MEAFS: Event-level Aggregation with MLLMs for Fake News Detection in Short Videos

Haisong Gong[1,2,*], Zhibo Liu[1,3,*], Qiang Liu[1,2], Shu Wu[1,2,†], Liang Wang[1,2]

[1]*New Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences*, Beijing, China
[2]*School of Artificial Intelligence, University of Chinese Academy of Sciences*, Beijing, China
[3]*School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences*, Beijing, China

Email: {gonghaisong2021, liuzhibo2025}@ia.ac.cn; {qiang.liu, shu.wu, wangliang}@nlpr.ia.ac.cn

***Abstract*—Short video platforms have become a popular medium for information sharing, yet they also facilitate the rapid spread of misinformation. Detecting fake news in short videos is challenging, as malicious creators can closely mimic legitimate content, making single-video analysis unreliable. Existing methods mainly rely on intra-video signals, limiting robustness against sophisticated forgeries. To address this, we propose MEAFS (MLLM-powered Event-level Aggregated Fake news detection in Short videos), a framework that incorporates event-level context. MEAFS employs multimodal large language models (MLLMs) to generate objective summaries of videos from the same event, which are aggregated into an event-level representation. A discrepancy feature is then derived by comparing the target video with the aggregated context, highlighting potential inconsistencies. Combined with multimodal features, these representations enable more reliable classification of real versus fake content. Experiments on public datasets show that MEAFS substantially improves detection accuracy over baseline approaches.**

***Index Terms*—Fake news detection, multimodal large language model, short video classification**

I. Introduction

Short video platforms have become a prominent medium for sharing life experiences, creative ideas, and skills [1]. However, the widespread dissemination of videos containing misinformation has raised significant societal concerns [2]. Often deliberately crafted with sophisticated editing to manipulate public opinion or provoke emotional reactions, these videos undermine social stability and disrupt online communities. Consequently, detecting misinformation within the vast volume of user-uploaded short videos has emerged as a critical and urgent challenge for platform governance.

Fake news detection fundamentally aims to identify and classify content that propagates false or misleading information [3]. In this work, we specifically focus on multimodal fake news detection within short video platforms. The primary objective is to determine the veracity of a specific video by analyzing its audiovisual content and textual metadata, distinguishing deceptive fabrications from authentic reportage.

In recent years, Multimodal Large Language Models (MLLMs), such as LLaVA [4] and Qwen2.5-VL [5], have demonstrated exceptional capabilities in handling complex reasoning tasks. Building on these advancements, researchers have begun to investigate the potential of leveraging MLLMs for fake news detection. Existing approaches typically involve either fine-tuning MLLMs directly for binary classification [6] or utilizing MLLM-generated descriptions to enrich features. These methods attempt to exploit the reasoning prowess of MLLMs to capture subtle deceptive cues, such as implicit malicious intentions or cross-modal inconsistencies between the visual track and the audio narrative [7], [8].

Despite the progress achieved by these methods, they share a critical limitation: they rely predominantly on *intra-video* signals—analyzing the target video in isolation. This paradigm becomes increasingly insufficient when confronting sophisticated malicious creators. These actors can closely mimic the stylistic patterns and production quality of legitimate content, rendering their videos indistinguishable from real news when viewed individually. In real-world scenarios, human fact-checkers rarely evaluate the veracity of a claim in isolation; instead, they cross-reference the target content with other reports covering the same event. By identifying discrepancies across a collection of related videos, auditors can construct a comprehensive view of the event, allowing for a more robust identification of misinformation.

Motivated by this human cognitive process, we propose MEAFS (MLLM-powered Event-level Aggregated Fake news detection in Short videos), a novel framework that leverages MLLMs to incorporate event-level context into detection. Instead of treating each video as an isolated instance, our approach contextualizes the target video within the broader narrative of its corresponding event. Specifically, given a target video, we retrieve other videos of the same event. Using customized prompts, an MLLM extracts objective descriptions from each video. These descriptions are embedded and fused via an attention mechanism to form an aggregated event feature. By contrasting the target video's content with this event-level representation, we derive a discrepancy feature that highlights inconsistencies between the target video and the broader event-level context. Finally, we synergize the event feature, the discrepancy feature, and the target video's intrinsic multimodal features to make the final prediction.

