Text-Guided Fine-grained Counterfactual Inference for Short Video Fake News Detection

Linlin Zong[1], Wenmin Lin[1], Jiahui Zhou[1], Xinyue Liu[1], Xianchao Zhang[1], Bo Xu[2*], Shimin Wu[1]

[1]Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, School of Software, Dalian University of Technology, Dalian, China

[2]School of Computer Science and Technology, Dalian University of Technology, Dalian, China

{llzong, xyliu, xczhang, xubo}@dlut.edu.cn, {wmlin, zjhjixiang, wushimin}@mail.dlut.edu.cn

Abstract

Detecting fake news in short videos is crucial for combating misinformation. Existing methods utilize topic modeling and co-attention mechanism, overlooking the modality heterogeneity and resulting in suboptimal performance. To address this issue, we introduce Text-Guided Fine-grained Counterfactual Inference for Short Video Fake News detection (TGFC-SVFN). TGFC-SVFN leverages modality bias removal and teacher-model-enhanced inter-modal knowledge distillation to integrate the heterogeneous modalities in short videos. Specifically, we use causality-based reasoning prompts guided text as teacher model, which then transfers knowledge to the video and audio student models. Subsequently, a multi-head attention mechanism is employed to fuse information from different modalities. In each module, we utilize fine-grained counterfactual inference based on a diffusion model to eliminate modality bias. Experimental results on publicly available fake short video news datasets demonstrate that our method outperforms state-of-the-art techniques.

Introduction

Fake news detection is essential for preventing misinformation and maintaining public trust (DiFonzo and Bordia 2007; Jin et al. 2017; Jankowski et al. 2020). Today, short video platforms are key channels for disseminating fake news. These platforms combine images, videos, audio, social content, and comments, each with different formats and characteristics. This diversity complicates the assessment of news authenticity.

The research on detecting fake news in short videos primarily focuses on the effective integration of multiple modalities using approaches like topic modeling (Choi and Ko 2021) and co-attention mechanism (Qi et al. 2023a). While these methods have shown success, they often overlook the **heterogeneity** among modalities. Greater differences among modalities increase the difficulty of effective integration. In text-image fake news detection, inter-modal knowledge distillation is used to balance modality differences (Wei et al. 2022; Hu et al. 2024). However, applying these methods directly to short video fake news detection could introduce two issues: modality bias amplification and teacher model inadequacy.

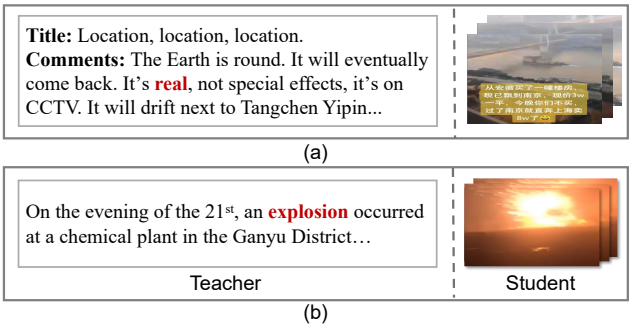


Figure 1: (a) Case of modality bias. (b) Modality bias amplification, the real news is that there was a fire, not an explosion. The reference to an explosion is misleading.

Modality bias amplification. Modality bias as the phenomenon in which certain factors or keywords exhibit spurious correlations with the labels during the model's learning and prediction process, leading to biased outcomes (Qian et al. 2021). Figure 1(a) shows a spurious correlation between the keyword "real" in the comment and the label, causing the model to misclassify the news as real. Inter-modal knowledge distillation based fusion methods often overemphasize certain modalities, leading to modality bias in short videos. As shown in Figure 1(b), the visual modality may present both "explosion" and "fire" information simultaneously. If the teacher model disproportionately highlights the "explosion" element, which is a piece of false information, it may be transferred to the visual modality, thereby obscuring the fact of "fire". In short videos, which typically integrate multiple modalities, biases can be further amplified. Modality bias amplification undoubtedly complicates the detection of fake news in short videos. This bias complicates accurate content analysis and fake information detection. Existing debiasing methods, such as those using counterfactual inference for images (Chen et al. 2023) or focusing on specific text (Zhang et al. 2024), are often too coarse-grained, making it difficult to identify key semantic features for recognizing fake news.

Teacher model inadequacy. Existing inter-modal knowl-

*Corresponding author

edge distillation (Wei et al. 2022; Hu et al. 2024) typically treat the entire text or image as the teacher model, emphasizing the transfer of the complete feature sequence. However, given the inherent heterogeneity between different modalities, each modality may contain various irrelevant information. Utilizing the full feature sequence as the teacher modality can introduce extraneous and disruptive signals, potentially misguiding the training process. Simply altering the teacher model's modalities does not address this issue. In the context of short videos, which often incorporate multiple modalities, this challenge is further exacerbated.

