning outputs for factual consistency and logical rationality, assigning scores from 0 to 5. Outputs scoring below 4 in either metric are reprocessed by the Gemini. Finally, perception generations are verified by 20 news creation experts after 10 hours annotation training; understanding and reasoning outputs are reviewed and revised by 10 detection experts (with 5+ years of video news censor experience in online public opinion regulatory authority) and 5 domain experts(with 5+ years of specialized field research experience). This iterative process continues until all task data is fully processed.

4.3 Task definition

According to practical detection process, evaluation tasks can be categorized into three types: 1) perception FND, which requires model to accurately capture fine-grained characteristics in videos. 2) understanding FND, which needs to be solved by observing the global contents in the video. 3) reasoning, which necessitates the joint utilization of general knowledge and information within the video. In this chapter, we present several key task definitions, while the remaining tasks are introduced in the Appendix B.2.

4.3.1 Perception

1) Key Elements Perception, KEP. The key elements perception task requires MLLMs identify

Table 1: Details for Tasks and Datasets. Avg EHR: the average elements hit rate. Avg AD: the average absolute distance. Avg ENHR: the average entities hit rate. Avg RHR: the average rationale hit rate. Avg KHR: average knowledge hit rate. FC: the factual consistency. TR: the theme relevance. CO: Completeness. CAC: the creator-added content. OSF: the original shooting footage.

Task	Task Format (output)	Dataset(Train/Test)	Category	Target Feature	Evaluation metrics
Key Elements Perception (KEP)	Open-Ended with Specific Format (text)	2.87k/0.75k	Perception	OSF	Avg EHR
Distinguish CAC and OSF (DCS)	Open-Ended with Specific Format (text)	2.87k/0.75k	Perception	CAC and OSF	ROUGE-L
Creator-added Text Color Perception(CCP)	Single Selection (text)	5.46k/1.15k	Perception	CAC	Accuracy
Creator-added Text 2D-position Perception (CPP)	Single Selection (text)	5.19k/1.08k	Perception	CAC	Accuracy
Key Elements Temporal Grounding (KEG)	Open-Ended with Specific Format (digit, text)	2.87k/0.75k	Perception	OSF	IoU
Shooting Angles Counting (SAC)	Open-Ended with Specific Format (digit)	2.87k/0.75k	Perception	OSF	Avg AD
Human Identity Recognition (HIR)	Multi Selection (options)	2.87k/0.75k	Perception	OSF	Accuracy
News Video Type Understanding (NTU)	Open-Ended (text)	2.87k/0.75k	Understanding	CAC and OSF	Accuracy
News Video Theme Understanding (NEU)	Single Selection (option)	2.87k/0.75k	Understanding	CAC and OSF	FC,TR,CO
Final Detection Reasoning (FDR)	Open-Ended (text)	2.87k/0.75k	Reasoning	CAC and OSF	Avg ENHR
Final Detection Reasoning (FDR)	Open-Ended (text)	2.87k/0.75k	Reasoning	External Knowledge	Avg KHR
Final Detection Reasoning (FDR)	Open-Ended (text)	2.87k/0.75k	Reasoning	Rationale	Avg RHR
Final Detection Reasoning (FDR)	Open-Ended (text)	/	Reasoning	Real/Fake	Accuracy

all elements that are crucial for identifying the veracity of the news video. These elements include on-site footage, the task is formatted as open-ended question, all questions are annotated by the process in Figure 4

2) Distinguish Creator-added Content and Original Shooting Footage, DCS. The distinguish task involves recognizing which elements in the video are author-added content and which correspond to the originally captured footage. The question is also annotated by pipeline in the Figure 4. And the format is open-ended generation with a specified format.

3) Creator-added Text Color Perception, CCP. In this task, MLLMs are required to perceive the font color of the provided creator-added text in the input key frame. According to our statistics, the value range of font colors in the video is limited; thus, this task involves fixed-format output with answers constrained within a finite scope. We employ accuracy as the evaluation metric.

4) Creator-added Text 2D-position Perception, CPP. The format of this question is defined as a single-choice type. The input includes 30 position exact descriptions and MLLMs need to recognize and select position of the text within screen based on the key video frame and target text. The performance is measured by accuracy. We create 6.3k Question-Answer pair by sampling key frames and human annotations.

4.3.2 Reasoning

1) Final Detection Reasoning, FDR. In this task, MLLMs need to utilized the multi-modal information and general knowledge to verify the veracity of videos. This task adopts an open-ended format while adhering to a fixed output structure:

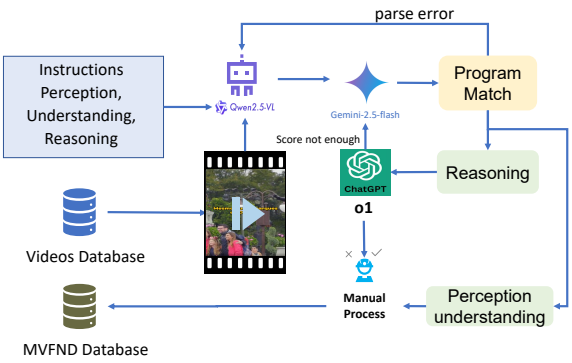


Figure 4: Pipeline of dataset construction

the reasoning process should be presented within the <think></think>tags, and the final binary classification result should be outputted within the <result></result> tags. We employ accuracy, entity hit rate, rationale hit rate, knowledge hit rate, theme relevance, factual consistency to evaluate the final detection reasoning result.

5 Experiments

5.1 Setting

We conduct comprehensive experiments on 10 tasks with 6 MLLMs in our benchmark, encompassing both open-source and proprietary models. For image MLLMs, which are primarily fine-tuned using image-related data, we leverage their inherent image process capacities to utilize sampled frames from original news video. For video MLLMs, fine-tuned with video-related instructions, we utilize a uniform sampling method or models' predefined sampling strategy for video processing. We employ models' official implementations or available APIs for evaluation. We assess both backbone and fine-tuned models with zero-shot manner. More

Table 2: Evaluation Result of MLLMs on Perception and Understanding in Fake News Detection

Model	Perception							Understanding			
	KEP	DCS	CCP	CPP	KEG	SAC	HIR	NTU	NEU(FC)	NEU(TR)	NEU(CO)
Gemini2.5-Flash	66.95	67.34	47.47	59.80	0.3819	1.88	78.72	49.72	4.36	4.40	3.8
GPT-4o-mini	55.59	34.78	45.26	28.51	0.1573	4.22	58.93	42.37	3.59	3.60	2.13
Qwen2.5-VL-72B	59.61	60.26	39.15	50.36	0.2346	2.03	77.27	44.86	4.04	4.08	2.63
Qwen2.5-VL-32B	54.30	57.03	31.71	35.73	0.2218	2.16	70.47	38.67	3.62	3.66	2.27
Qwen2.5-VL-7B	52.86	40.86	28.33	24.82	0.1589	2.71	19.38	30.84	3.37	3.51	1.87
InternVL3-78B	57.39	58.65	44.61	40.48	0.1553	2.24	59.22	41.12	3.58	3.63	2.30
InternVL3-38B	53.26	58.37	30.36	31.28	0.1663	2.28	61.39	36.95	3.45	3.41	2.08
InternVL3-8B	51.35	38.17	25.36	45.74	0.1312	2.86	16.52	28.15	3.28	3.32	1.62

Model	Final Detection Reasoning			Accuracy
	Avg ENHR	Avg RHR	Avg KHR	
Gemini1.5-Flash	62.67	79.57	47.11	78.61
GPT-4o-mini	31.74	41.26	24.75	71.05
Qwen2.5-VL-72B	46.62	68.76	37.59	77.00
Qwen2.5-VL-32B	30.14	61.36	29.85	74.90
Qwen2.5-VL-7B	25.35	40.47	22.73	70.12
InternVL3-78B	25.32	33.09	23.61	75.96
InternVL3-38B	21.17	30.05	21.51	73.43
InternVL3-8B	18.73	26.73	18.64	70.00

Table 3: Final Detection Reasoning results of different models.

evaluation-related information can be accessed in the Appendix B.3.

