Graphing the Truth: Harnessing Causal Insights for Advanced Multimodal Fake News Detection

Anonymous ACL submission

Abstract

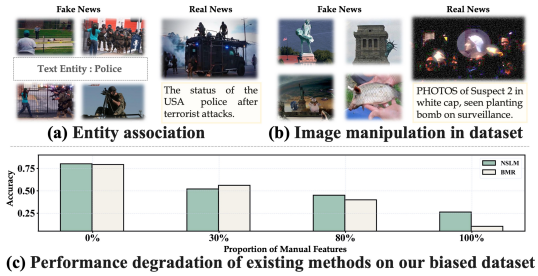
Fake news data, often sampled from the same communities, results in the veracity of news being highly correlated with certain textual and visual entities. This correlation leads fake news classification models to be prone to shortcut learning, quickly overfitting by capturing only shallow spurious correlations between labels and features. Consequently, neural networks trained on such data suffer from poor generalization and potential misclassification under distribution shifts. In this paper, we propose a **DI**sentanglement-based **C**ausality-awar**E** fake news detection method (DICE). DICE constructs multimodal news into a graph neural network and effectively models causal relationships between multimodal features and veracity labels through the use of node and edge mask disentanglers. To reinforce this disentanglement process, we designed a loss function aimed at minimizing extrapolation risk, which supervises the training and results in disentangled causal and biased representations of news. Extensive experiments demonstrate that DICE achieves superior performance on five large-scale fake news detection benchmarks. Additionally, our evaluation on a heavily biased fake news dataset demonstrates DICE's strong generalization, suggesting its potential to inform a new paradigm in causal fake news detection.

1 Introduction

With the advancement of generative artificial intelligence technology, fake news has increasingly become a tool on social media for influencing public opinion and manipulating information. In the context of hot topics, conflicting parties manufacture and disseminate false information to confuse the public, tarnish competitors, and mislead the masses, ultimately achieving manipulation of online public opinion (Aïmeur et al., 2023; Yin et al., 2024). The proliferation of well-crafted multimodal fake information has heightened the necessity and urgency of developing effective methods for identifying multimodal fake news, as such content is more likely to capture readers' attention.

Existing paradigms for deep learning-based multimodal fake news detection methods focus on encoding textual and visual information into task-relevant representations in a latent feature space (Wang et al., 2018; Chen et al., 2019). Early methods concentrated on the joint modeling of posts and images (Khattar et al., 2019; Zhang et al., 2021; Wu et al., 2023; Zhang et al., 2024a) or estimating consistency (Qi et al., 2021; Chen et al., 2022). Later approaches introduced multi-view fake news modeling (Wu et al., 2021; Qian et al., 2021; Ying et al., 2023) or incorporated logical connections between news and labels (Liu et al., 2023; Dong et al., 2024). However, due to the complex news environment, model bias remains a critical unresolved issue in this field. This bias often arises from distributional shifts present in the training data, causing the model to overly rely on shallow associations between textual and visual features and the labels. For instance, while features such as clickbait (Bourgonje et al., 2017), image manipulation (Wang et al., 2024), or sensationalism (Subbiah et al., 2023) are often correlated with fake news, the authenticity of news cannot always be determined solely based on these surface-level patterns. These features may represent causal relationships in certain contexts but can also result in spurious correlations due to biased data distribution. Models trained without awareness of these nuances risk relying on these spurious correlations, hampering their generalization ability across unseen scenarios. It is thus crucial to disentangle causally relevant patterns from spurious biases to ensure robust detection.

Furthermore, deep neural networks are highly sensitive to distributional shifts (Gawlikowski et al., 2023). Shortcut learning, where models rely on shallow correlations (e.g., specific visual features



(c) Performance degradation of existing methods on our biased dataset

Figure 1: Models habitually learn spurious associations between features and labels through shortcut learning.

or text patterns) rather than causal relationships, exacerbates this issue (Fan et al., 2024). As noted by Liu et al. (2023), fake news in existing datasets is often correlated with sensational language, leading to biased predictions. Similarly, Zhu et al. (Zhu et al., 2022) highlights how entity bias creates spurious correlations that hinder model generalization. These biases often arise from the temporal or contextual sampling of training data, where older events or entities dominate, resulting in distributional shifts between the training and test datasets.

Such issues are vividly exemplified in widely-used datasets. For instance, in the Pheme dataset (Zubiaga et al., 2017), which is constructed from real-world data sampled during numerous breaking news events, the frequent co-occurrence of "police" with fake news labels in training data leads to misclassification (Figure.1 a). Similarly, on benchmark Twitter (Detection and visualization of misleading content on Twitter, 2018), which are extensively adopted in fake detection studies, fake news often involves manipulated images, causing the model to associate spurious visual features with labels (Figure.1 b). However, real news can be manipulated to emphasize certain image aspects, resulting in classification errors.

These manipulations highlight the duality of certain features, such as manipulated visuals or sensational language: while they might reflect genuine patterns of fake news in certain cases, their presence in real news due to adversarial intent or biased data distribution can lead the model to establish false causal relationships. This reliance on shallow patterns results in misclassification of real news as fake news. To evaluate the effects of these biases, we constructed a dataset with opposing biases between the training and test sets (see Experiment for details). As shown in Figure 1(c), models relying on biased features exhibit a sharp decline in performance when tested on data with shifted distributions, confirming their inability to generalize effectively. This observation highlights the critical importance of addressing biases inherent in training datasets to ensure robust fake news detection.

