FakingRecipe: Detecting Fake News on Short Video Platforms from the Perspective of Creative Process

Yuyan Bu
Institute of Computing Technology, Chinese Academy of Sciences
University of Chinese Academy of Sciences
buyuyan22s@ict.ac.cn

Qiang Sheng
Institute of Computing Technology, Chinese Academy of Sciences
shengqiang18z@ict.ac.cn

Juan Cao
Institute of Computing Technology, Chinese Academy of Sciences
University of Chinese Academy of Sciences
caojuan@ict.ac.cn

Peng Qi
National University of Singapore
peng.qi@nus.edu.sg

Danding Wang
Institute of Computing Technology, Chinese Academy of Sciences
wangdanding@ict.ac.cn

Jintao Li
Institute of Computing Technology, Chinese Academy of Sciences
jtli@ict.ac.cn

ABSTRACT

As short-form video-sharing platforms become a significant channel for news consumption, fake news in short videos has emerged as a serious threat in the online information ecosystem, making developing detection methods for this new scenario an urgent need. Compared with that in text and image formats, fake news on short video platforms contains rich but heterogeneous information in various modalities, posing a challenge to effective feature utilization. Unlike existing works mostly focusing on analyzing *what is presented*, we introduce a novel perspective that considers *how it might be created*. Through the lens of the creative process behind news video production, our empirical analysis uncovers the unique characteristics of fake news videos in material selection and editing. Based on the obtained insights, we design **FakingRecipe**, a creative process-aware model for detecting fake news short videos. It captures the fake news preferences in material selection from sentimental and semantic aspects and considers the traits of material editing from spatial and temporal aspects. To improve evaluation comprehensiveness, we first construct FakeTT, an English dataset for this task, and conduct experiments on both FakeTT and the existing Chinese FakeSV dataset. The results show FakingRecipe's superiority in detecting fake news on short video platforms.

CCS CONCEPTS

• **Information systems** → **Multimedia information systems**.

KEYWORDS

misinformation video detection; multi-modal computing

MM'24, October 28 - November 1, 2024, Melbourne, Australia.

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
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 (MM'24)* ACM, New York, NY, USA, 13 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

In recent years, short-form video-sharing platforms like TikTok have been increasingly popular on mobile Internet and revolutionizing how people consume news [16, 31]. According to Pew Research Center, by 2023, 33% of U.S. adults have ever used TikTok [14], with nearly 43% of these users frequently sourcing their news from this platform [30]. However, the prevalence of news consumption on short video platforms also encourages the emergence and spread of fake news videos, posing new serious threats to the online information ecosystem [4, 46]. Consequently, customizing methods for detecting short video fake news is of urgent need.

Unlike fake news in text or image formats, fake news on short video platforms shows unique characteristics and is increasingly indistinguishable from real news, posing new challenges to developing effective detectors [4]. First, the easy-to-use video editing tools largely democratize news creation and enable almost everyone to edit a news video on par with professional journalists [32], making the edit traces widely exist in both real and fake news videos. Second, due to the public nature of short video platforms, even a real news video is likely to be repurposed or re-edited for news faking. However, existing methods for fake news video detection mostly follow ideas from the research line of text-image-based detection and focus on modeling *what is presented* via analyzing the authenticity of multimodal content (*e.g.*, detecting *deepfakes* [13]) and modeling cross-modal correlation [7, 36, 42], which are more likely to be misled by edited and repurposed contents. Faced with the more vague boundary between the real and the fake, it is necessary to find new perspectives and capture more effective clues for fake news video detection.

In this paper, we propose to switch the perspective from analyzing *what is presented in a fake news video* to considering *how it might be created*. Our idea is based on a straightforward assumption: Fake news creators on short video platforms often lack

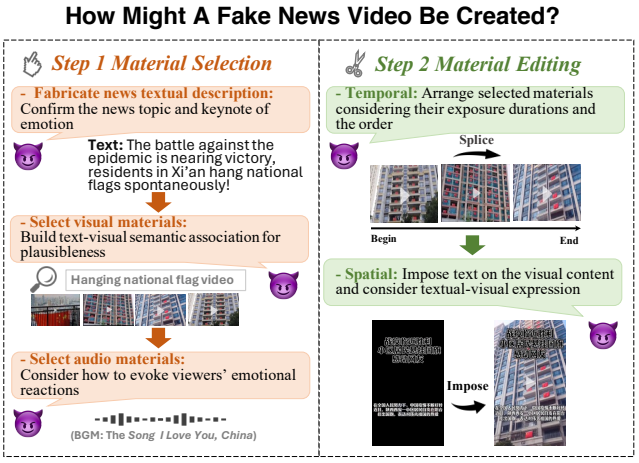


Figure 1: A fake news video about residents hanging national flags amid the COVID-19 pandemic in China, exhibited along with the speculated creative process. The text was translated into English.

first-hand, genuine news materials and professional editing skills while aiming to produce fake news for specific purposes intentionally [44]. This might leave unique characteristics of the resulting video. Figure 1 provides an intuitive example of the creative process of a fake news video about hanging national flags during the COVID-19 pandemic. The process typically unfolds in two main phases: **material selection** and **material editing**. For selecting the material, the creator first confirmed the news topic (i.e., residents hanging national flags during the pandemic) and the positive sentiment keynote and crafted an attractive narrative that diverges from the truth. Due to the lack of real visual materials (unlike real news), the creator had to repurpose historical materials collected from the Internet to make the fake video more convincing. Finally, an emotionally charged song is selected to impress audiences and achieve its underlying purpose. For the editing phase, the creator might consider arranging materials from the temporal and spatial views with the help of simple editing techniques. Due to the constraint of material sufficiency and editing skills, the collected visual materials were arranged with simple splicing temporally, and the text material was then spatially overlaid on the visual content for a straightforward textual-visual expression. Through this example, we intuitively find that the production of fake news videos may leave the nuances different from that of real ones in terms of material selection and editing. Therefore, modeling from the creative process perspective may help us capture more valuable instrumental clues for fake news video detection.

Inspired by the observation, in this paper, we first quantitatively examine how effective the clues from the creative process perspective are in distinguishing fake and real news videos via an empirical analysis (Section 2). The results validate that statistical discrepancies exist between real and fake news videos in material selection and editing. For instance, we find that compared with real ones, fake news videos exhibit a propensity for selecting more emotionally charged music, using a limited palette of colors, and adopting

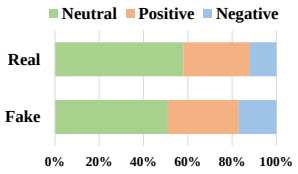


Figure 2: Sentiment analysis of audio material.

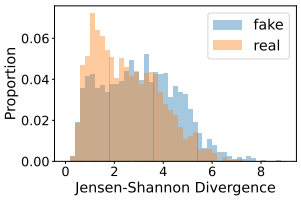


Figure 3: JS divergence between textual and visual materials.

a less dynamic on-screen text presentation. Based on the empirical analysis, we design **FakingRecipe**, a creative process-aware model for detecting fake news short videos.[1] FakingRecipe is a dual-branch network that models the characteristics of material selection and editing. In the two branches, the Material Selection-Aware Modeling (MSAM) module extracts multimodal features via attention to capture the sentiment resonance between audio and text and the semantic relevance between text and visual frames. The Material Editing-Aware Modeling (MEAM) module models typical spatial and temporal editing behaviors, via 1) analyzing the visual area and on-screen texts for the spatial editing; and 2) building hierarchical temporal structure that considers both intra- and inter-segment fusion for temporal editing. Ultimately, predictions from both branches are integrated through a late fusion function for the final prediction. Experiments on two real-world datasets demonstrate the superiority of the proposed FakingRecipe over seven baseline methods. Our main contributions are as follows:

- **Idea:** We for the first time consider the creative process as a pivotal aspect for detecting fake news on short video platforms and demonstrate the feasibility of this perspective through empirical analysis.
- **Method:** We propose FakingRecipe, a novel dual-branch model for fake news video detection that captures useful clues from the perspective of the creative process, *i.e.*, material selection and editing phases.
- **Resource & Effectiveness:** We construct FakeTT, an English short video dataset for fake news detection. Extensive experiments on both FakeTT and the public Chinese FakeSV dataset show the superiority of FakingRecipe over existing methods in fake news video detection. We will publicly release the new dataset to facilitate further research[2].

