Detecting Fake News in Short Videos Through Multi-View Aggregation

Nuo Li[1,2], Yuan Xiong[2], Chengliang Liu[3], Jie Wen[4], Chao Huang[2]*

[1]School of Automation, Nanjing University of Information Science and Technology
[2]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University
[3]Laboratory for Artificial Intelligence in Design, The Hong Kong Polytechnic University
[4]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
li.nuo.research@gmail.com, xiongy89@mail.sysu.edu.cn, liucl1996@163.com, jiewen_pr@126.com,
huangch253@mail.sysu.edu.cn

Abstract

The increasing prominence of short video platforms has positioned them as a primary channel for public awareness of current events, while also facilitating the widespread dissemination of fake news, thus highlighting the critical need for automated detection technologies. In contrast to fake news confined to text and images, short video news encompasses multiple modalities and extensive information, presenting heightened challenges. Most existing research emphasizes the analysis of news content or user comments alone, while overlooking the crucial role of publishers, leading to poor model performance when handling fake news lacking obvious false signals. Therefore, we propose a Publisher Profiling Module to identify new false signals. To enable a more comprehensive detection of misinformation, we design a Multi-View Aggregation (MVA) model, simultaneously evaluating news from three distinct perspectives: sentiment analysis, content understanding, and publisher profiling. Late fusion is applied at the decision level to leverage the complementary strengths of these perspectives, addressing the limitations of single-view methods. Our experiments conducted on the FakeSV and FVC datasets demonstrate the superior performance of the proposed method.

Introduction

Short video platforms have become a dominant force in the global internet ecosystem, as evidenced by their massive user engagement. In the United States, a 2023 Pew Research Center survey found that 33% of Americans actively use platforms like TikTok, reflecting significant adoption among the population (Gottfried 2024). On a global scale, the trend is equally significant. According to the DataReportal Digital 2025 Global Overview Report, TikTok ranked third in total time spent among mobile apps worldwide between September 1 and November 30, 2024, while also ranking first in both downloads and consumer spend (DataReportal 2025). These figures highlight TikTok's strong user engagement and its leading position in global app adoption and revenue generation, making it a key platform for content dissemination, including the spread of fake news. With this explosive growth in user base, short video platforms have gradually evolved from entertainment tools into a central medium for news dissemination, with a significant portion of users relying on them for current affairs information. However, this trend has also given rise to the proliferation of a new form of fake news—misleading content created through techniques such as editing, voiceovers, and AI-generated media (Bu et al. 2023), which spreads rapidly across platforms and poses a serious threat to public information security. Against this backdrop, developing fake news detection technologies tailored to the short video context has become a critical challenge in ensuring the credibility of digital content.

Figure 1: Two news use the same video material: in the real news, the video documents an event during Typhoon Lekima in 2019, while the fake news falsely claims it depicts a heavy rain event on July 6, 2020. Merely changing certain entities can turn the news into a fake one, yet the content still appears plausible.

Traditional fake news detection methods primarily focus on analyzing the authenticity of news content itself, such as identifying manipulated multimedia elements (e.g., deepfakes (Ganti 2022)) or verifying cross-modal semantic consistency (Choi and Ko 2021; Shang et al. 2021). While these approaches have shown effectiveness in detecting overtly fabricated content, they struggle when the falsification stems not from the perceptual integrity of the content itself but from the manipulation of contextual or meta-information (e.g., temporal inaccuracies, geographic misattributions, or flawed causal logic, as shown in Figure 1) or when the news itself is entirely fabricated. In such cases, content-centric methods fail to detect deceptive practices because both visual and textual elements appear plausible in isolation. This critical limitation highlights the inadequacy of relying solely on content analysis, as it overlooks the discrepancy be-

*Corresponding author.