In this paper, we propose a **DI**sentanglement-based **C**ausality-awar**E** fake news detection method (DICE). Instead of relying on potentially biased heuristics or pre-defined structures, DICE dynamically represents posts and images as a multimodal graph, where node and edge importance is adjusted through learnable "maskers". This approach disentangles causally relevant unbiased subgraphs from causally irrelevant biased subgraphs within the multimodal graph. Moreover, DICE incorporates a specialized loss function to adversarially encourage biased subgraphs to capture superficial shortcuts while compelling causal subgraphs to focus on features fundamentally linked to news authenticity, fostering a robust separation between bias-driven and causality-driven representations. Specifically, our contributions can be summarized as follows:

- We propose a novel fake news detection framework, DICE, which dynamically constructs multimodal graphs without predefined causal or biased relationships. DICE models intrinsic connections between multimodal features at the token level and incorporates a training process that minimizes feature-label loss variance across diverse bias environments. This approach fosters stable associations between features and news veracity, mitigating the influence of spurious correlations.
- To validate the robustness and generalization ability of our framework, we constructed a dataset with over 17,000 samples by sampling and restructuring data from widely used datasets. Handcrafted biases were introduced with varying types across splits, enabling systematic evaluation of the model's ability to maintain performance under distributional shifts.
- We tested our framework on five commonly used multimodal fake news datasets. Under the same experimental settings, our framework showed an average improvement of 2.68% in accuracy and 3.64% in F1-score.

2 Related Work

Multimodal Fake News Detection Multimodal fake news detection involves using multiple data modalities to detect fake news, requiring the integration and synergy of different modalities. On one hand, researchers enhance news representation by fusing multimodal features (Zhou et al., 2020; Zhang et al., 2021; Lao et al., 2024; Zhang et al.,

2024b,a; Wu et al., 2023). For instance, models like CAFE (Chen et al., 2022) have focused on semantic consistency as a key approach, conducting extensive research. Meanwhile, models like LogicDM (Liu et al., 2023) and NSLM (Dong et al., 2024) have incorporated symbolic logic to bring interpretability to multimodal fake news detection. However, as AI technology advances, the forms of fake news have become increasingly diverse. Models such as BMR (Ying et al., 2023) have adopted a multi-view learning approach, considering various levels such as textual sentiment, image manipulation, and semantic consistency to establish a unified framework for fake news detection (Wu et al., 2021; Qian et al., 2021). Yet, these methods often overlook the biases introduced by shortcut learning in models, resulting in suboptimal performance in real-world applications.

Causal Learning Causal learning has garnered attention in the machine learning field. Studies have shown that models tend to exploit biases as shortcuts for prediction (Mo et al., 2024). Leveraging causal relationships, many methods have achieved substantial success in various tasks (Liu et al., 2021; Wang et al., 2022). Some researchers have utilized the structure of graph neural networks to disentangle causal relationships, thereby providing ample interpretability for causal learning (Fan et al., 2022; Wu et al., 2022).

In the realm of fake news detection, models like ENDEF (Zhu et al., 2022) aim to mitigate entity bias from a causal perspective, extending fake news detection models to future datasets. Some works (Chen et al., 2023; Hu et al., 2022) employ counterfactual reasoning and causal intervention to eliminate psychological biases and image feature shifts. However, the shortcuts that models might rely on are diverse. These works address specific types of biases, necessitating a general framework to disentangle causal relationships in features.

3 METHODOLOGY

3.1 Problem Formulation and Causal Interpretation

Let (P, I) be the image-post pair in the dataset, where P and I denotes the news post and the image, respectively. We construct a multimodal graph $\mathcal{G}$ for each news sample, representing relationships between features extracted from the post and image. Our goal is to separate $\mathcal{G}$ into a causal subgraph $\mathcal{G}_c$, which contains features causally related to the news veracity label Y, and a biased subgraph $\mathcal{G}_b$, which contains features associated with Y but not causally related.

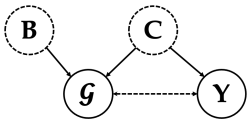


Figure 2: The SCM of the graph generation process. Dashed circles represent unobserved variables, while solid circles represent observed variables.

To achieve this, we disentangle the features within the original graph $\mathcal{G}$ into two sets of unobserved variables: causal variables C and non-causal variables B. In this decomposition, C influences both the multimodal news graph $\mathcal{G}$ and the downstream fake news detection task, while B only affects $\mathcal{G}$. Importantly, B and C are conditionally independent given $\mathcal{G}$, reflecting the structure of a collider $B \rightarrow \mathcal{G} \leftarrow C$. As illustrated in the Structural Causal Model (SCM) in Figure .2, where Y represents the ground truth, dashed circles denote unobserved latent variables, and solid circles denote observed variables.

In real-world testing scenarios, the variable B may change, but the conditional relationship $P(Y \mid C)$ remains invariant across different environments. The objective is to predict the label of a news sample as either 1 or 0 by identifying stable relationships between multimodal features and the news veracity label. Accordingly, we design a GNN model that prioritizes these stable associations while minimizing reliance on potentially unstable relationships involving B.

3.2 Model Design

Our framework, as illustrated in Figure.3, models complex multimodal interactions by disentangling causally relevant and irrelevant variables. Unlike prior causal inference works relying on predefined graph structures with fixed assumptions (Fan et al., 2022), our framework dynamically constructs fully connected graphs, allowing the model to learn node and edge importance directly from multimodal data during training. This design avoids reliance on prespecified features and adapts to the unique complexities of fake news detection.