2 EMPIRICAL ANALYSIS

We exhibit the manifestation differences between real and fake news videos in different phases of news video creation by conducting empirical analysis on real-world datasets, including the publicly available Chinese dataset FakeSV [36] and the newly constructed English dataset FakeTT. We identify the discrepancies between real and fake news production processes and provide plausible explanations for these phenomena, highlighting the nuances in the creative process to evaluate the news video veracity. Considering

[1]The creative process of faking a news video is metaphorically similar to cooking a dish following a recipe, so we use **FakingRecipe** to highlight the model's uniqueness.
[2]https://github.com/ICTMCG/FakingRecipe

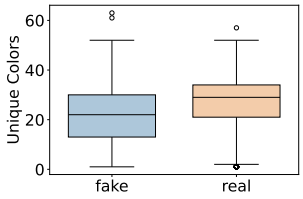


Figure 4: Color richness of on-screen text.

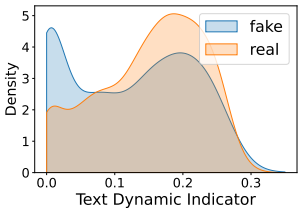


Figure 5: On-screen Text Dynamics.

that consistent results were observed across both datasets, we only present findings from FakeSV here due to space limitations and attach results on FakeTT in the supplementary material.

2.1 Phase I : Material Selection

Observation 1. When selecting audio materials, fake news tends to opt for more emotionally charged audio.

Considering background music (BGM) is a predominant option for short video news creators and the nature of BGM it serves primarily to evoke emotional responses, our analysis of audio selection behaviors focuses on the emotional aspect. We leverage the pre-trained wav2vec [40] that has been fine-tuned for audio emotion classification. Depicted in Figure 2, we can see that fake news videos exhibit an inclination towards using emotionally charged audio. Given that prior work [10] has indicated emotionality significantly boosts content sharing, we attribute this bias in audio selection to creators' intentions to maximize viewer engagement.

Observation 2. When selecting visual materials, fake news often employs clips that exhibit a relatively lower semantic consistency with the accompanying text.

We analyze creators' visual selection behaviors from the perspective of consistency between selected video materials and accompanying text. Specifically, we leverage the pre-trained text-image representation model CLIP [39] to extract textual and visual features. By normalizing these features and calculating the Jensen-Shannon (JS) Divergence between the attached text and each frame's visual content, we derive an average JS Divergence score across multiple frames as an indicator of text-visual consistency for the entire video. A lower indicator value signifies a higher semantic consistency between the video's textual and visual content. Figure 3 illustrates the distinct distributions of JS Divergence between textual and visual materials for fake and real news. The discrepancies have been statistically confirmed through the Kolmogorov-Smirnov (KS) test, with a p-value of less than 0.05. We find that fake news tends to utilize visual clips with noticeably lower semantic consistency with the accompanying text. We attribute the observed biases in video material selection to the nature that fabricated news inherently lacks access to a rich array of related video materials.

2.2 Phase II : Material Editing

We investigate two fundamental editing operations in video creation: spatial editing behaviors and temporal editing behaviors.

Observation 3. When spatially imposing text, fake news tends to display relatively plain textual visuals.

Spatial editing behaviors involve overlaying additional layers on top of the original visual materials. This can include adding animated stickers, text, and other elements. Among them, text imposition is a widely used operation in short news videos (85% in the FakeSV dataset), with variations reflected in decisions regarding the text's placement, color, typeface, and font. Here we quantified the color characteristics of the text visual areas in real and fake news videos respectively to explore the differences in color choice behaviors during text imposition. Figure 4 illustrates that real news videos tend to use a richer color palette for text presentation. We attribute this preference to that real news creators often follow conventional editorial norms and invest more effort to improve the presentation quality. Conversely, fake news creators often employ a monochromatic color scheme when imposing text, likely due to a lack of expertise in news production, leaving them unaware of the potential impact these details can have on viewers.

Observation 4. When temporally splicing materials, fake news tends to adopt a relatively simple arrangement.

Temporal editing behaviors, on the other hand, refer to the reorganization and splicing of multiple material segments. The duration and positioning of different segments can subtly influence viewers' perceptions of the news video. Here we examine the temporal editing behaviors related to text exposure, analyzing differences in the temporal arrangement of text segments between real and fake news. Specifically, we developed an indicator, I_D, to measure the dynamism of text presentation. By calculating the mean (μ) and standard deviation (σ) of exposure durations ($d_1, d_2, ...$) for different text phrases within a video, I_D is defined as $\sigma(1-\mu)$, based on the principle that shorter exposure times and greater variance among text exposure durations indicate stronger text temporal editing dynamism. Figure 5 shows the fitted sample density distribution of the on-screen text dynamic scores in the FakeSV dataset, revealing significant differences between the temporal editing behaviors of real and fake news, with real news exhibiting more dynamic text presentations. We ascribe this tendency to two factors: First, the disparity in video creation capabilities, wherein most creators of authentic news, endowed with professional media training, possess a nuanced understanding of effectively integrating text with visual elements. Second, the constraints posed by the availability of materials. Fabricated news, inherently characterized by its scant and biased content, often lacks the robust information necessary for dynamic presentations. This deficiency compels creators to resort to the static placement of limited information in specific areas of the screen.

3 METHOD

3.1 Overview

Drawing on the insights from our empirical analysis, we present FakingRecipe (Figure 6), a creative process-aware fake news video detection model. The model observes the given news video from the two pivotal phases of the creative process to unearth veracity indicating clues. Treating the feature from two phases as independent viewpoints, FakingRecipe is structured with dual branches

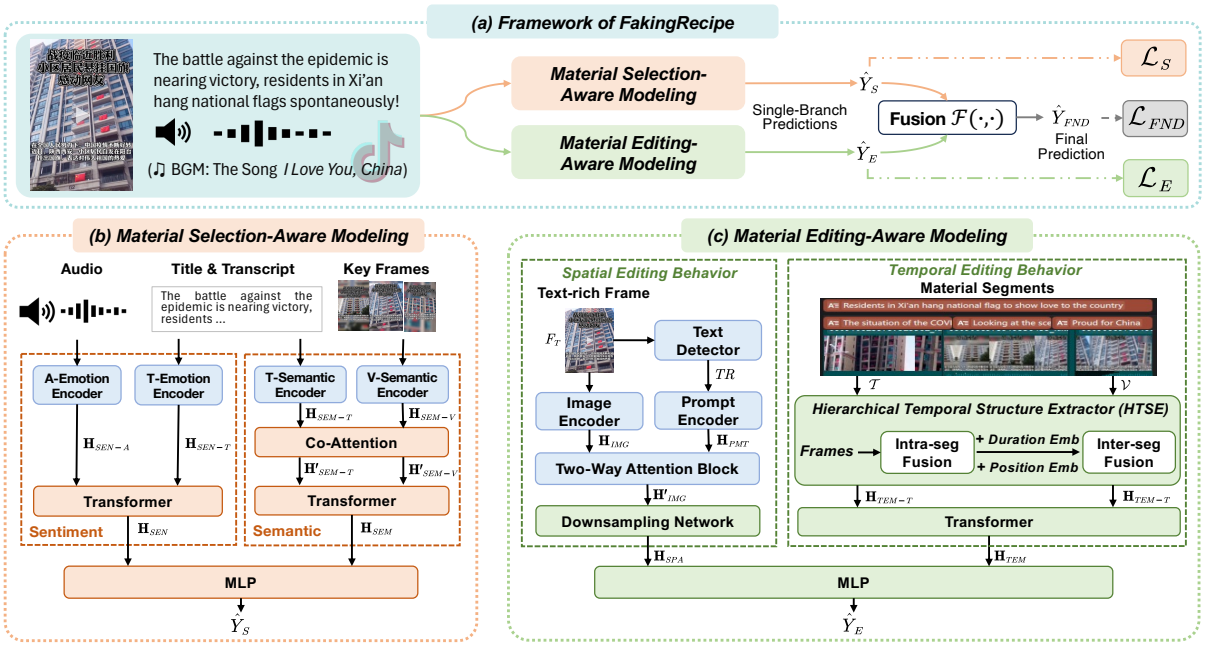


Figure 6: Overview of the proposed FakingRecipe model. (a) Overall framework: The news video is processed through dual perspectives, with a late fusion strategy employed to integrate clues for final prediction. (b) Material Selection-Aware Modeling (MSAM) module: Extracts clues from both sentimental and semantic aspects. (c) Material Editing-Aware Modeling (MEAM) module: Extracts clues based on spatial and temporal aspects. $\mathcal{F}(\cdot,\cdot)$ denotes the fusion function. The parameters in the modules in blue are frozen and others are trainable. The overall model is trained under the supervision of the loss functions $\mathcal{L}_{FND}$, $\mathcal{L}_S$, and $\mathcal{L}_E$. The text in this case is translated into English.

operating separately and employs a late fusion strategy to integrate predictions from these independent perspectives.