Our Approach. To address the above issues, we propose the Text-Guided Fine-grained Counterfactual Inference for Short Video Fake News detection (**TGFC-SVFN**). Generally, TGFC-SVFN uses text as the teacher model to guide the learning of video and audio student models, and then employ a multi-head attention mechanism to fuse information from different modalities. To address the **modality bias amplification problem**, we employ fine-grained counterfactual inference to eliminate bias. Specifically, we introduce a noise injection process based on a diffusion model to construct fine-grained virtual counterfactual scenarios at the semantic feature level. By enabling the model to learn the differences between counterfactual and factual scenarios, we effectively eliminated biases across various modalities in short video news. This approach not only prevents the amplification of modality bias but also aids in more fine-grained understanding and identification of key semantic features in news content. To address the **teacher model inadequacy problem**, we utilized the causality-based reasoning prompts to guide the teacher model's learning. The causality-based reasoning prompts generated by the large language model provide the teacher model with critical points for authenticity assessment, enabling it to focus on specific feature sequences, thereby reducing the interference of irrelevant information and enhancing its learning capability. The contributions of this paper are summarized as follows:

- We propose to eliminate modality bias and enhance the teacher model in inter-modality knowledge distillation, thereby effectively integrating the heterogeneous modalities of short videos.
- We realize fine-grained counterfactual inference based on a diffusion model and devise causality-based reasoning prompts, which effectively eliminate modality bias and improve the teacher model efficiency.
- Experiments on public short video fake news detection datasets show that our method significantly improves performance in fake news detection tasks.

Related Work

Fake Nwes Detection

Research on fake news detection has evolved through three distinct stages with the changing news platforms. **Text-modal detection** primarily focused on text or social interaction of fake news. Methods for detecting fake news have included analyzing news propagation paths (Liu and Wu 2018), using text stance detection (Kotonya and Toni 2019), examining language patterns (Przybyla 2020), applying Bayesian networks for truth and opinion dependencies (Yang et al. 2019), and employing TF-IDF and Word2Vec for text analysis (Zhang, Wang, and Tan 2018). **Text-image multi-modal detection** (Singhal et al. 2019; Choi and Ko 2021). (Ying et al. 2023) extracted representations from the perspectives of text, image pattern, and image semantics to predict the authenticity of the entire news. Researchers (Wu et al. 2021a; Song et al. 2021; Wu et al. 2021b) have also concentrated on fusing different modalities by utilizing the attention mechanism to capture modality interactions. **Short video detection** integrated multiple modalities. (Choi and Ko 2021) estimated video topic distributions using titles, descriptions, and comments through topic modeling. (Qi et al. 2023a) introduced the FakeSV dataset, which models multimodal features including video, audio, text, and user information, significantly improving fake news detection accuracy. Additionally, (Qi et al. 2023b) incorporated debunking videos to enhance detection efforts. (Zong et al. 2024) used diffusion models and prompt engineering to reveal the process of opinion evolution in fake news detection in short videos. Despite advancements in fake news detection, the challenge of integrating diverse modalities and addressing modality bias in short video fake news requires further research.

Counterfactual Inference

Recently, counterfactual inference has been widely applied in various fields (Ji, Chen, and Wu 2023; Zhou et al. 2022; Lin et al. 2024). In the field of fake news detection, (Chen et al. 2023) proposed a debiasing framework based on causal intervention and counterfactual inference. (Zhang et al. 2024) used counterfactual reasoning to predict key elements with significant impact. However, existing methods primarily focus on constructing counterfactual elements for specific information or evidence and are unable to address biases in localized elements across different modalities. To address this issue, we use a noise injection process based on diffusion models to simulate fine-grained counterfactual scenarios, allowing the model to make unbiased inferences in biased environments.

Knowledge Distillation

The knowledge distillation method, introduced by (Hinton, Vinyals, and Dean 2015), transfers knowledge by minimizing the KL divergence between the soft labels predicted by the teacher and student models, improving student model performance. It has since been used for knowledge transfer across modalities in multimodal research (Jin et al. 2021). Some studies (Wei et al. 2022; Hu et al. 2024) applied this to multimodal fake news detection. We perform distillation using inter-class and intra-class losses, with a text-based teacher network to enhance other modalities.

Counterfactual Inference

Counterfactual inference, originating from causal inference (Pearl, Glymour, and Jewell 2016), is a statistical method

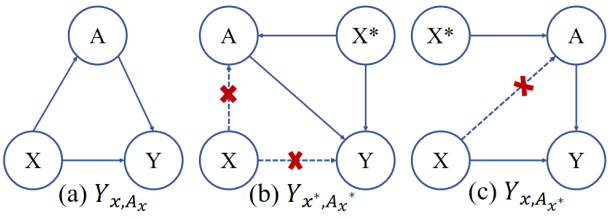


Figure 2: The counterfactual inference graphs for fake news detection. A is a "mediating variable" that lies between the causal factor X and the outcome Y, transmitting influence in the causal chain.

used to infer potential outcomes that differ from the actual reality (Pearl 2009). In short video fake news detection, counterfactual inference seeks to answer the question: How would the prediction outcome Y change if the short video fake news semantic feature X were altered (e.g., from state x to state x^*)? This can be quantified by learning the total indirect effect (TIE) of X on Y. The total indirect effect can be expressed as the difference between the total effect (TE) and the natural direct effect (NDE).