Multimodal Graphs Construction For each news sample k, we construct a cross-modal graph $\mathcal{G}^k = (\mathcal{H}^k, \mathcal{E}^k)$, where $\mathcal{H}^k$ represents nodes and $\mathcal{E}^k$ represents edges encoded by the adjacency matrix $\mathbf{A}^k$. Starting with a fully connected graph ensures no potential causal relationships are ex-

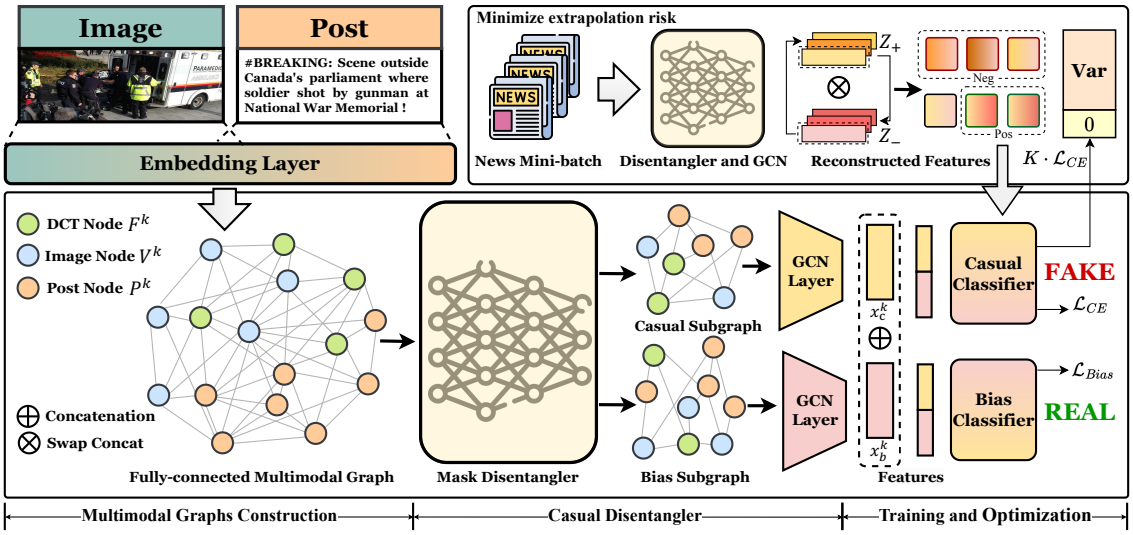


Figure 3: Overview of DICE. Initially, we extract semantic and frequency domain features of the images as well as the features of the tokens, constructing a fully connected multimodal graph. We employ a Mask Disentangler, comprising numerous MLPs, to determine the existence of nodes and edges, resulting in the derivation of causal subgraphs and bias subgraphs for classification. During training, we randomly combine causal and bias features from different samples within the mini-batch. Through risk extrapolation minimization and contrastive learning, we promote the disentanglement of causal features.

cluded prematurely. The causal disentangler refines this structure by learning the relative importance of connections, enabling the discovery of critical cross-modal causal paths.

For each post P^k, we tokenize it into m tokens and extract their features $\mathcal{T}^k = \{t_i^k \in \mathbb{R}^d\}_{i=1}^m$ using a frozen BERT (Devlin et al., 2019). Similarly, we split the image I^k into n patches and extract their semantic features $\mathcal{V}^k = \{v_j^k \in \mathbb{R}^d\}_{j=1}^n$ using a frozen ResNet (He et al., 2015). To further enhance image representation, especially for detecting image manipulation, we apply the Discrete Cosine Transform (DCT) (Liu and Li, 2003) to I^k, followed by a Multi-Head Self-Attention (MHSA) (Vaswani et al., 2017; Li et al., 2023) network, producing frequency domain features $\mathcal{F}^k = \{f_j^k \in \mathbb{R}^d\}_{j=1}^n$. The nodes of the graph $\mathcal{G}^k$ are defined as the union of token features, semantic image features, and frequency domain features, *i.e.*,

$$\mathcal{H}^k = \mathcal{T}^k \cup \mathcal{V}^k \cup \mathcal{F}^k.$$

These nodes are collectively represented as $H^k = [T^k, V^k, F^k] \in \mathbb{R}^{(m+2n)\times d}$, where T^k, V^k, and F^k are the matrix representations of $\mathcal{T}^k$, $\mathcal{V}^k$, and $\mathcal{F}^k$.

The edges $\mathcal{E}^k$ of the graph, which connect the nodes in $\mathcal{G}^k$, are represented by the adjacency matrix $\mathbf{A}^k$. To ensure that the model fully learns the causal patterns of multimodal fake news without introducing human bias, we adopt a fully connected strategy, where each element $A_{i,j}^k = 1$ indicates a connection between every pair of nodes.

Casual Disentangler When learning the multimodal graph $\mathcal{G}^k$, we leverage an MLP to quantify the degree of association si between each node and the veracity label Y, and the relationship e_{ij} between the features h_i^k and h_j^k of nodes i and j. These associations are modeled by concatenating the features and passing them through an MLP. We scale the outputs to the $[0, 1]$ range using the sigmoid function, *i.e.*,

$$s_i^k = \text{Sigmoid}(\text{MLP}(h_i^k)),$$
$$e_{ij}^k = \text{Sigmoid}(\text{MLP}([h_i^k, h_j^k])),$$

This produces the node mask $S^k \in \mathbb{R}^{m+2n}$ and the edge mask $E^k \in \mathbb{R}^{(m+2n)\times(m+2n)}$. These masks do not directly indicate causal relationships but guide the model to focus on features that align with stable associations.

