Disconfounding Fake News Video Explanation with Causal Inference

Lizhi Chen, Zhong Qian, Peifeng Li*, Qiaoming Zhu
School of Computer Science and Technology, Soochow University
20234027010@stu.suda.edu.cn, {qianzhong, pfli, qmzhu}@suda.edu.cn,

Abstract

The proliferation of fake news videos on social media has heightened the demand for credible verification systems. While existing methods focus on detecting false content, generating human-readable explanations for such predictions remains a critical challenge. Current approaches suffer from spurious correlations caused by two key confounders: 1) video object bias, where co-occurring objects entangle features leading to incorrect semantic associations; and 2) explanation aspect bias, where models over-rely on frequent aspects while neglecting rare ones. To address these issues, we propose CIFE, a causal inference framework that disentangles confounding factors to generate unbiased explanations. First, we formalize the problem through a Structural Causal Model (SCM) to identify confounding factors. We then introduce two novel modules: 1) the Interventional Video-Object Detector (IVOD), which employs backdoor adjustment to decouple object-level visual semantics; and 2) the Interventional Explanation Aspect Module (IEAM), which balances aspect selection during multimodal fusion. Extensive experiments on the FakeVE dataset demonstrate the effectiveness of CIFE, which generates more faithful explanations by mitigating object entanglement and aspect imbalance. Our code is available at https://github.com/Lieberk/CIFE.

1 Introduction

The detection of fake news videos aims to identify and flag misleading or fabricated news video content disseminated across media platforms. Given its profound impact on public opinion [Sundar *et al.*, 2021], public safety [Schoenherr and Thomson, 2020], and personal trust systems [Nakov and Martino, 2021], this task has garnered increasing attention. While early research primarily focused on binary accuracy classification, such approaches typically fail to provide transparent justifications for their predictions, limiting their practical utility in real-world scenarios. To bridge this gap, recent studies have shifted toward explainable fake news analysis [Zhang *et al.*, 2021; Wang *et al.*, 2024a], which generates human-comprehensible rationales to elucidate why given content is identified as misinformation. Building upon this paradigm, Chen et al. [2025] introduced the Fake News Video Explanation (FNVE) task, designed to produce explanations by analyzing inconsistencies between video content and associated textual (e.g., logical fallacies in videos [McCrae *et al.*, 2022] or visual-textual mismatches [Choi and Ko, 2021]). Despite promising results, this pioneering work remains vulnerable to data biases from two issues that undermine the reliability and generalizability of generated explanations.

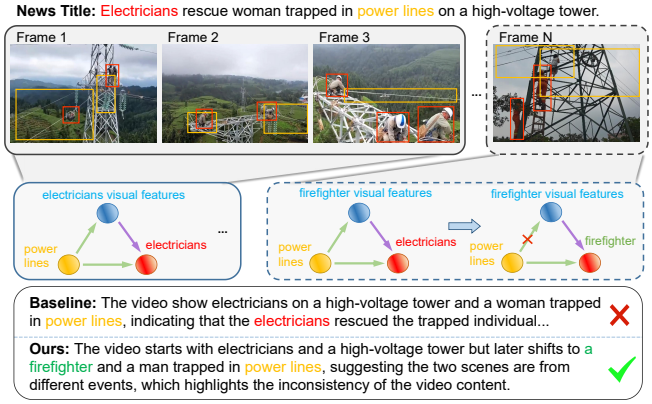

Figure 1: Examples of fake correlations in fake news videos, where the orange indicates visual semantic words that may lead to bias, and the generated correct and incorrect words are colored in green and red, respectively.

The first critical issue stems from video object bias, a phenomenon where spurious correlations between co-occurring objects distort feature representations. As illustrated in Figure 1, conventional FNVE systems employ pre-trained Faster R-CNN [Ren *et al.*, 2016] to extract region-based object features from video frames [Shang *et al.*, 2021]. However, our analysis reveals that when objects frequently co-occur, such models tend to encode entangled semantics. For instance, the visual features of electricians often absorb contextual attributes of power lines due to their high co-occurrence frequency in training data. This entanglement introduces a confounder effect: power lines act as object confounders,

*Corresponding author

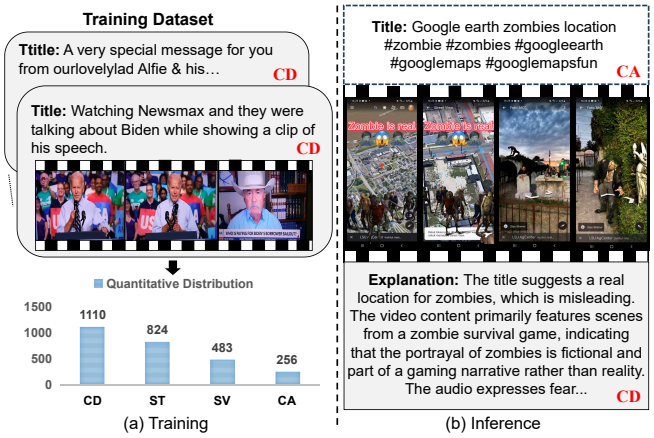

Figure 2: Explain spurious correlations between different aspects on the FNVE dataset.

creating a shortcut path [Geirhos *et al.*, 2020] that erroneously associates background elements (power lines) with foreground semantics (electricians). Consequently, when contextual noise dominates, the model may fail to recognize correct semantic concepts (e.g., identifying firefighters). Despite its prevalence, prior work has not explicitly addressed this bias in video representation learning, resulting in unreliable explanations.