Mathematically, the TE of $X = x$ on Y is given by:

$$\text{TE} = Y_{x,A_x} - Y_{x^*,A_{x^*}} \tag{1}$$

where $Y_{x,A_x} = Y\left(X = x, A = A\left(X = x\right)\right)$ (Figure 2(a)), $Y_{x^*,A_{x^*}} = Y\left(X = x^*, A = A\left(X = x^*\right)\right)$ indicates the effect of $X = x^*$ on Y (Figure 2(b)). When the mediating variable A is blocked, the NDE of X on Y is:

$$\text{NDE} = Y_{x,A_{x^*}} - Y_{x^*,A_{x^*}} \tag{2}$$

where $Y_{x,A_{x^*}} = Y\left(X = x, A = A\left(X = x^*\right)\right)$ indicates the impact on Y in a counterfactual scenario with X set to different values x and x^* (Figure 2(c)). Thus, the total indirect effect of X on Y can be expressed as:

$$\text{TIE} = \text{TE} - \text{NDE} = Y_{x,A_x} - Y_{x,A_{x^*}} \tag{3}$$

Methodology

Overview

Given a short video news dataset $\{D, L\}$ containing news D and the ground truth labels $L \in \{0, 1\}$. Short video news includes the title&transcript t, comments c, user information u, video key frames k and video clip segments v.

We propose the TGFC-SVFN method to integrate diverse modalities and eliminate modality bias. As shown in Figure 3, the text teacher model guides the learning of video and audio student models, and a multi-head attention mechanism fuses their semantic features for classification. Fine-grained counterfactual inference is used to remove biases, ensuring unbiased decisions under biased conditions. The text teacher model is enhanced with ChatGPT-3.5 to generate causality-based reasoning prompts. Each module is trained separately.

Causality-based Reasoning Prompt

To improve text information efficiency and the language model's focus, we customized a causality-based reasoning prompt template. This template provides concise judgment reasons and explanations through multi-step analysis, guiding the language model's learning. The method uses news title, transcript, user comments, and user information as $input$ for the LLM. The causality-based reasoning prompt template is as follows:

Step 1: This is the news information: News ID: {vid}, News context information:{$input$}. Please perform a preliminary information check: assess the reasonableness of the event description, the specificity of the provided information, the credibility of the publisher, and whether there is any obvious falsehood. Limit the response to 100 words, and output in the following format: {'News ID': , 'Analysis': }.

Step 2: Based on the analysis from Step 1, please further analyze whether the information conforms to common sense and logic. Do not return any analysis; strictly output in the following format: {'Conforms to common sense or logic / Does not conform to common sense or logic'}.

Step 3: Based on the analysis from Step 1 and Step 2, do you think this information is real or fake? Provide a concise reason and explanation for your judgment, limited to 100 words. Strictly output in the following format: {'News ID': , 'Prediction Result': 'Real/Fake', 'Reason and Explanation': ' '}.

Through the above three-step inference, each news instance will yield an inferred text g.

Feature Extraction

Textual features are extracted using the pre-trained Bert (Kenton and Toutanova 2019) model, audio features with VGGish (Hershey et al. 2017), static video frame features with VGG19 (Mohbey et al. 2022), and video clip segment features with C3D (Tran et al. 2015). The text features are represented as: $T = \{e^t, e^u, e^c, e^g\}$, video features as: $V = \{e^k, e^v\}$, and audio features as: $A = \{e^a\}$

Teacher Model Learning

Text is the dominant modality to detect fake news in the literature (Shu et al. 2017). Similarly, in short video fake news detection, we found that the detection capability of the language modality surpasses that of other non-language modalities. Therefore, we chose to use the all text modalities as the teacher.

Multiple text alignment. Firstly, since our model takes into account the temporal relationships between the tokens in the text, we utilize GPT-2 to analyze the temporal evolution of the long text features e^t, e^c, and e^g, and predict future action features (Zhong et al. 2023):

$$g^t, g^c, g^g = \text{GPT-2}(e^t, e^c, e^g) \tag{4}$$

Secondly, we use a Cross-modal Transformer (CMT) (Tsai et al. 2019) to perform semantic alignment for the long-text features g^t, g^c, and g^g, resulting in the aligned features f^t, f^c, and f^g. For f^t:

$$f^t = CMT_{c\to t}(g^t, g^c) \oplus CMT_{g\to t}(g^t, g^g) \tag{5}$$

Finally, we concatenate all textual features to obtain the fused textual feature F^{TG}:

