Unveiling Opinion Evolution via Prompting and Diffusion for Short Video Fake News Detection

Linlin Zong[1], **Jiahui Zhou**[1], **Wenmin Lin**[1], **Xinyue Liu**[1], **Xianchao Zhang**[1], **Bo Xu**[2*]
[1]Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province,
School of Software, Dalian University of Technology, Dalian 116620, China
[2]School of Computer Science and Technology, Dalian University of Technology, Dalian 116620, China
{llzong, xyliu, xczhang, xubo}@dlut.edu.cn, {zjhjixiang, wmlin}@mail.dlut.edu.cn

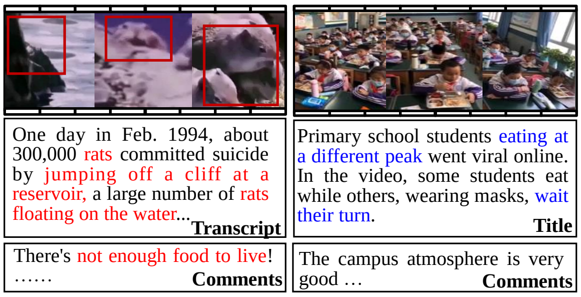
Abstract

Short video fake news detection is crucial for combating the spread of misinformation. Current detection methods tend to aggregate features from individual modalities into multimodal features, overlooking the implicit opinions and the evolving nature of opinions across modalities. In this paper, we mine implicit opinions within short video news and promote the evolution of both explicit and implicit opinions across all modalities. Specifically, we design a prompt template to mine implicit opinions regarding the credibility of news from the textual component of videos. Additionally, we employ a diffusion model that encourages the interplay among diverse modal opinions, including those extracted through our implicit opinion prompts. Experimental results on a publicly available dataset for short video fake news detection demonstrate the superiority of our model over state-of-the-art methods.

1 Introduction

Fake news has the potential to mislead the public, influencing their decisions and behaviors. Identifying and stopping the spread of fake news promptly is vital for maintaining societal stability. Initially, fake news primarily relied on textual content (Rashkin et al., 2017). However, with the rise of online video platforms, fake news dissemination expanded to include videos (Choi and Ko, 2021; Qi et al., 2023a,b). Short video news comprises multiple elements such as images, videos, audio, social content, comments, and various media sources. Fake news creators can employ various manipulative tactics on different dimensions of video content, thereby complicating the task of assessing the authenticity of a news story.

The modality features of fake news contain various aspects of information. In order to distinguish information that is not relevant to fake news judgments, we use opinion to refer to the judgement related to fake news contained in each modality, which is the result obtained from the modality features. Currently, approaches for identifying fake news in short videos have shown promising results by skillfully combining various modality features (Choi and Ko, 2021; Qi et al., 2023a,b). However, these methods mainly focus on incorporating single-modal features into multimodal ones for authenticity assessment, neglecting the implicit opinion on the authenticity of news and the dynamic evolution of opinions across different modalities.



One day in Feb. 1994, about 300,000 rats committed suicide by jumping off a cliff at a reservoir, a large number of rats floating on the water... **Transcript**

There's not enough food to live! …… **Comments**

Primary school students eating at a different peak went viral online. In the video, some students eat while others, wearing masks, wait their turn. **Title**

The campus atmosphere is very good … **Comments**

(a) Short video fake news: Case A (b) Short video fake news: Case B

Figure 1: Examples of short video fake news. (a) Short video fake news with consistent content across multiple modalities. (b) Short video fake news with partial textual misleading content.

Implicit opinions are instrumental in identifying fake news. This is because, in fake news, the creators often conceal their true opinions and mislead readers through seemingly objective statements. For instance, consider short video news depicted in Figure 1(a), the video, transcript, and comments are all discussing *the event of rats collectively committing suicide by jumping off cliffs*. This apparent consistency across multiple modalities may lead the public to believe the news is real. However, it's a deliberate fabrication by the publisher for sensationalism. Such deceptive content isn't easily discerned as fake at first glance; instead, its authen-

[*]Corresponding author

Findings of the Association for Computational Linguistics: ACL 2024, pages 10817–10826
August 11-16, 2024 ©2024 Association for Computational Linguistics

ticity requires assessing the implied opinions in the news. Therefore, identifying implicit opinions is of great significance for assessing news authenticity.

Opinion evolution is crucial since fake news often contains localized deceptive elements in only a few modalities. Opinions regarding news authenticity vary across modalities, and individual frames in videos or specific sentences in texts may not significantly influence other single-modal opinions, potentially leading to erroneous authenticity judgments. For example, as illustrated in Figure 1(b), a short video news is fake due to tampering in the blue section of the title, yet opinions from other modalities suggest the news is real. Integrating such diverse modalities might erroneously conclude the news is real, disregarding the presence of fake information. To accurately evaluate the authenticity of such news, there is a need for mutual reinforcement among different modal opinions to comprehensively diffuse deceptive elements. Therefore, uncovering opinion evolution is essential for detecting fake news in short videos.