The second critical issue arises from training bias in explanation aspects, a phenomenon unique to fake news explanation. Following Chen et al. [2025] taxonomy, we categorize explanations into four distinct aspects: Contextual Dishonesty (CD), Splice Tampering (ST), Synthetic Voiceover (SV), and Contrived Absurdity (CA). As shown in Figure 2 (a), our analysis of the FNVE dataset reveals significant class imbalance across these aspects, with CD-related samples substantially outnumbering CA examples. This skewed distribution mirrors real-world trends [Murdoch *et al.*, 2019] where certain manipulation techniques are more prevalent in fake news videos. During the inference phase (Figure 2 (b)), when presented with CA-aspect fake news videos, the model erroneously generates CD-oriented explanations due to overexposure to CD aspects during training.

To address these challenges, we formulate the Fake News Video Explanation (FNVE) generation process through a causal graph that explicitly models two key confounding factors: video object confounders and explanation aspect confounders. This leads to our proposed Causal Inference-based Fake news video Explanation (CIFE) framework, which primarily consists of two novel components: 1) the Interventional Video-Object Detector (IVOD), which integrates causal inference into Faster R-CNN [Ren *et al.*, 2016] to disentangle region-based visual semantics and effectively eliminate object-level confounding effects (e.g., separating the entangled features of co-occurring objects); and 2) the Interventional Explanation Aspect Module (IEAM), which applies causal intervention to Transformer-encoded [Lewis *et al.*, 2021] multimodal representations to mitigate aspect-related biases during explanation generation. The final explanations are produced by a Transformer decoder that operates on these debiased representations, yielding more reliable and robust outputs. Our contributions are summarized as follows:

- We propose CIFE, a fake news video explanation framework based on SCM, which explicitly decouples object-level semantic entanglement and explanation aspect distribution bias through causal intervention, overcoming the limitations of traditional methods that rely on correlation learning.
- We design the IVOD module to decouple visual features in dynamic scenes and the IEAM module to balance aspect preferences in multimodal explanation generation. Their synergy effectively eliminates spurious correlations for low-frequency forgery types.
- Our causal framework CIFE is compatible with existing multimodal models. On the FakeVE benchmark, it achieves improvements of 16.2% in BLEU-1 and 20.1% in ROUGE-L, with extensive experiments validating the critical role of causal intervention in ensuring explanation faithfulness.

2 Related Work

2.1 Multimodal Fake News Explanation

Multimodal fake news detection aims to verify the veracity of multimodal news posts. Current research has extensively explored diverse dimensions including linguistic patterns [Przybyla, 2020], image quality [Cao *et al.*, 2020], and cross-modal inconsistencies [Zhou *et al.*, 2020]. However, several recent works [Yao *et al.*, 2023; Qi *et al.*, 2024] have attempted to analyze veracity through natural language generation. For instance, Yao et al. [2023] proposed an end-to-end multimodal fact-checking and explanation generation framework that predicts veracity by retrieving relevant evidence and generates explanatory statements.

More recently, Chen et al. [2025] introduced the Fake News Video Explanation (FNVE) task, accompanied by an expert-annotated dataset supporting veracity analysis of news videos. Their approach adopted a generative language model [Lewis *et al.*, 2021] as backbone and incorporates a multimodal relation graph to comprehensively characterize cross-modal relationships. Despite significant progress in explanatory paradigms, the challenge of addressing semantic explanation biases induced by video-text confounders in FNVE remains largely underexplored.

2.2 Causal Intervention

Recently, researchers have begun exploring novel approaches to integrate causal reasoning into deep learning models. These efforts have significantly enhanced the performance of computer vision and natural language processing models across various tasks, including image classification [Lopez-Paz *et al.*, 2017], semantic segmentation [Yue *et al.*, 2020], visual feature representation [Wang *et al.*, 2020], image captioning [Liu *et al.*, 2022], and dialogue generation [Zhu *et al.*, 2020]. In the domain of fake news detection, Zhu et al. [2022] addressed entity bias from a causal perspective to improve model generalizability to future data, while Zeng et al.