*Equal contribution.
†Corresponding author.

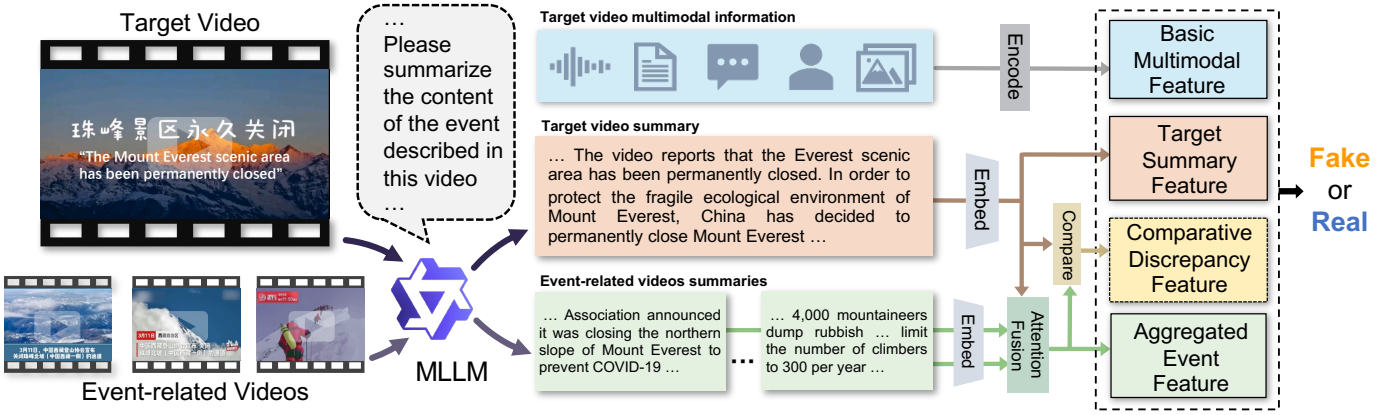

Fig. 1. Overview of MEAFS. Given a target video and a set of related videos associated with the same event, an MLLM is employed to generate objective textual summaries. These summaries are utilized to compute (i) a target summary feature, (ii) an aggregated event-level feature, and (iii) a comparative discrepancy feature. By combining these representations with basic multimodal features, the model predicts whether the target video is fake or real.

Extensive experiments on public datasets demonstrate that by effectively modeling event-level context, MEAFS significantly outperforms baselines in detection accuracy.

II. Related Works

A. Multimodal Fake News Detection

Early research on fake news detection primarily concentrated on unimodal settings, determining whether a single piece of information (mostly text posts) conveyed misinformation [9], [10]. These methods generally treated the problem as a binary text classification task, focusing on semantic content, emotional sentiment [11], or writing styles [12]. Some researchers extended this by incorporating user behavior, modeling social interactions such as comments and retweets to capture propagation patterns [13]. With the proliferation of multimedia content, attention has shifted toward multimodal fake news detection, particularly within the short video domain, where videos contain rich, intertwined auditory and visual signals. SVFEND [14] serves as a foundational benchmark, utilizing pre-trained encoders to extract unimodal features and employing fusion mechanisms for prediction. To improve generalization, FANVM [15] employs adversarial learning and topic modeling to extract topic-agnostic representations. Focusing on manipulation traces, FakingRecipe [16] designs targeted modules to detect modality inconsistencies and editing artifacts. More recently, MMVD [17] addresses cognitive biases via a multi-view causal reasoning strategy, while MMSFD [18] introduces a multi-grained fusion strategy to enhance cross-modal alignment.