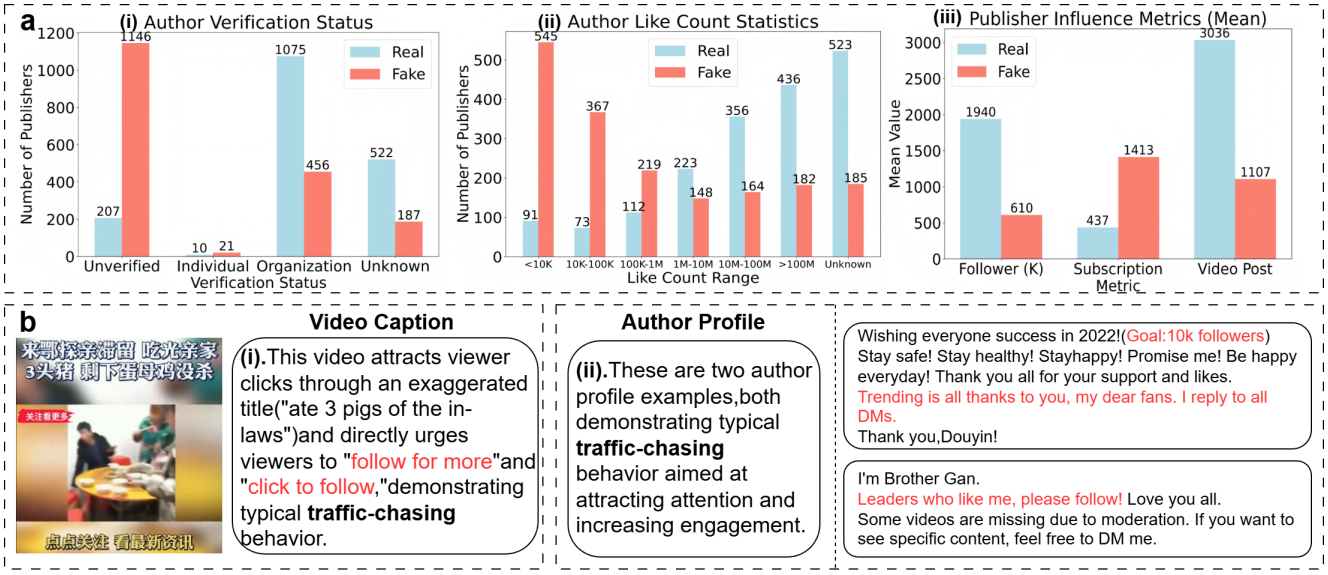

Figure 2: Analysis of Publisher Characteristics: (a) Statistics on Publisher Characteristics, showing that real news publishers are mostly verified with more followers, like counts and video posts, while fake news publishers are largely unverified with more subscriptions; (b) Behavioral Patterns in Video Captions and Author Profiles, revealing that many profit-driven fake news publishers use follower-seeking text in video captions and author profiles.

tween factual accuracy and contextual authenticity—a challenge exacerbated by the increasingly sophisticated context-aware manipulation tactics on short video platforms. To address this issue, we build upon the **Content Comprehension Module(CCM)** and introduce additional perspectives to compensate for its shortcomings.

Beyond semantic content-based detection methods, existing fake news detection research has explored the use of emotional signals (Vosoughi, Roy, and Aral 2018; Ajao, Bhowmik, and Zargari 2019; Giachanou, Rosso, and Crestani 2019) and social context (Qi et al. 2023a) as auxiliary features. However, these methods face critical limitations. The inherent subjectivity of emotional features often results in inferior performance compared to semantic features (Zhang and Ghorbani 2020), and they primarily target textual modality, making it difficult to extend to multimodal scenarios (Tufchi, Yadav, and Ahmed 2023). Similarly, while some studies leverage user comments as social context to enrich textual information (Qi et al. 2023a), this approach often introduces significant noise due to the subjectivity and irrelevance of user comments. Moreover, experimental data indicate that text features extracted solely from user comments exhibit significantly lower performance compared to other text features (Khattar et al. 2019; Singhal et al. 2019). Due to the complexity and heterogeneity of user behavior patterns, as well as significant data missing in many samples, conventional models struggle to extract effective information (Shu et al. 2020).

To address these challenges, we propose the **Emotion-Aware Module (EAM)** as an auxiliary view to improve the robustness of the model by capturing novel emotional signals from audio-modality (Sundar, Molina, and Cho 2021) emotional expressions in short video-based fake news, which often use emotionally charged background music more frequently (Bu et al. 2024). Additionally, we conducted a statistical analysis of the metadata in the FakeSV dataset as shown in Figure 2(a), focusing on publisher characteristics, and propose the **Publisher-Content Profiling Module (PCPM)** to capture clues such as the fact that real news publishers are predominantly verified, 83.9% vs. 29.4% for fake news, their interaction metrics are significantly higher, and fake news publishers exhibit abnormally high subscription counts, suggesting potential profit-driven amplification strategies. Furthermore, we observed that a majority of fake news events and some real news events involve highly repetitive content across samples, with publishers primarily being individual users chasing trending topics rather than organized malicious actors, and their behavioral patterns further support this conclusion as shown in Figure 2(b). Unlike prior work that directly extracts features from social metadata to infer publisher intent (Medina Serrano, Papakyriakopoulos, and Hegelich 2020; Qi et al. 2023a), we leverage large language models (LLMs) to meticulously analyze publisher behavior patterns and intent, designing prompt templates to extract user behavior features even in the presence of missing data, while also uncovering implicit clues in the news. The information obtained from the LLMs is then fused with the original input features to further enhance detection performance.

We adopt a late fusion strategy to perform weighted fusion of the fresults from the three views, balancing the contributions of different views to the model's final predictions. Experimental results demonstrate that the multi-view approach for news detection can effectively compensate for

the limitations of single-view methods, thereby improving the model's accuracy and robustness. Our contributions can be summarized as follows:

- We propose a multi-view aggregation model, MVA, for short video fake news detection, which integrates information from content understanding, sentiment analysis, and publisher profiling to effectively enhance model performance.
- We propose a novel method to detect fake news from the perspective of publisher intent, designing a lightweight prompt template to capture fake clues in publisher behavior and content.
- We conduct experiments on a public benchmark dataset. The results show that our method significantly improves detection accuracy compared to baseline methods.