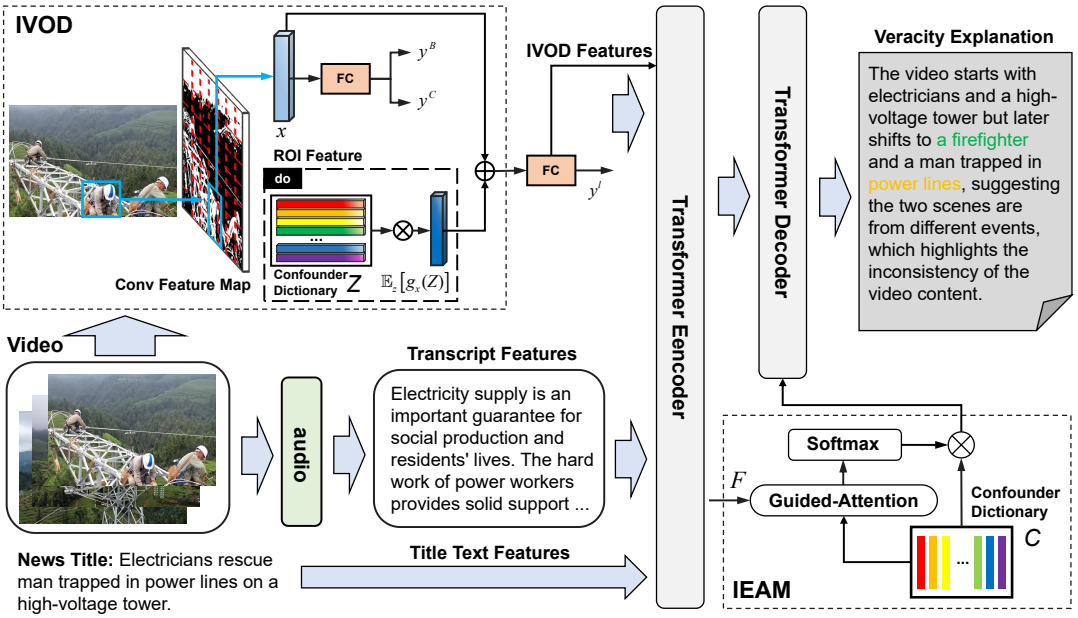

Figure 3: Illustration of the CIFE framework for fake news video explanation.

[2023] proposed a multimodal debiasing framework that pioneered the mitigation of various multimodal biases in fake news detection. However, these approaches remain insufficient in analyzing confounding factors specific to news video content.

3 Method

As illustrated in Figure 3, the CIFE framework employs a Transformer-based architecture comprising both encoder and decoder components, with causal reasoning strategically incorporated at two stages: 1) during the visual representation phase for video processing, where we implement intervention mechanisms to address object-level confounding effects; and 2) in the fusion module following multimodal encoder processing, where we apply causal adjustments to mitigate aspect-related biases in the combined representations.

3.1 Interventional Video-Object Detector

This section investigates causal inference in video object detection, aiming to address biased visual representations induced by object-level confounding factors.

Causal Intervention in Video-Object Detection

In causal graphs [Chalupka *et al.*, 2017; Wang *et al.*, 2024b], a variable is defined as a confounder when it serves as a common cause of two other variables. As illustrated in Figure 4(a), we construct causal relationships among region-based visual features X, visual confounder Z, and category labels Y based on Structural Causal Models (SCM) [Chalupka *et al.*, 2017], where straight edges denote direct causal connections between variables. The confounding mechanism operates through two pathways: 1) The causal effect $Z \rightarrow X$ occurs because visual features extracted during Faster R-CNN's classifier training inevitably incorporate contextual influences

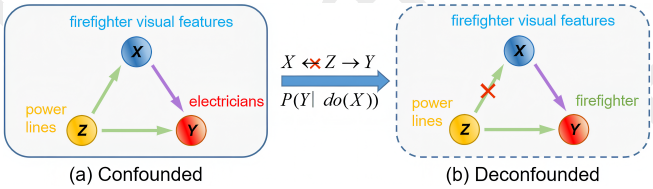

Figure 4: The causal intervention in object detection.

from real-world visual scenes; 2) Simultaneously, the causal effect $Z \rightarrow Y$ emerges as visual context systematically biases the classifier's probability outputs. Consequently, under dataset bias conditions, Faster R-CNN learns spurious associations between X and Y induced by Z - primarily by over-utilizing coincidental co-occurrences between visual contexts and category labels - resulting in biased visual representations of image regions.

As shown in Figure 4, conventional object detectors, such as Faster R-CNN, essentially use the likelihood $P(Y \mid X)$ as the training objective for the classifier, but this is usually affected by the confounder Z, which leads to false associations. To see this, we formulate $P(Y \mid X)$ as follows:

$$P(Y \mid X) = \sum_{z} P(Y \mid X, Z = z) P(Z = z \mid X), \quad (1)$$

where the confounding factor Z introduces observational bias through $P(z \mid X)$. As shown in Figure 4(a), when $P(z = power\ lines \mid X = electricians)$ is significantly large while $P(z = firefighter \mid X = electricians)$ remains relatively small, according to Equation (1), $P\left(Y = l_{electricians} \mid X, z = power\ lines\right)$ dominates over $P\left(Y = l_{electricians} \mid X, z = firefighter\right)$ ($l_{electricians}$ denotes the category label of electricians). Consequently, the classifier erroneously learns spurious associ-

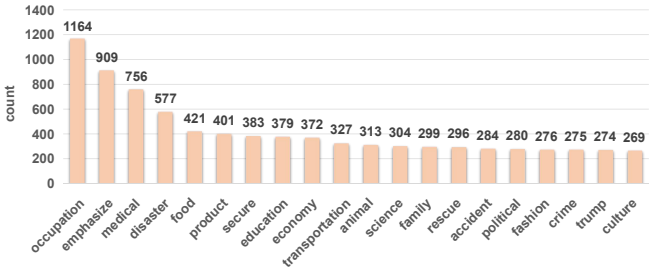

Figure 5: Frequency statistics of objects in the top 20 categories.

ations between visual features of power lines and the class label of electricians - meaning the learned RoI features of electricians actually represent visual characteristics of their surrounding power lines rather than the authentic features of electricians themselves.