3.2 Material Selection-Aware Modeling (MSAM)

Based on prior analysis, we examine the creators' material selection behavior from two aspects (*i.e.*, sentiment and semantic). The dominant role of different modalities varies in conveying information: Audio primarily expresses emotion, text renders emotional tones while conveying semantic information, and visuals generally complement the text to communicate semantic content. Therefore, we strategically select combinations of modalities for multifaceted feature extraction, subsequently fusing multimodal features from multiple viewpoints.

Specifically, for the sentimental aspect, we consider audio and textual content as the primary sources. We utilize fine-tuned versions of HuBERT [18] and XLM-RoBERTA [9] as encoders to extract audio sentimental features $\mathbf{H}_{SEN-A}$ and textual sentimental features $\mathbf{H}_{SEN-T}$, respectively. These sentimental features from different modalities are then concatenated and fed into a standard Transformer layer [48]. By leveraging self-attention, the transformer layer fuses multimodal sentimental features to produce a unified sentimental feature representation $\mathbf{H}_{SEN}$.

In the semantic aspect, visual and textual contents take precedence, while the audio mainly serves as background music, playing a minimal role. Keyframes are extracted from videos, serving as the basis for visual analysis. Utilizing CLIP [39], we encode text and keyframes to token/frame-level text semantic features $\mathbf{H}_{SEM-T}$ and visual semantic features $\mathbf{H}_{SEM-V}$. Interaction between text and visual content is facilitated through a co-attention transformer [28], resulting in visually enhanced textual features $\mathbf{H'}_{SEM-T}$ and textually enhanced visual features $\mathbf{H'}_{SEM-V}$. These features are then averaged, concatenated, and input into a transformer layer mirroring the structure used in the sentimental analysis. This process integrates semantic features from various modalities into a singular semantic feature representation $\mathbf{H}_{SEM}$.

The sentimental feature $\mathbf{H}_{SEN}$ and semantic feature $\mathbf{H}_{SEN}$ are then concatenated and fed into a two-layer MLP to derive the fake news predicted score $\hat{Y}_S$ from the material selection analysis perspective:

$$\hat{Y}_S = \text{MLP}([\mathbf{H}_{SEN}; \mathbf{H}_{SEM}]). \tag{1}$$

3.3 Material Editing-Aware Modeling (MEAM)

In mining detecting clues from the perspective of creator editing behaviors, we focus on spatial and temporal aspects, identified as critical in our empirical analysis.

Spatially, we examine the prevalent practice of imposing text. Given a video V, we select a representative text-rich frame F_T, identified based on the size of the text presence area, as our starting point. We first employ an OCR spotting model, CRAFT [1], to delineate text regions $TR = \{\text{box}_1, \text{box}_2, ...\}$ within F_T. These regions are subsequently transformed into prompt embeddings $\mathbf{H}_{PMT}$ employing a methodology inspired by the prompt encoder

in Segment Anything Model (SAM) [24]. In parallel, F_T undergoes processing by a pre-trained Vision Transformer (ViT) [11] to produce initial encodings $\mathbf{H}_{IMG}$. Both $\mathbf{H}_{IMG}$ and $\mathbf{H}_{PMT}$ are then fed into a Two-Way Attention block to refine the initial visual encoding $\mathbf{H}_{IMG}$, ensuring it focuses more accurately on text regions within the frame. Suppose the attention mechanism is described as $\text{Att}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}(\mathbf{Q'}\cdot\mathbf{K'}/\sqrt{d})\mathbf{V'}$, where $\mathbf{Q'} = \mathbf{W}_Q\mathbf{Q}$, $\mathbf{K'} = \mathbf{W}_K\mathbf{K}$, $\mathbf{V'} = \mathbf{W}_V\mathbf{V}$, and d is the dimensionality. We format self-attention as $\text{SA}(\mathbf{X}) = \text{Att}(\mathbf{X},\mathbf{X},\mathbf{X})$ and cross-attention as $\text{CA}(\mathbf{X},\mathbf{Y}) = \text{Att}(\mathbf{X},\mathbf{Y},\mathbf{Y})$. The Two-Way Attention block functions in two directions as follows:

$$\mathbf{H}\text{-}\mathbf{H}_{PMT} = \text{LN}(\mathbf{H}_{PMT} + \text{SA}(\mathbf{H}_{PMT})), \tag{2}$$

$$\mathbf{H'}\text{-}\mathbf{H}_{PMT} = \text{LN}(\mathbf{H}\text{-}\mathbf{H}_{PMT} + \text{CA}(\mathbf{H}\text{-}\mathbf{H}_{PMT}, \mathbf{H}_{IMG})), \tag{3}$$

$$\mathbf{H''}\text{-}\mathbf{H}_{PMT} = \text{LN}(\mathbf{H'}\text{-}\mathbf{H}_{PMT} + \text{MLP}(\mathbf{H'}\text{-}\mathbf{H}_{PMT})), \tag{4}$$

$$\mathbf{H'}_{IMG} = \text{LN}(\mathbf{H}_{IMG} + \text{CA}(\mathbf{H}_{IMG}, \mathbf{H''}\text{-}\mathbf{H}_{PMT})), \tag{5}$$

where LN represents layer normalization. This block uses an embedding dimension of 256, and all attention layers use 8 heads. Following SAM, we adopt two layers of such block to obtain the updated image feature $\mathbf{H'}_{IMG}$ that focuses on text regions.
Following attention processing, the updated $\mathbf{H'}_{IMG}$ undergoes downsampling via two convolutional layers and then flattened to derive the spatial pattern feature $\mathbf{H}_{SPA}$:

$$\mathbf{H}_{SPA} = \text{GeLU}(\text{Conv}(\text{GeLU}(\text{LN}(\text{Conv}(\mathbf{H'}_{IMG}))))). \tag{6}$$

Temporally, we examine the splicing practice of text segment and video segment. The audio track is omitted due to the observation that most audio tracks consist of continuous background music. For a Video V, preprocessing extracts a sequence of text content $\mathcal{T} = \{(t_1,d_1),(t_2,d_2),...(t_n,d_n)\}$ and a sequence of visual content $\mathcal{V} = \{(v_1,d_1),(v_2,d_2),...(v_m,d_m)\}$, with n and m indicating the counts of text and video segments respectively. t_i represents the i-th textual segment, v_i denotes the middle frame of the i-th video clip, and $d_i = [\text{FrameIdx}_i^{begin}, \text{FrameIdx}_i^{end}]$ marks the time interval of the i-th segment's appearance. The input also incorporates frame rate (fps) and the total frame count (vframes) of the video to contextualize duration. Each modality's temporal structure is initially modeled separately, followed by an interaction phase to derive overall temporal editing features. Specifically, we design a Hierarchical Temporal Structure Extractor (HTSE) for extracting temporal structure features applicable to both modalities. HTSE first performs intra-segment fusion for content occurring in the same time span to derive segment content features $\mathbf{Seg}_i$. For text, $\mathbf{Seg}_i^T$ is obtained by concatenating multiple segments and encoding them collectively, while for visuals, it applies the self-attention (SA) mechanism for integration:

$$\mathbf{Seg}_i^V = \text{MEAN}(\text{SA}([v_1, v_2, ..., v_k])), \tag{7}$$

where k is the frame count within segment i, and $\text{MEAN}(\cdot)$ denotes the mean pooling. To model the subtle influences of the duration and temporal position of different segments, we introduce each segment's temporal position and exposure duration information. Positional encoding (**PE**) is generated using sine and cosine functions to reflect each segment's temporal position, akin to that leveraged by Transformer [48]:

$$\text{PE}_i^{(ei)} = \begin{cases} \sin(w_k i), & \text{if } ei = 2k \\ \cos(w_k i), & \text{if } ei = 2k+1, \end{cases} \tag{8}$$

where $w_k = 1/(10000^{2k/\dim_{PE}})$ represents the frequency of the sinusoid for each dimension and $\mathbf{PE}_i$ is the i-th segment's positional embedding. For duration encoding (**DE**), we employ an equi-frequency binning approach, determining duration groups through empirical analysis and assigning a learnable embedding to each group. Considering the audience's perception of exposure time in reality, both absolute and relative durations are evaluated:

$$\text{Dura}_i^{abs} = (\text{FrameIdx}_i^{begin} - \text{FrameIdx}_i^{end})/\text{fps},$$
$$\text{Dura}_i^{rel} = (\text{FrameIdx}_i^{begin} - \text{FrameIdx}_i^{end})/\text{vframes}.$$

Here Dura_i^{abs} and Dura_i^{rel} denote the absolute (in seconds) and relative (the proportion of the total duration) durations of segment i, respectively. The i-th segment's duration embedding is:

$$\mathbf{DE}_i = [\text{Emb}(\text{Group}(\text{Dura}_i^{abs})); \text{Emb}(\text{Group}(\text{Dura}_i^{rel}))], \tag{9}$$

where $\text{Group}(\cdot)$ maps a duration to its designated group and $\text{Emb}(\cdot)$ retrieves the corresponding embedding for that group.