Related Work

Early research on fake news detection primarily focused on text and image modalities, with relatively limited work on video modality detection. Papadopoulou et al. (Papadopoulou et al. 2018) pioneered the exploration of video fake news detection by constructing a Support Vector Machine (SVM) classifier based on heterogeneous features such as video metadata, titles, and user comments, laying the foundation for subsequent research. Hou et al. (Hou et al. 2019) further expanded the multimodal analysis dimension by innovatively introducing audio features, enhancing detection signals through cross-modal alignment. To address the issue of topic bias in multi-source features, Choi and Ko (Choi and Ko 2021) proposed the FANVM, which models the topic differences between titles, comments, and video frames to drive the model to learn topic-agnostic features. TikTec (Shang et al. 2021) extracts key information from video captions and learns common misinformation information from both video and audio modalities. Qi et al. (Qi et al. 2023a) constructed the largest Chinese short video fake news dataset and proposed the SV-FEND model, which extracts multimodal features and integrates them using Transformer. Additionally, they introduced NEED (Qi et al. 2023b), a model that leverages a Graph Attention Network (GAT) to obtain debunking information from additional event samples, significantly enhancing the detection capability. MMVD (Zeng et al. 2024) proposed a multi-view debiasing framework that categorizes multimodal features into social, static, and dynamic perceptual views, applying different debiasing methods to reduce the impact of noise. Bu et al. (Bu et al. 2024) proposed a novel approach by detecting fake news videos from the perspective of their creation process and designed FakingRecipe, providing new insights for detection tasks. OpEvFake (Zong et al. 2024) explored implicit opinions in news and implemented a multimodal opinion fusion evolution process based on diffusion models (Ho, Jain, and Abbeel 2020).

Methodology

Framework Overview

The proposed Multi-View Aggregation (MVA) framework, as shown in Figure 3, builds upon a semantics-based **Content Comprehension Module**, complemented by an **Emotion-Aware Module** to enhance model robustness, and integrates a **Publisher-Content Profiling Module** to address the limitations of traditional semantics-based detection approaches, thereby improving overall accuracy. Specifically, an input news video is decomposed into multiple modalities (e.g., text, images, and audio), which are then processed by three distinct sub-modules. Each sub-module independently analyzes its corresponding data, producing individual predictions and loss values. Subsequently, a late fusion strategy is employed to aggregate the outputs of the sub-modules, resulting in the final prediction.

Content Comprehension Module

As the most commonly used carrier of news content, we start with the text and visual modalities to primarily construct the Content-Comprehension Module. Specifically, we use the pre-trained BERT-Base-Chinese (Devlin et al. 2019) model to extract token-level text features $\boldsymbol{e_{ct}}$ from news headlines and OCR-extracted subtitles; and use the pre-trained CLIP (Radford et al. 2021) visual encoder to extract visual features $\boldsymbol{e_{cv}}$ from video frames. The obtained semantic features of text and video will undergo interactive fusion through a Co-Attention Transformer (Lu et al. 2016; Vaswani et al. 2017), obtaining visually enhanced text features $\boldsymbol{e_{ct-v}}$ and text-enhanced visual features $\boldsymbol{e_{cv-t}}$. Subsequently, they will undergo semantic alignment and use simple methods to obtain comprehensive language features, which are then input into the classification head to obtain the predicted value Y_c and loss of the CCM model.

To simplify the notation of the Co-Attention mechanism, we define a general attention function Attn as:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V, \tag{1}$$

where Q, K, and V are features projected by learnable matrices W, and d is the feature dimension.

The Co-Attention mechanism enhances visual and text features as follows:

$$\boldsymbol{e_{cv-t}} = \text{LN}\left(\boldsymbol{e_{cv}} + \text{Attn}(W^q_{v2t}\boldsymbol{e_{cv}}, W^k_{v2t}\boldsymbol{e_{ct}}, W^v_{v2t}\boldsymbol{e_{ct}})\right) \tag{2}$$

$$\boldsymbol{e_{ct-v}} = \text{LN}\left(\boldsymbol{e_{ct}} + \text{Attn}(W^q_{t2v}\boldsymbol{e_{ct}}, W^k_{t2v}\boldsymbol{e_{cv}}, W^v_{t2v}\boldsymbol{e_{cv}})\right) \tag{3}$$

where W is a learnable projection matrix, d is the feature dimension, and LN denotes Layer Normalization.

The classification head outputs the prediction:

$$y_c = \text{softmax}\left(W_c\left(\text{mean}(\boldsymbol{e_{cv-t}}) + \text{mean}(\boldsymbol{e_{ct-v}})\right) + b_c\right), \tag{4}$$

where $\text{mean}(\cdot)$ denotes the mean pooling operation over the sequence dimension, W_c and b_c are the weight matrix and bias of the classification layer, respectively.

Emotion-Aware Module

To construct the Emotion-Aware Module (EAM), we utilize text and audio modalities as sources of emotional signals. Text contains a wealth of emotionally charged vocabulary, while background music in audio is similarly rich in emotional content and highly evocative. Unlike previous work,