Recently, several studies have applied causal inference to deep learning [Wang *et al.*, 2020; Liu *et al.*, 2022; Chen *et al.*, 2023] with success. Inspired by this, we introduce causal intervention $P(Y \mid do(X))$ into the object detection task to block the backdoor path $X \leftarrow Z \rightarrow Y$, where the do computation $do(\bullet)$ is responsible for cutting the path $Z \rightarrow X$. As shown in Figure 4, the backdoor adjustment [Wang *et al.*, 2020; Chalupka *et al.*, 2017] is calculated as follows:

$$P(Y \mid do(X)) = \sum_{z} P(Y \mid X, Z = z)P(Z = z). \quad (2)$$

where the do-operator $P(Y \mid do(X))$ compels X to uniformly consider all confounders z in the confounding set during prediction of Y. The resulting visual representations consequently achieve higher quality by capturing genuine semantic features rather than artifact correlations, with the intervention effectively '*borrowing*' counterfactual scenarios to establish robust causal dependencies.

However, when Eq. (2) is applied to the deep object detection network, the sampling cost will become large, and the overhead of training time will also increase, resulting in the infeasibility of evaluating $P(Y \mid do(X))$. Fortunately, by applying the approximate representation of Normalized Weighted Geometric Mean (NWGM) [Wang *et al.*, 2020; Xu *et al.*, 2015], Eq. (2) can be approximated as follows:

$$P(Y \mid do(X = x)) \approx P(Y|concat(x, \frac{1}{n}\sum_{i=1}^{n} P(y_i^c \mid x)z_i)), \quad (3)$$

where $concat(\bullet)$ denotes the concatenation of vectors, y_i^c is the i class label, and $P(y_i^c \mid x)$ is the probability output of the pre-trained classifier, indicating that x belongs to class y_i^c.

For constructing the video object confounder set Z, we employ MLLM [OpenAI, 2023] to extract diverse keywords covering topics, objects, and events from the dataset. These keywords are clustered into 200 semantically similar and mutually exclusive categories (e.g., '*occupation*', '*disaster*'), and a representative generic word is generated for each category. These generic words summarize the features of each class in the data. Figure 5 displays the representations of the top 20 generic category words. Using Faster R-CNN detected visual object features, we compute the mean Region of Interest (ROI) features across all samples within each category z_i, ultimately forming an $N \times d$ dimensional confounder dictionary matrix $Z = [z_1, z_2, \ldots, z_N]$, where $N = 200$ denotes the vocabulary size. This confounder dictionary supports dynamic expansion through: 1) semantic similarity-based assignment of novel words to existing categories, and 2) incremental updating of corresponding feature means, thereby maintaining representational completeness while accommodating new visual concepts.

IVOD Architecture

In Figure 3, we propose a novel IVOD network to extract the visual features of video decoupling, where Faster R-CNN [Ren *et al.*, 2016] is used as the visual backbone network. In IVOD, we use the same bounding box regressor as Faster R-CNN to specify each RoI on the feature map. As shown in Figure 3, the RoI feature x is then fed into two parallel branches to predict the class probability output y^C and the bounding box y^B, respectively. Finally, based on the RoI features x, the category probability output y^C, and the predefined confounder dictionary Z, we perform calculus to implement the interventional category predictor and output the final object category label y^l. In this way, the RoI feature x can be effectively decoupled and, when adopted, can facilitate the transformer decoder to generate unbiased FNVE.

3.2 Multi-Modal Transformer Encoder

For the visual representation in news videos, we utilize IVOD to extract disentangled visual object semantic word features from any Region of Interest (RoI) proposals (referred to as IVOD features $X_I \in \mathbb{R}^{K \times d}$, where K denotes the size of visual semantic words). For the audio representation in news videos, we consider that audio can convey the emotional semantics of the video creators and live reports, encompassing global contextual relationships within the video. Therefore, we convert the audio into a transcript denoted as $A\,[a_1, a_2, \cdots, a_M]$, where a_i represents a token in the audio transcript. Additionally, we extract the textual embedding features of the news video's title, denoted as $T = [t_1, t_2, \cdots, t_L]$, where $t_m \in T$ represents the feature of a word, and L is the length of the news title text.

To encode the above features, we turn to transformer encoders, which have shown convincing success on various natural language processing tasks. We first concatenate them denoted $X = \{T, X_I, A\}$, and then feed X into the transformer encoder TE as follows:

$$F = TE(X), \quad (4)$$

where $F \in \mathbb{R}^{S \times D}$ is the coded representation matrix, S is the total number of tokens in X. It is obvious that the TE encoder generates more discriminative visual features for decoding by deep stacking.

3.3 Interventional Explanation Aspect Module

To mitigate the spurious correlations between visual features and their corresponding explanation aspects, we propose a novel IEAM to address the confounding factors between visual and explanation aspects in the FNVE task. First, based

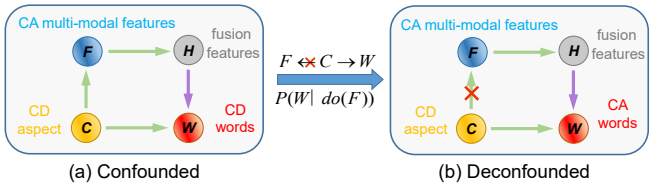

Figure 6: The causal intervention $P(W \mid do(F))$ in explanation aspect.