Integrating positional and duration encodings, the segment features are updated to $\mathbf{SEG}_i$, serving as the input for inter-segment fusion, which captures the relationships between different segments using a similar self-attention mechanism, generating temporal pattern features for each modality:

$$\mathbf{SEG}_i = \mathbf{Seg}_i + \mathbf{PE}_i + \mathbf{DE}_i, \tag{10}$$

$$\mathbf{H}_{TEM-M} = \text{MEAN}(\text{SA}([\mathbf{SEG}_1^M, \mathbf{SEG}_2^M, ...])). \tag{11}$$

Utilizing HTSE, we derive temporal editing features for both text ($\mathbf{H}_{TEM-T}$) and visual ($\mathbf{H}_{TEM-V}$) modalities, and they are subsequently processed by a standard Transformer layer to produce the consolidated temporal editing feature $\mathbf{H}_{TEM}$. The spatial editing feature $\mathbf{H}_{SPA}$ and the temporal editing feature $\mathbf{H}_{TEM}$ are then concatenated and fed into a two-layer MLP to compute the fake news predicted score $\hat{Y}_E$ from the material editing analysis perspective:

$$\hat{Y}_E = \text{MLP}([\mathbf{H}_{SPA}; \mathbf{H}_{TEM}]. \tag{12}$$

3.4 Predication and Optimization

3.4.1 Prediction. Building on the predicted scores $\hat{Y}_S$ from material selection modeling and $\hat{Y}_E$ from material editing modeling, we adopt a late fusion strategy to get the final score $\hat{Y}_{FND}$:

$$\hat{Y}_{FND} = \mathcal{F}(\hat{Y}_S, \hat{Y}_E) = \hat{Y}_S * \tanh(\hat{Y}_E), \tag{13}$$

where $\mathcal{F}(\cdot,\cdot)$ is the fusion function. Inspired by previous works [6, 51], we adopt the $\tanh(\cdot)$ function, which introduces non-linearity to enhance the fusion strategy's representational capacity.

3.4.2 Optimization. Following previous works [7, 36, 42], we utilize cross-entropy loss to optimize our model:

$$\mathcal{L}_{FND} = \text{Cross-Entropy}(\hat{Y}_{FND}, Y), \tag{14}$$

where Y is the ground-truth label for each short video news.

To further supervise the material selection and material editing modeling, the final loss $\mathcal{L}$ incorporates the loss for $\hat{Y}_S$ and $\hat{Y}_E$:

$$\mathcal{L} = \mathcal{L}_{FND} + \alpha\mathcal{L}_S + \beta\mathcal{L}_E, \tag{15}$$

Table 1: Statistics of two datasets for evaluation.

Dataset	Time Range	Avg Duration (s)	#Fake	#Real	#All
FakeSV	2017/10-2022/02	39.88	1,810	1,814	3,624
FakeTT	2019/05-2024/03	47.69	1,172	819	1,991

where α and β are hyperparameters that balance the impacts on the back-propagation of the three branches. $\mathcal{L}_S$ and $\mathcal{L}_E$ denote the Cross-Entropy($\hat{Y}_S$, Y) and Cross-Entropy($\hat{Y}_E$, Y), respectively.

4 EXPERIMENTS

In this section, we conduct extensive experiments on two real-world datasets to verify the effectiveness of FakingRecipe by comparing it with seven representative baselines and the FakingRecipe variants.

4.1 Datasets

To validate the generalizability of the proposed FakingRecipe, we conduct experiments on two datasets of different languages:

FakeSV[3]: The largest publicly available Chinese dataset for fake news detection on short video platforms, featuring samples from *Douyin* and *Kuaishou*, two popular Chinese short video platforms. Each sample in FakeSV contains the video itself, its title, comments, metadata, and publisher profiles. We do not use the last three values to focus on understanding the content itself.

FakeTT: Our newly constructed English dataset for a comprehensive evaluation in English-speaking contexts [4]. Curated from TikTok, this dataset follows a similar collection process to [36], focusing on videos related to events reported by the fact-checking website Snopes[5]. Each video was rigorously annotated for authenticity by at least two independent annotators, resulting in a collection of 1,172 fake news videos and 819 real news ones from May 2019 to March 2024, with video, audio, and text description (title) available. See more details in the supplementary material.

Table 1 shows the statistics of the two datasets. To simulate real-world scenarios, we adopt a temporal split strategy for our experiments, dividing the data chronologically into training, validation, and testing sets with ratios of 70%, 15%, and 15%, respectively. Such a data split reflects the potential of applying compared methods in reality.

4.2 Experimental Setup

We compare the proposed FakingRecipe with a range of state-of-the-art baselines, including handcrafted features-based baselines, neural networks-based baselines, and (multimodal) large language model ((M)LLMs) baselines:

Handcraft Feature-based Baselines: (1) **HCFC-Hou [17]** utilizes linguistic features from speech, acoustic emotion features, and user engagement statistics with a linear kernel SVM for classification. (2) **HCFC-Medina [41]** extracts TF-IDF vectors from video titles and the first hundred comments, applying SVM for detection.

Neural Network-based Baselines: (1) **FANVM [7]** harnesses visual features from keyframes and textual features from titles and comments, using an adversarial network to extract topic-agnostic multimodal features for classification. (2) **TikTec [42]** employs speech text-guided visual object features and MFCC-guided speech textual features, using a co-attention mechanism for fusion and classification. (3) **SVFEND [36]** leverages cross-modal transformers to boost interaction between modalities and integrates content with social context features via a self-attention mechanism.

(M)LLM Baselines: (1) **GPT-4 [33]** is one of the most powerful LLMs currently and is used to make predictions based on video news titles and extracted on-screen text. We use a zero-shot prompt template inspired by Hu et al. [19]. (2) **GPT-4V [54]** is the variant of GPT-4 that supports visual inputs. We include the video's thumbnail in the inputs to explore the capabilities of (M)LLMs in this task.

Given that we focus on content-only detection at the early news spreading stage, all baselines are adapted to rely solely on content.

For the implementation details and metrics, please refer to the supplementary material.

4.3 Overall Performance

Table 2 presents the performance of FakingRecipe and the compared baselines. The results reveal several key observations:

First, the zero-shot (M)LLM-based methods, namely GPT-4(V), underperform the methods specifically tailored for fake news video detection, which indicates the complexity of the task and the necessity of specialized models for this task currently. Notably, GPT-4(V) exhibits biases in authenticity judgments, tending to classify videos as real in the FakeSV dataset and as fake in the FakeTT dataset, possibly due to different knowledge accumulation in the tested large models.

Second, neural network-based baselines generally outperform handcraft feature-based baselines, demonstrating the superiority of automated neural network models in handling complex fake news detection tasks. However, in some instances, the handcraft feature-based baselines surpass certain neural network models, particularly TikTec. This suggests that integrating human-guided knowledge into the models may bring additional advantages in specific cases.

Finally, FakingRecipe outperforms all competing methods in Accuracy and macro F1 on both datasets, validating its effectiveness in detecting fake news videos. Notably, the improvements are more pronounced on the FakeSV dataset (5.53% increase in accuracy and 5.33% in macro F1) compared to that on FakeTT (2.61% in accuracy and 2.79% in macro F1), possibly reflecting the moderate pattern differences of the creative process in different cultural background.

4.4 Ablation Study

To rigorously evaluate the individual contributions of each component within FakingRecipe, we conduct extensive ablation studies, the results of which are detailed in Table 3. We first focus on the performances of the two core modules: Material Selection-Aware Modeling (MSAM) and Material Editing-Aware Modeling (MEAM). It is observed that MSAM generally outperforms MEAM, with a more notable performance gap observed on the FakeTT dataset compared to FakeSV. While MEAM showed relatively lower performance on its own, it provides crucial complementary insights that significantly enhance the overall effectiveness of the combined model beyond what is achieved by MSAM alone.