B. Multimodal Large Language Models (MLLMs)

The advent of MLLMs, represented by LLaVA and Qwen2.5-VL, marks a paradigm shift in understanding inputs beyond text. These models demonstrate exceptional capabilities in visual perception and complex reasoning. Consequently, a growing line of work has begun to leverage MLLMs for short-video fake news detection. One category of approaches utilizes MLLMs to detect internal inconsistencies. For instance, CA-FVD [7] uses MLLMs to assess cross-modal consistency. Another category employs MLLMs for semantic enrichment and interpretability. OpEvFake [8] prompts MLLMs to mine explicit and implicit opinions from videos, and VMID [19] integrates model-generated explanations as high-level semantic features. Furthermore, ExMRD [20] employs MLLMs to analyze video content and perform knowledge retrieval. More recently, TGFC-SVFN [21] designs causality-driven prompting strategies to guide reasoning, whereas 3MFact [22] adopts a stepwise approach, generating summaries and verifiable questions for fact-checking.

While these studies showcase the diverse potential from feature enrichment to reasoning guidance, they remain limited to extracting signals from *individual* videos. In contrast, our proposed MEAFS explicitly incorporates *event-level, cross-video context*. By leveraging MLLMs to aggregate information from multiple videos covering the same event, we capture critical discrepancies that are essentially undetectable through single-video analysis.

III. Methodology

A. Task Formulation

Let $\mathcal{D} = \{x_i\}_{i=1}^{N}$ denote a dataset of short videos. Each instance is represented as a tuple $x_i = \{x_i^v, x_i^t, x_i^a, x_i^e\}$, where x_i^v, x_i^t, and x_i^a correspond to the visual, textual (e.g., caption, OCR), and acoustic modalities, respectively. Additionally, x_i^e represents the event identifier, grouping videos that report on the same specific real-world event. The goal of fake news detection is to learn a mapping function $\mathcal{M} : x_i \rightarrow \hat{y}_i \in \{0, 1\}$, where 1 indicates the video contains misinformation and 0 indicates legitimate content.

To facilitate a clear exposition of MEAFS, we designate the specific video currently under classification as the **target video**, denoted as x_0. All other videos sharing the same event tag x_0^e are referred to as auxiliary event videos.

B. Video Content Comprehension

First, we employ an MLLM to derive objective descriptions from individual videos. Specifically, we prompt the MLLM to act as an experienced short-video evaluator, tasked with summarizing the video content in a factual and unbiased manner. The generated summary is required to cover key aspects of the event described in the video. We design a prompt template T for this purpose:

> **System Instruction:** You are an expert fact-checker for short videos. Your task is to extract objective factual information.
> **Input:**
> - *Video Frames*: {{*Sampled Frames*}}
> - *Metadata*:
> Title: {{*Video Title*}};
> Description: {{*Video Description*}}
>
> **Task:** Summarize the event described. Include key elements: time, location, entities, cause, and progression.
> **Constraint:** Strictly avoid subjective opinions, emotional language, or personal bias. Output a concise paragraph.

Formally, for any given video x_j (including the target x_0), we generate a semantic summary c_j:

$$c_j = \text{MLLM}\left(T(x_j^v, x_j^t)\right). \tag{1}$$

Next, each textual video description c_j is embedded into a dense vector representation $\mathbf{h}_j$ through a pretrained embedding model $\text{Embed}(\cdot)$, this embedding serves as the basis of the following event-level aggregation and discrepancy feature:

$$\mathbf{h}_j = \text{Embed}(c_j) \in \mathbb{R}^d. \tag{2}$$

C. Event Context Modeling

1) Event Feature Aggregation: To capture the "consensus view" of an event, we aggregate information from the set of videos associated with the same event tag as the target x_0, including x_0 itself. Let $\mathcal{H}_0$ denote the set of embeddings for this event cluster:

$$\mathcal{H}_0 = \{\mathbf{h}_j \mid x_j^e = x_0^e,\ \forall j\}, \tag{3}$$

where each $\mathbf{h}_j$ is obtained from x_j through Equation 1 and 2.