To address the aforementioned issues, we analyze the implicit opinions embedded within short video news and facilitate the evolution of both explicit and implicit opinions across all modalities within the news. We propose the model OpEvFake, unveiling the Opinion Evolution via prompting and diffusion for short video Fake news detection. Firstly, to conduct a thorough analysis of news authenticity, we create a customized chain-of-thought prompt template that is specifically designed for mining implicit opinions regarding the credibility of news from their textual content. Subsequently, recognizing that multimodal interaction involves a continuous updating and summarizing of opinions by each modality according to specific rules, we design a diffusion model to facilitate interaction among the various modal opinions including the generated implicit opinion prompts. Finally, we employ the evolved opinions for the purpose of classification. The contributions of this paper can be summarized as follows:

- We introduce implicit opinion learning and opinion evolution in the task of short video fake news detection. This introduces new perspectives and deeper insights to authenticate short video news.
- We devise an implicit opinion prompting template and use the diffusion model to achieve multimodal opinion interaction, which strengthens the opinion in each modality.
- We conduct experiments on a publicly available dataset for short video fake news detection. Experimental results demonstrate a significant performance improvement in fake news detection tasks.

2 Related Work

2.1 Multimodal Fake News Detection

Early multimodal fake news detection is designed for text and images. These methods typically identify fake news from two perspectives: modality interaction and modality similarity. For example, HMCAN (Qian et al., 2021) used a hierarchical multimodal contextual attention network to handle the interaction of inter-modality and intra-modality. MCAN (Wu et al., 2021b) used co-attention networks to better fuse textual and visual features. CAFE (Chen et al., 2022) used a cross-modal fusion module to capture the cross-modal correlations, then aggregated single-modal features and cross-modal correlations. MMICF (Zeng et al., 2023) divided multimodal inconsistency into local and global inconsistency. Although these methods have achieved good results in text-image fake news detection, they cannot be directly applied to short video fake news detection due to the different ways of fabricating short video fake news and text-image fake news.

With the rise of online video platforms, the dissemination of fake news has expanded from text and images to videos. Multimodal fake news detection methods for three or more modalities have been proposed. According to Bu et al. (Bu et al., 2023), most of the misinformation video detection methods use concatenation (Choi and Ko, 2021), attention (Shang et al., 2021; Qi et al., 2023a) and multitask (Choi and Ko, 2021) for clue integration. For example, Choi et al. (Choi and Ko, 2021) proposed a topic-agnostic fake news detection model based on adversarial learning and topic modeling. They employed linear combination to combine feature vectors of comments, title and video. Qi et al. (Qi et al., 2023a) used two cross-modal transformers to mine the relationship between text and audio, text and video, respectively. Meanwhile, they proposed a short video fake news detection dataset FakeSV containing a variety of information. Subsequently, Qi et al. (Qi et al., 2023b) designed the NEED framework to identify fake news by the

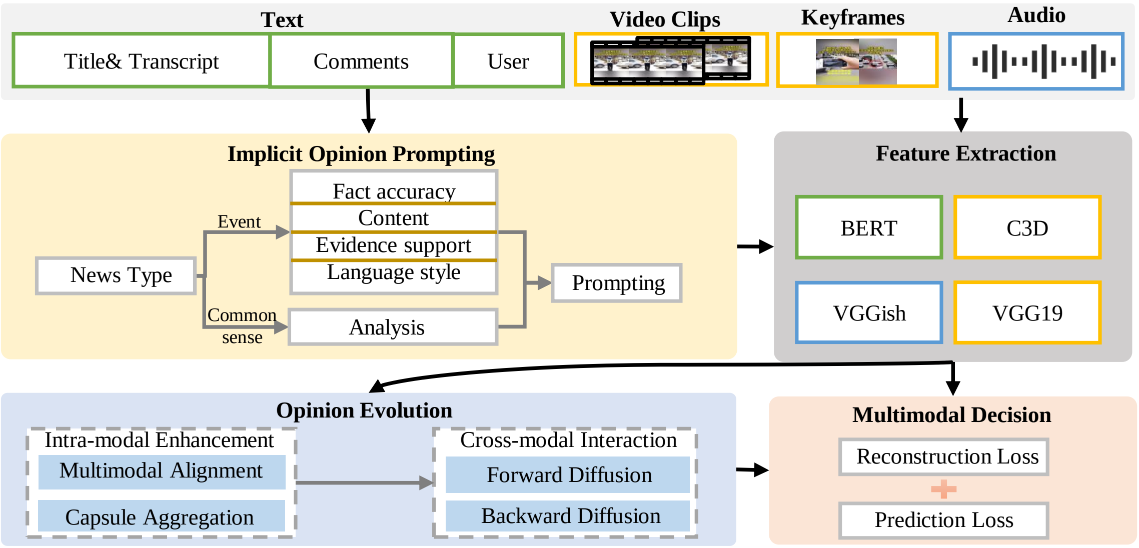


Figure 2: The main architecture of OpEvFake.

debunking relationship between debunking news and fake news.