5.2 Main Results

The overall evaluation results for all open-source and proprietary MLLMs in the test set are shown in Table 2 and 3, from the results, we have derived several conclusions for the main results and attach them in the Appendix B.4.

5.3 Discussions

We evaluated the performance of MLLMs across various capabilities on the MVFND task, utilizing both process-based and result-based tasks within the benchmark. Through further analysis of the evaluation process, we identified the following insights and discussed their underlying intrinsic causes.

Insight 1. For MVFND task, the performance bottlenecks of the model vary across different types of news videos. As presented in Figure 5, reasoning relies on different information types across distinct news video categories. It can be observed that knowledge-based news videos rely more heavily on knowledge during the reasoning process, while exhibiting low utilization of entities in the video (e.g., screen text and footage). In contrast, the event-based news videos are less sensitive to external knowledge, and they utilize far more entities derived from the video contents.

Discussion 1. We attribute this bias to the fact that knowledge-based news, false information is typically embedded in the video scripts, whereas the video contents mostly are irrelevant to the authenticity of the news. Instead, the factual claims and logical relationships within the scripts constitute the primary targets for verification; the verification for veracity of event-based news typically focuses on determining whether the target event occurred. This process relies on the complete recording of the target event in the video; otherwise, the subsequent reasoning process lacks a valid evidence.

Insight 2. For MVFND task, dense frame sampling is not requisite for video content perception. Video frame sampling strategies should take into account both video duration and model input load. As shown in Figure 6, we take the performance of InternVL3-78B as an example. For each video duration group, there exists an optimal number of sampling frames. And this optimal number tends to increase as the video duration extends. But dense frame sampling does not guarantee a continuous improvement in accuracy; instead, it may even compromise the model performance. Dynamic frame sampling strategy yields higher accuracy, especially for shorter news videos.

Discussion 2. We ascribe the result to the news videos are typically short in duration, with key information scattered and numerous redundant frames. For authenticity detection, global perception is more important but fine-grained. Once the number of sampled frames is sufficient for the model to capture the video's key information, the bottleneck of accuracy will shift from the model's multi-modal perception capability to its FND reasoning capability. Continuing to increase the num-

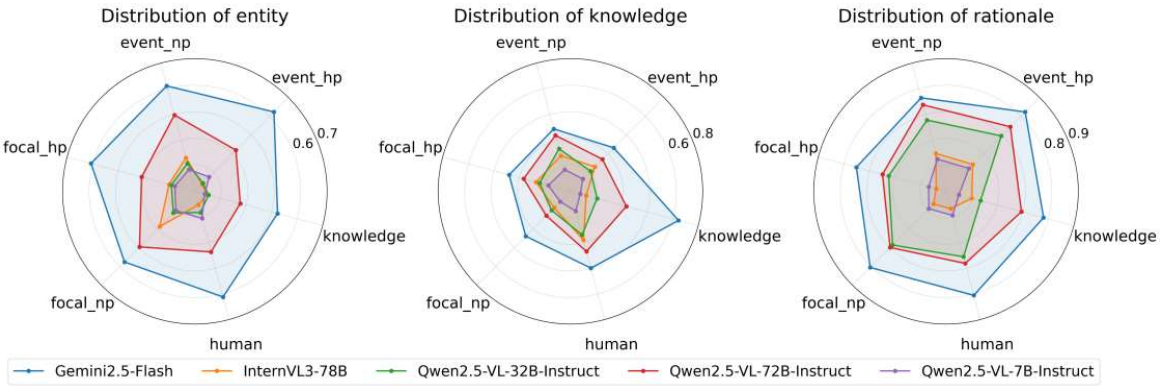


Figure 5: Comparison of entity hit rate, knowledge hit rate and rationale hit rate in reasoning output.

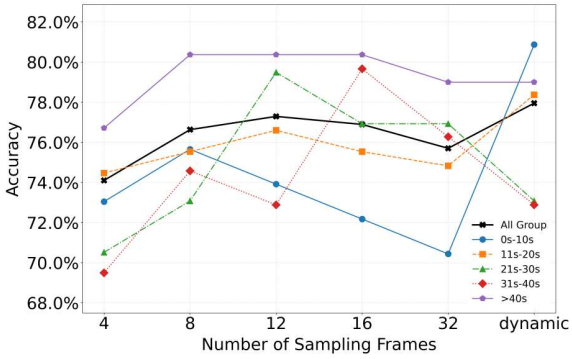


Figure 6: Comparison of Accuracy with Different Number of Sampling Frames and Duration of News Video. Dynamic Sampling FPS=1 and the Max Number of Frames is 16.

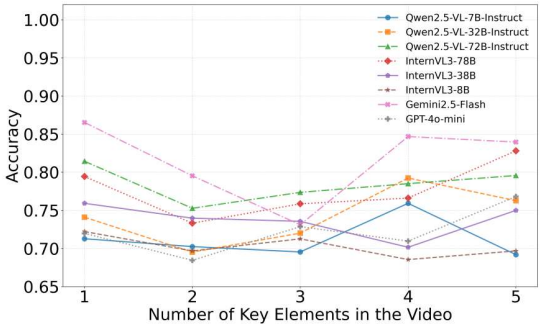


Figure 7: Comparison of Accuracy with Different Number of Key elements in the Video

ber of frames at this point will only impose greater input pressure on the model, resulting in performance degradation.

Insight 3. For MVFND task, better performance can only be achieved with sufficient key elements in the video and the model's temporal localization capability. As demonstrated in Figure 7, when the number of key elements is one, all models achieve relatively favorable performance. But as the number of key elements in the video increases, only some more advanced models exhibit improved performance.

Discussion 3. We attribute this phenomenon to the fact that, the number of key elements generally increases as the news video duration grows. Thus, MLLMs can handle most samples containing only one element. However, as the number of elements increases, less capable models fail to capture the key information within these longer videos. As presented in Figure 21, model's temporal localization capability decrease with the number of key elements needed to be localized. However, MLLMs that capture key information can obtain more descriptions of news content, thereby improving their accuracy.

Limitations

Our work has explored the utilization of MLLMs for VFND, however, there are also some limitations Firstly, Fake news instances of the same type may exhibit similar deceptive characteristics, as the nature of the facts they aim to obscure is comparable. This aspect, however, remains underexplored in our current work. Secondly, While we have made every effort to maintain a balanced distribution of news video types within the dataset, there remains room for improvement. type balance should be incorporated as a key evaluation metric for dataset construction in the future works.

References

Mary Angela Bock, Robert J. Richardson, Christopher T. Assaf, and Dariya Tsyrenzhapova. 2023. Improvisation, economy, and mtv moves: Online news and video production style. *Electronic News*, 17(3):146–163.

Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. 2024. Fakingrecipe: Detecting fake news on short video platforms from the perspective of creative process. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1351–1360.

Davide Bucciarelli, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Personalizing multimodal large language models for image captioning: an experimental analysis. In *European Conference on Computer Vision*, pages 351–368. Springer.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970.

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.

Arianna D'ulizia, Maria Chiara Caschera, Fernando Ferri, and Patrizia Grifoni. 2021. Fake news detection: a survey of evaluation datasets. *PeerJ Computer Science*, 7:e518.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. 2024. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. 2024. Video-of-thought: Step-by-step video reasoning from perception to cognition. *arXiv preprint arXiv:2501.03230*.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2025. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*.

Julie Firmstone. 2024. *Routines and Practices: Studying the Making of News*, pages 63–84. Springer International Publishing, Cham.

Lei Guo, Chris Chao Su, and Hsuan-Ting Chen. 2025. Do news frames really have some influence in the real world? a computational analysis of cumulative framing effects on emotions and opinions about immigration. *The International Journal of Press/Politics*, 30(1):211–231.

Hui Han, Siyuan Li, Jiaqi Chen, Yiwen Yuan, Yuling Wu, Yufan Deng, Chak Tou Leong, Hanwen Du, Junchen Fu, Youhua Li, and 1 others. 2025. Videobench: Human-aligned video generation benchmark. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18858–18868.

Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang W Chen. 2024. Et bench: Towards open-ended event-level video-language understanding. *Advances in Neural Information Processing Systems*, 37:32076–32110.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. 2023. Videobench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*.

Olga Papadopoulou, Markos Zampoglou, Symeon Papadopoulos, and Ioannis Kompatsiaris. 2019. A corpus of debunked and verified user-generated videos. *Online information review*, 43(1):72–88.

Andrea Pereira, Elizabeth Harris, and Jay J. Van Bavel. 2023. Identity concerns drive belief: The impact of partisan identity on the belief and dissemination of true and false news. *Group Processes & Intergroup Relations*, 26(1):24–47.

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

Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. 2024a. Momentor: Advancing video large language model with fine-grained temporal reasoning. *arXiv preprint arXiv:2402.11435*.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. 2024b. Streaming long video understanding with large language models. *Advances in Neural Information Processing Systems*, 37:119336–119360.

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

Shangkun Sun, Xiaoyu Liang, Songlin Fan, Wenxu Gao, and Wei Gao. 2025. Ve-bench: Subjective-aligned benchmark suite for text-driven video editing quality assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7105–7113.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, and 1 others. 2025a. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*.

Yunlong Tang, Gen Zhan, Li Yang, Yiting Liao, and Chenliang Xu. 2025b. Cardiff: Video salient object ranking chain of thought reasoning for saliency prediction with diffusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7302–7310.

Andong Wang, Bo Wu, Sunli Chen, Zhenfang Chen, Haotian Guan, Wei-Ning Lee, Li Erran Li, and Chuang Gan. 2024a. Sok-bench: A situated video reasoning benchmark with aligned open-world knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13384–13394.

Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2024b. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM Web Conference 2024*, pages 2452–2463.

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, and 1 others. 2025a. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. 2024c. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. 2025b. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*.

Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. 2024. Videogrounding-dino: Towards open-vocabulary spatio-temporal video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18909–18918.

Wenjie Wei, Yanyue Zhang, Jinyan Li, Panfei Liu, and Deyu Zhou. 2025. Cross-domain fake news detection based on dual-granularity adversarial training. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9407–9417.

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

Lianwei Wu, Yuzhou Long, Chao Gao, Zhen Wang, and Yanning Zhang. 2023. Mfir: Multimodal fusion and inconsistency reasoning for explainable fake news detection. *Information Fusion*, 100:101944.

Yuan Yuan, Zhaojian Li, and Bin Zhao. 2025. A survey of multimodal learning: Methods, applications, and future. *ACM Computing Surveys*, 57(7):1–34.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Zicheng Zhang, Ziheng Jia, Haoning Wu, Chunyi Li, Zijian Chen, Yingjie Zhou, Wei Sun, Xiaohong Liu, Xiongkuo Min, Weisi Lin, and 1 others. 2025. Q-bench-video: Benchmark the video quality understanding of lmms. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3229–3239.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. 2023. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597.

Jiaqi Zheng, Xi Zhang, Sanchuan Guo, Quan Wang, Wenyu Zang, and Yongdong Zhang. 2022. Mfan: Multi-modal feature-enhanced attention networks for rumor detection. In *IJCAI*, volume 2022, pages 2413–2419.

Jianlong Zhou, Xianglin Miao, Feijuan He, and Yalin Miao. 2022. Effects of font style and font color in news text on user cognitive load in intelligent user interfaces. *IEEE Access*, 10:10719–10730.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, and 1 others. 2025. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13691–13701.

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

A Empirical Analysis

A.1 Data process

For empirical analysis, we also divide the data into creator-added contents and original shooting footage. Because creator-added contents carry more author's intents, so it has less credibility. In order to census more fine-grained data, we construct two data process pipelines to collect creator-added contents and original shooting scenes in details. As shown in Figure 10, The content in the yellow box is the text added by the author, while the content in the red box is the text present in the original footage.

Most of author-added elements in news video are text, the creators utilize specialized software tools to insert textual elements as complementary information. For example, adding captions to explain event progression and introduce key figures, sometimes authoritative medias may exhibit distinctive graphical identifiers. Apart from discrepancies in the Hue, 2D-position of creator-added contents(CAC), there are also a few characteristics of video news have demonstrate discrepancy. We add other analysis towards to fake and real video news in the following parts. As the Figure 8 shown,For creator-added contents process pipeline, we uniformly sampled 16 frames per video using FFmpeg, maintaining temporally equidistant intervals between extracted frames.After that, we leverage the OCR function of Paddle to extract text regions of each frame. The outputs of Paddle include bounding box and screenshots. However there are also a few text regions extracted from original shooting footage, such as bus stop marker. So we manually removed these texts and get all creator-added text regions. Then, we design prompt templates and employ qwen2.5vl-72b to recognize and generate semantic description of text content, color, 2D-position and aspect ratios. Finally, experts refine the descriptions generated by MLLMs. Figure 9 demonstrates the overall process of extraction, generation and refinement.

Textual information is usually placed on video screen by authors to deliver key descriptions about the news. However, in order to minimize text occlusion with the original shooting scenes and maximize the transmission efficiency of both textual and visual information, the text region should be designed empirically with above considerations. To analyze the outline discrepancy of text regions in fake/real news video, we census the area size and aspect ratio of them. As shown in Figure 12, real videos are more likely to utilize a moderate region size, while fake videos have a larger proportion in both relatively large and small parts. Such comparison result aligns with practical observations, where larger text regions are often employed by false information creators to divert viewers' attention toward misleading text content or to occlude authentic details in the original captured scene. Smaller text regions typically emerge when fake news producers have limited relevant news material of the news. The Figure 13 demonstrates that aspect ratio of text regions in news videos serves as a distinct delimiter for differentiating between authentic and fabricated content. Specifically, text regions' aspect ratio in fake news videos tend to be smaller than 15, whereas those in real news videos typically have an aspect ratio larger than 15.

A.1.1 Color Distribution

In color theory, hue, saturation, and value (HSV) are the fundamental attributes of each color. Color is typically employed to emphasize key information in video news. Saturation describes the intensity or purity of a color, where varying its value yields distinct visual effects even within the same hue. Value (or brightness) represents the perceived luminance, with higher values indicating a greater proportion of white light in the color. The combined numerical representation of these three dimensions provides a unique discriminative signature for each color. We utilize these factors to describe creator-added text in pixel level. We have already show the difference of hue in fake and real video news with the Figure 1, and we will continue to demonstrate the discrepancy of saturation and value in Figure 14 and Figure 15. As shown in Figure 14, fake video news have lower saturation compared with fake news video when saturation lower than 0.35. And the relationship of fake and real video news has reversed when saturation larger than 0.35. However, the difference between real and fake video news is not clear.

A.1.2 Emotion of audience while watching the video

Fake news intend to change audience's emotion, as a result, the cognition may change along with the corresponding emotion. So we utilize the MLLMs to emulate audience watching video news, assessing audience's arousal and dominance level with scres. As demonstrated in Figure 16, real