We employ an attention-based fusion mechanism to synthesize these individual perspectives into a unified event representation. This mechanism allows the model to adaptively weigh the importance of different videos—filtering out noise from low-quality uploads while emphasizing consistent factual patterns. The aggregated event feature $\mathbf{f}_0^e$ is computed as:

$$\mathbf{f}_0^e = \texttt{Pool}\left(\texttt{SelfAttention}(\mathcal{H}_0)\right) \in \mathbb{R}^d. \tag{4}$$

Here, we utilize multi-head self-attention [23] followed by mean pooling (`Pool`). This process ensures that $\mathbf{f}_0^e$ encapsulates the comprehensive, globally calibrated context of the event, serving as a reliable reference for verification.

2) Comparative Discrepancy Feature: While aggregated event features provide an overall view of the event, it is also crucial to characterize how the target video aligns or deviates from this context. To capture such signals, we design a comparative discrepancy feature between the target summary feature $\mathbf{h}_0$ and aggregated event feature $\mathbf{f}_0^e$.

Inspired by natural language inference [24], we employ a comparison strategy that captures both similarity and difference between two representations. Specifically, we concatenate the raw embeddings, their element-wise product, and their absolute difference as input to a feed-forward network:

$$\begin{aligned}\mathbf{d}_0^e &= \texttt{Compare}(\mathbf{h}_0, \mathbf{f}_0^e) \\ &= \texttt{FFN}\left([\mathbf{h}_0; \mathbf{f}_0^e; \mathbf{h}_0 \odot \mathbf{f}_0^e; |\mathbf{h}_0 - \mathbf{f}_0^e|]\right) \in \mathbb{R}^d,\end{aligned} \tag{5}$$

where $[\cdot]$ denotes concatenation, $\odot$ represents element-wise multiplication, and $|\cdot|$ signifies absolute difference. The Feed-Forward Network (`FFN`) then projects these interaction signals into a compact discrepancy embedding.

D. Feature Fusion and Prediction

The final classification relies on a holistic analysis that integrates high-level semantic reasoning with low-level perceptual signals. The proposed framework extracts three distinct semantic features: the target summary feature $\mathbf{h}_0$ (local semantic), the aggregated event feature $\mathbf{f}_0^e$ (global context), and the discrepancy feature $\mathbf{d}_0^e$ (comparative reasoning).

However, semantic summaries alone may miss non-textual cues such as visual tampering artifacts or specific audio patterns. To address this, we incorporate a baseline multimodal feature vector $\mathbf{b}_0$, extracted using standard pre-trained encoders (e.g., VGGish [25] for audio, Bert [26] for text and similar encoders for other signals) following the protocol proposed by Qi et al. [14].

We fuse these heterogeneous features using a co-attention mechanism to model their complementary relationships. The fused representation is then fed into a classifier to produce the probability of the target video being fake:

$$\hat{y} = \sigma\left(\texttt{MLP}\left(\texttt{Attention}(\{\mathbf{f}_0^e, \mathbf{d}_0^e, \mathbf{h}_0, \mathbf{b}_0\})\right)\right), \tag{6}$$

where σ is the sigmoid function and `MLP` is a multi-layer perceptron. The model is trained end-to-end using the binary cross-entropy loss function, minimizing the divergence between the prediction $\hat{y}$ and the ground-truth label y.

IV. Experiments

A. Dataset

To comprehensively evaluate the effectiveness of MEAFS, we conduct experiments on the FakeSV dataset [14], a widely recognized benchmark in short-video fake news detection. FakeSV comprises 3,624 videos associated with 614 distinct public events. Following established protocols [8], [18], we adopt two rigorous data split strategies to assess different capabilities of the model:

- ***Event-level split***, where five-fold cross-validation is performed with a 4:1 train–test ratio in each fold, ensuring

TABLE I
PERFORMANCE COMPARISON OF BASELINE MODELS AND MEAFS ON FAKESV DATASET.