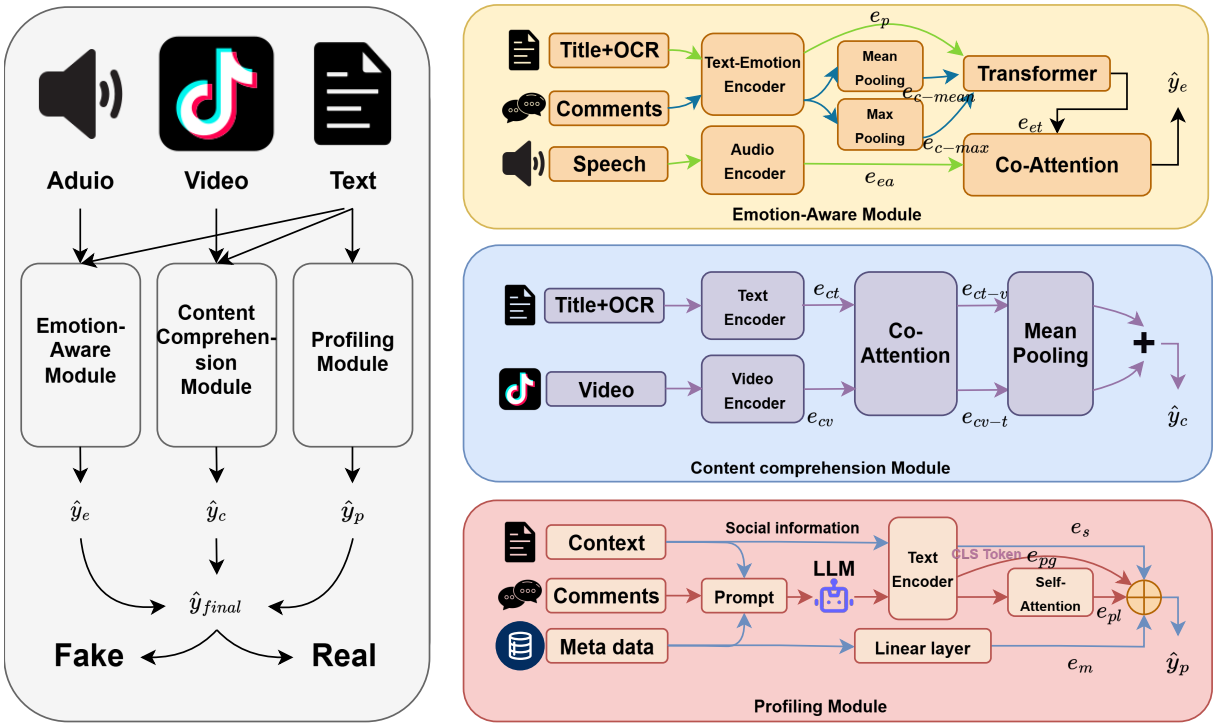

Figure 3: Overview of the proposed MVA model.The news video will be split into different modalities and types of data, which are then fed into three sub-models: EAM, CCM, and PCPM. Subsequently, a late fusion strategy is employed to combine the predicted outputs from these sub-models into the final prediction result.

we do not solely rely on news headlines and subtitles as sources of textual emotion. Instead, we extract emotional features from user comments. According to (Zhang et al. 2021), there exists a phenomenon of "emotion resonance" or "emotion dissonance" between publisher emotion and social emotion, and this dual emotion can be indicative of the news veracity.

Specifically, we use fine-tuned XLM-RoBERTa (Conneau et al. 2019) and HuBERT (Hsu et al. 2021) models as emotional encoders for text and audio, respectively. Both models have been fine-tuned on emotion classification tasks, and we extract their last hidden states as embedded representations of emotional signals. Features extracted from headlines and subtitles are denoted as publisher emotion $\boldsymbol{e_p}$, while features extracted from comments are processed through mean pooling and max pooling to generate social emotion features $\boldsymbol{e_{c-mean}}$ and $\boldsymbol{e_{c-max}}$, respectively. Subsequently, these features are fed into a standard Transformer layer, where a self-attention mechanism fuses the dual emotions and captures the phenomena of emotion resonance and dissonance, yielding a unified textual emotion feature $\boldsymbol{e_{et}}$. This feature is then further fused with the audio-derived emotion feature $\boldsymbol{e_{ea}}$ via a Co-Attention mechanism to produce a comprehensive emotion feature, which is input into the classification head to obtain the predicted probabilities Y_E and loss of the EAM.

The text emotion features are then concatenated and fused using a Transformer encoder:

$$\boldsymbol{E_{text}} = [\boldsymbol{e_p}, \boldsymbol{e_{c-mean}}, \boldsymbol{e_{c-max}}], \tag{5}$$

$$\boldsymbol{E}^{\text{fused}}_{\boldsymbol{text}} = \text{Transformer}(\boldsymbol{E_{text}}), \tag{6}$$

$$\boldsymbol{e_{et}} = \text{mean}(\boldsymbol{E}^{\text{fused}}_{\boldsymbol{text}}, \text{dim} = 1), \tag{7}$$

where $\text{Transformer}(\cdot)$ denotes the Transformer encoder, and $\text{mean}(\cdot, \text{dim} = 1)$ computes the mean along the sequence dimension.

Subsequently, the fused text emotion feature $\boldsymbol{e_{et}}$ and the audio emotion feature $\boldsymbol{e_{ea}}$ are further fused through a Co-Attention mechanism:

$$\boldsymbol{e}^{\text{out}}_{\boldsymbol{ea}} = \text{LN}\left(\boldsymbol{e_{ea}} + \text{Attn}(W^q_{t2a}\boldsymbol{e_{et}}, W^k_{t2a}\boldsymbol{e_{ea}}, W^v_{t2a}\boldsymbol{e_{ea}})\right) \tag{8}$$

$$\boldsymbol{e}^{\text{out}}_{\boldsymbol{et}} = \text{LN}\left(\boldsymbol{e_{et}} + \text{Attn}(W^q_{a2t}\boldsymbol{e_{ea}}, W^k_{a2t}\boldsymbol{e_{et}}, W^v_{a2t}\boldsymbol{e_{et}})\right) \tag{9}$$

where W is a learnable projection matrix, and d is the feature dimension.

Finally, the comprehensive emotion feature is obtained by:

$$\boldsymbol{e_{fused}} = \boldsymbol{e}^{\text{out}}_{\boldsymbol{et}} + \boldsymbol{e}^{\text{out}}_{\boldsymbol{ea}}. \tag{10}$$

This fused feature is then input into the classification head to obtain the predicted value and loss of the EAM.