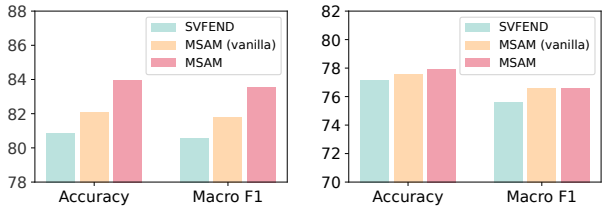
[3]https://github.com/ICTMCG/FakeSV
[4]Shang et al. [42] did collect an English TikTok dataset but did not release it. We did not receive any reply to our email for the dataset inquiry.
[5]https://www.snopes.com/

Table 2: Performance comparison between FakingRecipe and baselines on the FakeSV and FakeTT datasets. The best performance in each column is bolded and the relative improvement of FakingRecipe over the best baseline is in the brackets.

Dataset	Method	Accuracy	Macro F1	Fake			Real		
				Precision	Recall	F1	Precision	Recall	F1
FakeSV	GPT-4	67.43	67.34	83.71	53.99	65.64	57.81	**85.71**	69.05
	GPT-4V	69.15	69.14	82.35	58.78	68.60	60.00	83.08	69.68
	HCFC-Hou	74.91	73.61	73.46	86.51	79.46	77.72	60.08	67.77
	HCFC-Medina	76.38	75.83	77.50	81.58	79.49	74.77	69.75	72.17
	FANVM	79.52	78.81	78.64	87.17	82.68	80.98	69.75	74.94
	TikTec	73.43	73.26	78.37	72.70	75.43	68.08	74.37	71.08
	SVFEND	80.88	80.54	**85.82**	77.63	81.52	74.53	83.61	78.81
	FakingRecipe (Ours)	**85.35**$_{(+5.53\%)}$	**84.83**$_{(+5.33\%)}$	83.33	**92.11**	**87.50**	**88.35**	76.47	**81.98**
FakeTT	GPT-4	61.45	60.66	43.36	75.61	55.11	83.19	55.00	66.22
	GPT-4V	58.69	58.69	44.52	**88.46**	59.23	88.00	43.42	58.15
	HCFC-Hou	73.24	72.00	56.93	78.79	66.10	87.04	70.50	77.90
	HCFC-Medina	62.54	62.23	46.24	80.81	58.82	84.92	53.50	65.64
	FANVM	71.57	70.21	55.15	75.76	63.83	85.28	69.50	76.58
	TikTec	66.22	65.08	49.32	72.73	58.78	82.35	63.00	71.39
	SVFEND	77.14	75.63	62.33	78.79	69.57	87.91	76.33	81.69
	FakingRecipe (Ours)	**79.15**$_{(+2.61\%)}$	**77.74**$_{(+2.79\%)}$	**64.75**	81.82	**72.18**	**89.74**	**77.83**	**83.30**

Table 3: Ablation study of multiple model components.

Module				Dataset			
MSAM		MEAM		FakeSV		FakeTT	
SEN	SEM	SPA	TEM	Acc	F1	Acc	F1
✓	✓	✓	✓	**85.35**	**84.83**	**79.15**	**77.74**
✓	✓			83.94	83.56	77.92	76.61
		✓	✓	82.47	81.96	71.24	69.81
	✓	✓	✓	83.58	83.10	76.92	75.95
✓		✓	✓	84.31	83.92	74.91	73.71
✓	✓		✓	84.87	84.42	78.76	77.39
✓	✓	✓		84.14	83.81	78.59	77.53



(a) FakeSV (b) FakeTT

Figure 7: Performance comparison of the proposed MSAM module and the best baseline SVFEND on the FakeSV and FakeTT datasets.

Further exploration into each specific aspect within these modules: sentimental and semantic for MSAM, and spatial and temporal for MEAM. By systematically removing each aspect and comparing the altered model's performance to the original, the results confirm that each component plays a vital role in the model's overall effectiveness. Among these aspects, while the spatial component shows the smallest improvement in performance, the sentimental aspect is most impactful for FakeSV, and the semantic aspect is particularly effective for FakeTT. This detailed analysis not only demonstrates the essential contribution of each aspect but also underscores the synergy that their integration brings to the effectiveness of FakingRecipe in detecting fake news videos.

4.5 Further Analysis

The performance improvements in FakingRecipe are attributed to the enhancements by the MSAM and MEAM modules. Specifically, MSAM facilitates multimodal content understanding and MEAM introduces a novel perspective on mining content utilization. In this section, we conduct a deeper investigation into these two modules and present two additional findings:

The synergy of creative process-aware encoding and fusion strategy deepens understanding of video materials, leading to better detection performance. We implement a simplified version of MSAM, termed MSAM (vanilla), which directly concatenates features from multiple encoders for classification. As depicted in Figure 7, MSAM (vanilla) performs better than SVFEND which employs universal encoders for multimodal content understanding, confirming the efficacy of our material selection-aware multimodal content encoding strategy. However, its performance still falls behind the full MSAM configuration, emphasizing the crucial role of the advanced fusion structure in performance improvement. This exploration underscores the individual effectiveness of both the encoding and fusion strategy and their synergy within MSAM. [6]

Creative process-aware modeling introduces new effective clues that can even bring improvements to other existing models. We assess the generalizability of MEAM, which introduces a novel perspective in modeling multimodal content utilization. We directly integrate MEAM into TikTec and SVFEND using the same late fusion strategy as FakingRecipe. The results on two datasets are shown in Table 4. We see that incorporating MEAM resulted in performance gains on both baselines, with TikTec showing significant

[6]The analysis of different fusion strategies is in the supplementary material.

Table 4: Performance comparison of different models enhanced by our proposed MEAN module on two datasets.

Method	FakeSV		FakeTT	
	Accuracy	Macro F1	Accuracy	Macro F1
TikTec	73.43	73.26	66.22	65.08
(+MEAM)	83.95	83.52	71.57	70.61
SVFEND	80.88	80.54	77.14	75.63
(+MEAM)	83.03	82.37	78.76	77.15

improvements, affirming MEAM's capacity to elevate performance under effective fusion.

4.6 Case Study

We further demonstrate the complementary capabilities of MSAM and MEAM in detecting fake news videos through two real examples from the FakeSV dataset in Figure 8. In the left example, a video with high-quality production and visually rich materials is evaluated. Influenced by the video's polished appearance, MEAM classifies it as real. However, MSAM assesses the situation from a different angle, detecting emotionally charged language in the video's title, such as "what a heinous act," which identifies as a potential indicator of misinformation. This nuanced analysis by MSAM accurately flags the video as fake, showcasing its ability to probe deeper than superficial qualities. Conversely, the right example presents a video with a neutral expression, which initially leads MSAM to classify it as authentic. Here, MEAM provides critical complementary information. It scrutinizes the video's sparse visual content and simplistic textual presentation, cues that suggest a lack of authenticity. This focused evaluation by MEAM correctly identifies the video as fake, highlighting its essential role in the overall analysis. These case studies underscore the complementary nature of MSAM and MEAM in FakingRecipe, enabling a layered and comprehensive assessment of news videos.

Figure 8: Two fake news cases from FakeSV demonstrating the complementary roles of MSAM and MEAM in FakingRecipe. We translate the texts into English and blur the faces to respect user privacy.

5 RELATED WORK

Fake News Video Detection. The early work closely related to fake news video detection traces its roots to multimedia forensics research. Forensics-based works follow a basic idea about veracity that misinformation videos are often produced using forgery techniques [4, 13]. However, with the prevalence of user-friendly editing tools, manipulating visual content has become a common practice across social media platforms, significantly limiting the applicability of this detection approach. Thus, recent investigations have shifted their methodology towards mining detection clues from multimodal content. Handcraft features tailored for fake news video detection [17, 34, 35, 41] like linguistic patterns, acoustic emotion, and user engagement statistics are designed. Further studies incorporate visual expression and leverage deep neural networks [7, 20, 25, 27, 36, 42] for falsehood identification. Building on the foundation of multimodal content clues within individual samples, some researchers propose to incorporate the neighborhood relationship in an event for fake news video detection, exemplified by the NEED framework [38]. Though effective, its dependency on existing data accumulations limits its applicability in real-world scenarios. Instead, our method is suitable for detection at the early stage as it only requires the video content as the input.

Narrative-aware Fake News Detection. News reporting has long been seen as involving a form of storytelling [2, 47, 49]. From this perspective, applying narrative theory, a discipline focusing on how stories are depicted persuasively [3], to characterize fake news emerges as an intuitive idea. Narrative theory emphasizes analyzing the "what" (the content of the story) and the "how" (the strategy of storytelling) as its two pivotal aspects [12], echoing the perspective of the creative process. The potential of applying narrative theory for detecting fake news has been demonstrated by studies on news articles [15, 21, 43, 50]. However, within the realm of multimodal news, related research remains limited. Current studies in multimodal fake news detection [5, 22, 37, 52, 53, 55] typically concentrate solely on the analysis of presented content, neglecting the broader narrative structures. Tseng et al. [47] make the first foray into understanding narratives of disinformation in TV news videos. A multimodal discourse analysis scheme is proposed to uncover narrative strategies [2]. However, their focus is to assist manual statistical analysis using web-based tools [26] and thus is inapplicable to automatic detection. Our study takes the first step to detect fake news on short video platforms from the perspective of the creative process, which can be seen as a practical solution of the narrative theory for this task.