Model	*Event-Level Split*				*Temporal Split*			
	Acc.	**F1**	**Prec.**	**Rec.**	**Acc.**	**F1**	**Prec.**	**Rec.**
Bert	77.30	77.27	77.43	77.30	80.81	80.32	80.80	80.07
FANVM	75.49	75.46	75.71	75.53	82.66	82.21	82.72	81.94
SVFEND	79.31	79.24	79.62	79.31	81.05	81.02	81.24	81.05
FakingRecipe	79.60	79.59	79.67	79.60	84.69	84.30	84.80	84.01
MMVD	82.64	82.63	82.63	82.73	-	-	-	-
MMSFD	81.83	81.81	82.02	81.84	-	-	-	-
CA-FVD	-	-	-	-	85.79	85.28	86.57	84.78
OpEvFake	-	-	-	-	88.01	87.80	87.90	87.71
SVFEND-NEED-D	82.76	82.75	82.78	82.75	88.01	87.66	88.47	87.26
ExMRD	80.48	80.46	80.67	80.51	86.90	86.52	87.31	86.13
GPT-4o-m	67.10	67.08	67.15	67.21	66.42	65.88	65.90	65.87
LLaVA-OV	58.14	54.76	60.44	57.51	57.54	50.71	61.57	55.94
Qwen2.5-VL	67.78	78.38	58.19	66.79	70.16	79.46	54.02	64.31
MEAFS (Ours)	**84.11**	**84.07**	**84.30**	**84.10**	**89.83**	**89.53**	**90.47**	**89.07**

TABLE II
PERFORMANCE ON TEMPORAL SPLIT OF FAKETT DATASET.

Model	Acc.	F1	Prec.	Rec.
SVFEND	78.26	75.99	75.54	76.61
CA-FVD	81.61	80.26	79.50	82.17
ExMRD	84.28	83.13	82.27	**85.19**
Qwen2.5-VL	71.41	76.68	73.87	75.25
MEAFS	**84.94**	**83.37**	**82.78**	84.15

that videos from the same event do not appear across folds;

- ***Temporal split***, where videos are ordered chronologically, with the earliest 70% used for training, the middle 15% for validation, and the most recent 15% for testing, simulating real-world deployment.

Furthermore, to verify the transferability and generalization of our framework, we conduct supplementary experiments on the FakeTT dataset [16]. FakeTT shares a similar data collection methodology with FakeSV and is evaluated under its standard temporal split. Performance is measured using four standard metrics: Accuracy, Macro-F1 score, Precision, and Recall.

B. Baseline Models

We benchmark MEAFS against a comprehensive set of existing approaches for short-video fake news detection. *Feature-based Baselines*: We adopt Bert [26] as a unimodal text baseline. For multimodal baselines, we select SVFEND [14], FANVM [15], FakingRecipe [16], MMSFD [18], and MMVD [17]. These models are specifically tailored to capture multimodal artifacts within short videos. *Event-Augmented Baseline*: We compare our method with SVFEND-NEED-D, a variant of SVFEND-NEED [27] which utilizes graph attention networks to model video interactions within an event. We append the suffix "-D" to clarify that we exclude the debunking videos originally used in their method, ensuring a fair comparison where no external debunking information is available to any model. *MLLM-Enhanced Baselines*: We include ExMRD [20], which employs 3-step Chain-of-Thought (CoT) prompting and knowledge retrieval; CA-FVD [7], which uses MLLMs to assess cross-modal consistency; and OpEvFake [8], which mines explicit and implicit opinions. Finally, we evaluate *MLLM Zero-Shot Baselines* using GPT-4o-mini [28], LLaVA-OV [29], and Qwen2.5-VL [5]. These models perform inference directly on the raw video and metadata, simply prompted to determine the veracity of the target video without training.

C. Implementation Details

Qwen2.5-VL-7B is adopted as the MLLM backbone in the video content comprehension stage. The embedding model is Qwen3-Embedding-0.6B, followed by a trainable linear projector that maps embeddings to a fixed dimension of $d = 128$. To balance performance and efficiency, the size of $\mathcal{H}_0$ is restricted to at most 8. All attention modules are implemented with 2 layers and 4 heads. The model is optimized using Adam (learning rate 1e-3, batch size of 32, dropout rate 0.3, maximum 50 epochs) with early stopping.