Publisher-Content Profiling Module

To comprehensively extract behavioral pattern features of publishers, the Profiling Module utilizes the following data:

(1) *context*: including news content text and social media context text; (2) *comments*: user comments reflect social interactions with the publisher, thereby revealing certain deceptive behavioral patterns; (3) *metadata*: including the publisher's verification level, number of followers, and number of subscribers.

It should be noted that, due to potential data missing in samples, such as videos lacking comments or author profiles without descriptions, and given that such cases are not uncommon, conventional methods may struggle to handle them effectively. Therefore, we encapsulate all the above data into $\mathcal{D}_{\text{news}}$ and input it into a Large Language Model (LLM) via a carefully designed prompt to obtain comprehensive analysis results. Specifically, the news content text helps the LLM understand the news event, the social media context and metadata provide critical information for the LLM to analyze the publisher's behavior, and user comments offer insights into the interactions between the publisher and users. To this end, we design the following prompt template:

> Assume you are a professional expert in identifying fake news. Based on your expertise, analyze the following news sample. First, analyze the behavior of the news publisher and describe their profile. Then, conduct the analysis from the following perspectives and provide justifications:
>
> 1. Does the author likely have the intent to attract traffic or gain followers? (Focus on checking for behaviors that encourage viewers to like, follow, or share.)
> 2. Is the news content easily verifiable?
> 3. Considering the comments, does the news exhibit inflammatory characteristics?
> 4. Does the news contain content that can be refuted by scientific knowledge?
> 5. Taking into account both the author and the news content, how credible is the event?

In fact, the prompt related to news content can be independently utilized within the CCM, and both prompts could be further detailed and specified. However, for the sake of computational resource efficiency, we have opted to merge them and integrate them into the PCPM. The analysis results generated by the LLM are processed through a BERT model to extract features. Specifically, the $[CLS]$ token is used as the global feature $\boldsymbol{e_{pg}}$, while the remaining tokens are treated as local features and integrated through a self-attention layer to obtain $\boldsymbol{e_{pl}}$. Meanwhile, the social information in the context is also processed by BERT to extract features $\boldsymbol{e_s}$, enabling the model to learn publisher-specific patterns. The metadata is mapped to $\boldsymbol{e_m}$ through a linear layer. Subsequently, the four features are concatenated and fed into the classification head to obtain the predicted probabilities Y_P and loss of the PCPM:

$$\boldsymbol{e}_{\text{final}} = [\boldsymbol{e_{pg}}, \boldsymbol{e_{pl}}, \boldsymbol{e_s}, \boldsymbol{e_m}], \tag{11}$$

$$\hat{y}_{PCPM} = \text{softmax}(W_c \boldsymbol{e}_{\text{final}} + b_c), \tag{12}$$

where W_c and b_c are the weight matrix and bias of the classification head, respectively.

Multi-View Late Fusion

In the Multi-View Late Fusion stage, we integrate the outputs from three different views to produce the final prediction. To ensure the independence of each view, we avoid using methods like Transformers for multi-view feature interaction. Instead, we employ a weighted averaging approach on the predictions to enable complementary information sharing across views. Moreover, since each view can make decisions independently, this approach enhances the interpretability of the results, allowing us to understand the "fakeness" of news by examining the predictions of individual views. Specifically, Let $\hat{y}_{CCM}$, $\hat{y}_{EAM}$, and $\hat{y}_{PCPM}$ denote the predicted probabilities from the CCM, EAM, and PCPM, respectively. The final prediction $\hat{y}_{\text{final}}$ is computed as:

$$\hat{y}_{\text{final}} = \lambda_1 \hat{y}_{CCM} + \lambda_2 \hat{y}_{EAM} + \lambda_3 \hat{y}_{PCPM}, \tag{13}$$

where λ_1, λ_2, and λ_3 are hyperparameters set manually, adjusted through experiments to balance the contributions of each module, satisfying $\lambda_1 + \lambda_2 + \lambda_3 = 1$.

To optimize the model, each view employs the cross-entropy loss function as its loss metric. The cross-entropy loss is defined as:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i), \tag{14}$$

where C is the number of classes, y_i is the ground-truth label for class i, and $\hat{y}_i$ is the predicted probability for class i. Let $\mathcal{L}_{CCM}$, $\mathcal{L}_{EAM}$, and $\mathcal{L}_{PCPM}$ denote the cross-entropy losses for the CCM, EAM, and PCPM, respectively. The final loss $\mathcal{L}_{\text{final}}$ is computed as a weighted sum of the individual losses, using the same weights as in the prediction fusion:

$$\mathcal{L}_{\text{final}} = \lambda_1 \mathcal{L}_{CCM} + \lambda_2 \mathcal{L}_{EAM} + \lambda_3 \mathcal{L}_{PCPM}, \tag{15}$$

Experiments

Experimental Setup

Dataset Our experiments are primarily conducted on the **FakeSV** (Qi et al. 2023a) dataset, which is the largest Chinese short video fake news dataset and currently the only benchmark dataset for short video fake news detection. The dataset is collected from Douyin and Kuaishou, two of the most popular short video platforms in China, including video, title content, and social context. Additionally, we conducted experiments on the **FVC** (Papadopoulou et al. 2018) dataset, which was developed in the InVIDeo Veritas (InVID) project, and is the largest dataset for Fake Videos detection. This study mainly focusing on short videos on YouTube.