$$F^{TG} = f^t \oplus f^c \oplus e^u \oplus f^g \tag{6}$$

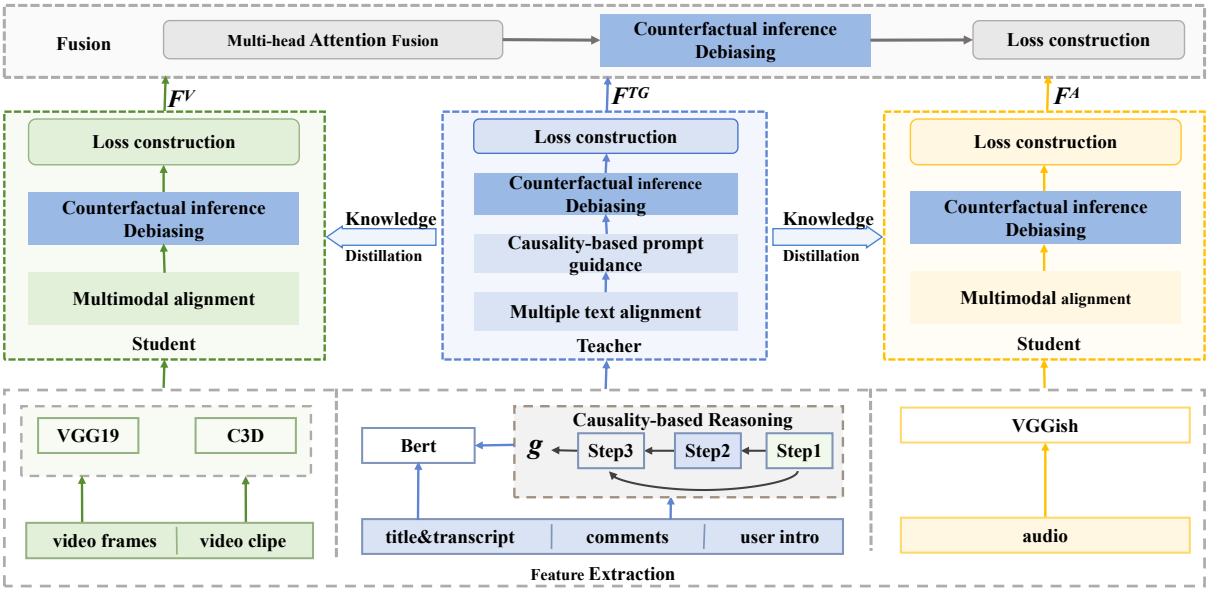


Figure 3: The main architecture of TGFC-SVFN.

where $\oplus$ denotes feature concatenation.

Causality-based prompt guidance. To implement the guidance of the teacher model using causality-based reasoning prompts generated by ChatGPT-3.5, we designed a feature similarity constraint $\mathcal{L}_g$. We first compute the similarity between f^g and F^{TG}, and obtain the corresponding probability distributions using the softmax function:

$$P = \text{softmax}\left(\frac{f^g \cdot f^{gT}}{\tau}\right); Q = \text{softmax}\left(\frac{f^g \cdot F^{TG^T}}{\tau}\right) \tag{7}$$

where $\tau > 0$ is the temperature parameter controlling the smoothness of the distribution. The feature similarity constraint $\mathcal{L}_g$ is defined as the KL divergence between these two probability distributions:

$$\mathcal{L}_g = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} P_{ij}\log\left(\frac{P_{ij}}{Q_{ij}}\right) \tag{8}$$

Counterfactual inference Debiasing. To eliminate biases in the teacher model and ensure unbiased knowledge transfer to the student models, we apply counterfactual inference techniques. This involves constructing counterfactual scenarios based on factual scenarios, allowing the model to learn the differences and then remove biases.

(1) *Counterfactual scenario Simulation.* As shown in Figure 4(a), we generate counterfactual scenario using the forward noise process of the diffusion model (Ho, Jain, and Abbeel 2020) to construct fine-grained local bias elements. The forward diffusion process is defined as $q(x_t \mid x_0)$, where noise is gradually added to the factual feature x_0 (where x_0 represents F^{TG}), constructing the counterfactual feature x_t. The forward noise injection process is expressed as:

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I\right) \tag{9}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t}\alpha_i$, $\alpha_t = 1-\beta_t$, and β is noise scheduling parameter. The counterfactual feature x_t generated at any time step t in the counterfactual scene is expressed as:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(0, I) \tag{10}$$

This method constructs T counterfactual scenes for fine-grained inference.

To enhance the effectiveness of counterfactual inference, we retain the reverse denoising process and add a diffusion constraint $\mathcal{L}_d$. The reverse diffusion process is as follows:

$$p_\theta\left(x_{t-1}|x_t\right) = \mathcal{N}\left(x_{t-1}; \mu_\theta\left(x_t, t\right), \sigma_t^2 I\right) \tag{11}$$

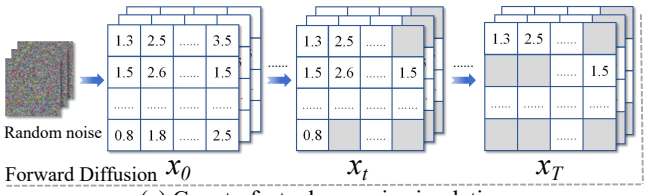
$$\mu_\theta\left(x_t, t\right) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta\left(x_t, t\right)\right) \tag{12}$$

where σ_t^2 is the variance of β_t. $\mu_\theta\left(x_t, t\right)$ is the conditional mean predicting x_{t-1} given x_t and time step t. By predicting x_{t-1}, the model can gradually remove noise and recovers the original data $\hat{x}_0$. The mean squared error measures the difference between the denoised and original features, optimizing the model for counterfactual scene construction. $\mathcal{L}_d$ is defined as:

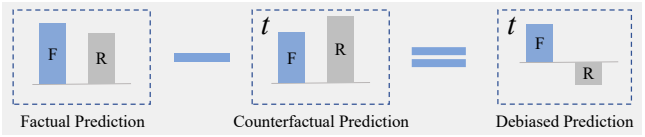
$$\mathcal{L}_d = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{x}_0^{(i)} - x_0^{(i)}\right)^2 \tag{13}$$

(2) *Bias removal.* As shown in Figure 4(b), during training, we perform debiasing by using the difference between factual and counterfactual predictions, while in the inference, the factual prediction is used as the debiased result. The total indirect effect of feature F^{TG} guides the debiasing learning. The total indirect effect at time step t is expressed as:

$$\hat{p} = \text{TIE} = Y_{x_0} - Y_{x_t} = p^{x_0} - p^{x_t} \tag{14}$$



(a) Counterfactual scenario simulation



(b) Bias removal

Figure 4: (a) The noise injection process of the diffusion model simulates the construction of T counterfactual scenarios, (b) Remove bias at time step t.