The above methods overlook the implicit opinions in the news and are difficult to spread the influence of local fake factors to the global level. Therefore, we design an opinion evolution process to expand the global influence of local fake factors under the instruction of Large Language Models (LLMs).

2.2 Opinion Dynamics

In a group setting, individuals neither wholly accept nor completely ignore the opinions of others but consider these opinions in the process of updating their opinions. Through the interactive process within the group, individuals continuously update and integrate their opinions on the same issue, ultimately leading to the formation of three stable states: consensus, polarization, or division (Dong et al., 2018). Research on opinion dynamics originated in France (French Jr, 1956), and various intriguing opinion dynamic models with different opinion formats and fusion rules have been proposed, such as the DeGroot model (DeGroot, 1974), voter model (Ben-Naim et al., 1996), continuous opinion and discrete action model (Martins, 2008, 2014).

The process of multimodal fusion in short video fake news detection can be seen as a form of opinion dynamics, where each modality (fake or real) continuously updates and aggregates information based on certain rules, similar to the evolution of opinions. We introduce opinion dynamics to enable multiple modalities to simultaneously participate in the opinion evolution process.

2.3 Diffusion Model

The diffusion model (Ho et al., 2020) is a neural generative model based on the thermodynamics-inspired stochastic diffusion process. This process involves gradually adding noise to samples from the data distribution and then training a neural network to reverse this process by gradually removing the noise. Recent developments in the diffusion model have primarily focused on generative tasks, such as image generation (Ho et al., 2020; Dhariwal and Nichol, 2021), natural language generation (Austin et al., 2021), and audio generation (Popov et al., 2021).

We simulate the propagation of opinions through a diffusion model, reconstructing single-modal opinion representations under the guidance of multimodal opinion representations, including the implicit opinion interaction with each modal opinion.

3 Methodology

3.1 Problem Formulation

The primary objective of the short video fake news detection task is to accurately identify fake news within a given short video news dataset. The dataset, denoted as D, comprises N news items. Each news item encompasses seven modalities: video transcript, title, user information, comments, audio, keyframes, and video motion. Since both video transcript and title indicate the video content in textual, we concatenate the title and transcript. Therefore, these modalities are captured as $F = \{f^t, f^u, f^c, f^a, f^k, f^v\}$, with f^t representing the text obtained by concatenating the title and transcript, and f^u, f^c, f^a, f^k, f^v denoting user information, comments, audio, keyframes, and video motion, respectively. Our task is to predict the label y of F, where $y \in \{0, 1\}$, with 0 indicating real news and 1 indicating fake news.

3.2 Overview of the Proposed Method

To mine the implicit opinions embedded within short video news and facilitate the evolution of opinions across all modalities within the news, we propose a multimodal fake news detection method, named OpEvFake, based on the prompting and diffusion model. The proposed framework is illustrated in Figure 2. Specifically, our model first designs a reasoning chain prompt template tailored for fake news, utilizing an LLM for implicit opinion analysis. Subsequently, features are extracted from each modality. Furthermore, we incorporate the concept of opinion dynamics to facilitate intra-modal opinion enhancement and cross-modal opinion interaction, leveraging a diffusion model to simulate the spreading of opinions. Finally, the refined opinions from each modality are utilized for classification.

3.3 Implicit Opinion Prompting

In short video news, there are typically two types of news: those centered on events and those grounded in common sense. Identifying fake news within these categories relies on implicit opinions on distinct clues. These opinions play a crucial role in effectively recognizing fake news. Therefore, we propose leveraging an LLM-based prompting learning technique to efficiently uncover the implicit opinions on clues of fake news across various news types. Specifically, to generate implicit opinion prompts tailored to short video fake news detection, we devise a prompt template as follows.

Prompt-1: Suppose you are a professional news detection expert. Based on the short video news {*news_input*}, please output the content type in the following format: {*News id: event/common sense*}.

Prompt-1-1: The short video is of the event-type. Please make a comprehensive evaluation and specific analysis of the implicit opinion on the credibility of the content from four perspectives: fact accuracy (high/medium/low), content source (credible/suspicious/untrustworthy), evidence support (strong/medium/weak), and language style (appropriate/exaggerated). Please follow the following format output: {*news id, fact accuracy, content source, evidence support, language style*}.

Prompt-1-2: The short video is of the common sense-type. Please analyze the implicit opinion on the credibility of the content based on scientific knowledge. Please output in the following format: {*News id: analysis*}.