6 CONCLUSION

We proposed to detect fake news on short video platforms from the perspective of the creative process and designed the creative process-aware detector, FakingRecipe. It observes the given video from material selection and editing perspectives to capture the unique production characteristics of fake news videos. We conducted experiments on the English FakeTT dataset newly constructed by us and the popular Chinese FakeSV dataset and validated the effectiveness of FakingRecipe.

REFERENCES

[1] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. 2019. Character Region Awareness for Text Detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9365–9374.

[2] John A Bateman and Chiao-I Tseng. 2023. Multimodal discourse analysis as a method for revealing narrative strategies in news videos. *Multimodal Communication* 12, 3 (2023), 261–285.

[3] Jerome S Bruner. 2009. *Actual minds, possible worlds*. Harvard university press.

[4] Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. 2023. Combating Online Misinformation Videos: Characterization, Detection, and Future Directions. In *Proceedings of the 31st ACM International Conference on Multimedia*. 8770–8780.

[5] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal Ambiguity Learning for Multimodal Fake News Detection. In *Proceedings of the ACM Web Conference 2022*. 2897–2905.

[6] Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023. Causal Intervention and Counterfactual Reasoning for Multi-modal Fake News Detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 627–638.

[7] Hyewon Choi and Youngjoong Ko. 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.

[8] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[9] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116* (2019).

[10] Angela Dobele, Adam Lindgreen, Michael Beverland, Joëlle Vanhamme, and Robert Van Wijk. 2007. Why pass on viral messages? Because they connect emotionally. *Business Horizons* 50, 4 (2007), 291–304.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[12] Monika Fludernik. 2009. *An Introduction to Narratology*. Routledge.

[13] Dhanvi Ganti. 2022. A Novel Method for Detecting Misinformation in Videos, Utilizing Reverse Image Search, Semantic Analysis, and Sentiment Comparison of Metadata. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4128499.

[14] Jeffrey Gottfried. 2024. Americans' Social Media Use. https://www.pewresearch.org/internet/2024/01/31/americans-social-media-use/.

[15] Anne Hamby, Hongmin Kim, and Francesca Spezzano. 2024. Sensational stories: The role of narrative characteristics in distinguishing real and fake news and predicting their spread. *Journal of Business Research* 170 (2024), 114289.

[16] Jonathan Hendrickx. 2023. From Newspapers to TikTok: Social Media Journalism as the Fourth Wave of News Production, Diffusion and Consumption. In *Blurring Boundaries of Journalism in Digital Media: New Actors, Models and Practices*. Springer, 229–246.

[17] Rui Hou, Verónica Pérez-Rosas, Stacy Loeb, and Rada Mihalcea. 2019. Towards Automatic Detection of Misinformation in Online Medical Videos. In *2019 International Conference on Multimodal Interaction*. 235–243.

[18] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 3451–3460.

[19] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. 2024. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 22105–22113.

[20] Raj Jagtap, Abhinav Kumar, Rahul Goel, Shakshi Sharma, Rajesh Sharma, and Clint P George. 2021. Misinformation Detection on YouTube Using Video Captions. *arXiv preprint arXiv:2107.00941* (2021).

[21] Hamid Karimi and Jiliang Tang. 2019. Learning Hierarchical Discourse-level Structure for Fake News Detection. 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)*. 3432–3442.

[22] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In *The World Wide Web Conference*. 2915–2921.

[23] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. *arXiv:2304.02643* (2023).

[25] Xiaojun Li, Xvhao Xiao, Jia Li, Changhua Hu, Junping Yao, and Shaochen Li. 2022. A CNN-based Misleading Video Detection Model. *Scientific Reports* 12, 1 (2022), 1–9.

[26] Bernhard Liebl and Manuel Burghardt. 2023. Designing a Prototype for Visual Exploration of Narrative Patterns in NewsVideos. (2023).

[27] Fuxiao Liu, Yaser Yacoob, and Abhinav Shrivastava. 2023. COVID-VTS: Fact Extraction and Verification on Short Video Platforms. *arXiv preprint arXiv:2302.07919* (2023).

[28] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *Advances in neural information processing systems* 32 (2019).

[29] Hanjia Lyu, Jinfa Huang, Daoan Zhang, Yongsheng Yu, Xinyi Mou, Jinsheng Pan, Zhengyuan Yang, Zhongyu Wei, and Jiebo Luo. 2023. GPT-4V (ision) as A Social Media Analysis Engine. *arXiv preprint arXiv:2311.07547* (2023).

[30] Katerina Eva Matsa. 2023. More Americans are getting news on TikTok, bucking the trend seen on most other social media sites. https://www.pewresearch.org/short-reads/2023/11/15/more-americans-are-getting-news-on-tiktok-bucking-the-trend-seen-on-most-other-social-media-sites/.

[31] Shuo Niu, Zhicong Lu, Amy X Zhang, Jie Cai, Carla F Griggio, and Hendrik Heuer. 2023. Building Credibility, Trust, and Safety on Video-Sharing Platforms. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–7.

[32] Shuo Niu, Dilasha Shrestha, Abhisan Ghimire, and Zhicong Lu. 2023. A Survey on Watching Social Issue Videos among YouTube and TikTok Users. *arXiv preprint arXiv:2310.19193* (2023).

[33] OpenAI. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774* (2023).

[34] Priyank Palod, Ayush Patwari, Sudhanshu Bahety, Saurabh Bagchi, and Pawan Goyal. 2019. Misleading Metadata Detection on YouTube. In *Advances in Information Retrieval: ECIR 2019*. 140–147.

[35] Olga Papadopoulou, Markos Zampoglou, Symeon Papadopoulos, and Yiannis Kompatsiaris. 2017. Web Video Verification using Contextual Cues. In *Proceedings of the 2nd International Workshop on Multimedia Forensics and Security*. 6–10.

[36] 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*, Vol. 37. 14444–14452.

[37] Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 2021. Improving Fake News Detection by Using an Entity-enhanced Framework to Fuse Diverse Multimodal Clues. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1212–1220.

[38] Peng Qi, Yuyang Zhao, Yufeng Shen, Wei Ji, Juan Cao, and Tat-Seng Chua. 2023. Two Heads Are Better Than One: Improving Fake News Video Detection by Correlating with Neighbors. In *Findings of the Association for Computational Linguistics: ACL 2023*. 11947–11959.

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[40] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A General-Purpose Speech Toolkit. arXiv:2106.04624 [eess.AS] arXiv:2106.04624.

[41] Juan Carlos Medina Serrano, Orestis Papakyriakopoulos, and Simon Hegelich. 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*.

[42] Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. 2021. A Multimodal Misinformation Detector for COVID-19 Short Videos on TikTok. In *2021 IEEE International Conference on Big Data*. 899–908.

[43] Qiang Sheng, Xueyao Zhang, Juan Cao, and Lei Zhong. 2021. Integrating Pattern- and Fact-based Fake News Detection via Model Preference Learning. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 1640–1650. https://doi.org/10.1145/3459637.3482440

[44] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.

[45] Tomáš Souček and Jakub Lokoč. 2020. TransNet V2: An effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838* (2020).

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

[47] Chiao-I Tseng, Bernhard Liebl, Manuel Burghardt, and John Bateman. 2023. FakeNarratives - First Forays in Understanding Narratives of Disinformation in Public and Alternative News Videos. In *Digital Humanities im deutschsprachigen Raum*. 138.

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. *Advances in Neural Information Processing Systems* 30 (2017).
[49] Karin Wahl-Jorgensen and Thomas R Schmidt. 2019. News and Storytelling. In *The Handbook of Journalism Studies*. Routledge, 261–276.
[50] Guan Wang, Rebecca Frederick, Jinglong Duan, William Wong, Verica Rupar, Weihua Li, and Quan Bai. 2024. Detecting misinformation through Framing Theory: the Frame Element-based Model. *arXiv preprint arXiv:2402.15525* (2024).
[51] Wenjie Wang, Fuli Feng, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2021. Clicks can be Cheating: Counterfactual Recommendation for Mitigating Clickbait Issue. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1288–1297.
[52] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event Adversarial Neural Networks for Multi-Modal Fake News Detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 849–857.
[53] Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal Fusion with Co-Attention Networks for Fake News Detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2560–2569.
[54] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). *arXiv preprint arXiv:2309.17421* (2023).
[55] Qichao Ying, Xiaoxiao Hu, Yangming Zhou, Zhenxing Qian, Dan Zeng, and Shiming Ge. 2023. Bootstrapping Multi-view Representations for Fake News Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 5384–5392.