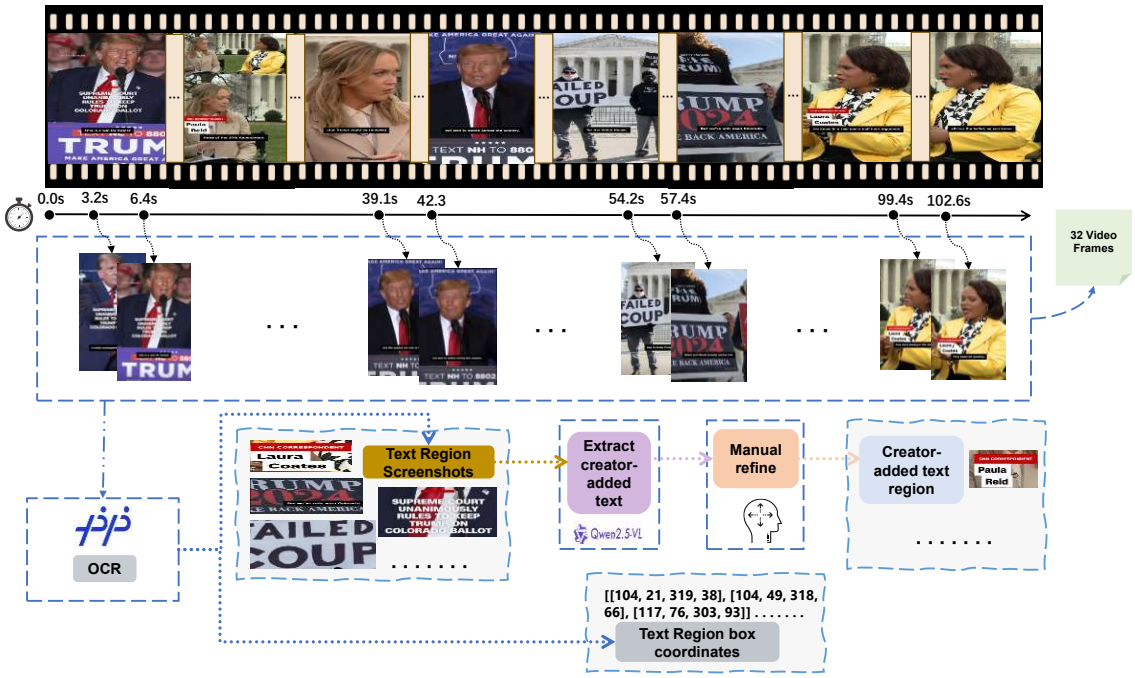


Figure 8: Creator-added Content Extraction Process

video news arouse high arousal and low dominance, which proves real news can utilize moderate emotional level to stimulate higher arousal, because real news have more valuable information that have clear emotional influence. However, fake news give audience higher dominance that reprensents human' s feeling over sense of control, People are more willing to make decision when dominance is high. So we discover fake news video creates an illusion of high control for the audience. so much so that the audience believes the content in the video is true.

A.2 Original shooting footage

Original shooting footage usually is the first hand description of event, especially sudden incident. Most of the footage is dynamic, so we utilize Gemini to process nascent videos. For key event perception, first we leverage MLLMs to process video with prompt template, after that experts manually optimize the outputs. For key event grounding, we employ MLLMs to locate key events with outputs of the perception, then experts adjust the time ranges of key events generated by MLLMs. For human identity perception, we use MLLMs to retrieve figures related to news event. Then experts manually revise the results.

A.2.1 Subject Identity Distribution

News videos frequently feature human subjects to convey key event information through specific visual formats, such as face-to-face interviews and press conference footage. These individuals include involved parties, victims, law enforcement personnel, self-media creators, and others. However, the reliability of the information provided by these identities varies significantly (Pereira et al., 2023). To investigate this, we analyzed the proportional distribution of different identities across the entire duration of the videos. Detailed results are presented in the Appendix. We find the perpetrator, victim and office personnel appear more frequently in real news video. However, larger proportion of fake news videos contain no people and the we-media appears frequently in the fake news. For the distribution of subject identity shown in Figure 17, we find the perpetrator, victim and office personnel appear more frequently in real news video. Especially law enforcement officers who appear in real news videos more than twice as frequently as in fake news videos. We attribute this significant disparity to real news producers can access law enforcement personnel involved in reported events. Moreover, the inclusion of such content is an essential procedure to alleviate audience skepticism. Conversely, fake news producers actively avoid this element. However, larger proportion of

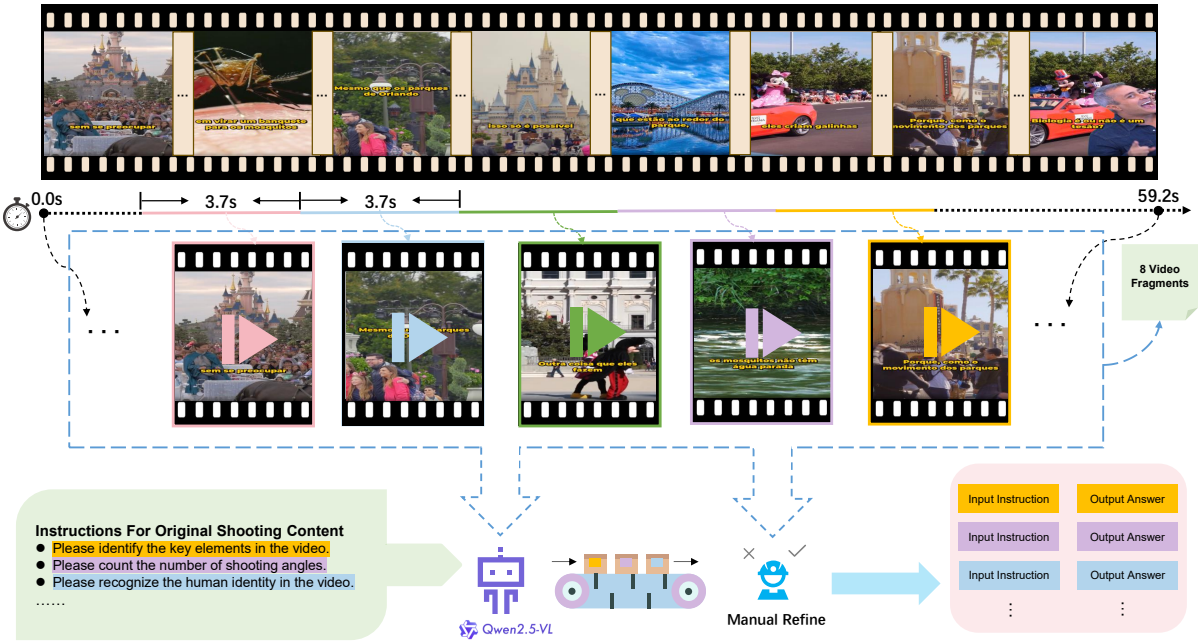


Figure 9: Original Shooting Dataset Extraction Process

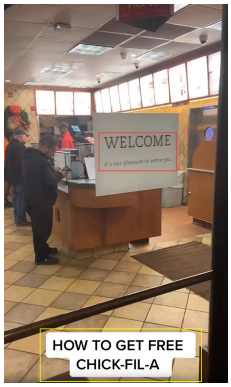


Figure 10: Text region in video screen

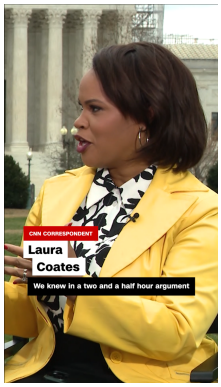


Figure 11: Individual with identity

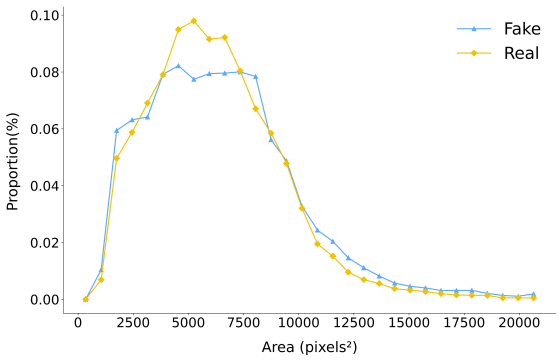


Figure 12: Distribution of area in fake/real news video

fake news videos contain no people and the we-media appears frequently in the fake news. The result align with our practical observation, because fake news generally has less relevant subjects to enforce the veracity of the video news.

A.2.2 Relevant shooting angles

Finding In terms of shooting angles, fake news tends to exhibit fewer camera shot transitions, whereas real news shows the opposite pattern.

Analysis To provide a comprehensive and objective depiction of news events, creators often employ multiple shooting angles within their videos. This approach not only enables audiences to gain a thorough understanding of the event's occurrence and progression but also enhances the credibility and persuasiveness of news reports (Bock et al., 2023). We attribute the discrepancy to the reasons, 1) fake news producers typically have no enough shooting materials about the events. 2) less shooting footage prevent audience from understanding the fact. 3) producers of fake news lack the professional skills required for camera shot transitions in the process of news video production.