here, $\hat{p}$ is the debiased prediction probability, p^{x_0} is the prediction probability of factual features F^{TG}, and p^{x_t} is the prediction probability of counterfactual features constructed from F^{TG}. The counterfactual constraint $\mathcal{L}_C$ is calculated using the cross-entropy loss between the debiased prediction $\hat{p}$ and the true label y.

$$\mathcal{L}_C = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log\left(\hat{p}_i\right) + \left(1-y_i\right)\log\left(1-\hat{p}_i\right)\right] \quad (15)$$

The total debiasing loss through counterfactual inference is:

$$\mathcal{L}_{debiase} = \mu_1 \mathcal{L}_C + \mu_2 \mathcal{L}_d \quad (16)$$

μ_1 and μ_2 are weight parameters.

Training loss. The teacher model loss $\mathcal{L}_{teacher}$ is:

$$\mathcal{L}_{teacher} = \mathcal{L}_{debiase} + \alpha(\mathcal{L}_{CE} + \mathcal{L}_g) \quad (17)$$

where $\mathcal{L}_{CE}$ represents the cross-entropy loss between the predicted values of F^{TG} and the ground truth.

Student Model Learning

We independently train the video and audio student models during knowledge distillation due to their significant modality differences.

Multimodal alignment. We use a cross-modal Transformer (CMT) to achieve early semantic alignment of heterogeneous multimodal data.

$$F^V = \{CMT_{v\to k}(e^k, e^v) \oplus CMT_{TG\to k}(e^k, F^{TG})\} \oplus \{CMT_{k\to v}(e^v, e^k) \oplus CMT_{TG\to v}(e^v, F^{TG})\} \quad (18)$$

$$F^A = CMT_{TG\to a}(e^a, F^{TG}) \quad (19)$$

Knowledge transfer. We transfer knowledge by leveraging the correlation between teacher and student model predictions. Instead of KL divergence, we use the Pearson correlation coefficient to measure inter-class and intra-class relationships between the teacher and student models. This approach captures the linear correlation better. These relationships are then used to compute the distillation loss $\mathcal{L}_{dis}$.

$$Y^t_{i,:} = \text{softmax}\left(X^t_{i,:}/\tau\right); Y^s_{i,:} = \text{softmax}\left(X^s_{i,:}/\tau\right) \quad (20)$$

$$\mathcal{L}_{inter} = \frac{\tau^2}{B}\sum_{i=1}^{B} d\left(Y^s_{i,:}, Y^t_{i,:}\right) \quad (21)$$

$$\mathcal{L}_{intra} = \frac{\tau^2}{C}\sum_{i=1}^{C} d\left(Y^{s\,T}_{i,:}, Y^{t\,T}_{i,:}\right) \quad (22)$$

$$\mathcal{L}_{dis} = \mathcal{L}_{inter} + \mathcal{L}_{intra} \quad (23)$$

where B is batch size, C is the number of classes, $s \in \{F^V, F^A\}$, t is F^{TG}, $X^s \in R^{B\times C}$ is the student's prediction matrix, $X^t \in R^{B\times C}$ is the teacher's prediction matrix, d is Pearson correlation coefficient, T denotes the transpose of a matrix, $\mathcal{L}_{inter}$ is the inter-class loss and $\mathcal{L}_{intra}$ is the intra-class loss. The parameter $\tau > 0$ controls the smoothness of the soft labels.

Counterfactual inference Debiasing. We treat the audio student model's semantic features as factual features F^V and the video student model's as F^A, then construct their corresponding counterfactual features. By calculating the total indirect effect for both student models, we eliminate biases using Equations (9-14).

Training loss. The loss for the audio and video student models is defined as:

$$\mathcal{L}_{audio} = \mathcal{L}_{video} = \mathcal{L}_{debiase} + \alpha(\mathcal{L}_{CE} + \mathcal{L}_{dis}) \quad (24)$$

where $\mathcal{L}_{CE}$ represents the cross-entropy loss between the predicted values of student model and the ground truth, and α is weight parameters.

Multimodal Fusion and Classification

Multi-modal fusion. Our method uses a multi-head attention mechanism (MHA) to fuse the semantic features for classification from both the teacher model and the student model.

$$F^M = \text{MHA}(F^V \oplus F^A \oplus F^{TG}) \quad (25)$$

Counterfactual inference Debiasing. We use Equations (9-14) to construct counterfactual features from F^M, calculate the total indirect effect, and eliminate biases in the fusion model.

Training loss. The total loss of the fusion model is:

$$\mathcal{L} = \mathcal{L}_{debiase} + \alpha\mathcal{L}_{CE} \quad (26)$$

where $\mathcal{L}_{CE}$ represents the cross-entropy loss between the predicted values of F^M and the ground truth, and α is weight parameters.

Experiments

Dataset

We experiment on the FakeSV dataset (Qi et al. 2023a), the only benchmark for short video fake news detection. FakeSV includes rich content like videos, audio, comments, titles, and media information. Following the original study, the dataset is segmented in two ways: by event and by time.