In this template, *news_input* is the textual content of a short video, including the title, transcript, user information, and comments. { } denotes content to be replaced with real data. After implicit opinion prompting, each news F will correspond to a generated implicit opinion prompt denoted as f^p. By harnessing LLMs to uncover implicit opinions of fake news, the generated opinion representation introduces intermediate reasoning cues, thereby fortifying the model's capability to distinguish different types of short video fake news.

3.4 Feature Extraction

The descriptive implicit opinion prompts f^p serve as one of the criteria for discerning the authenticity of short video news. We incorporate f^p as a new type of feature into F. Now, a short video fake news is represented as $F = \{f^t, f^u, f^c, f^a, f^k, f^v, f^p\}$.

This module takes F as its input. Textual features, including transcript&title, user information, implicit opinion prompts, and comments are extracted using the pre-trained Bert (Kenton and Toutanova, 2019) model, moreover, the comment features are followed by a weighted sum based on the number of likes for all comments. Audio features are extracted using the pre-trained VGGish (Hershey et al., 2017) model. For video, we extracted information from two levels to obtain a more comprehensive video representation. At the frame level, we used a pre-trained VGG19 (Mohbey et al., 2022) model to extract keyframe features. At the clip level, we selected 16 frames centered around each time step as the video clip and employed a pre-trained C3D (Tran et al., 2015) model to extract video motion features. The resulting features are denoted as $E = \{e^t, e^u, e^c, e^a, e^k, e^v, e^p\}$.

3.5 Diffusion Model based Opinion Evolution

We analyze the authenticity of news by mining opinions from its content using different modalities like text, visuals, and audio. Specifically, we focus on extracting textual information from title&transcripts, visual information from video clips, and audio information from audio recordings, to explore how opinions interact across these modalities. Furthermore, we recognize the importance of implicit opinion prompts within the news, considering them as an extra modality to understand how opinions about news authenticity interact across different modalities.

3.5.1 Intra-modal Opinion Enhancement

Multimodal Alignment. Considering that the multimodal features are not aligned, we use a multimodal transformer to achieve alignment and initial interaction of core modality features. As illustrated in Equation (1), z^t, z^p, z^a, z^v are the aligned features by multi-layer Multimodal Transformer (MT),

$$z^t, z^p, z^a, z^v = MTs(e^t, e^p, e^a, e^v) \tag{1}$$

MTs consists of multiple multimodal transformers. Taking title&transcripts for example, as shown in Equation (2), a Multimodal Transformer (MT) contains three cross-modal transformers (CT) (Tsai et al., 2019).

$$\begin{aligned} MT^t &= CT^{v\to t}(e^t, e^v) \\ &\oplus CT^{a\to t}(e^t, e^a) \\ &\oplus CT^{p\to t}(e^t, e^p) \end{aligned} \tag{2}$$

Since both title&transcript and the implicit opinion prompts indicate the textual information, we integrate them into a unified textual feature, obtaining semantically rich textual information $z^{tp} = FC(z^t \oplus z^p)$, where $\oplus$ is the feature concatenation, FC is the fully connected layer.

Capsule Aggregation. In short video news, opinions on the authenticity of news are dispersed in each element of the modality features. In order to enhance the long-range correlated opinion in each modality, we attempt to utilize capsule networks (Wu et al., 2021a) for further aggregation of intra-modal opinions. Specifically, for each modality z^m, $m \in \{tp, a, v\}$, we initiate a set of capsules to explore intra-modal opinions from diverse perspectives. The capsules are formulated as:

$$Cap^m_{i,j} = w^m_{i,j} z^m[i,:] \tag{3}$$

where the $w^m_{i,j}$ is a trainable parameter, the $Cap^m_{i,j}$ represents the capsule created with information from the i-th row of z^m. Then, $Cap^m_{i,j}$ is used for aggregating information to obtain opinion representation x^m of each modality in Equation (4),

$$x^m[j,:] = \textstyle\sum_i Caps^m_{i,j} \times r^m_{i,j} \tag{4}$$

where $x^m[j,:]$ is the j-th row of x^m, $r^m_{i,j}$ is the result of normalizing the routing coefficient $b^m_{i,j}$. During the training process, the routing coefficient $b^m_{i,j}$ is dynamically updated based on the similarity between $Cap^m_{i,j}$ and $x^m[j,:]$ as follows:

$$b^m_{i,j} \leftarrow b^m_{i,j} + Caps^m_{i,j} \odot x^m[j,:] \tag{5}$$

$$r^m_{i,j} = \frac{exp(b^m_{i,j})}{\sum_j exp(b^m_{i,j})} \tag{6}$$

Iterating multiple times using dynamic routing shown in Equations (4-6) allows for the updating of $r^m_{i,j}$ to obtain the enhanced single-modal opinion representation x^m.