Subsequently, we derive the unbiased subgraph $\mathcal{G}^k{}_c = \{A^k \odot E^k, H^k \odot S^k\}$ and the biased subgraph $\mathcal{G}^k{}_b = \{A^k \odot (1 - E^k), H^k \odot (1 - S^k)\}$ through a Hadamard product. These subgraphs are learned independently using two separate Graph Convolutional Networks (GCNs) (Kipf and Welling, 2017), defined as:

$$H_c^k = \text{ReLU}(\tilde{A}^k(H^k \odot S^k)W),$$

where $\tilde{A}^k = (D^k)^{-\frac{1}{2}}(A^k \odot E^k)(D^k)^{-\frac{1}{2}}$, D^k is the degree matrix, and W represents the learnable

parameters. The final representation x_c^k is obtained through a linear transformation, *i.e.*,

$$x_c^k = W_c H_c^k,$$

Using another independent GCN and linear layer, we can obtain the biased representation x_b^k. During training, we concatenate the biased and causal representations and input them into two MLP-based classifiers, $\mathrm{CLS_c}$ and $\mathrm{CLS_b}$ respectively, to obtain the final prediction results, *i.e.*,

$$y_c^k = \mathrm{CLS_c}([x_c^k, x_b^k]),$$
$$y_b^k = \mathrm{CLS_b}([x_c^k, x_b^k]).$$

3.3 Training and Optimization

To achieve robust and disentangled representations for accurate fake news detection, we carefully design our training and optimization procedure with a three-fold focus.

Amplifying Bias in Subgraph Learning To train an accurate causal relevant model, we first need to rely more on shortcut learning during the learning process of the biased subgraph. In the classification process of $\mathrm{CLS_b}$, we use $\frac{C_b^k}{C_c^k}$ as the coefficient of the cross-entropy loss, *i.e.*,

$$L_b = \sum_{k \in Y} \frac{C_b^k}{C_c^k} \hat{y}^k \log y_b^k,$$

where C_b^k and C_c^k represent the softmax outputs of $\mathrm{CLS_b}$ and $\mathrm{CLS_c}$ and their probabilities of belonging to the target class y^k, respectively, and $\hat{y}^k$ is the ground truth. The GCN and classifier generating the biased subgraph will highly trust and rely on the more convenient paths, so the classification loss can be represented as

$$L_{cls} = \sum_{k \in Y} \hat{y}^k \log y_c^k + L_b.$$

Minimize Extrapolation Risk Although we have generated biased and causal news representations, there will be statistical correlations between causal variables inherited from the biased observation graph and the biased variables. To further reduce the connection between them and obtain more robust disentangled representations, we achieve this goal by randomly combining biased and causal representations from multiple samples. Specifically, for a mini-batch, we randomly sample U other samples and concatenate the causal representation x_c^k of the current news with the biased representations $\{x_b^u\}_{u=1}^U$ from other news samples. This simulates news representations in different environments, allowing the model to perform ideal extrapolation. To ensure the generation process of the causal subgraph is not disturbed by external environments, we minimize the variance of the cross-entropy loss to achieve this purpose. The variance minimization ensures that the learned causal features are robust across different environments, *i.e.*,

$$L_c = \mathbb{V}(\sum_{u=1}^{U} \sum_{k \in Y} \hat{y}^k \log(\mathrm{CLS_c}([x_c^k, x_b^u]))).$$

Enforcing Representation Orthogonality To promote disentanglement between biased and causal representations in the latent space, we incorporate a contrastive learning loss. Specifically, for each news sample, we sample U other samples from the dataset to construct two sets of latent variables: $Z_+ = \{z_+^u\}_{u=1}^U$, obtained by swapping biased representations while keeping x_c^k fixed, and $Z_- = \{z_-^u\}_{u=1}^U$, obtained by swapping causal representations $\{x_c^u\}_{u=1}^U$ while keeping x_b^k fixed. These sets simulate controlled environments to disentangle the contribution of biased and causal features. To enforce orthogonality, we optimize the latent variable z^k of the current news sample to maximize cosine similarity with positive samples z_+^u and minimize cosine similarity with negative samples z_-^u, formulated as:

$$L_d = \sum_{\langle z^k, z^u \rangle} D_{KL}(s(z^k, z^u) \parallel \mathbb{1}_{z^u \in Z_+}),$$

where $D_{KL}(\cdot \parallel \cdot)$ denotes the Kullback-Leibler divergence, and $\mathbb{1}_{z^u \in Z_+}$ is an indicator function specifying whether z^u is a positive sample. This encourages the model to learn representations where biased and causal components are distinct and less likely to interfere. The final optimization objective integrates the classification loss L_{cls}, the causal loss L_c, and the contrastive loss L_d, *i.e.*,

$$L = L_{cls} + \alpha L_c + \beta L_d,$$

where α and β are hyperparameters controlling the contributions of the causal and contrastive losses.

4 Experiment

In this section, we present a comprehensive evaluation of the proposed DICE framework. The experiments are designed to validate three key aspects:

Method	Twitter		Weibo		Fakeddit		Pheme		Weibo21	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
BERT (Devlin et al., 2019)	0.733	0.717	0.823	0.823	0.860	0.859	0.815	0.781	0.852	0.848
ResNet (He et al., 2015)	0.644	0.633	0.710	0.708	0.721	0.632	0.755	0.662	0.728	0.689
DCT (Li et al., 2023)	0.741	0.728	0.619	0.619	0.790	0.789	0.551	0.462	0.700	0.667
Vanilla	0.784	0.757	0.837	0.837	0.873	0.873	0.831	0.796	0.853	0.849
$CCD^{\dagger}_{Vanilla}$ (Chen et al., 2023)	0.800	0.785	0.855	0.855	0.889	0.889	0.831	0.808	0.871	0.868
SpotFake (Singhal et al., 2019)	0.771	0.776	0.839	0.838	0.891	0.875	0.812	0.804	0.851	0.847
MCAN (Wu et al., 2021)	0.874	0.856	0.852	0.851	0.894	0.893	0.834	0.782	0.895	0.893
CAFE (Chen et al., 2022)	0.869	0.851	0.855	0.855	0.898	0.898	0.831	0.810	0.882	0.881
EMSFM (Zeng et al., 2023)	0.804	0.784	0.834	0.805	0.820	0.807	0.782	0.768	0.843	0.832
BMR (Ying et al., 2023)	0.872	0.851	0.884	0.884	0.901	0.891	0.849	0.808	0.900	0.895
NSLM† (Dong et al., 2024)	0.810	0.782	0.866	0.866	0.895	0.895	0.850	0.815	0.853	0.849
DICE	**0.930**	**0.926**	**0.897**	**0.897**	**0.926**	**0.926**	**0.870**	**0.846**	**0.917**	**0.916**

Table 1: Comparison of DICE with the latest and most commonly used multimodal fake news detection approaches on five datasets. **Bold** indicates the best performance. † indicates that the code is not open-source or that the open-source code could not be reproduced, so we reproduced the code independently.