on the SCM, we establish causal relationships among the multimodal features F fused by the Transformer encoder, the explanation aspect C, the fused features H, and the predicted word W, as illustrated in Figure 6(a). Specifically, the causal effect $H \rightarrow W$ indicates that the encoded multimodal features trigger the generation of the corresponding word. The causal effect $C \rightarrow F$ reflects the influence of C on F, as the FNVE generation model, when trained, causes the multimodal features fused by the Transformer encoder to be heavily affected by frequently occurring explanation contexts. The effect $F \rightarrow H$ represents the use of the fused features H to infer the next word W. Consequently, when the observational likelihood $P(W|F)$ is used as the objective, the decoder may learn spurious explanation associations between F and W due to the presence of the confounder C. To elucidate the principle of causal intervention in FNVE, we formulate $P(W|F)$ as follows:

$$P(W \mid F) = \sum_{c} P(W \mid F, C)P(C \mid F), \tag{5}$$

where the confounding variable C usually introduces observation bias through $P(C \mid F)$. Similar to intervention manipulation (IVOD), we replace the traditional FNVE training objective with causal intervention $P(W \mid do(F))$, which aims to eliminate the causal influence of C on F, as shown in Figure 6 (b). Thus, a backdoor path $F \leftarrow C \rightarrow W$ is blocked, thus eliminating spurious correlations. Assuming that the confounding variable C can be stratified, the $P(W \mid do(F))$ can be calculated according to the backdoor adjustment [Wang *et al.*, 2020; Chalupka *et al.*, 2017] formula as follows:

$$P(W \mid do(F)) = \sum_{c} P(W \mid F, C)P(C). \tag{6}$$

Therefore, based on the interventional probability in Eq.(6), the FNVE device is forced to learn the true causal effect $F \rightarrow W$ instead of the spurious correlation caused by the explanation aspect confounding variable C.

To construct the explanation aspect confounder dictionary C, we first follow the theory of Chen et al. [Chen *et al.*, 2025] to categorize news video explanations into four distinct aspects (CD, ST, SV, and CA). Then, leveraging the powerful semantic generation capability of MLLM, we automatically annotate the explanation aspect for each news video and generate corresponding logical rationales. These are concatenated to form the initial explanation aspect space $W_e \in \mathbb{R}^{U \times d_e}$, where U represents the size of the entire dataset. To ensure compatibility with model fusion requirements, we further transform W_e into a confounder dictionary C in a common d-dimensional space using a learnable linear projection matrix $P_w \in \mathbb{R}^{d_e \times d}$. This allows seamless integration with models of varying dimensions.

3.4 Veracity Explanation Generation

We input the corresponding entries from the explanation aspect confounder dictionary C, obtained by IEAM, for each batch into the guided attention module [Sun *et al.*, 2023] to derive the fused intervention representation, which we define as E^I. Finally, we input $F + E^I$ into the pre-trained Transformer decoder TD. The decoder operates in an autoregressive manner, generating the next word by considering all previously decoded outputs, as follows:

$$\widehat{\mathbf{y}}_t = TD\left(F + E^I, \hat{\mathrm{Y}}_{<t}\right), \tag{7}$$

where $t \in [1, N_y]$ and $\widehat{\mathbf{y}}_t \in \mathbb{R}^{|\mathcal{V}|}$ are the t-th token probability distributions of the truthfulness explanation. $\hat{Y}_{<t}$ refers to the previously predicted $t-1$ labeling.

To optimize the generation of CIFE, we again employ the standard cross-entropy loss function as follows:

$$\mathcal{L}_{Gen} = -1/N_y \sum_{i=1}^{N_y} \log\left(\widehat{\mathbf{y}}_i[t]\right), \tag{8}$$

where $\widehat{\mathbf{y}}_i[t]$ is the element of $\widehat{\mathbf{y}}_i$ corresponding to the i token of the generated explanation, and N_y is the total number of tokens in the generated veracity explanation Y.

Split	#of News	Avg.Title	Avg. Dur (s)	Avg. Exp
Train	2138	21.40	61.23	49.76
Val	267	16.17	63.45	50.50
Test	267	15.37	60.32	50.08
Total	2672	20.27	61.78	49.86

Table 1: Statistics of the FakeVE dataset, where "Avg.". "Dur." and "Exp." refer to "Average", "Duration" and "Explanation", respectively.

4 Experiments

4.1 Datasets

We conducted experimental research on FakeVE [Chen *et al.*, 2025], currently the largest and most comprehensive publicly available FNVE dataset, with brief statistical details presented in Table 1. FakeVE comprises 2,672 fake news video samples collected from three major social platforms: Twitter, YouTube, and TikTok. The dataset innovatively categorizes fake content into four distinct aspects, with each news video sample featuring precise frame-level multimodal content analysis, comprehensively covering 12 news topics including politics, health, and disasters.

4.2 Comparative Models

The CIFE framework can be applied to any fake news video explanation method with news headlines and videos as input. Therefore, we apply the CIFE framework to the following three strong baselines: 1) HAAV [Kuo and Kira, 2023],

Method	BLEU				M	Rouge			Sent-B
	B@1	B@2	B@3	B@4		R-1	R-2	R-L	
MLLM-based approach									
Qwen2-VL	25.94	11.92	7.76	5.66	75.89	28.84	8.66	12.94	69.23
LLaVA	29.32	15.42	9.13	6.21	79.75	31.34	9.27	16.60	74.02
GPT-4o	32.59	17.33	11.23	6.70	82.84	32.57	9.78	17.8	77.79
Fine-tuning approach									
HAAV	34.45	16.90	9.23	4.78	84.67	28.12	8.90	18.45	75.67
W/ CIFE	**39.32**	**22.34**	**14.36**	**10.68**	**86.21**	**34.58**	**13.26**	**22.56**	**76.23**
AMFM	35.67	18.89	10.01	6.12	86.68	33.45	10.89	22.67	78.12
W/ CIFE	**39.23**	**23.45**	**16.81**	**10.83**	**87.82**	**38.65**	**15.37**	**25.34**	**80.72**
MRGT	39.39	22.44	13.08	8.21	91.62	38.07	12.34	27.21	84.66
W/ CIFE	**45.77**	**28.65**	**18.54**	**12.57**	**93.70**	**45.64**	**18.14**	**32.66**	**85.22**