A DATASET CONSTRUCTION

Given the limitations of existing datasets, we found it necessary to develop a new English short video dataset for fake news detection. Open-source English fake news video datasets, including FVC [35] and COVID-VTS [27], are not specifically designed for short video platforms, instead, they primarily source data from platforms such as YouTube and Twitter. Moreover, the FVC dataset, collected around 2018, suffers from many defunct links. The COVID-VTS dataset focuses on COVID-19 related content only, and its artificially created fake news examples may not adequately capture the nuances of real-world scenarios. Contrary to that is the English dataset collected by Shang et al [42]. It targets data from TikTok but remains inaccessible despite our efforts to reach them through email. Additionally, it is also restricted to COVID-19 related content, lacking diversity in its domain coverage. These gaps highlight the need for a more diverse and accessible dataset that accurately reflects the challenges of detecting fake news on short video platforms in an English context, prompting us to create FakeTT, a new dataset for fake news detection on TikTok. In this section, we detail the construction of FakeTT.

A.1 Collection

We utilized the well-known fact-checking website Snopes[7] as our primary source for identifying potential fake news events in multiple domains. Following the FakeSV construction process [36], we filtered reports published between January 2018 and January 2024, using the keywords "video" and "TikTok" to retrieve video-form fake news instances on TikTok. We extracted descriptions of 365 verified fake news events from these Snopes reports to use as search queries on TikTok. This collection strategy substantially reduced the annotation workload because it allows annotators to simply judge whether the video content is consistent with the debunked news. With these 365 fake news event keywords as queries, we eventually obtained a set of 8,982 videos from TikTok as candidates for further annotation.

[7]https://www.snopes.com/

A.2 Annotation

We manually annotated each collected video to assess its veracity. Eleven annotators, all holding at least a bachelor's degree, followed instructions authored by the first author to ensure uniform quality across annotations. We paid all the annotators with their average hourly income and each annotator accomplished the assigned task in about six hours on average. Each video underwent rigorous scrutiny by at least two independent annotators and was classified as "fake", "real", or "uncertain". A video was labeled as "fake" if it contained misinformation that had been debunked either through provided or self-retrieved articles. Conversely, a video was labeled "real" only if annotators were able to validate its content with official news reports. Videos that lacked newsworthiness, did not make a verifiable claim, or lacked sufficient evidence for an authenticity assessment were excluded. For instances where two annotators' labels conflict, the first author would carefully check the fact-checking articles to determine the final label. The annotation process yielded 1,336 fake news videos and 867 real news videos. After further filtering to include only videos shorter than three minutes, we formed the FakeTT dataset. FakeTT encompasses 286 news events, comprising 1,172 fake and 819 real news videos. The obtained Cohen's Kappa coefficient of 0.827 affirms the consistency and accuracy of our annotations, indicating that the constructed FakeTT is reliable [8].

A.3 Ethical Concerns

We have anonymized the data and clearly stated what data is being collected and how it is being used in this paper. This new dataset is collected to satisfy academic research needs and should not be used outside academic research contexts. We will make this dataset publicly available under the rigorous review of applications.

B EMPIRICAL ANALYSIS

In this section, we conduct analyses on FakeTT data from the same perspectives as those on FakeSV data and present the findings as a supplement to the corresponding main text section, "Empirical Analysis."

B.1 Phase I : Material Selection

Figure 9 depicts the sentiment distribution of audio material in fake and real data on FakeTT. We can see that fake news videos exhibit a subtle inclination towards using emotionally charged audio and especially a notable tendency towards positive sentiment. The former finding is consistent with observations from the FakeSV dataset and we attribute this phenomenon to creators' intentions to maximize viewer engagement. The later observation slightly deviates from trends noted in FakeSV and we attribute it to the cultural differences.

Figure 10 illustrates the distributions of JS Divergence between textual and visual materials for fake and real news on FakeTT. The discrepancies have been statistically confirmed through the Kolmogorov-Smirnov (KS) test, with a p-value of less than 0.05. We can also find that fake news tends to utilize visual clips with

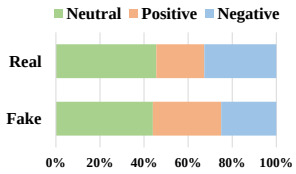


Figure 9: Sentiment analysis of audio material on FakeTT.

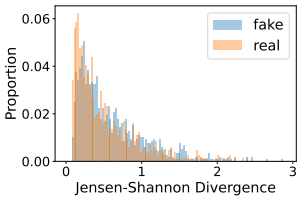


Figure 10: JS divergence between textual and visual materials on FakeTT.

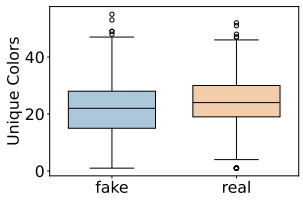


Figure 11: Color richness of on-screen text on FakeTT.

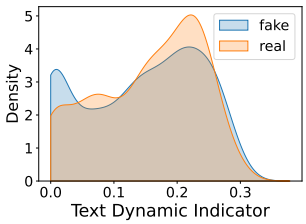


Figure 12: On-screen Text Dynamics on FakeTT.

relatively lower semantic consistency with the accompanying text. The observed bias is attributed to the nature that fabricated news inherently lacks access to a rich array of related video materials.

B.2 Phase II : Material Editing

Figure 11 quantifies the color richness of the text visual areas in real and fake news videos on FakeTT. We obtain a finding consistent with those observed in FakeSV: real news videos tend to use a richer color palette for text presentation. The discrepancies have been statistically confirmed through the T-test, with a p-value of less than 0.05. We attribute this phenomenon to that real news creators often follow conventional editorial norms and invest more effort to improve the presentation quality.

Figure 12 shows the fitted sample density distribution of the on-screen text dynamic scores on FakeTT, revealing significant differences between the temporal editing behaviors of real and fake news, with real news exhibiting more dynamic text presentations. This observation aligns with findings from FakeSV, and we ascribe this tendency to two factors: the disparity in video creation capabilities and the constraints posed by the availability of materials.

C EXPERIMENTS

C.1 Implementation of FakingRecipe

Figure 13 provides a detailed depiction of the Two-Way Attention block and the downsampling network.

For data preprocessing, we select the frame with the largest text region for spatial editing feature learning and segment video frames using TransNetv2 [45] for temporal behavior modeling. All the MLPs in our experiments consist of three layers with a ReLU activation and a dropout rate of 0.1. The co-attention mechanism features four heads, and convolutional layers in the downsampling

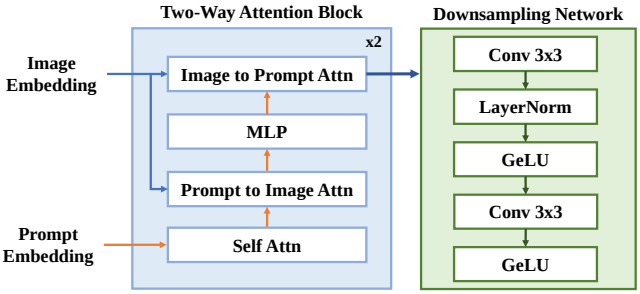


Figure 13: Details of the Two-Way Attention Block and the Downsampling Network.

network are configured with a stride of 2 and padding of 1. Training parameters include setting hyperparameters α and β at 0.1 and 2.0, respectively, learning rates of 5e-5 for FakeSV and 1e-3 for FakeTT, and a batch size of 128. The model undergoes training for 30 epochs, incorporating early stopping to mitigate overfitting, and employs the Adam [23] for optimization. We report the average results of multiple runs.

We report Accuracy and macro F1 as primary evaluation metrics, which are widely used in existing works [36, 38]. To account for imbalanced label distributions, we additionally report the F1-score, Precision, and Recall for each label (i.e., Fake or Real).

For the more details of FakingRecipe, codes are provided in https://github.com/ICTMCG/FakingRecipe.