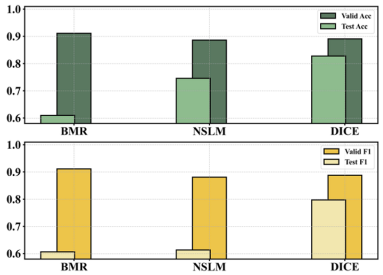


Figure 4: Results of DICE on Cross-Domain News.

Dataset Type		Train	Val	Test
Real News	Sample Size	6311	1175	889
	Bias Type	Text	Text	Image
Fake News	Sample Size	6820	1143	1193
	Bias Type	Image	Image	Text

Table 2: Dataset Partitioning of the Handcrafted Bias Dataset.

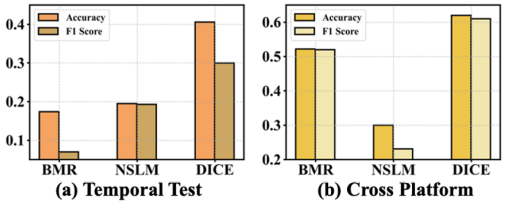


Figure 5: (a) Temporal Test. (b) Cross-platform Test.

(1) the overall performance of DICE on multimodal fake news detection benchmarks, (2) its robustness to distributional shifts and biased data, and (3) the interpretability of its disentangled representations through case studies and visualizations.

4.1 Experiment Settings

Public Datasets We evaluated DICE on five widely-used fake news detection benchmarks, including Twitter (Detection and visualization of misleading content on Twitter, 2018), Weibo (Jin et al., 2017), Weibo21 (Nan et al., 2022), Fakeddit (Nakamura et al., 2020), and Pheme (Zubiaga et al., 2017). Additional details about datasets can be found in Appendix A.

Handcrafted Bias Dataset To validate DICE's causal learning capability in more challenging scenarios, we sampled from three English datasets—Twitter, Fakeddit, and Pheme—to create a new cross-platform fake news dataset. In the training set, we simulated news environment biases by adding specific short texts to real news posts and manipulating images of fake news. In the test set, we manipulated images of real news and added corresponding short texts to fake news posts. For all experiments, we divided the data into training, valid, and test sets according to the dataset requirements. The validation set follows the same data distribution as the training set to select checkpoints. Detailed information about the datasets is presented in Table.2, and specific short texts and manipulation methods are provided in Appendix D.

4.2 Baselines

We compared our approach with six classic or state-of-the-art multimodal fake news detection methods: Spotfake, MCAN, CAFE, EMSFM, BMR, and NSLM. Among these, BMR and NSLM are considered the latest and most interpretable best-performing methods. Additionally, to ensure the effectiveness of DICE, we conducted experiments using BERT and ResNet individually, where Vanilla refers to the combination of BERT and ResNet. To further demonstrate DICE's generalization performance, we implemented the bias correction model CCD on Vanilla for a fair evaluation of DICE's capabilities.

To ensure fairness, we replaced the backbones of the latest strong baselines BMR and NSLM with BERT and ResNet, keeping the parameters consistent. BMR suggests using MAE as a more suitable

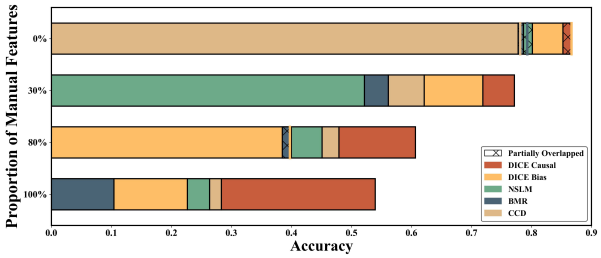


Figure 6: Performance of the different models on the test set as a function of the proportion of manual bias features. "Partial Overlappe" indicates a small advantage in accuracy of the current model, embedded in the previous color to show it. The colored thick line indicates the actual accuracy of the model corresponding to the previous color.

backbone for fake news detection; thus, we did not change its backbone. For the current state-of-the-art multi-modal fake news debiasing detection model, CCD, we utilize the "NRC Emotion Intensity Lexicon" (Mohammad, 2018) to mitigate text bias and employ Vanilla as the base model for fair comparison. Detailed descriptions of the datasets, baselines, and the implementation of the handcrafted dataset are provided in the Appendix.

4.3 Performance Analysis

4.3.1 Comparisons

Table.1 presents a comprehensive comparison of DICE with popular baseline methods in terms of accuracy and F1-score. The results consistently indicate that DICE outperforms other models across all five datasets. On average, DICE achieves an accuracy improvement of 2.68% and an F1-score improvement of 3.64% compared to the baselines, demonstrating its superiority.

From the experiments, we observed that BMR's results were second only to DICE, benefiting from its multi-angle learning of news features. However, further experiments revealed that BMR's learning model captured many biased features, resulting in non-robust representations. The goal of EMSFM is to interpretably fuse global and local alignment features for multimodal fake news verification. NSLM aims to expose fake news patterns through logical reasoning but is easily disrupted by entities in news posts, leading to prediction errors. Regarding DICE's excellent performance, traditional models struggle to improve due to reliance on biased information. While the debiasing model CCD removes biases from psychological factors and image semantics, real-world biases are complex. Our model effectively judges the causal relationship between multimodal features and authenticity labels, successfully identifying challenging samples.