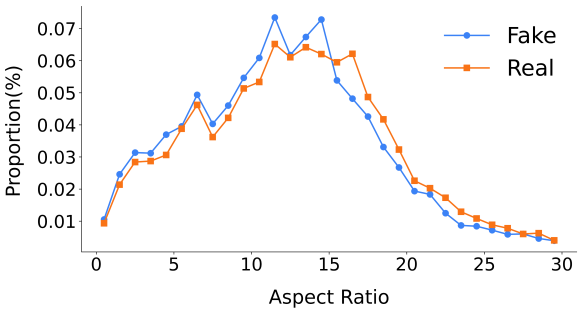


Figure 13: Distribution of aspect ratio in fake/real news video

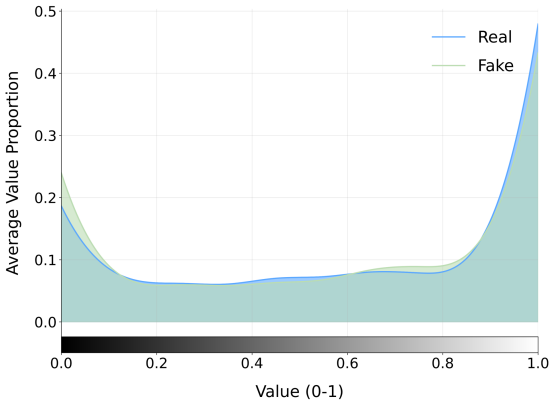


Figure 15: Value distribution in creator-added content of fake/real news video

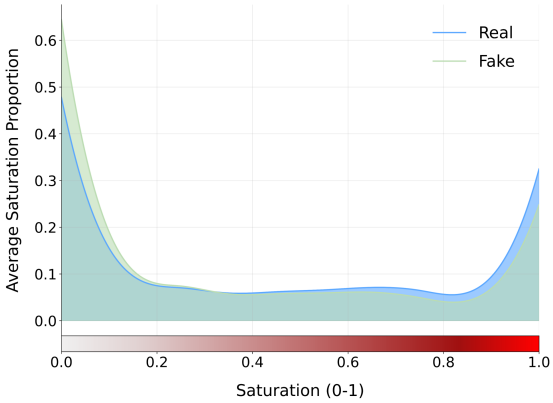


Figure 14: Saturation distribution in creator-added content
of fake/real news video

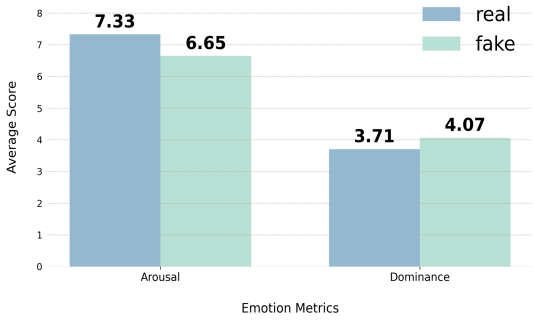


Figure 16: Arousal and Dominance scores of materials distribution

B Experiments

B.1 Data Curation

B.1.1 Details of Videos Category

event_hp: This category of news primarily reports on specific incidents and their associated individuals. Typical cases include fatalities caused by electric shock, major traffic accidents, and other similar events, all of which are attributed to human factors.

event_np: This category of news primarily focuses on reporting specific incidents without involving information about associated individuals. Typical examples include sudden natural disasters such as typhoons, earthquakes, and flash floods. These incidents have no clear direct correlation with human activities, and the core content of the reports is centered on the description and introduction of the incidents themselves.

focal_hp: This category of news primarily focuses on reporting events related to social hotspots, with the main content centered on human subjects. A typical example is the withdrawal of a star player from a match due to injury during the FIFA World Cup—such events are closely associated with the topics under discussion in social media public opinion at the corresponding time.

focal_np: This category of news primarily reports on events associated with social hotspots, and the content of such reports does not involve specific individuals or groups. A typical instance is the sell-out of goods in a certain supermarket prior to the arrival of a typhoon.

human: This category of news primarily focuses on profiling an individual, with its content centered on introducing a specific aspect of the individual rather than targeting a single event associated with them. A typical example is the summary of a worker's work experience. Generally, this type of news does not prioritize timeliness, as it is more inclined to present a summary of the individual's past experiences.

knowledge: This category of news primarily focuses on disseminating common knowledge, with its core content dedicated to the introduction

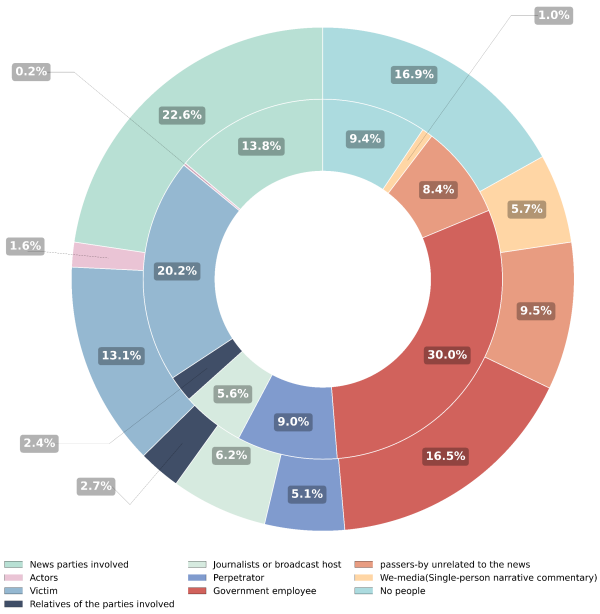


Figure 17: Distribution of subject identity in fake(external)/real(internal) news video.

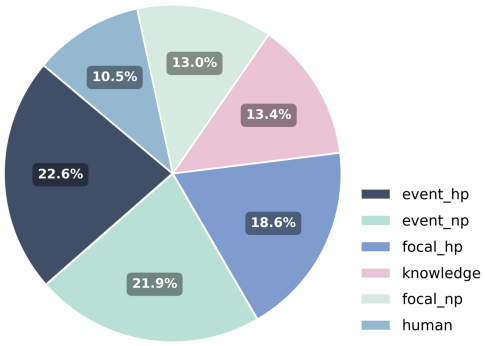


Figure 18: Distribution of category-test-split

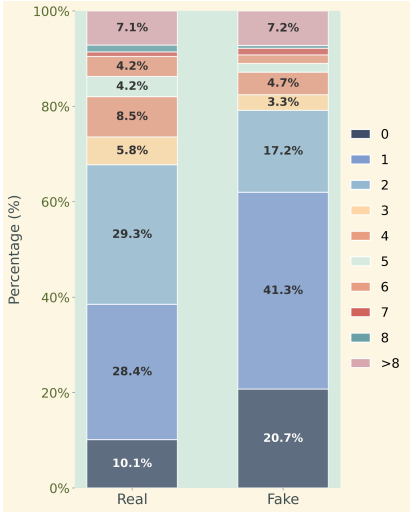


Figure 19: Number of shooting angles in the fake and real video news

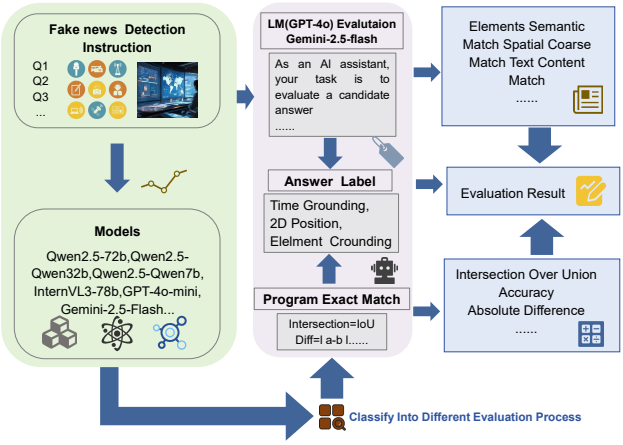


Figure 20: Evaluation process of MLLMs' outputs

and elaboration of such knowledge. Typical examples include the popularization of scientific knowledge, health maintenance guidance, and daily life tips. Generally, this type of news features a relatively high knowledge density, and its main body relies on verbal narration for content delivery rather than video demonstrations.