Dataset	Time				Event			
Models	**F1%**	**Rec%**	**Pre%**	**Acc%**	**F1%**	**Rec%**	**Pre%**	**Acc%**
Key frames	68.62	69.94	70.20	68.63	73.89	73.80	74.15	73.89
Video clip	68.62	69.90	70.11	68.63	72.50	72.45	72.65	72.50
Audio	67.76	67.74	67.78	68.27	72.22	71.98	72.95	72.22
User	78.83	78.40	80.48	79.70	77.50	77.39	78.05	77.50
Comments	63.61	63.78	65.82	65.87	57.64	57.37	57.91	57.64
Title&Transcript	79.23	79.03	79.57	79.70	80.56	80.51	80.82	80.56
FANVM	82.32	81.97	83.12	82.84	75.81	75.82	76.30	77.38
CAFE	78.30	78.12	78.60	78.78	73.87	74.05	74.78	74.05
MultiEMO	82.05	81.87	82.30	82.58	79.09	79.15	79.34	79.11
SV-FEND	81.69	81.78	81.63	81.92	79.19	79.26	79.51	79.24
SV-FEND-SNEED	81.67	81.03	84.65	82.66	76.06	76.46	78.22	76.48
MMCAN	85.32	84.72	86.64	85.98	79.67	79.60	79.95	79.64
ChatGPT-3.5-direct	42.25	49.76	49.60	45.94	42.25	49.76	49.60	45.94
ChatGPT-4-direct	71.40	71.66	71.37	71.59	74.92	73.35	77.27	75.36
ChatGPT-3.5-CR	75.86	76.16	75.64	76.02	73.95	74.29	75.68	74.30
TGFC-SVFN	**91.50**	**90.68**	**92.93**	**91.99**	**81.66**	**81.68**	**81.80**	**81.73**

Table 1: Comparative experiments on datasets partitioned by time and events.

Method	F1%	Rec%	Pre%	Acc%
W/o GPT-2	86.70	85.91	88.19	87.50
W/o CMT	85.77	84.87	87.70	86.72
W/o CI	85.19	84.08	88.01	86.33
W/o CR	82.01	81.73	82.39	82.81
W/o KD	88.21	87.51	89.4	88.87
TGFC-SVFN	**91.50**	**90.68**	**92.93**	**91.99**

Table 2: Component performance study on time-partitioned datasets. CI: counterfactual inference, CR: causality-based reasoning prompt guidance, KD: knowledge distillation.

Method	F1%	Rec%	Pre%	Acc%
MultiEMO	82.05	81.87	82.30	82.58
MultiEMO+CI	82.42	81.41	85.15	83.79
MultiEMO+CR	83.05	82.49	84.04	83.98
MultiEMO+KD	81.57	80.60	84.24	83.00
MultiEMO+All	84.22	83.26	86.50	85.35
SV-FEND	81.69	81.78	81.63	81.92
SV-FEND+CI	87.60	86.94	88.71	88.28
SV-FEND+CR	82.48	82.00	83.29	83.40
SV-FEND+KD	84.34	83.27	87.11	85.55
SV-FEND+All	89.05	87.88	91.70	90.04

Table 3: Applicability study. ALL: CI + CR + KD.

Implement Details

The experiments were conducted on an RTX 3090 Ti GPU using the PyTorch framework, with Python version 3.9.18. The initial learning rate (lr) was set to 0.0004 for the teacher model, 0.0005 for the video model, and 0.01 for the audio model, with a batch size of 64. The model parameters were optimized using the Adam optimizer (Kingma and Ba 2014).

Baselines

We compare our method with single-modal, multimodal, and LLM baselines. For **single-modal baselines**, conducted a comparative study on all single modalities. We combined the title and transcript into a single feature due to some empty title fields. **Multimodal baselines** include FANVM (Choi and Ko 2021), CAFE (Chen et al. 2022), MultiEMO (Shi and Huang 2023), SV-FEND (Qi et al. 2023a), SV-FEND-SNEED (Qi et al. 2023b), and MMCAN (Hu et al. 2024). For SV-FEND-SNEED, we only calculate the similarity between the candidate and debunking video text due to the lack of key frames in NEED. **LLM baselines** use the video title, transcript, user information, and comments. ChatGPT-3.5-direct and ChatGPT-4-direct predict the news authenticity directly, while ChatGPT-3.5-CR uses causality-based reasoning prompts.

Performance Comparison and Analysis

We use F1-score ($F1$), macro recall (Rec), macro precision (Pre) and accuracy (Acc) as evaluation metrics. Results in Table 1, highlighting the following achievements: (1) The overall performance of the language modality is superior to that of the non-language modality; (2) Our proposed causality-based reasoning prompt template outperforms the direct prediction prompt template; (3) Our method surpasses the current state-of-the-art across two different dataset partitioning strategies, achieving an accuracy of 91.99% on datasets partitioned by time. Additionally, our method better integrates multimodal information. Overall, the results demonstrate the significant advantages of using causality-based reasoning prompts to guide text learning and applying counterfactual inference for debiasing.

Ablation Study and Analysis

Component performance study. We conducted five ablation experiments to assess the impact of each component on the model's performance, as shown in Table 2.

The results demonstrate that all components contributed significantly to performance improvement. The causality-based reasoning prompts having the greatest impact, followed by counterfactual inference. This is likely because the causality-based reasoning prompts guided the language

model's learning, which in turn influenced the performance of the non-language model. These prompts helped the language model identify key points from large amounts of text, highlighting their importance in language model learning. Additionally, counterfactual inference significantly helped in eliminating model biases and enhancing the model's generalization capabilities.