Baselines *Unimodal Baselines:* Traditional methods for fake news detection often focus on extracting expressive unimodal features. In this work, we employ several well-established single-modal models to analyze their respective modalities: BERT (Devlin et al. 2019) for textual features, VGGish (Hershey et al. 2017) for audio features,

Type	Method	FakeSV				FVC			
		Acc	F1	Pre.	Rec.	Acc	F1	Pre.	Rec.
Unimodal	Text(BERT)	76.82	76.80	76.89	76.82	76.37	76.35	76.39	76.33
	Keyframes(VGG19)	69.40	69.33	69.64	69.40	65.79	65.81	65.49	66.08
	Audio(VGGish)	66.78	66.63	67.07	66.78	58.44	58.61	58.48	58.63
	Video(C3D)	69.05	68.93	69.36	69.05	71.81	71.72	71.89	71.85
Multimodal	FANVM (Choi and Ko 2021)	79.52	78.81	79.81	78.46	85.81	85.32	85.20	85.44
	TikTec (Shang et al. 2021)	76.43	73.26	73.22	73.53	77.02	73.95	74.24	73.67
	SV-FEND (Qi et al. 2023a)	80.88	80.54	80.17	80.62	84.71	85.37	84.25	86.53
	MMVD (Zeng et al. 2024)	82.64	82.63	82.63	82.73	89.28	90.36	90.27	90.46
	FakingRecipe (Bu et al. 2024)	85.35	84.83	85.84	84.29	85.60*	85.07*	85.86*	85.45*
LLMs	QWQ-32b (Qwen Team 2024)	74.37	72.48	75.82	72.18	78.20	68.07	76.26	66.22
	Deepseek-R1 (Guo et al. 2025)	76.21	74.32	78.52	73.96	79.21	77.36	79.45	75.27
	MVA(Ours)	**88.38**	**87.94**	**89.43**	**87.36**	**95.19**	**94.80**	**93.89**	**94.33**

Table 1: The performance comparison of different methods on the FakeSV and FVC datasets is presented, with the best results emphasized in bold. The second-best results are underlined. The results marked with * indicate that the model excluded a portion due to the absence of the required on-screen text in the dataset.

VGG19 (Simonyan and Zisserman 2014) for visual features from static frames, and C3D (Tran et al. 2015) for spatiotemporal features from video sequences. *Multimodal Baselines:* We use several state-of-the-art (SOTA) models as multimodal baselines: FANVM (Choi and Ko 2021), TikTec (Shang et al. 2021), SV-FEND (Qi et al. 2023a), MMVD (Zeng et al. 2024), FakingRecipe (Bu et al. 2024) and OpEvFake (Zong et al. 2024). *Large Language Models Baselines:* To explore the potential of LLMs in fake news detection tasks, we select the widely adopted **deepseek-R1** and the lightweight deep reasoning model **QWQ**, which, despite having only 32 billion parameters, achieves performance comparable to DeepSeek-R1-671b.

Implementation Details We implement our model using PyTorch and conduct all experiments on an NVIDIA RTX3090 GPU with 24GB of memory. Both Deepseek-671b and QwQ-32b are accessed via API calls. The learning rate is set to 1e-3 for all modules, and we train the model for up to 30 epochs using the AdamW optimizer (Loshchilov and Hutter 2019) with a weight decay of 0.01, applying early stopping with a patience of 5 epochs. The weights λ_1, λ_2, and λ_3 for late fusion and loss computation are initialized as 0.35, 0.2, and 0.45, respectively, and are optimized during training. To ensure reproducibility, the random seed was set to 2025 using PyTorch 2.0 for all experiments. For evaluation, we adopt standard metrics for fake news detection, including Accuracy, macro-averaged Precision, macro-averaged Recall, and macro-averaged F1-Score, to comprehensively assess the model's performance.

Overall Performance

The experiments are conducted on the FakeSV dataset, and the performance of our method compared to the baseline models is presented in Table **??**. As shown in the table, single-modal methods, due to their limited capacity to capture information, generally perform significantly worse than multimodal methods, underscoring the importance of information interaction across multiple modalities for fake news detection. Among the multimodal methods, our model achieves an accuracy of 88.38%, outperforming all baseline models under comparison, thus validating the effectiveness of our approach for short video fake news detection.

Large language models (LLMs) show suboptimal performance in fake news detection despite their deep reasoning abilities, likely due to reliance on text-only prompts without multimodal support. Moreover, LLMs often overanalyze irrelevant details, impairing decision-making. Thus, we believe fine-tuning general-purpose LLMs may enhance their capability in fake news detection tasks.

Ablation Study

Emotion	Profile	Content	Acc.	F1
✓			81.18	80.62
	✓		83.21	82.70
		✓	84.87	84.17
✓	✓		85.06	84.86
✓		✓	85.24	84.56
	✓	✓	88.01	87.59
✓	✓	✓	88.38	87.94

Table 2: Ablation study on the FakeSV dataset. We evaluate the performance of different view combinations in MVA. A checkmark (✓) indicates the view is included.