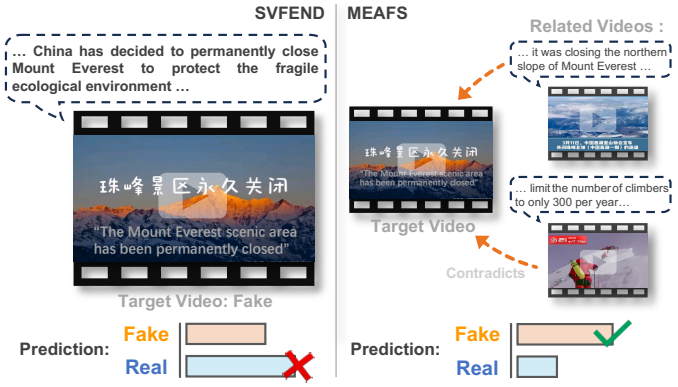

Fig. 2. An example of event-level context aiding fake news detection.

D. Results and Analysis

1) Overall Performance: Table I presents the comparative results on the FakeSV dataset. MEAFS achieves consistent superiority, outperforming all baselines. Specifically, our method surpasses the strongest specialized baseline (SVFEND-NEED-D) by 1.35% and 1.82% in accuracy on the two splits, respectively. It is worth noting that while SVFEND-NEED-D also incorporates event-level structural information via graphs, MEAFS achieves better results. The key distinction lies in our utilization of MLLM-derived semantic summaries, which offer significantly richer interpretability and semantic density compared to simple node embeddings. Compared to sophisticated MLLM-enhanced frameworks like ExMRD and OpEvFake, our model maintains a clear lead. This confirms that merely using MLLMs for reasoning (CoT) or retrieval is insufficient; explicitly modeling the *event-level context* and *cross-video discrepancies* is crucial for identifying sophisticated misinformation.

To further examine generalization, we report supplementary results on the FakeTT dataset in Table II. Compared with representative baselines, MEAFS continues to demonstrate clear advantages.

2) Case Study: To intuitively demonstrate the efficacy of our approach, Fig. 2 visualizes a representative case. The target video falsely claims that "Mount Everest is completely closed." Viewed in isolation, the video appears visually authentic, misleading single-video baselines. However, MEAFS retrieves related videos from the same event timeline, which clarify that "only the northern slope is closed" and "climber numbers are restricted." By aggregating this context, our model identifies the logical conflict between the absolute claim ("completely closed") and the nuanced reality, correctly classifying it as fake.

3) Ablation Study: To dissect the contribution of each module, we perform ablation studies on the FakeSV dataset (Table III).

Impact of Event Context: The removal of context-aware features causes a sharp performance degradation. Excluding the *Aggregated Event Feature* ("w/o Aggr.") leads to a drop of 1.1% and 2.4% in accuracy on the two splits. More critically, removing the *Comparative Discrepancy Feature* ("w/o Comp.") results in even larger losses. When both are removed ("w/o Context"), the model effectively regresses to a single-video baseline, with performance plummeting by roughly 3.8% (Event-level) and 5.5% (Temporal). This strongly validates our core hypothesis: detecting fake news requires looking beyond the video itself to the broader event consensus.

TABLE III
ABLATION STUDY ON DIFFERENT COMPONENTS OF MEAFS ON FAKESV.

Variant	Acc.	F1	Prec.	Rec.
Event-Level Split				
MEAFS (Full)	**84.11**	**84.07**	**84.30**	**84.10**
w/o Summary	83.91	83.89	83.99	83.89
w/o Aggr.	83.01	82.99	83.15	83.00
w/o Comp.	81.98	81.96	82.07	81.97
w/o Basic.	80.88	80.85	81.02	80.88
w/o Context	80.29	80.24	80.56	80.28
Temporal Split				
MEAFS (Full)	**89.83**	**89.53**	**90.47**	**89.07**
w/o Summary	89.27	89.06	89.30	88.89
w/o Aggr.	87.43	87.01	88.16	86.52
w/o Comp.	86.87	86.70	86.65	86.75
w/o Basic.	86.13	85.99	85.88	86.18
w/o Context	84.28	83.66	85.25	83.13

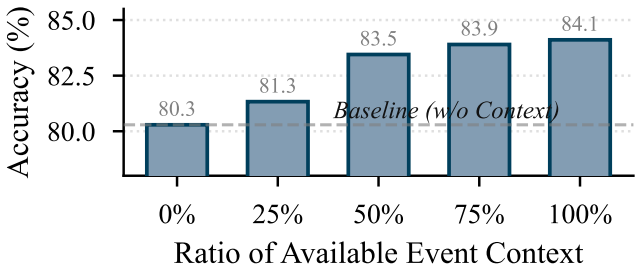

Fig. 3. Early detection performance.