To validate our method's applicability, we incorporated our components into two strong multimodal baselines on the time-segmented dataset. Table 3 shows that our method performed well across different approaches. Notably, MultiEMO, which uses less information, showed smaller improvements compared to SV-FEND, highlighting the importance of fully utilizing all modality information in short video news detection.

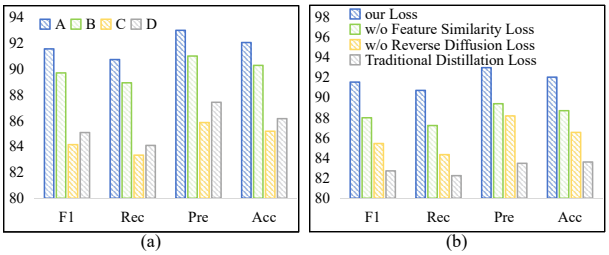


Figure 5: (a) Counterfactual inference strategy study; (b) Loss constraint strategies study.

Counterfactual inference strategy study. We studied four counterfactual inference strategies: "A: Subtract during learning, keep constant during inference", "B: Subtract during learning, add during inference", "C: Keep constant during learning, add during inference", and "D: Keep constant in both phases". "Subtract" means subtracting the counterfactual prediction from the factual, "add" means adding them, and "keep constant" means using the factual prediction. Figure 5(a) shows that strategy "A" is the most effective, which is the debiasing strategy used in this paper.

Loss constraint strategies study. We examined the impact of reverse diffusion loss $\mathcal{L}_d$, feature similarity loss $\mathcal{L}_g$, and knowledge distillation loss $\mathcal{L}_{dis}$. Figure 5(b) shows that: (1) **Reverse diffusion loss** improves model performance, likely because the reverse diffusion process better captures underlying patterns and relationships in the data, helping to eliminate modality biases and enhance generalization capabilities. (2) Omitting **feature similarity loss** harms performance, highlighting its importance. (3) Compared to traditional KL divergence loss, our **distillation loss**, which calculates intra-class and inter-class similarities, is more effective, demonstrating its superiority in knowledge transfer.

Case Study

As shown in Figure 6, cases (a) and (b) were incorrectly predicted without using improved knowledge distillation and counterfactual inference debiasing. However, after applying these methods, they were accurately identified. In case (a), the long text made it difficult for conventional detection

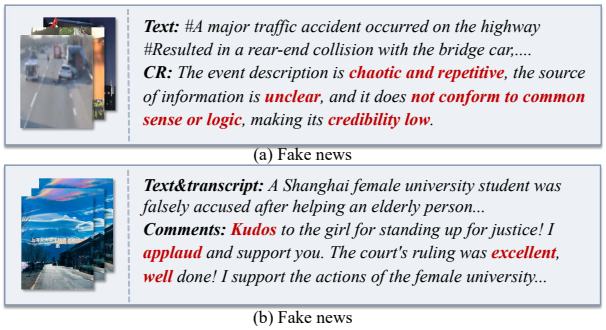


Figure 6: (a) indicates adding knowledge distillation, CR: causality-based reasoning prompts guidance; (b) indicates adding counterfactual inference.

methods to spot anomalies, but the causality-based reasoning prompts generated by ChatGPT-3.5 revealed its falsity, guiding the model to learn and ultimately classify it correctly as fake news. In case (b), the majority of the text is positive. Without debiasing, the model incorrectly classifies it as real news due to the positive sentiment. However, after applying counterfactual inference for debiasing, the model accurately predicts it as fake news.

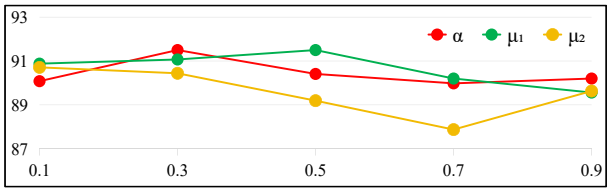


Figure 7: Hyperparameter Study. The y-axis is F1-score.

Hyperparameter Study

We studied the impact of the three hyperparameters μ_1, μ_2, and α in Equation (26) using the controlled variable method. In each experiment, two hyperparameters were held constant while the other was adjusted, with initial values set to $\mu_1 = 0.5$, $\mu_2 = 0.2$, and $\alpha = 0.3$. As shown in Figure 7, when μ_1 and α vary within the range [0.1, 0.5], and μ_2 within [0.1, 0.3], the model remains relatively stable, indicating that the model is insensitive to these parameters within this range.

Conclusion

We leverage the analytical capabilities of large language models (LLMs) to assist in the fake news detection task, and design fine-grained counterfactual inference and text-guided multimodal fusion methods to achieve heterogeneous structured multimodal information interaction. The proposed model can better eliminate biases caused by local elements and enhance the contributions of each modality in detecting fake video news. Experimental results show that our model outperforms existing fake news detection methods on publicly available video fake news detection datasets.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 62476040), the Ministry of Education Humanities and Social Science Project (No.22YJC740110), the Fundamental Research Funds for the Central Universities (DUT24MS003), and the Liaoning Provincial Natural Science Foundation Joint Fund Program(2023-MSBA-003).

References

Chen, Y.; Li, D.; Zhang, P.; Sui, J.; Lv, Q.; Tun, L.; and Shang, L. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM Web Conference 2022*, 2897–2905.