Category	Settings	Accuracy	F1-Score
Full Model	**DICE**	**0.926**	**0.926**
Component	w/o causal	0.905	0.905
	w/o L_c	0.907	0.906
	w/o L_d	0.923	0.923
Graph Learning	GAT	0.916	0.916
	Factor GCN	0.919	0.919

Table 3: Ablation study of the DICE. The tests were conducted on the Fakeddit dataset. Results on other datasets are provided in the Appendix.

4.3.2 Verification of Causal Learning

To verify DICE's capability in causal learning, we conducted evaluations from multiple perspectives.

Cross-Domain Performance. We evaluated DICE's cross-domain performance using **Weibo21**, which includes domain labels for news. We selected "entertainment" and "business" news as the training data and used the checkpoint from this training. For testing, we chose news from the "natural disaster" and "military" domains, which are least related to the training set. Figure.4 illustrates our cross-domain detection results. Although BMR achieved better results on the same domain validation set compared to DICE, it performed worse on test set. This demonstrates DICE's strong generalization capability.

Temporal Test. For the temporal test, we used the **Weibo** data as the training set and employed the checkpoint from the main experiment. We then tested on the **Weibo21** dataset, where the data collection period differs by 4 years, showcasing temporal evolution and dynamic characteristics. As shown in Figure.5(a), although performance testing on datasets with temporal variations did not outperform random selection, the DICE method demonstrates unique advantages in multiple aspects. DICE excels in feature extraction and stability, capturing the complex characteristics of temporal data more effectively. Moreover, its accuracy and F1-scores are the best among all tested methods. DICE consistently maintains superior performance across different testing scenarios, further proving its research and application value in handling complex temporal data tasks. These attributes give it a significant edge in the study and application of temporal data.

Cross-Platform Evaluation. To validate DICE's robustness across different social media environments, we trained on the **Fakeddit** (sampled from Reddit) and tested on the **Pheme** (sampled from Twitter). Due to significant data repetition in

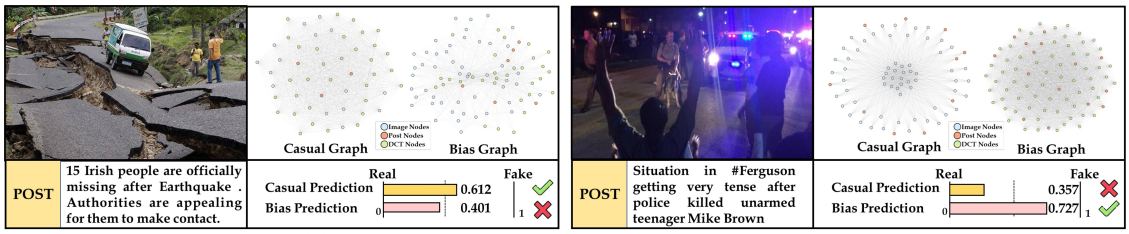


Figure 7: Case studies on fake news detection using DICE.

the Twitter, its analysis results are not representative. Figure.5 (b) illustrates DICE's robustness in cross-platform detection, highlighting its practical significance in diverse real-world applications.

Performance on Severely Biased Datasets. To assess DICE's robustness to bias, we evaluate its performance on a handcrafted dataset with varying levels of injected bias (Figure. 6). While this synthetic bias may not fully encapsulate the complexities of real-world news bias, it allows us to systematically analyze our model's debiasing capabilities under controlled conditions. As shown in Figure. 6, BMR and NSLM exhibit significant performance degradation with the introduction of just 30% biased data. In stark contrast, DICE's causal learning module maintains relatively stable performance across all levels of contamination, demonstrating its robustness even when the specific nature of the bias is unknown a priori.

For further evidence of DICE's disentangling capabilities, a detailed t-SNE visualization of the learned representations under 80% handcrafted bias is provided in Appendix F. This visualization highlights the clear separation between causal and biased representations achieved by DICE.

4.4 Ablation Study

We conducted further analysis to examine the role of each module in our proposed model. The corresponding results are shown in Table.3. To validate the reliability of our approach, we included the following variations:"w/o causal" represents the results of using graph convolution without causal disentanglement. "w/o L_c" represents the results without performing sample swapping during optimization for correlation adjustment. "w/o L_d" represents the results without using contrastive learning to promote orthogonal representations of biased and causal features. Additionally, we compared our method with GAT (Velickovic et al., 2017) and Factor GCN (Yang et al., 2020), a classic method for causal learning, to demonstrate that our causal learning approach outperforms attention-based modeling methods and other graph learning methods. The experimental results further emphasize the advantages of our approach.

4.5 Case Studies

Figure.7 presents case studies of DICE, where causal prediction and bias prediction represent the prediction results of DICE-Causal and DICE-Bias, respectively. Figure.7 (a) depicts a fake news instance from the Twitter dataset, which, according to MediaEval's description, is highly likely to involve image manipulation. In testing, DICE successfully disentangled numerous frequency domain feature nodes within the causal subgraph and connected them based on their inherent causal relationships, leading to a correct classification. Interestingly, Figure.7 (b), a case from the Pheme dataset, shows that the disentangled causal subgraph does not contain any frequency domain feature nodes, yet DICE-Causal still achieves an accurate classification. Conversely, DICE-Bias, relying on a subgraph with numerous frequency domain feature nodes, erroneously classifies this sample as fake news. This highlights DICE's ability to focus on causally relevant features rather than incorporating all features into the inference process, thereby improving interpretability. For additional insights into the effectiveness of feature disentanglement and causal learning, refer to Appendix G, where we present further studies and analyses.