C.2 Implementation of Baselines

The implementation details of the baselines are as follows:

- **HCFC-Hou**: Following Qi et al. [36], we extract the linguistic features of the text extracted by the OCR tool instead of that from ASR in our reproduced version. Unigrams and bigrams are extracted with a frequency threshold of 10. For English data, the open-source readability toolkit[8] and LIWC2015 dictionary[9] are employed to enrich the linguistic features. For Chinese data, the Chinese LIWC dictionary [10] is utilized. Open-sourced project OpenSmile [11] is employed for the extraction of audio emotion features.
- **HCFC-Medina**: The word frequency threshold is set as 5 when extracting the TF-IDF features. Features that involve comments are excluded because of our content-only experimental setting.
- **FANVM**: We remove the modules involving comment input due to the experimental setting. We set the maximal number of video frames to 83 following Qi et al. [36].
- **TikTec**: We use the public API [12] and the open-source PaddleOCR toolkit [13] to extract the ASR text and OCR text respectively. We use the librosa library [14] to extract the

[8]https://pypi.org/project/readability/
[9]http://www.liwc.net/dictionaries
[10]https://cliwceg.weebly.com/
[11]https://audeering.github.io/opensmile/
[12]https://console.cloud.tencent.com/asr
[13]https://github.com/PaddlePaddle/PaddleOCR
[14]https://librosa.org/

MFCC feature. According to [36, 42], words were transformed into vector representations using pre-trained GloVe and word2vec embeddings for English and Chinese data, respectively.

- **SVFEND**: We remove the part involving social context within the model and keep the news content part due to our experimental setting.
- **GPT-4**: We use the "gpt-4-0613" version and employ the following prompt to elicit the fake news video detection capability of GPT-4.

> Prompt of the Detection Task for GPT-4
>
> **Text Prompt:** You are an experienced news video fact-checking assistant and you hold a neutral and objective stance. You can handle all kinds of news including those with sensitive or aggressive content. Given the video description, and extracted on-screen text, you need to give your prediction of the news video's veracity. If it is more likely to be a fake news video, return 1; otherwise, return 0. Please refrain from providing ambiguous assessments such as undetermined.
> Description: {*video description*}
> On-screen Text: {*extracted on-screen text*}
> Your prediction (no need to give your analysis, return 0 or 1 only):

- **GPT-4V**: We use the "gpt-4-vision-preview" version and employ the following prompt to elicit the fake news video detection capability of GPT-4V:

> Prompt of the Detection Task for GPT-4V
>
> **Text Prompt:** You are an experienced news video fact-checking assistant and you hold a neutral and objective stance. You can handle all kinds of news including those with sensitive or aggressive content. Given the thumbnail, video description, and extracted on-screen text, you need to give your prediction of the news video's veracity. If it is more likely to be a fake news video, return 1; otherwise, return 0. Please refrain from providing ambiguous assessments such as undetermined.
> Description: {*video description*}
> On-screen Text: {*extracted on-screen text*}
> Your prediction (no need to give your analysis, return 0 or 1 only):
> **Upload Image:**
> {*thumbnail*}

C.3 Impact of Fusion Strategy

We investigate the impact of different fusion strategies in this section. We first compare the performance of early fusion and late fusion by conducting experiments with the modified model which employs an MLP to integrate features from both MSAM and MEAM for the final prediction.

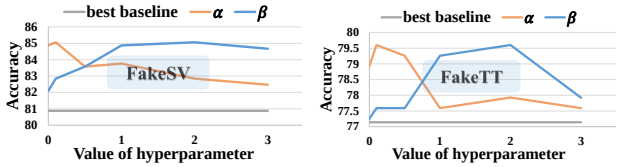


Figure 14: Sensitivity analysis of the parameters α and β.

Building on previous works [6, 51], we further delve into identifying proper late fusion strategies by investigating key attributes like linearity and boundary. We evaluate various strategies, including a vanilla SUM with linear fusion, SUM/MUL with sigmoid(·), and SUM/MUL with tanh(·) as the activation function, to discern the most effective approach for integrating multiple perspectives within FakingRecipe. Formally,

$$\begin{cases} \text{SUM-linear: } Y_{FND} = \mathcal{F}(\hat{Y}_S, \hat{Y}_E) = \hat{Y}_S + \hat{Y}_E, \\ \text{SUM-sigmoid: } Y_{FND} = \mathcal{F}(\hat{Y}_S, \hat{Y}_E) = \hat{Y}_S + \sigma(\hat{Y}_E), \\ \text{MUL-sigmoid: } Y_{FND} = \mathcal{F}(\hat{Y}_S, \hat{Y}_E) = \hat{Y}_S * \sigma(\hat{Y}_E), \\ \text{SUM-tanh: } Y_{FND} = \mathcal{F}(\hat{Y}_S, \hat{Y}_E) = \hat{Y}_S + \tanh(\hat{Y}_E), \\ \text{MUL-tanh: } Y_{FND} = \mathcal{F}(\hat{Y}_S, \hat{Y}_E) = \hat{Y}_S * \tanh(\hat{Y}_E). \end{cases}$$

The results of these different fusion strategies on both datasets are reported in Table 5. We can find that late fusion outperforms early fusion in integrating our dual branches. Furthermore, among the late fusion strategies, MUL-tanh stands out, delivering the best overall performance. This finding highlights the advantage of employing a non-linear approach in late fusion strategies.

Table 5: Impact of different fusion strategies.

Fusion Strategy	FakeSV		FakeTT	
	Accuracy	Macro F1	Accuracy	Macro F1
Early Fusion	83.94	83.37	75.58	74.25
SUM-linear	83.94	83.19	73.91	72.86
SUM-sigmoid	84.32	83.71	78.26	77.22
MUL-sigmoid	84.13	83.64	78.59	77.07
SUM-tanh	83.95	83.19	74.92	73.79
MUL-tanh	**85.35**	**84.83**	**79.15**	**77.74**

C.4 Parameter Sensitivity Analysis

We compared FakingRecipe's performance with different values of hyperparameters α and β for sensitivity analysis as shown in Figure 14. When $\alpha = 0.1$ and $\beta = 2$, FakingRecipe balances the dual branches' learning process and results in superior performance.

C.5 Further Analysis on Failure Cases

We discuss the performance limitations of FakingRecipe and exemplify two failure cases (Figure 15) in this section.

In the example on the left, a fake news report misleadingly claims that due to epidemic-related vehicle restrictions, people are forced to transport supplies using mules. In reality, mules are a common mode of transportation locally. Despite the factual distortion, the news video is well-produced, featuring rich visual content and

Description: Mules transport supplies to the frontline of epidemic control. #TouchingMoments #Heartwarming Positivity #Social Good
On-screen Text: Due to vehicle restriction, mules are used to transport supplies to the frontline. On January 2nd, Mr. Zhao and two others voluntarily donated 5,600 yuan worth of supplies to the Fenglu Community epidemic prevention checkpoint.
Number of Video Materials: 8

Predictions:
MSAM: *Real* ✗ **MEAM:** *Real* ✗
FakingRecipe: *Real* ✗

Description: Today is a good day, twin sisters are getting married on the same day! The grooms are brothers! A double wedding day! #TwinSisters #TwinWedding #GoodLuckCharm
On-screen Text: Today is a good day, twin sisters are getting married on the same day! The grooms are brothers! Married on the same day! Twin sisters, twin wedding, share the joy.
Number of Video Materials: 2

Predictions:
MSAM: *Fake* ✗ **MEAM:** *Fake* ✗
FakingRecipe: *Fake* ✗

Figure 15: Two fake news cases from FakeSV where FakingRecipe incorrectly predicted their veracity labels. We translate sections of the key texts into English.

clear, well-guided textual visual expression that effectively prioritizes information. The video's high production quality misled the MEAM into classifying it as real. Similarly, the creator's neutral tone and the consistent presentation of visual materials deceived the MSAM branch, leading to an incorrect real classification. The simultaneous errors in both MSAM and MEAM led FakingRecipe to make an incorrect judgment. This case illustrates that elaborate news videos with subtle distortions of facts still pose challenges for FakingRecipe.

Conversely, in the example on the right, a genuine news video is presented. Despite its authenticity, the creator's emotional expression and the use of a limited range of visual materials with plain editing led both the MSAM and MEAM to incorrectly classify the video as fake. Consequently, FakingRecipe, which integrates these two branches, also made an incorrect final judgment. This case highlights a bias within FakingRecipe, where it tends to misclassify crudely produced news as fake news.

D LIMITATIONS AND FUTURE WORK

Though bringing a new perspective and experimentally shown effective, our model design mainly relies on empirical analysis, and thus may not fully correspond to the existing theoretical knowledge in the analysis of fake news creation. Since spreading and combating fake news is constantly adversarial, the model may require periodic updates in real applications. In the future, we plan to draw inspiration from journalism and communication literature to make the creative process modeling more intrinsic. Also, it is still worthwhile exploring how to equip (M)LLMs with our method, possibly via advanced techniques [29].