Table 2: Results of comparison among different models on FakeVE dataset, where the best results are in bold. B@1, B@2, B@3, B@4, M, R-1, R-2, R-L and Sent-B are short for BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-1, ROUGE-2, ROUGE-L and SentBERT.

which independently encodes each view in the multimodal input through a shared encoder; 2) AMFM [Zhang *et al.*, 2024], which dynamically enhances pre-trained visual features by learning latent visual relationships between frame-level and video-level embeddings; and 3) MRGT [Chen *et al.*, 2025], which builds upon BART as its backbone and introduces a multimodal relation graph to capture intrinsic connections between visual and semantic elements.

We also chose MLLMs for comparison: 1) Qwen2-VL [Wang *et al.*, 2024b], a multimodal series of large language models developed by Alibaba Cloud's Qwen team with advanced image and video understanding capabilities; 2) LLaVA [Liu *et al.*, 2023] an end-to-end trained large multimodal model designed to comprehend and generate content based on visual inputs and textual instructions; and 3) GPT-4o [OpenAI, 2023], the latest iteration of GPT-4 Omni developed by OpenAI, which facilitates rapid deployment and integration.

Model	B@1	B@4	M	R-L	Sent-B
MRGT w/ CIFE	**45.77**	**12.57**	**93.70**	**32.66**	**85.22**
w/o IVOD	41.53	9.13	92.54	29.73	84.76
w/o IEAM	43.23	11.46	92.23	30.42	84.94
w/o CIFE	39.39	8.21	91.62	27.21	84.66

Table 3: Experiments on the impact of ablative causal inference.

4.3 Experimental Setup

Following the previous work [Chen *et al.*, 2025], we selected several standard metrics for evaluating the performance of generated explanations. These include BLEU [Papineni *et al.*, 2002] (BLEU-1, BLEU-2, BLEU-3, BLEU-4), which measures the similarity between the generated text and the reference text. ROUGE [Chin-Yew, 2004] (ROUGE-1, ROUGE-2, ROUGE-L), which evaluates the similarity between the generated summary and the reference summary. METEOR [Banerjee and Lavie, 2005], for explanation quality assessment of synonym matching strategies. SentBERT [Reimers and Gurevych, 2019], which uses Sentence-BERT to embed semantic similarity between reference and generated explanations in the space.

When integrating the CIFE framework into baseline models, IVOD and IEAM are incorporated in a flexible manner rather than directly replacing existing modules. Specifically, IVOD serves as a visual intervention feature extractor that is inserted during the feature extraction stage to derive causally-intervened visual features. Meanwhile, IEAM operates prior to the decoding generation phase, where it fuses the explanation-intervened features with the baseline encoded features for subsequent decoding. During training, we uniformly sample video frames with a maximum sequence length of 55 frames per video, applying pooling operations to each frame as the visual source representation. We employ AdamW [Loshchilov and Hutter, 2017] as the optimizer with a learning rate of 1e-4, a batch size of 16, and train the models for a maximum of 15 epochs.

4.4 Experimental Results

As shown in Table 2, the experimental results demonstrate the significant advantages of the CIFE framework in fake news video explanation tasks. Compared with MLLMs, general-purpose models like GPT-4o exhibit acceptable performance on basic metrics (e.g., BLEU-1 reaching 32.59) but show notable limitations in fine-grained explanation generation, highlighting the necessity of domain adaptation. In contrast, CIFE-enhanced fine-tuned models (e.g., MRGT+CIFE) achieve substantial improvements across key metrics, validating the value of causal intervention in specialized tasks. Particularly noteworthy is CIFE's outstanding improvement in ROUGE-2 (e.g., +4.48 for AMFM+CIFE), indicating its effectiveness in capturing key factual units, while the consistent enhancement in SentBERT scores demonstrates that the framework maintains global semantic coherence while optimizing local features.

The superiority of CIFE stems from its proposed debiasing modules: IVOD disentangles visual confounders, enabling accurate identification of key objects, while IEAM balances aspect distributions in multimodal representations to prevent

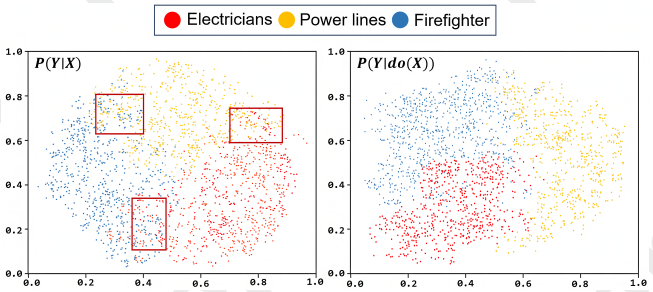

Figure 7: The t-SNE visualization [Van der Maaten and Hinton, 2008] of some object features extracted by Faster R-CNN (left) and IVOD (right).

bias in explanation generation. This synergistic effect leads to consistent improvements across different baseline models, with the most significant enhancement observed for MRGT (ROUGE-L +5.45). The results confirm that explicitly modeling and intervening on confounders not only addresses the limitations of general MLLMs in specialized domains but also generates more accurate and explanatory results while preserving semantic coherence.