5 Conclusion

The goal of this work is to enhance the performance of fake news detectors by disentangling the causal relationships between features and news veracity. We explore the challenges in fake news detection from a new perspective, modeling causal relationships by solving the intrinsic connections between different modal information. By sampling other examples within a mini-batch, we aim to minimize extrapolation risk. We designed comprehensive experiments to validate our method's robust detection capabilities across varying news environments.

6 Limitations

Our proposed DICE framework demonstrates strong performance and robustness in detecting multimodal fake news, but we acknowledge certain areas that could benefit from further exploration. First, the handcrafted bias dataset, while designed to simulate real-world distributional shifts, may not perfectly replicate the full complexity of naturally occurring biases in large-scale social media data. Future work could further validate our approach using diverse real-world datasets with annotated causal and biased features. Second, while DICE provides insights into its decision-making process through case studies and heatmap visualizations, the current analysis primarily relies on qualitative evaluation. Developing more systematic and quantitative methods to assess causal disentanglement could strengthen the interpretability of the model. Finally, as with many approaches leveraging pre-trained models, DICE's reliance on feature extractors such as BERT and ResNet could reflect some inherent biases in these models. Addressing these aspects in future work may further enhance the generalizability and robustness of the proposed framework.

7 Ethics Statement

This work seeks to enhance the detection of multimodal fake news, a pressing societal challenge, particularly in mitigating the spread of misinformation on social media platforms. While our approach aims to improve the robustness and generalization of fake news detection models, we recognize the potential misuse of such technologies, including the unjust labeling of truthful content as false or the surveillance of user-generated content. To minimize these risks, we advocate for the ethical deployment of our methods, ensuring transparency and fairness through human oversight. Additionally, all datasets utilized in this study are publicly available, and we have adhered to the original licensing agreements and ethical guidelines governing their use. No personally identifiable information (PII) was included in the data, and efforts were made to ensure that the constructed datasets reflect only artificial biases for research purposes. We encourage further research to address the broader ethical implications of multimodal misinformation detection technologies, promoting their development as tools for social good.

References

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Soc. Netw. Anal. Min.*, 13(1):30.

Peter Bourgonje, Julian Moreno Schneider, and Georg Rehm. 2017. From clickbait to fake news detection: An approach based on detecting the stance of headlines to articles. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 84–89, Copenhagen, Denmark. Association for Computational Linguistics.

Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Tun Lu, and Li Shang. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 2897–2905. ACM.

Yixuan Chen, Jie Sui, Liang Hu, and Wei Gong. 2019. Attention-residual network with CNN for rumor detection. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 1121–1130. ACM.

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 627–638. Association for Computational Linguistics.

Detection and visualization of misleading content on Twitter. 2018. Boididou, christina and papadopoulos, symeon and zampoglou, markos and apostolidis, lazaros and papadopoulou, olga and kompatsiaris, yiannis. *International Journal of Multimedia Information Retrieval*, 7(1):71–86.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Yiqi Dong, Dongxiao He, Xiaobao Wang, Youzhu Jin, Meng Ge, Carl Yang, and Di Jin. 2024. Unveiling implicit deceptive patterns in multi-modal fake news via neuro-symbolic reasoning. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 8354–8362. AAAI Press.

Shaohua Fan, Xiao Wang, Yanhu Mo, Chuan Shi, and Jian Tang. 2022. Debiasing graph neural networks via learning disentangled causal substructure. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Shaohua Fan, Xiao Wang, Chuan Shi, Peng Cui, and Bai Wang. 2024. Generalizing graph neural networks on out-of-distribution graphs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 46(1):322–337.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna M. Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiaoxiang Zhu. 2023. A survey of uncertainty in deep neural networks. *Artif. Intell. Rev.*, 56(S1):1513–1589.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Linmei Hu, Ziwei Chen, Ziwang Zhao, Jianhua Yin, and Liqiang Nie. 2022. Causal inference for leveraging image-text matching bias in multi-modal fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11141–11152.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2915–2921. ACM.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

An Lao, Qi Zhang, Chongyang Shi, Longbing Cao, Kun Yi, Liang Hu, and Duoqian Miao. 2024. Frequency spectrum is more effective for multimodal representation and fusion: A multimodal spectrum rumor detector. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18426–18434. AAAI Press.

Xinyu Li, Yanyi Zhang, Jianbo Yuan, Hanlin Lu, and Yibo Zhu. 2023. Discrete cosin transformer: Image modeling from frequency domain. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5468–5478.

Dugang Liu, Pengxiang Cheng, Hong Zhu, Zhenhua Dong, Xiuqiang He, Weike Pan, and Zhong Ming. 2021. Mitigating confounding bias in recommendation via information bottleneck. In *RecSys '21: Fifteenth ACM Conference on Recommender Systems, Amsterdam, The Netherlands, 27 September 2021 - 1 October 2021*, pages 351–360. ACM.

Hui Liu, Wenya Wang, and Haoliang Li. 2023. Interpretable multimodal misinformation detection with logic reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9781–9796. Association for Computational Linguistics.

Yan Liu and Hong-Dong Li. 2003. Image and video processing techniques in the dct domain. *Journal of Image and Graphics*, 8(2):121–128.

Yanhu Mo, Xiao Wang, Shaohua Fan, and Chuan Shi. 2024. Graph contrastive invariant learning from the causal perspective. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 8904–8912. AAAI Press.

Saif M. Mohammad. 2018. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 6149–6157. European Language Resources Association.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2022. MDFEND: multi-domain fake news detection. *CoRR*, abs/2201.00987.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 2021. Improving fake news detection by using an entity-enhanced framework to fuse

diverse multimodal clues. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 1212–1220. ACM.

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.

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *Fifth IEEE International Conference on Multimedia Big Data, BigMM 2019, Singapore, September 11-13, 2019*, pages 39–47. IEEE.