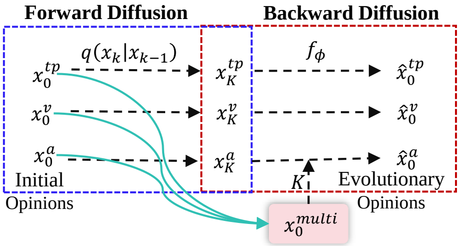


Figure 3: Diffusion model for opinion interaction.

3.5.2 Cross-modal Opinion Interaction

Adhering to the principles of opinion dynamics, we treat the three modalities (text, audio, and video actions), each carrying rich semantic information, as three distinct entities with unique opinions. To simulate the diffusion and propagation of these opinions, a diffusion model as in Figure 3 is employed, this process unfolds in two steps.

Forward Diffusion. We use the forward noise injection to update the initial opinions shown in x^m. For each modality, noise sampled from a Gaussian distribution is progressively incorporated into the true distribution. We denote the initial distribution x^m as x^m_0. According to the properties of the Markov Chain, the distribution of state x^m_k is most relevant at state x^m_{k-1}, where $k \in [1, K]$ represents the k-th step of the K-length Markov process. Therefore, the procedure for forward noise injection is as follows:

$$q(x^m_k | x^m_{k-1}) = \mathcal{N}(x^m_k; \sqrt{1-\beta^m_k} x^m_{k-1}, \beta^m_k I) \tag{7}$$

$$q(x_{1:K}^m|x_0^m) = \prod_{k=1}^{K} q(x_k^m|x_{k-1}^m) \tag{8}$$

where β_k^m is a weight that represents the proportion of noise added at k-th step, and x_k^m is sampled through the following process:

$$x_k^m = \sqrt{\bar{\alpha}_k^m} x_0^m + \sqrt{1-\bar{\alpha}_k^m}\epsilon \tag{9}$$

$$q(x_k^m|x_0^m) = \mathcal{N}(x_k^m; \sqrt{\bar{\alpha}_k^m} x_0^m, (1-\bar{\alpha}_k^m)I) \tag{10}$$

where ϵ represents the sampled noise from $\mathcal{N}(0,1)$, $\bar{\alpha}_k^m = \prod_{i=1}^{k} \alpha_i^m$, $\alpha_i^m = 1 - \beta_i^m$.

Backward Diffusion. We conduct opinion evolution through the interaction between each modal opinion and multimodal opinion. Initially, a basic linear layer is utilized to fuse the three single-modal opinion representations into preliminary multimodal opinion representations:

$$x_0^{multi} = FC(x_0^{tp} \oplus x_0^a \oplus x_0^v) \tag{11}$$

Then, the noise-injected opinion representations obtained in the first step are used to predict the distribution of single-modal opinion representations in the denoising model f_ϕ with the guidance from multimodal opinion representations, thus each modality dynamically accepts the opinions of other modalities. At the same time, to recapture the changes caused by the noise in the first step, the denoising model is more likely to capture the key parts of the single-modal opinion representations:

$$\hat{x}_0^m = f_\phi(x_0^{multi}, x_K^m, K) \tag{12}$$

We define the reconstruction loss for each modality as l^m, $l^m[i,j] = (x_0^m[i,j] - \hat{x}_0^m[i,j])^2$, $x_0^m, \hat{x}_0^m \in \mathbb{R}^{d_1 \times d_2}$, $i \in [1, d_1]$, $j \in [1, d_2]$, where d_1 and d_2 are dimensions of x_0^m and $\hat{x}_0^m$. The single-modal opinion representations acquired at this stage represent the fused single-modal opinions after opinion evolution. During this reconstruction process, with multimodal opinion representations serving as guides and single-modal opinion representations as targets, we achieve interaction among multimodal opinions.

3.6 Multimodal Opinion Decision

To fully leverage the information from various modalities in the news, we concatenate the primary representation of the evolved opinion with the auxiliary information from keyframes, comments and users, and input them into a multi-layer perception for classification.

$$\hat{y} = MLP(\hat{x}_0^{tp} \oplus \hat{x}_0^a \oplus \hat{x}_0^v \oplus FC(e^u) \oplus FC(e^c) \oplus FC(e^f)) \tag{13}$$

The loss function of the model consists of two parts, one is the reconstruction loss $\mathcal{L}_R$ in the process of opinion evolution, the other is the prediction loss $\mathcal{L}_P$, and the final loss is obtained by weighting the two parts. The calculation process is as follows, where λ is the weight coefficient and B is the batch size.

$$\mathcal{L}_P = -\sum_{n=1}^{B}(y_n log\hat{y}_n + (1-y_n) log(1-\hat{y}_n)) \tag{14}$$

$$\mathcal{L}_R = -\sum_{n=1}^{B}\sum_{i=1}^{d_1}\sum_{j=1}^{d_2}(l^{tp} + l^a + l^v) \tag{15}$$

$$\mathcal{L} = \mathcal{L}_P + \lambda\mathcal{L}_R \tag{16}$$

4 Experiments and Results

4.1 Experimental Settings

4.1.1 Dataset

We experimented on the FakeSV dataset (Qi et al., 2023a), which is the only benchmark dataset for short video fake news detection. FakeSV is collected from popular short video platforms in China, such as Douyin, encompassing rich content such as videos, audio, comments, titles, and media information. The dataset comprises a total of 3624 data entries, with 2536 in the training set, 546 in the validation set, and 542 in the test set.