B.2 Task definition

In order to assess the MLLMs with more balanced test data, we construct a new test-split, we call it category-test-split, and the dataset shown in Figure 18. Apart from perception sub-tasks ,we also define a few understanding and reasoning sub-tasks, the definition as following:

B.2.1 Perception

1) Shooting angles counting, SAC. More angles can record news from multi aspects, especially in accident scene footage. MLLMs were required to recognize and output the number of shooting angles after browsing the news video.

2) Human identity recognition, HIR. Given a video, MLLMs were required to retrieve all identities that have appeared in the video. The task is multi-choice question with limited options. **3) Key Elements Temporal Grounding, KEG** The Key elements temporal grounding task requires MLLMs to answer the time range in which the target elements appear in the video. The description of target elements is provided in the input, and the task is annotated by the process in the Figure 4. we

use the Intersection over Union (IoU) metric, calculated as the ratio of the intersection to the union of predicted and labeled time intervals.

B.2.2 Understanding

1) News video type understanding, NTU. Video's category depends on the theme of video, the task is a single-choice question, MLLMs' are instructed to evaluate the theme of the news and output the type. All options can checked in Figure 18

2) News video Theme understanding, NEU. In news video, almost all information is added and managed to reflect theme of the news video. Understanding theme of the video guide the detection to pay attention to the elements relevant to the key contents. MLLMs' were guided to generate semantic description of video theme.

B.3 Metrics

To evaluate the capabilities of MLLMs across the 10 tasks, we define multiple metrics tailored to the task type and output format.

1) Avg EHR(element hit rate): The metric measures MLLMs' element hit rate in KEP task. Prediction is $\hat{y} = \{\hat{e_0}, \hat{e_1}...\hat{e_n}\}$, corresponding label is $y = \{e_0, e_1...e_m\}$. n is the number of elements in prediction, m is the number of elements in label. To ensure the highest possible accuracy of the results, each label element e_i is evaluated by the GPT-4. The result of matching $r_i = M(e_i, \hat{y}), r_i \in \{0, 1\}$, M is GPT-4. $EHR_j = \sum_{i=0}^{m} r_i/m, 0 \leq j < t$, t is number of video. $AvgEHR = \sum_{j=0}^{t} EHR_j/t$.

2) Avg ENHR(entity hit rate): The metric measures MLLMs' entity hit rate in FDR task. Reasoning output is $\hat{y}$, corresponding label is $y = \{e_0, e_1...e_m\}$, m is the number of label entities. To ensure the highest possible accuracy of the results, each label entity e_i is evaluated by the GPT-4. The result of matching $r_i = M(e_i, \hat{y}), r_i \in \{0, 1\}$, M is GPT-4. $ENHR_j = \sum_{i=0}^{m} r_i/m, 0 \leq j < t$, t is number of video. $AvgENHR = \sum_{j=0}^{t} ENHR_j/t$.

3) Avg KHR(knowledge hit rate): The metric measures MLLMs' Knowledge hit rate in FDR task. Reasoning output is $\hat{y}$, corresponding label is $y = \{e_0, e_1...e_m\}$, m is the number of label knowledge. To ensure the highest possible accuracy of the results, each label knowledge e_i is evaluated by the GPT-4. The result of matching $r_i = M(e_i, \hat{y}), r_i \in \{0, 1\}$, M is GPT-4. $KHR_j = \sum_{i=0}^{m} r_i/m, 0 \leq j < t$, t is number of video. Avg KHR $= \sum_{j=0}^{t} KHR_j/t$.

4) Avg RHR(rationale hit rate): The metric measures MLLMs' rationale hit rate in FDR task. Reasoning output is $\hat{y}$, corresponding label is $y = \{e_0, e_1...e_m\}$, m is the number of label rationales. To ensure the highest possible accuracy of the results, each label entity e_i is evaluated by the GPT-4. The result of matching $r_i = M(e_i, \hat{y}), r_i \in \{0, 1\}$, M is GPT-4. $RHR_j = \sum_{i=0}^{m} r_i/m, 0 \leq j < t$, t is number of video. Avg RHR $= \sum_{j=0}^{t} RHR_j/t$.

5) Avg AD(Absolute Distance): The metric measures MLLMs' performance in SAC task. $\hat{d}$ is the number of shooting angles predicted by MLLMs, d is ground truth. Absolute distance of each video is $AD_i = |\hat{d} - d|, 0 \leq i < n$,n is the number of video. $AvgAD = \sum_{i=0}^{t} AD_i/n$

6) Avg IoU(Intersection over Union): The metric measures MLLMs' temporal grounding capacity in KEG task. Prediction time range is $\hat{y} = \{\hat{t_0}, \hat{t_1}...\hat{t_n}\}$, corresponding label time range is $y = \{t_0, t_1...t_m\}$. n is the number of time range in prediction, m is the number of time range in label. $IoU_i(\hat{t_0}, t_0) = \frac{|\hat{t_0} \cap t_0|}{|\hat{t_0} \cup t_0|}, 0 \leq i < m$, m is the number of video, Avg IoU $= \sum_{j=0}^{m} IoU_j/m$.

For tasks involving creator-added content recognition, position detection, and color perception, we measure performance using accuracy based on evaluation results generated by GPT-4o. For tasks such as counting shooting angles and reasoning on three metrics (importance, emotional level, and relevance), we calculate the average absolute distance between the candidate and correct answers. To assess theme understanding, author-added content, and video shooting reasoning, we employ three metrics: Factual Consistency (FC), Theme Relevance (TR), and Completeness (CO). These metrics capture the degree of hallucination, alignment with the theme, and comprehensiveness of the output, respectively. For tasks involving human identity recognition (multiple-choice) and video type classification (single-choice), we evaluate performance using selection accuracy. To assess the temporal grounding of key elements, we use the Intersection over Union (IoU) metric, calculated as the ratio of the intersection to the union of predicted and labeled time intervals. All scores generated by the language model are normalized to a scale of 0 to 5.

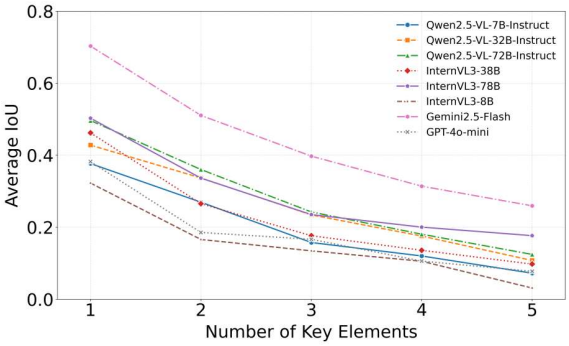


Figure 21: Comparison of IoU with Different Number of Key elements in the Video

For final detection, in addition to using accuracy to evaluate the binary output, six evaluation metrics are defined for the reasoning process: entity hit rate, rationale hit rate, knowledge hit rate, theme relevance, factual consistency, and ROUGE-L. Among these metrics, the first three fall into the category of exact match, while the latter three belong to semantic match. Specifically, entity hit rate and theme relevance reflect the degree of correlation between the reasoning process and the video content; rationale hit rate, knowledge hit rate, and factual consistency reflect the reasoning capabilities of MLLMs; and ROUGE-L measures the semantic quality of the reasoning output.

B.4 Main Results

1) The Gemini-2.5-Flash model exhibits superior performance across all evaluated tasks, primarily due to its advanced video processing capabilities. Unlike the GPT-4o-mini model, which relies on frame-based segmentation, Gemini-2.5-Flash processes video streams directly, preserving temporal coherence by avoiding the extraction of individual frames. This approach mitigates the loss of temporal context inherent in frame-based methods, which often disrupt the continuity of video data. In the creator-added content extraction processes (DCS, CCP, and CPP), Gemini-2.5-Flash outperforms the second-best model, Qwen2.5VL-72B, by a margin of up to 10%. Additionally, it surpasses Qwen2.5VL-72B by approximately 8% in Knowledge Extraction Performance (KEP), though its advantage in Human Interaction Recognition (HIR) is narrower, at roughly 1%.