Impact of Multimodal Features: We observe that removing the *Target Summary* ("w/o Summary") has a relatively minor impact, suggesting that the event-level context can partially compensate for the target's semantic representation. However, removing the *Basic Multimodal Feature* ("w/o Basic.") causes a substantial decline. This reflects a crucial design choice: while Qwen2.5-VL excels at high-level semantic reasoning, it may overlook low-level signals (e.g., specific acoustic frequencies or OCR text details). The basic feature extractor (using VGGish/BERT) complements the MLLM by capturing these fine-grained perceptual cues, proving that a hybrid approach is superior to relying solely on MLLMs.

4) Early Detection Analysis: In real-world social media governance, prompt identification of misinformation is

paramount to preventing its widespread propagation. To evaluate MEAFS in such time-sensitive scenarios, we simulate an early detection setting by restricting the availability of event-level context. Specifically, we vary the ratio of accessible videos in the auxiliary event set $\mathcal{H}_0$ from 0% (isolated target video) to 100% (full event context) on the FakeSV event-level split. Fig. 3 illustrates the performance trajectory. We observe that MEAFS exhibits a consistent "cold-start" capability. Even with only 25% of the event context available, the accuracy improves by roughly 1.0% compared to the isolated baseline (0%). As information accumulates, the detection accuracy steadily climbs, peaking at 84.11% when the full context is utilized.

V. Conclusion

In this paper, we focus on detecting fake news in the domain of short videos. We highlight the challenge posed by sophisticated forgeries that are difficult to identify from individual videos. To address this, we propose MEAFS, which leverages MLLMs to incorporate event-level context by generating video summaries, aggregating them into event representations, and deriving discrepancy features. Combined with basic multimodal features, MEAFS consistently outperforms baseline models on public short video datasets. Ablation studies further confirm that the event-level components are the key contributors to its success.

Acknowledgments

This work is supported by National Natural Science Foundation of China (62372454, 62576339, 62141608).

References

[1] C. Violot, T. Elmas, I. Bilogrevic, and M. Humbert, "Shorts vs. regular videos on youtube: A comparative analysis of user engagement and content creation trends," in *Proceedings of the 16th ACM Web Science Conference*, 2024, pp. 213–223.

[2] Y. Bu, Q. Sheng, J. Cao, P. Qi, D. Wang, and J. Li, "Combating online misinformation videos: Characterization, detection, and future directions," in *ACM MM*, 2023, pp. 8770–8780.

[3] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys*, pp. 1–40, 2020.

[4] B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, and L. Yuan, "Video-llava: Learning united visual representation by alignment before projection," in *EMNLP*, 2024, pp. 5971–5984.

[5] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2. 5-vl technical report," *arXiv*, 2025.

[6] X. Lu, T. Zhang, C. Meng, X. Wang, J. Wang, Y.-F. Zhang, S. Tang, C. Liu, H. Ding, K. Jiang *et al.*, "Vlm as policy: Common-law content moderation framework for short video platform," in *KDD*, 2025, pp. 4682–4693.

[7] J. Wang, J. Liu, N. Zhang, and Y. Wang, "Consistency-aware fake videos detection on short video platforms," in *International Conference on Intelligent Computing*. Springer, 2025, pp. 200–210.

[8] L. Zong, J. Zhou, W. Lin, X. Liu, X. Zhang, and B. Xu, "Unveiling opinion evolution via prompting and diffusion for short video fake news detection," in *ACL Findings*, 2024, pp. 10 817–10 826.