4.1.2 Implement Details

In our model, the intra-modal opinion enhancement module employs a multimodal transformer with three cross-modal transformers, each consisting of 2 attention heads and 5 encoder layers. The capsule aggregation section utilizes dynamic routing with 2 iterations. During the model training process, we employed the AdamW optimizer with a batch size of 16, a learning rate of 5e-6, and a weight decay set to 0.99. The hyperparameter λ used to calculate the loss function is set to 6e-6.

4.1.3 Compared Baselines

Single-modal Baselines. We examined single modalities by considering the original features. A total of six experimental groups were included, where each feature was processed in the same way as in the feature extraction stage. The extracted features were encoded using a Bidirectional Long Short-Term Memory network (BiLSTM) further, and a linear layer was used for binary classification.

Multimodal Baselines. The multimodal fusion baselines include FANVM (Choi and Ko, 2021), CAFE (Chen et al., 2022), MultiEMO (Shi and Huang, 2023), SV-FEND (Qi et al., 2023a), and simplified NEED on SV-FEND (SV-FEND-SNEED) (Qi et al., 2023b). Since we can not obtain the keyframes of the debunking news used in NEED, we only calculated the similarity between the candidate video text and the debunking video text for classification in NEED.

LLMs Baselines. This investigation aimed to explore the performance of LLMs in the task of detecting fake news. We fed short video news titles, transcribed texts, user information, and comments into the GPT-3.5, and devised three prompts tailored for GPT-3.5. **Prompt-1** use the prompts in section 3.3 to predict whether the input news is real or fake; **Prompt-2** ignore the type of news and use the prompt-1-1 in section 3.3 to predict whether the input news is real or fake; **Prompt-3** directly predict whether the input news is real or fake.

4.2 Performance Comparison and Analysis

We use accuracy (Acc), F1-score ($F1$), macro recall (Rec), and macro precision (Pre) as evaluation metrics. Results in Table 1, highlighting the following achievements: (1) Comments modality yields the poorest results, while the Title&Transcript modality achieves the best performance. This indicates that various modalities contribute distinct clues, and it's important to use them all to catch fake news effectively. (2) OpEvFake outperforms CAFE, MultiEMO, FANVM and SV-FEND. Even without incorporating debunking data, our model surpasses SV-FEND-SNEED, highlighting the benefits of our diffusion model based opinion evolution. (3) GPT-3.5 with prompt-1 achieves the best results, implying that telling LLMs to classify news and then analyze it from different views is good for detecting fake news. But OpEvFake still outperforms these results. We suspect this could be because LLMs may focus on language patterns from big datasets, which might stop them from finding fake information. Overall, the results demonstrate that OpEvFake achieves notable improvements compared to the state-of-the-art method. This underscores the effectiveness of our implicit opinion prompting and opinion evolution learning strategy.

Models	$F1$	Rec	Pre	Acc
Keyframes	68.62	69.94	70.20	68.63
Video motion	68.62	69.90	70.11	68.63
Audio	67.76	67.74	67.78	68.27
User	78.83	78.40	80.48	79.70
Comments	63.61	63.78	65.82	65.87
Title&Transcript	79.23	79.03	79.57	79.70
CAFE	78.30	78.12	78.60	78.78
MultiEMO	82.05	81.87	82.30	82.58
FANVM	82.32	81.97	83.12	82.84
SV-FEND	81.69	81.78	81.63	81.92
SV-FEND-SNEED	81.67	81.03	84.65	82.66
GPT-3.5+Prompt-1	49.27	52.66	54.35	56.46
GPT-3.5+Prompt-2	44.43	49.87	49.82	46.68
GPT-3.5+Prompt-3	42.25	49.76	49.60	45.94
OpEvFake(our)	**87.80**	**87.71**	**87.90**	**88.01**

Table 1: Comparative experiments on FakeSV dataset.

4.3 Ablation Study and Analysis

We conducted three ablation experiments, investigating the influence of prompting strategy, intra-modal opinion enhancement and cross-modal opinion interaction.