2) The Qwen2.5VL-72B model achieves the highest performance among open-source MLLMs. Within the same architectural framework, Qwen2.5VL-72B significantly outperforms its smaller counterparts, Qwen-32B and Qwen-7B, primarily due to its larger model scale. Surprisingly, it also consistently surpasses InternVL-78B across all evaluated tasks, despite the latter's greater parameter count. This superior performance is attributed to Qwen2.5VL-72B's native video processing capabilities and its dynamic FPS sampling mechanism, which enables the model to effectively process and comprehend video data at varying sampling rates.

3) All models demonstrated suboptimal performance in the Content Creation Process (CCP), with the highest achieving only 47.47% (on a 0%–100% scale). This limited performance can be attributed to the fact that most multimodal large language models (MLLMs) are derived from language models, which primarily focus on semantic relationships and object descriptions during pretraining. In contrast, the color of fonts in visual elements is particularly critical for Fine-Grained Narrative Detection (FND). Other general video tasks, such as captioning, prioritize semantic alignment over such visual details.

4) Temporal grounding across multiple time ranges in news videos presents greater challenges than traditional grounding tasks. This difficulty arises because news videos typically contain fewer key elements and are shorter in duration compared to videos used for visual question answering (VQA). We define temporal grounding as the identification of multiple time ranges within a video. However, only the Gemini-2.5-Flash model achieved a performance score exceeding 0.3, the standard threshold commonly used as a metric for video grounding tasks.

5) Insights contribute to the Fake News Detection. As demonstrated in Table 2 and 4, accuracy of fake news detection and other sub-tasks metrics have the same trend. The Gemini-2.5-Flash model demonstrates superior capabilities in processing creator-added content (CAC) and original shooting footage (OSF), achieving the highest detection accuracy and Macro-F1 score. As shown in Table 4, the CAC-MVFND-CoT framework integrates MVFND-CoT with prompts specific to creator-added content, including color, position, and text content. Similarly, the OSF-MVFND-CoT framework combines MVFND-CoT with prompts related to original shooting footage, such as key elements, shooting angles, and human identities. The ALL-MVFND-CoT framework encompasses CAC,

Table 4: Evaluation Result of CoT on Fake News Detection

Model	CAC-MVFND-CoT		OSF-MVFND-CoT		ALL-MVFND-CoT		MVFND-CoT		Instruct-Tuning	
	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1	ACC	M-F1
Gemini-2.5-flash	78.88	78.81	79.02	78.87	78.75	79.66	78.61	77.87	-	-
GPT-4o-mini	72.24	78.81	72.00	78.87	71.31	74.66	71.05	72.35	-	-
Qwen2.5vl-72b	77.95	74.08	78.35	74.19	77.56	78.12	77.00	74.23	-	-
Qwen2.5vl-32b	74.90	69.16	74.76	74.68	75.43	70.69	74.90	76.05	-	-
Qwen2.5vl-7b	69.85	65.66	70.38	71.77	70.25	69.75	70.12	71.51	-	-
InternVL3-78b	76.49	77.04	76.89	75.58	76.62	74.15	75.96	72.45	-	-
InternVL3-38b	73.43	70.64	73.97	70.66	73.71	71.56	73.43	71.65	-	-
InternVL3-8b	70.51	69.14	70.39	68.51	70.65	69.16	70.00	71.55	-	-
Qwen2.5-7b-Instruct	57.23	53.71	57.50	55.39	57.76	55.36	58.43	58.23	81.14	80.26

OSF, and CoT components. The MVFND-CoT, designed based on insights proposed in this study, yields promising results compared to zero-shot settings across all models. Furthermore, performance improves significantly when CoT is combined with either CAC or OSF in the Gemini-2.5-Flash and Qwen2.5VL-72B models. However, as model scale decreases, these performance gains diminish. We attribute this decline to the reduced capacity of smaller models to effectively process prompt information, such as extracting relevant details from prompts. When critical information (CAC or OSF) cannot be accurately extracted and utilized, it becomes redundant or erroneous data, adversely impacting the final judgment.

C Answers For Submission

C.1 Potential Risks

1) Ignoring Adversarial Evolution of Fake Videos. A benchmark's value lies in its ability to reflect real-world deployment challenges. We may reduce our work's relevance if we overlook these risks, fake news creators continuously adapt to detection tools—for example, using adversarial deepfakes (designed to evade LLMs), mixing fake and real footage, or leveraging AI to generate contextually consistent manipulations. If our benchmark uses static, outdated fake samples :1. Our results will quickly become obsolete as new manipulation techniques emerge.2. The benchmark will not test models' ability to generalize to "future" fake video types, limiting its long-term utility for industry or policy.

2) Lack of Cross-Domain Validation. Fake news videos appear in diverse domains (e.g., politics, healthcare, entertainment), each with unique characteristics (e.g., medical deepfakes use specialized terminology; political fakes rely on event context). If our benchmark focuses on a single domain: 1. Models optimized for our benchmark may fail in critical other domains (e.g., a model good at detecting political fakes may miss fake medical advice videos). 2. We cannot support cross-domain comparisons, a key requirement for organizations (e.g., health agencies, social media platforms) that need tools for multiple use cases.

2) Vague or Narrow Definition of "Fake News Video". Fake news videos encompass a spectrum of manipulations (e.g., deepfakes, selective editing, context misattribution) and intents (malicious disinformation vs. accidental misinformation). If we use an imprecise or overly narrow definition (e.g., only labeling face swaps as "fake"), we create inconsistency: Models optimized for specific fake types (e.g., deepfake detection) will be unfairly penalized for failing to detect other valid fake cases (e.g., a real video paired with misleading audio).

C.2 The License For Artifacts

In our research on multimodal large model-based fake news video detection, we recognize that clarifying the license and usage terms of core artifacts is critical to ensuring academic compliance. For the Fake News Short Video Dataset (FakeSV-2023) — a key resource supporting our benchmark construction — we have engaged in in-depth discussions with the dataset authors and reached a formal agreement on its use and distribution terms. This agreement not only specifies that the dataset is strictly limited to non-commercial research purposes, prohibiting any commercial exploitation such as resale or use in commercial product development, but also mandates mandatory citation of its original AAAI 2023 publication (in the specified BibTeX format) in all our related outputs. Additionally, it outlines clear responsibilities: we and our affiliated institution assume full liability for any consequences of

dataset use and must indemnify the authors against claims arising therefrom, while the authors retain the right to terminate our access if necessary. We have formally consented to all these provisions, and the agreement details have been systematically documented in our research records to ensure every step of our work aligns with the dataset's usage specifications.

C.3 Intended Use

Regarding the use of the Fake News Short Video Dataset (FakeSV-2023) in our research, we confirm that our usage fully aligns with the intended use specified in the license agreement reached with the dataset authors. As explicitly stipulated in the agreement, the Dataset is restricted to non-commercial research purposes only—a scope that precisely matches our intended application of the Dataset in constructing and validating the multi-modal large model-based fake news video detection benchmark. Throughout the entire research process, we have strictly adhered to this boundary: the Dataset and any derived data generated from it (e.g., preprocessed video clips, annotated sub-datasets for model training) have been exclusively used for academic research activities, including algorithm development, performance testing, and result analysis related to fake news video detection. No part of the Dataset or its derivatives has been employed in any scenarios outside of research, such as commercial product development, commercial service provision, or any other non-academic uses that are prohibited by the agreement. This compliance ensures that our use of the Dataset remains consistent with both the authors' intended design for the Dataset and the original access conditions outlined in the license terms.

C.4 Model Size And Budget

Regarding the reporting of model parameters, total computational budget, and computing infrastructure used in our research, we provide the following details: For the state-of-the-art multimodal large language models (LLMs) employed, we distinguish between closed-source and open-source models due to differences in parameter disclosure and deployment methods: 1. Closed-source models: We accessed these via cloud service APIs. Specifically, we made a total of 137,927 API calls to the Gemini 2.5-Flash model, and a cumulative 315,061 API calls to GPT series models (including GPT-4o-, GPT-4o-mini, and GPT-o1). Note that the exact number of parameters for these closed-source models is not publicly disclosed by their developers, so we do not report proprietary parameter details. 2. Open-source models: These were deployed on our in-house computing infrastructure, which consists of 8 H800 GPUs. The total computational budget for training and inference with these open-source multimodal LLMs amounted to approximately 1,500 GPU hours.