[9] G. Liu, J. Zhang, Q. Liu, J. Wu, S. Wu, and L. Wang, "Uni-modal event-agnostic knowledge distillation for multimodal fake news detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 12, pp. 9490–9503, 2024.

[10] Q. Liu, J. Wu, S. Wu, and L. Wang, "Out-of-distribution evidence-aware fake news detection via dual adversarial debiasing," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 11, pp. 6801–6813, 2024.

[11] X. Zhang, J. Cao, X. Li, Q. Sheng, L. Zhong, and K. Shu, "Mining dual emotion for fake news detection," in *Proceedings of the web conference 2021*, 2021, pp. 3465–3476.

[12] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, "A stylometric inquiry into hyperpartisan and fake news," in *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)*, 2018, pp. 231–240.

[13] M. Zhang, H. Gong, Q. Liu, S. Wu, and L. Wang, "Breaking event rumor detection via stance-separated multi-agent debate," *arXiv preprint arXiv:2412.04859*, 2024.

[14] P. Qi, Y. Bu, J. Cao, W. Ji, R. Shui, J. Xiao, D. Wang, and T.-S. Chua, "Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms," in *AAAI*, 2023, pp. 14 444–14 452.

[15] H. Choi and Y. Ko, "Using topic modeling and adversarial neural networks for fake news video detection," in *Proceedings of the 30th ACM international conference on information & knowledge management*, 2021, pp. 2950–2954.

[16] Y. Bu, Q. Sheng, J. Cao, P. Qi, D. Wang, and J. Li, "Fakingrecipe: Detecting fake news on short video platforms from the perspective of creative process," in *ACM MM*, 2024, pp. 1351–1360.

[17] Z. Zeng, M. Luo, X. Kong, H. Liu, H. Guo, H. Yang, Z. Ma, and X. Zhao, "Mitigating world biases: A multimodal multi-view debiasing framework for fake news video detection," in *ACM MM*, 2024, pp. 6492–6500.

[18] S. Ren, Y. Liu, Y. Zhu, W. Bing, H. Ma, and W. Wang, "Mmsfd: Multi-grained and multi-modal fusion for short video fake news detection," in *2024 7th International Conference on Data Science and Information Technology (DSIT)*. IEEE, 2024, pp. 1–11.

[19] W. Zhong, Y. Xiao, M. Xu, and X. Cheng, "Vmid: A multimodal fusion llm framework for detecting and identifying misinformation of short videos," *ArXiv preprint*, vol. abs/2411.10032, 2024. [Online]. Available: https://arxiv.org/abs/2411.10032

[20] R. Hong, J. Lang, J. Xu, Z. Cheng, T. Zhong, and F. Zhou, "Following clues, approaching the truth: Explainable micro-video rumor detection via chain-of-thought reasoning," in *WWW*, 2025, pp. 4684–4698.

[21] L. Zong, W. Lin, J. Zhou, X. Liu, X. Zhang, B. Xu, and S. Wu, "Text-guided fine-grained counterfactual inference for short video fake news detection," in *AAAI*, 2025, pp. 1237–1245.

[22] K. Niu, D. Xu, B. Yang, W. Liu, and Z. Wang, "Pioneering explainable video fact-checking with a new dataset and multi-role multimodal model approach," in *AAAI*, 2025, pp. 28 276–28 283.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[24] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," in *EMNLP*, 2017, pp. 670–680.

[25] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *ICASSP*. IEEE, 2017, pp. 131–135.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019, pp. 4171–4186.

[27] P. Qi, Y. Zhao, Y. Shen, W. Ji, J. Cao, and T.-S. Chua, "Two heads are better than one: Improving fake news video detection by correlating with neighbors," in *ACL Findings*, 2023, pp. 11 947–11 959.

[28] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *ArXiv preprint*, vol. abs/2303.08774, 2023. [Online]. Available: https://arxiv.org/abs/2303.08774

[29] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu *et al.*, "Llava-onevision: Easy visual task transfer," *arXiv preprint arXiv:2408.03326*, 2024.