Chen, Z.; Hu, L.; Li, W.; Shao, Y.; and Nie, L. 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.

Choi, H.; and Ko, Y. 2021. Using topic modeling and adversarial neural networks for fake news video detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2950–2954.

DiFonzo, N.; and Bordia, P. 2007. Rumor, gossip and urban legends. *Diogenes*, 54(1): 19–35.

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

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, L.; Zhao, Z.; Qi, W.; Song, X.; and Nie, L. 2024. Multimodal matching-aware co-attention networks with mutual knowledge distillation for fake news detection. *Information Sciences*, 664: 120310.

Jankowski, J.; Bartkow, P.; Pazura, P.; and Bortko, K. 2020. Evaluation of the costs of delayed campaigns for limiting the spread of negative content, panic and rumours in complex networks. In *Computational Science–ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Proceedings, Part IV 20*, 291–304. Springer.

Ji, X.; Chen, J.; and Wu, X. 2023. Counterfactual Inference for Visual Relationship Detection in Videos. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 162–167. IEEE.

Jin, W.; Sanjabi, M.; Nie, S.; Tan, L.; Ren, X.; and Firooz, H. 2021. Msd: Saliency-aware knowledge distillation for multimodal understanding. *arXiv preprint arXiv:2101.01881*.

Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; Wang, Y.; and Luo, J. 2017. Detection and analysis of 2016 us presidential election related rumors on twitter. In *Social, Cultural, and Behavioral Modeling: 10th International Conference, SBP-BRiMS 2017, Washington, DC, USA, July 5-8, 2017, Proceedings 10*, 14–24. Springer.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, 2.

Kingma, D.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *Computer Science*.

Kotonya, N.; and Toni, F. 2019. Gradual argumentation evaluation for stance aggregation in automated fake news detection. In *Proceedings of the 6th Workshop on Argument Mining*, 156–166.

Lin, W.; Zhuang, Z.; Yu, L.; and Wang, L. 2024. Boosting Multiple Instance Learning Models for Whole Slide Image Classification: A Model-Agnostic Framework Based on Counterfactual Inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3477–3485.

Liu, Y.; and Wu, Y.-F. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Mohbey, K. K.; Sharma, S.; Kumar, S.; and Sharma, M. 2022. COVID-19 identification and analysis using CT scan images: Deep transfer learning-based approach. In *Blockchain Applications for Healthcare Informatics*, 447–470. Elsevier.

Pearl, J. 2009. Causal inference in statistics: An overview.

Pearl, J.; Glymour, M.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Przybyla, P. 2020. Capturing the style of fake news. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 490–497.

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

Qi, P.; Zhao, Y.; Shen, Y.; Ji, W.; Cao, J.; and Chua, T. 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*, 11947–11959.

Qian, C.; Feng, F.; Wen, L.; Ma, C.; and Xie, P. 2021. Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5434–5445.

Shi, T.; and Huang, S.-L. 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)*, 14752–14766.

Shu, K.; Sliva, A.; Wang, S.; Tang, J.; and Liu, H. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1): 22–36.

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

Song, C.; Ning, N.; Zhang, Y.; and Wu, B. 2021. A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Information Processing & Management*, 58(1): 102437.

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

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, 6558. NIH Public Access.

Wei, Z.; Pan, H.; Qiao, L.; Niu, X.; Dong, P.; and Li, D. 2022. Cross-modal knowledge distillation in multi-modal fake news detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4733–4737. IEEE.

Wu, Y.; Zhan, P.; Zhang, Y.; Wang, L.; and Xu, Z. 2021a. Multimodal Fusion with Co-Attention Networks for Fake News Detection. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, 2560–2569.

Wu, Y.; Zhan, P.; Zhang, Y.; Wang, L.; and Xu, Z. 2021b. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, 2560–2569.

Yang, S.; Shu, K.; Wang, S.; Gu, R.; Wu, F.; and Liu, H. 2019. Unsupervised fake news detection on social media: A generative approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 5644–5651.

Ying, Q.; Hu, X.; Zhou, Y.; Qian, Z.; Zeng, D.; and Ge, S. 2023. Bootstrapping multi-view representations for fake news detection. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 37, 5384–5392.

Zhang, S.; Wang, Y.; and Tan, C. 2018. Research on text classification for identifying fake news. In *2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, 178–181. IEEE.

Zhang, Y.; Kong, L.; Tian, S.; Fei, H.; Xiang, C.; Wang, H.; and Wei, X. 2024. Multi-view Counterfactual Contrastive Learning for Fact-checking Fake News Detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, 385–393.

Zhong, Z.; Schneider, D.; Voit, M.; Stiefelhagen, R.; and Beyerer, J. 2023. Anticipative feature fusion transformer for multi-modal action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 6068–6077.

Zhou, S.; Pfeiffer, N.; Islam, U. J.; Banerjee, I.; Patel, B. K.; and Iquebal, A. S. 2022. Generating counterfactual explanations for causal inference in breast cancer treatment response. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, 955–960. IEEE.

Zong, L.; Zhou, J.; Lin, W.; Liu, X.; Zhang, X.; and Xu, B. 2024. Unveiling opinion evolution via prompting and diffusion for short video fake news detection. In *Findings of the Association for Computational Linguistics ACL 2024*, 10817–10826.