Models	$F1$	Rec	Pre	Acc
w/o Prompt	83.73	83.78	83.69	83.95
our with Prompt-2	86.90	86.94	86.87	87.08
w/o Transformer	84.08	83.77	84.70	84.50
w/o Capsule	83.40	83.20	83.72	83.76
w/o Enhance	80.85	80.56	81.45	81.37
w/o OpiEvo	85.58	85.22	86.30	85.98
w/o Prompt&OpiEvo	83.48	83.14	84.21	83.95
OpEvFake(our)	**87.80**	**87.71**	**87.90**	**88.01**

Table 2: Ablation Study on FakeSV dataset.

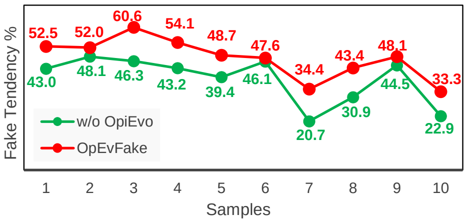


Figure 4: Study on the validity of opinion evolution. The news with fake tendency greater than 50% is predicted as fake.

4.3.1 Prompting

"w/o Prompt" indicates our model OpEvFake without implicit opinion prompting, "our with Prompt-2" indicates deleting prompt-1 from OpEvFake but using prompt-2. Results in Table 2 indicate that removing the implicit opinion prompts right away makes the model much less accurate. Adding prompt-2 helps a bit, but it's still not as good as OpEvFake. This validates implicit opinion prompts

Video	Title & Transcript	User	Comment	Prompting Text	Prediction
(a)	**Title:** Water spread in the home, a man calmly eat lunch. The man went viral on the Internet **Transcript:** Food is the life of the people. A man's home is soaked in water. The water is half man high.	Agency Certification	"Once experienced it, you will not care anymore." "It's already flooded. We can't change anything. Might as well have a meal."	**Event** **factual accuracy**: Medium. The video suggests the water level is around the height of half a person, but its exact level is unclear, causing doubt. **content source**: Suspicious. More checks are required to determine if this group is a reliable media source. **evidence support**: Medium. The cases mentioned in the comments don't fully confirm the video's situation, leading to only a moderate level of evidence. **language style**: Exaggerated. The news contains some hyperbolic words.	OpEvFake: √ w/o Prompt: × FANVM: × CAFE:× MultiEMO : √ SV-FEND : × GPT-3.5: ×
(b)	**Title:** Kosher salt can break a car window. Look at the experiment. **Transcript:** The experiment shows that kosher salt can break car windows, but it's dangerous and shouldn't be tried.	Persona Certification	"Can you remove the spark plug during an emergency?" "Kosher salt is only effective when ingested and mixed with saliva.."	**Common Sense** **Analysis:** Using coarse salt and saliva to break a car window is not a reliable method because car windows are made of strong glass that regular salt and saliva can't easily shatter.	OpEvFake: √ w/o OpiEvo: × FANVM: √ CAFE: × MultiEMO: × SV-FEND : × GPT-3.5: ×

Figure 5: Two cases from FakeSV in short video fake news detection.

are useful for detecting fake news in short videos. Also, the prompts we made, which are tailored to different news types, even improve the model's performance more.

4.3.2 Intra-modal Opinion Enhancement

"w/o Transformer", "w/o Capsule" and "w/o Enhance" indicate deleting the multimodal transformer, capsule aggregation, and the whole intra-modal opinion enhancement from OpEvFake, respectively. Results in Table 2 show that taking out both the multimodal transformer and capsule aggregation makes the model less good. But if taking out the whole module, the model gets even worse. This means that every part of the module helps the model detect fake news in short videos.

4.3.3 Cross-modal Opinion Interaction

"w/o OpiEvo" indicates deleting the cross-modal opinion interaction from OpEvFake, and "w/o Prompt&OpiEvo" removes prompting from "w/o OpiEvo" further. Results in Table 2 show that taking out both the cross-modal opinion interaction and prompting makes the model much weaker. Adding prompting helps a bit, but it still doesn't match the original model's performance. This means that even with prompting, the interaction between different modal opinions is good for detecting fake news. Also, as shown in Figure 4, after using OpEvFake to evolve opinions, 10 fake news samples that are wrongly classified as true news with "w/o OpiEvo" have a higher chance of being identified as fake. Four of these samples are even correctly classified, which shows how useful and important the opinion evolution process is.

4.4 Case Study

We present two cases in Figure 5 to illustrate how OpEvFake improves fake news detection performance by capturing implicit opinions and opinion evolution. In case (a), the visual content corresponds to the majority of the text content. These implicit opinions of fake are difficult to identify without implicit opinion prompting. Our prompts effectively capture the fake elements in the text resulting from exaggerated descriptions. In case(b), fake information is manifested in localized text. Before opinion evolution, the model mistakenly considers this fake news as real. After opinion evolution, our model amplified the impact of local fake factors on the global fake tendency, leading to correct classification.