4.5 Ablation Study

Effectiveness of Each Component

We used the FakeVE test sets in combination with the powerful baseline model MRGT to deeply explore the bias effect of each module in CIFE. As shown in Table 3, we tested the performance of CIFE for removing the IVOD part (w/o IVOD) and CIFE for removing the IEAM part (w/o IEAM). In the w/o IVOD model, we ignore the confounder of semantic words related to object and rely only on the regional features of the original video frames to generate explanations. Although the w/o IEAM model debiases the object semantics, it does not debiases the explanation aspect. Experimental results show that the performance degrades when removing the IVOD part, which proves the effectiveness of IVOD in removing the object semantic word bias in videos. Similarly, removing the IEAM part also leads to a performance slippage, which further validates the positive role of the CIFE framework in mitigating the explanation aspect bias.

Analysis on Biases

Figure 7 shows some visual features of FNVE video frames extracted by Faster R-CNN (left) and proposed IVOD (right). We can see that our IVOD can learn more discriminative feature representations compared to Faster R-CNN. For example, electricians and firefighter features and power lines when using regular likelihood (left). The features are entangled in the red box. They are clearly untangled after the causal intervention (right), meaning that IVOD actually obfuscates visual confounders when extracting visual features.

4.6 Qualitative Analysis

Figure 8 shows two examples of good interpretation of CIFE compared to MRGT. In Figure 8(a) SpaceX case, CIFE effectively decoupage the rocket body features from the visual elements of the explosion scene through the IVOD module, accurately identifying the lack of the unique design of the

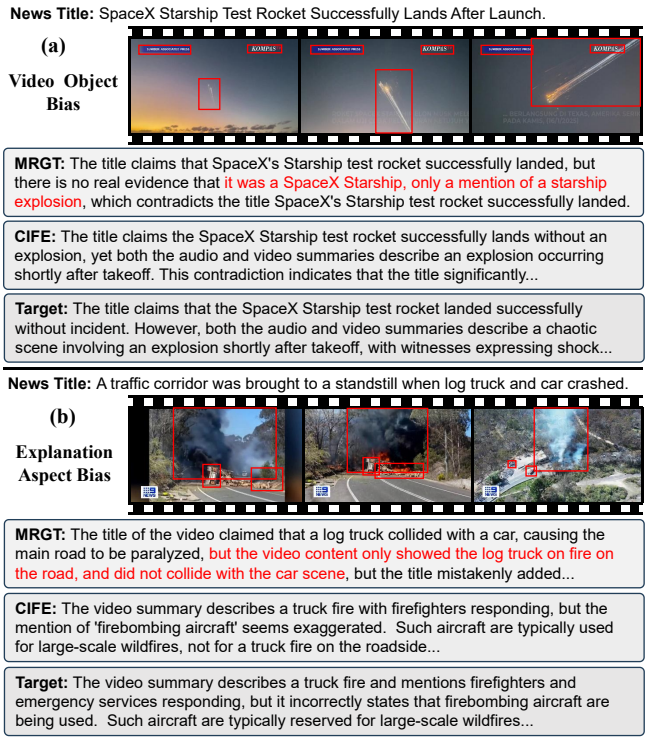

Figure 8: Visualization of fake news video explanations generated by CIFE and the MRGT baseline under both video object and aspect conditions. Target denotes the human-annotated ground-truth explanation. Red highlights indicate segments that led to biased explanations in the generated outputs.

Starship wreckage, thus generating a more accurate explanation: '*The video shows a rocket exploding, but the debris signatures are not consistent with the SpaceX Starship.*' In contrast, MRGT can only detect surface contradictions. In Figure 8(b) health policy case, CIFE's IEAM module successfully detected a lack of "legislative process", which, combined with IVOD's character identification capability, clearly stated that '*the speaker is a representative of civil society organizations and has no legislative power*', while MRGT was dominated by the high-frequency '*tax policy*' aspect and failed to detect this key contradiction.

5 Conclusion

In this paper, we propose CIFE, a novel causal intervention framework for generating explainable fake news video explanations. Through structural causal modeling, CIFE effectively addresses two key challenges: resolving object-level confounding via the Intervened Video-Object Detector (IVOD), and mitigating explanation-aspect bias through the Intervened Explanation-Aspect Module (IEAM) to ensure balanced reasoning across different falsification aspects. Extensive experiments on the FakeVE benchmark demonstrate CIFE's superiority in effectively eliminating biases and enhancing fake news video explanation generation. Future work will focus on optimizing the dictionary of confounding variables by introducing knowledge graphs.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Nos. 62276177 and 62376181), Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions, and Natural Science Foundation of the Jiangsu Higher Education Institutions of China (No. 24KJB520036).

References

[Banerjee and Lavie, 2005] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[Cao *et al.*, 2020] Juan Cao, Peng Qi, Qiang Sheng, Tianyun Yang, Junbo Guo, and Jintao Li. Exploring the role of visual content in fake news detection. *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*, pages 141–161, 2020.