The results of the MVA ablation study are presented in Table 2. To assess each view's contribution, we isolate the three views and train them as independent models. The Content View, the most fundamental perspective, performs best. The Profile View, enhanced by LLM reasoning, achieves 83.21% accuracy with text alone. Though the Emotion View lags behind, it still outperforms single-modal methods and

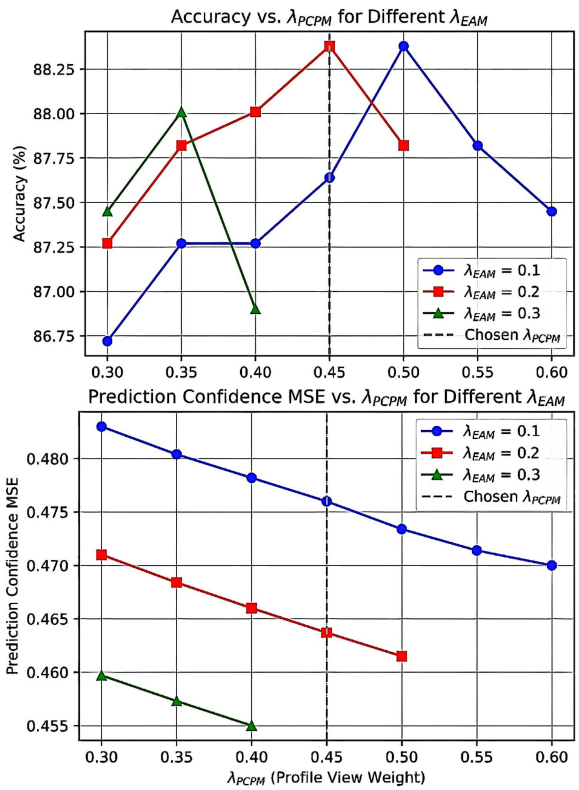

Figure 4: Influence of hyperparameters.

some multimodal approaches. Additionally, we remove individual views from the full model to study their roles in multi-view interactions. Results show removing the Emotion View causes a minimal 0.37% drop, supporting our hypothesis that emotional features mainly boost robustness. Conversely, removing Profile and Content Views reduces accuracy to 85.24% and 85.06%, respectively, highlighting the significant information complementarity between the Profile and Content Views.

Hyperparameters Analysis

The Emotion View in MVA is an auxiliary perspective, with its weight (λ_{EAM}) limited to a lower range to avoid over-dominance. We fix λ_{EAM} at 0.1, 0.2, and 0.3, varying λ_{PCPM} from 0.3 to 0.6 (step size 0.05), with $\lambda_{CCM} = 1 - \lambda_{EAM} - \lambda_{PCPM}$ in $[0.3, 0.6]$. We evaluate Accuracy and Prediction Confidence MSE for prediction stability. Results are shown in Figure 4, with line plots for Accuracy and MSE across hyperparameter settings.

As illustrated in Figure 4, a key trade-off is observed: peak accuracy occurs at λ_{EAM} values of 0.1 or 0.2, while a further increase lowers the MSE. Guided by this trend, we selected $\lambda_{EAM} = 0.2$ and $\lambda_{PCPM} = 0.45$ to best balance this trade-off. This configuration delivers superior accuracy while preserving stability through a minimized MSE.

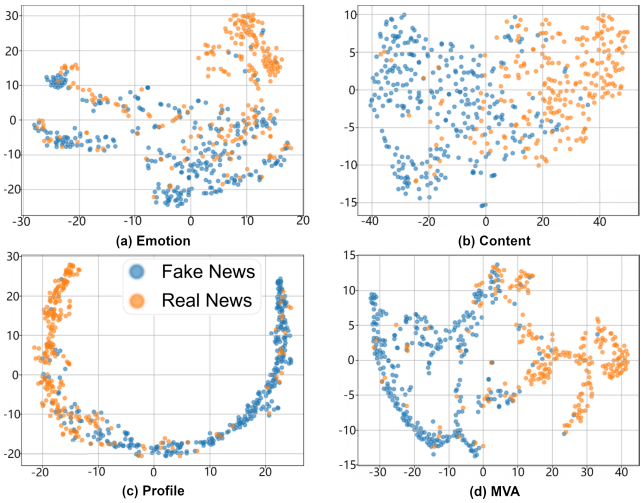

Figure 5: T-SNE Visualization of FakeSV dataset.

Visualization

To assess the discriminative capability of our MVA framework, we used t-SNE for visualization on the test set of the FakeSV dataset, with 542 samples (304 fake news, 238 real news). As shown in Figure 5, subfigures (a), (b), and (c) represent features from the three individual views, while (d) illustrates a four-dimensional vector, constructed by concatenating the probabilities of positive samples predicted by the three views and the fused probability. The Emotion view exhibits a clustering effect only for a subset of positive samples. The Content view effectively separates positive and negative samples into the left and right regions, with high dispersion. The Profile view forms a "C"-shaped structure, with positive and negative samples segregated at the arc ends, but overlapping in the -10 to 10 range. Finally, the MVA framework segregates samples at dimensions 0 and 10 with compact clustering, with only a small number of indistinguishable samples in the 0-10 range. This shows the multi-view aggregation approach enables interaction among views, improving discriminative capability and robustness, although it may introduce noise due to erroneous judgments from certain views.

Conclusion

In this paper, we proposed MVA, a multi-view aggregation framework for short video fake news detection. By integrating Content, Profile, and Emotion Views, MVA effectively addresses the limitations of traditional content-based detection methods, offering a robust and comprehensive solution for identifying fake news on short-video platforms. The incorporation of large language models enhances MVA's capability to capture nuanced publisher behavior patterns, improving detection accuracy. Our comprehensive evaluation on the FakeSV dataset demonstrates MVA's superior performance compared to existing methods, highlighting its potential as a reliable tool for combating misinformation in short video contexts.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No.62301621), Shenzhen Science and Technology Program (No. 20231121172359002, 2023A008), Shenzhen General Research Project (No. JCYJ20241202125904007), and Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515011398).