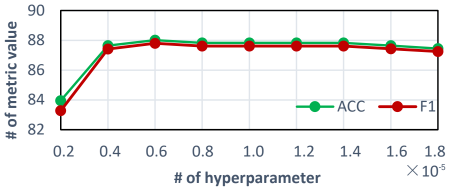


Figure 6: Influence of hyperparameter λ.

4.5 Hyperparameter Analysis

We use hyperparameter λ to control the proportion of reconstruction loss in the training target, as shown in Equation (16). In the experiment, we find the optimal value of λ through grid search. As shown in Figure 6, our method is relatively stable when λ varies in the range $[0.2 \times 10^{-5}, 1.8 \times 10^{-5}]$, indicating that our model is insensitive to this parameter within a certain range.

5 Conclusion

We leverage the analytical capabilities of LLMs to assist in fake news detection tasks and devise opinion evolution based on a diffusion model to achieve cross-modal interaction. The proposed model can better strengthen the interaction of opinion in each modality in detecting fake videos. Experimental results demonstrate that our model outperforms existing fake news detection methods on a publicly available dataset for short video fake news detection. We analyze the effectiveness of opinion evolution and discuss how to utilize LLMs to assist in fake news detection tasks.

6 Limitations

Our framework utilizes Large Language Models (LLMs) to produce analyses of news text content. The effectiveness of these generated analyses relies on the reasoning and analytical capabilities inherent in the LLM itself. However, our proposed framework currently lacks an evaluation of the quality of analyses generated by LLMs, particularly in the context of fake news detection. In future work, we intend to develop an evaluation framework focusing on the quality of analyses generated by LLMs for fake news detection. This framework aims to enhance results by filtering out analyses with higher quality based on predefined evaluation metrics.

7 Ethical Consideration

We utilize the publicly available datasets created by previous researchers, adhering to all pertinent legal and ethical standards during their acquisition and utilization. Given that analyses derived from news text using LLMs could potentially influence individuals or communities, we take precautionary measures. To support fellow researchers in the realm of fake news detection, we furnish only the prompt templates, refraining from disclosing the specific content of analyses generated by LLMs. This approach ensures that the generated analyses do not inadvertently contribute to misinformation or negatively impact the public.

Acknowledgements

This work was supported in part by the Ministry of Education Humanities and Social Science Project (No.22YJC740110), the Fundamental Research Funds for the Central Universities (DUT23YG136, DUT24MS003), and the Liaoning Provincial Natural Science Foundation Joint Fund Program(2023-MSBA-003).

References

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993.

Eli Ben-Naim, Laurent Frachebourg, and Paul L Krapivsky. 1996. Coarsening and persistence in the voter model. *Physical Review E*, 53(4):3078.

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, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, pages 8770–8780. ACM.

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*, pages 2897–2905.

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*, pages 2950–2954.

Morris H DeGroot. 1974. Reaching a consensus. *Journal of the American Statistical association*, 69(345):118–121.

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794.

Yucheng Dong, Min Zhan, Gang Kou, Zhaogang Ding, and Haiming Liang. 2018. A survey on the fusion process in opinion dynamics. *Inf. Fusion*, 43:57–65.

John RP French Jr. 1956. A formal theory of social power. *Psychological review*, 63(3):181.

Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 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*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

André CR Martins. 2008. Continuous opinions and discrete actions in opinion dynamics problems. *International Journal of Modern Physics C*, 19(04):617–624.

André CR Martins. 2014. Discrete opinion models as a limit case of the coda model. *Physica A: Statistical Mechanics and its Applications*, 395:352–357.

Krishna Kumar Mohbey, Savita Sharma, Sunil Kumar, and Meenu Sharma. 2022. Covid-19 identification and analysis using ct scan images: Deep transfer learning-based approach. In *Blockchain Applications for Healthcare Informatics*, pages 447–470. Elsevier.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. 2021. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR.

Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. 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, pages 14444–14452.

Peng Qi, Yuyang Zhao, Yufeng Shen, Wei Ji, Juan Cao, and Tat-Seng Chua. 2023b. 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, Toronto, Canada, July 9-14, 2023*, pages 11947–11959.

Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 153–162.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2931–2937.

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 (Big Data), Orlando, FL, USA, December 15-18, 2021*, pages 899–908. IEEE.

Tao Shi and Shao-Lun Huang. 2023. Multiemo: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14752–14766.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6558–6569. Association for Computational Linguistics.

Jianfeng Wu, Sijie Mai, and Haifeng Hu. 2021a. Graph capsule aggregation for unaligned multimodal sequences. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 521–529.

Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021b. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2560–2569.

Zhi Zeng, Mingmin Wu, Guodong Li, Xiang Li, Zhongqiang Huang, and Ying Sha. 2023. Correcting the bias: Mitigating multimodal inconsistency contrastive learning for multimodal fake news detection. In *IEEE International Conference on Multimedia and Expo, ICME 2023, Brisbane, Australia, July 10-14, 2023*, pages 2861–2866. IEEE.