C.5 Recruitment And Payment

We hired a total of 35 data annotators on a part-time basis, with hourly remuneration ranging from HK60$toHK$100. This payment rate is higher than the local statutory minimum wage. Details of their specific circumstances are provided in Section 4.2.

C.6 Use Of Ai Assistants

Assisting with program bug fixes: During the development and debugging of code for model deployment (e.g., integrating open-source multi-modal LLMs with our experimental framework) and data preprocessing (e.g., optimizing video frame extraction scripts), we used AI assistants to identify syntax errors, logical inconsistencies, and performance bottlenecks in the code. Their role was limited to providing diagnostic suggestions and potential solution references, with final bug fixes and code validation completed by our research team to ensure accuracy and alignment with our experimental objectives.

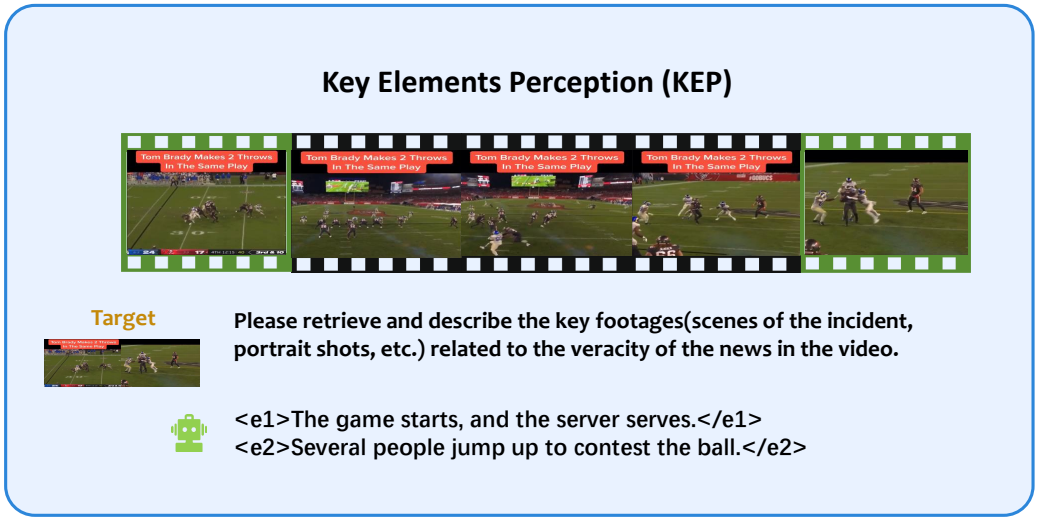


Figure 22: KEP case

Figure 23: DCS case

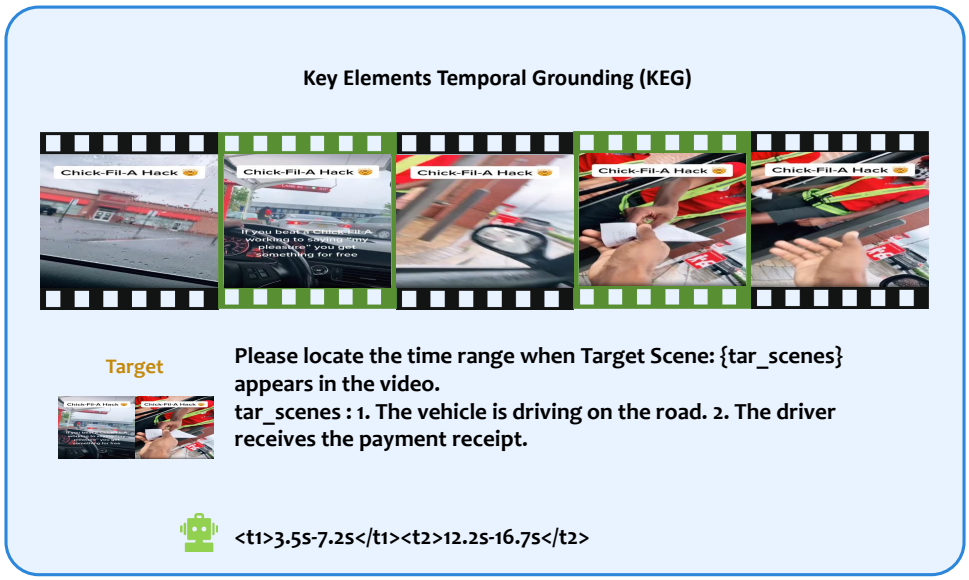


Figure 24: KEG case

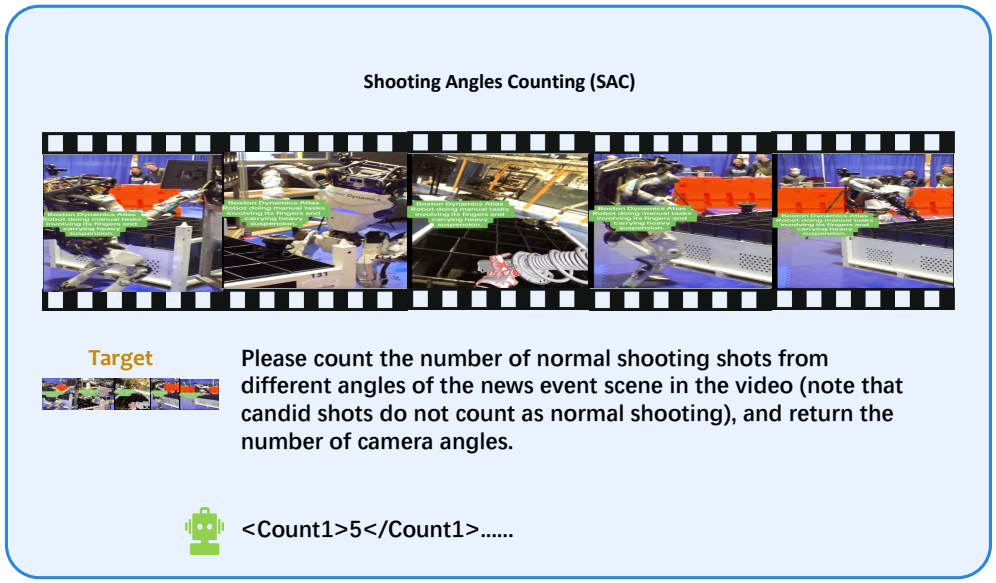


Figure 25: SAC case

News Video Type Understanding (NTU)

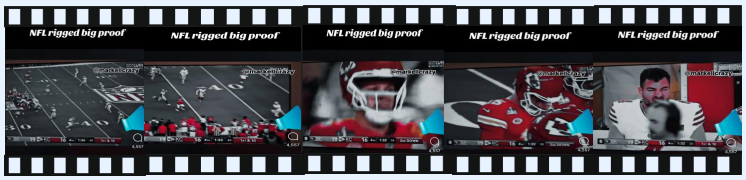
Target

Please analyze the content of the video and determine the type of news it tells. Choose from the following options ():
A. Breaking news events unrelated to people;
B. Breaking news events related to people;
……

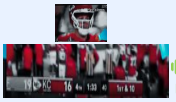
<Type> F. News strongly related to basic knowledge;</Type>

Figure 26: NTU case

News Video Theme Understanding (NEU)



Target



Please describe the theme of the news and analyze which elements are important for supporting the veracity of the news.
/aifs4su/hansirui_2nd/cuiyakun/datasets/fakett_detail/video/7335108541226798378.mp4

<Response>
The theme of the news video is allegations of rigging in the NFL (National Football League), specifically referencing a game or event where there are claims of unfairness or manipulation. The video appears to focus on a particular moment during a football game, highlighting specific plays and reactions that are presented as evidence of rigging.
1. Title and Text Overlay……2. Game Footage……3. Player Reactions……4. Scoreboard and Game Clock……5. Social Media Attribution……
Conclusion :
</ Response>While the video presents a compelling in the video.

Figure 27: NEU case