[Chalupka *et al.*, 2017] Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Causal feature learning: an overview. *Behaviormetrika*, 44:137–164, 2017.

[Chen *et al.*, 2023] Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 627–638, 2023.

[Chen *et al.*, 2025] Lizhi Chen, Zhong Qian, Peifeng Li, and Qiaoming Zhu. Multimodal fake news video explanation generation. *CoRR*, abs/2501.08514, 2025.

[Chin-Yew, 2004] Lin Chin-Yew. Rouge: A package for automatic evaluation of summaries. In *In Proceedings of the Workshop on Text Summarization Branches Out*, pages 74–81, 2004.

[Choi and Ko, 2021] Hyewon Choi and Youngjoong Ko. Using topic modeling and adversarial neural networks for fake news video detection. In *In Proceedings of the 30th ACM international conference on information knowledge management*, pages 2950–2954, 2021.

[Geirhos *et al.*, 2020] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[Kuo and Kira, 2023] Chia-Wen Kuo and Zsolt Kira. Haav: Hierarchical aggregation of augmented views for image captioning. In *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11039–11049, 2023.

[Lewis *et al.*, 2021] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2021.

[Liu *et al.*, 2022] Bing Liu, Dong Wang, Xu Yang, Yong Zhou, Rui Yao, Zhiwen Shao, and Jiaqi Zhao. Show, deconfound and tell: Image captioning with causal inference. In *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18041–18050, 2022.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[Lopez-Paz *et al.*, 2017] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. Discovering causal signals in images. In *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6979–6987, 2017.

[Loshchilov and Hutter, 2017] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017.

[McCrae *et al.*, 2022] Scott McCrae, Kehan Wang, and Avideh Zakhor. Multi-modal semantic inconsistency detection in social media news posts. In *In International Conference on Multimedia Modeling*, pages 331–343, 2022.

[Murdoch *et al.*, 2019] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019.

[Nakov and Martino, 2021] Preslav Nakov and Giovanni Da San Martino. Fake news, disinformation, propaganda, media bias, and flattening the curve of the COVID-19 infodemic. In *In Proceedings of the 27th ACM SIGKDD conference on knowledge discovery data mining*, pages 4054–4055, 2021.

[OpenAI, 2023] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *In Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[Przybyla, 2020] Piotr Przybyla. Capturing the style of fake news. In *In Proceedings of the AAAI conference on artificial intelligence*, pages 490–497, 2020.

[Qi *et al.*, 2024] Peng Qi, Zehong Yan, Wynne Hsu, and Mong-Li Lee. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13052–13062, 2024.

[Reimers and Gurevych, 2019] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3980–3990, 2019.

[Ren *et al.*, 2016] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[Schoenherr and Thomson, 2020] Jordan Richard Schoenherr and Robert Thomson. Health information seeking behaviour, risk communication, and mobility during COVID-19. In *In 2020 IEEE International Symposium on Technology and Society (ISTAS)*, pages 283–289, 2020.

[Shang *et al.*, 2021] Lanyu Shang, Ziyi Kou, Yang Zhang, and Dong Wang. A multimodal misinformation detector for COVID-19 short videos on tiktok. In *In 2021 IEEE international conference on big data (big data)*, pages 899–908, 2021.

[Sun *et al.*, 2023] Tiening Sun, Zhong Qian, Peifeng Li, and Qiaoming Zhu. Graph interactive network with adaptive gradient for multi-modal rumor detection. In *In Proceedings of the 2023 ACM international conference on multimedia retrieval*, pages 316–324, 2023.

[Sundar *et al.*, 2021] S. Shyam Sundar, Maria D. Molina, and Eugene Cho. Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps? *J. Comput. Mediat. Commun.*, 26(6):301–319, 2021.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[Wang *et al.*, 2020] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense R-CNN. In *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10760–10770, 2020.

[Wang *et al.*, 2024a] Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. Explainable fake news detection with large language model via defense among competing wisdom. In *In Proceedings of the ACM Web Conference 2024*, pages 2452–2463, 2024.

[Wang *et al.*, 2024b] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191, 2024.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *In Proceedings of the 32nd International Conference on Machine Learning*, pages 2048–2057, 2015.

[Yao *et al.*, 2023] Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2733–2743, 2023.

[Yue *et al.*, 2020] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. *Advances in neural information processing systems*, 33:2734–2746, 2020.

[Zhang *et al.*, 2021] Zijian Zhang, Koustav Rudra, and Avishek Anand. Explain and predict, and then predict again. In *In Proceedings of the 14th ACM international conference on web search and data mining*, pages 418–426, 2021.

[Zhang *et al.*, 2024] Yunzuo Zhang, Yameng Liu, and Cunyu Wu. Attention-guided multi-granularity fusion model for video summarization. *Expert Systems with Applications*, 249:123568, 2024.

[Zhou *et al.*, 2020] Xinyi Zhou, Jindi Wu, and Reza Zafarani. SAFE: similarity-aware multi-modal fake news detection. In *In Pacific-Asia Conference on knowledge discovery and data mining*, pages 354–367, 2020.

[Zhu *et al.*, 2020] Qingfu Zhu, Weinan Zhang, Ting Liu, and William Yang Wang. Counterfactual off-policy training for neural response generation. *CoRR*, abs/2004.14507, 2020.

[Zhu *et al.*, 2022] Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. Generalizing to the future: Mitigating entity bias in fake news detection. In *In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2120–2125, 2022.