References

Ajao, O.; Bhowmik, D.; and Zargari, S. 2019. Sentiment aware fake news detection on online social networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2507–2511. IEEE.

Bu, Y.; Sheng, Q.; Cao, J.; Qi, P.; Wang, D.; and Li, J. 2023. Combating online misinformation videos: Characterization, detection, and future directions. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8770–8780.

Bu, Y.; Sheng, Q.; Cao, J.; Qi, P.; Wang, D.; and Li, J. 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*. Association for Computing Machinery.

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

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Unsupervised Cross-Lingual Representation Learning at Scale. *arXiv preprint arXiv:1911.02116*.

DataReportal. 2025. Digital 2025: Global Overview Report. Technical report, DataReportal. Available at: https://datareportal.com/reports/digital-2025-global-overview-report.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Ganti, D. 2022. A novel method for detecting misinformation in videos, utilizing reverse image search, semantic analysis, and sentiment comparison of metadata. *Utilizing Reverse Image Search, Semantic Analysis, and Sentiment Comparison of Metadata (June 5, 2022)*.

Giachanou, A.; Rosso, P.; and Crestani, F. 2019. Leveraging emotional signals for credibility detection. In *Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval*, 877–880.

Gottfried, J. 2024. Americans' Social Media Use. Technical report, Pew Research Center, Washington, DC, USA.

Guo, D.; Yang, D.; Zhang, H.; et al. 2025. Deepseek-r1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*.

Hershey, S.; Chaudhuri, S.; Ellis, D. P.; Gemmeke, J. F.; Jansen, A.; Moore, R. C.; Plakal, M.; Platt, D.; Saurous, R. A.; Seybold, B.; et al. 2017. CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, 131–135. IEEE.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. Curran Associates, Inc.

Hou, R.; Pérez-Rosas, V.; Loeb, S.; and Mihalcea, R. 2019. Towards Automatic Detection of Misinformation in Online Medical Videos. In *2019 International Conference on Multimodal Interaction*, 235–243. ACM.

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3451–3460.

Khattar, D.; Goud, J. S.; Gupta, M.; and Varma, V. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, 2915–2921.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical Question-Image Co-Attention for Visual Question Answering. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 289–297. Curran Associates, Inc.

Medina Serrano, J. C.; Papakyriakopoulos, O.; and Hegelich, S. 2020. NLP-based Feature Extraction for the Detection of COVID-19 Misinformation Videos on YouTube. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Online: Association for Computational Linguistics.

Papadopoulou, O.; Zampoglou, M.; Papadopoulos, S.; and Kompatsiaris, I. 2018. A corpus of debunked and verified user-generated videos. *Online Information Review*.

Qi, P.; Bu, Y.; Cao, J.; Ji, W.; Shui, R.; Xiao, J.; Wang, D.; and Chua, T.-S. 2023a. Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 14444–14452.

Qi, P.; Zhao, Y.; Shen, Y.; Ji, W.; Cao, J.; and Chua, T.-S. 2023b. Two Heads Are Better Than One: Improving Fake News Video Detection by Correlating with Neighbors. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 11947–11959. Toronto, Canada: Association for Computational Linguistics.

Qwen Team. 2024. Qwq: Reflect Deeply on the Boundaries of the Unknown. https://qwenlm.github.io/blog/qwq-32b-preview/.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.;

Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 8748–8763. PMLR.

Shang, L.; Kou, Z.; Zhang, Y.; and Wang, D. 2021. A multimodal misinformation detector for covid-19 short videos on tiktok. In *2021 IEEE international conference on big data (big data)*, 899–908. IEEE.

Shu, K.; Bhattacharjee, A.; Alatawi, F.; Nazer, T. H.; Ding, K.; Karami, M.; and Liu, H. 2020. Combating disinformation in a social media age. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(6): e1385.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singhal, S.; Shah, R. R.; Chakraborty, T.; Kumaraguru, P.; and Satoh, S. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, 39–47. IEEE.

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

Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; and Paluri, M. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 4489–4497.

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

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, ; and Polosukhin, I. 2017. Attention is All You Need. In *Advances in Neural Information Processing Systems 30*.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *science*, 359(6380): 1146–1151.

Zeng, Z.; Luo, M.; Kong, X.; Liu, H.; Guo, H.; Yang, H.; Ma, Z.; and Zhao, X. 2024. Mitigating World Biases: A Multimodal Multi-View Debiasing Framework for Fake News Video Detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, 6492–6500. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706868.

Zhang, X.; Cao, J.; Li, X.; Sheng, Q.; Zhong, L.; and Shu, K. 2021. Mining Dual Emotion for Fake News Detection. In *Proceedings of the Web Conference 2021*, 3465–3476.

Zhang, X.; and Ghorbani, A. A. 2020. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57(2): 102025.

Zong, L.; Zhou, J.; Lin, W.; Liu, X.; Zhang, X.; and Xu, B. 2024. Unveiling Opinion Evolution via Prompting and Diffusion for Short Video Fake News Detection. In Ku, L.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, 10817–10826. Association for Computational Linguistics.