Melanie Subbiah, Amrita Bhattacharjee, Yilun Hua, Tharindu Kumarage, Huan Liu, and Kathleen McKeown. 2023. Towards detecting harmful agendas in news articles. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 110–128, Toronto, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio', and Yoshua Bengio. 2017. Graph attention networks. *ArXiv*, abs/1710.10903.

Bing Wang, Shengsheng Wang, Changchun Li, Renchu Guan, and Ximing Li. 2024. Harmfully manipulated images matter in multimodal misinformation detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 2262–2271, New York, NY, USA. Association for Computing Machinery.

Minjie Wang, Lingfan Yu, Da Zheng, Quan Gan, Yu Gai, Zihao Ye, Mufei Li, Jinjing Zhou, Qi Huang, Chao Ma, Ziyue Huang, Qipeng Guo, Hao Zhang, Haibin Lin, Junbo Zhao, Jinyang Li, Alexander J. Smola, and Zheng Zhang. 2019. Deep graph library: Towards efficient and scalable deep learning on graphs. *CoRR*, abs/1909.01315.

Wenjie Wang, Xinyu Lin, Fuli Feng, Xiangnan He, Min Lin, and Tat-Seng Chua. 2022. Causal representation learning for out-of-distribution recommendation. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*, pages 3562–3571. ACM.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 849–857. ACM.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Lianwei Wu, Pusheng Liu, and Yanning Zhang. 2023. See how you read? multi-reading habits fusion reasoning for multi-modal fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13736–13744.

Yang Wu, Pengwei Zhan, Yunjian Zhang, LiMing Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2560–2569. Association for Computational Linguistics.

Yingxin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. 2022. Discovering invariant rationales for graph neural networks. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Yiding Yang, Zunlei Feng, Mingli Song, and Xinchao Wang. 2020. Factorizable graph convolutional networks. *Advances in Neural Information Processing Systems*, 33:20286–20296.

Shu Yin, Peican Zhu, Lianwei Wu, Chao Gao, and Zhen Wang. 2024. Gamc: an unsupervised method for fake news detection using graph autoencoder with masking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 347–355.

Qichao Ying, Xiaoxiao Hu, Yangming Zhou, Zhenxing Qian, Dan Zeng, and Shiming Ge. 2023. Bootstrapping multi-view representations for fake news detection. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 5384–5392. AAAI Press.

Zhi Zeng, Mingmin Wu, Guodong Li, Xiang Li, Zhongqiang Huang, and Ying Sha. 2023. An explainable multi-view semantic fusion model for multimodal fake news detection. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1235–1240. IEEE.

Litian Zhang, Xiaoming Zhang, Ziyi Zhou, Feiran Huang, and Chaozhuo Li. 2024a. Reinforced adaptive knowledge learning for multimodal fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16777–16785.

Litian Zhang, Xiaoming Zhang, Ziyi Zhou, Feiran Huang, and Chaozhuo Li. 2024b. Reinforced adaptive knowledge learning for multimodal fake news detection. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 16777–16785. AAAI Press.

Wenjia Zhang, Lin Gui, and Yulan He. 2021. Supervised contrastive learning for multimodal unreliable news detection in COVID-19 pandemic. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 3637–3641. ACM.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. Safe: Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 354–367. Springer.

Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. 2022. Generalizing to the future: Mitigating entity bias in fake news detection. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 2120–2125. ACM.

Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2017. Exploiting context for rumour detection in social media. In *Social Informatics - 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part I*, volume 10539 of *Lecture Notes in Computer Science*, pages 109–123. Springer.

A Details of Public Datasets

Twitter was released in 2015 at MediaEval, comprising 17673 news. Weibo is the most extensively used Chinese dataset with 9528 news exposing fake news. Pheme is designed for detecting fake news spread on social media and consists of five breaking news stories, encompassing a total of 3670 news. Fakeddit is a dataset sampled from reddit, containing 31,011 news samples for training and 6,181 for testing. Weibo-21 is a newly released dataset containing a total of 4640 real news items and 4487 fake news items, and we adopt the dataset partitioning approach adopted by Ying et al(Ying et al., 2023).

B Baslines

To verify DICE's ability to detect fake news, we compared DICE with the following six strong baselines:

SpotFake (Singhal et al., 2019) utilize the pre-trained language model to learn the textual information, and employ the pre-trained visual model to obtain image features.

CAFE (Chen et al., 2022) adaptively aggregates features based on the inherent cross-modal ambiguity, addressing misclassification issues arising from differences between different modalities.

EMSFM (Zeng et al., 2023) interpretably fuses global and local alignment features of multimodal news to exploit cross-modal consistency and inconsistency for fake news verification.

MCAN (Wu et al., 2021) integrates pattern features into the co-attention network. It conducts detection by incorporating multiple views that fuse text, image semantics, and image pattern features.

CCD (Chen et al., 2023) proposes a framework combining causal reasoning and counterfactual reasoning to enhance the accuracy and robustness of multimodal fake news detection.

BMR (Ying et al., 2023) models news features from multiple views through bootstrap multi-view representations. It utilizes the Mixture of Experts network for the fusion of multi-view features.

NSLM (Dong et al., 2024) predefines three typical deception patterns and effectively captures different deception modes in fake news through logical symbol reasoning.

C Implementation

We utilized PyTorch (Paszke et al., 2017), DGL (Wang et al., 2019), scikit-learn (Pedregosa et al., 2011), and Transformers (Wolf et al., 2020) to implement DICE. In this case, in order for the biased model to converge quickly and learn the biased representation, we set its learning rate slightly higher than that of the causal model. For the model's backbone, "bert-base-case" was employed for English datasets, while "bert-base-chinese" was used for Weibo and Weibo-21. The experiments were conducted on a RTX3090 GPU. In the experiment, different learning rates were used for the Chinese and English datasets. Table.4 outlines the hyperparameter settings for easy replication of experimental results.

Table 4: Specific Hyperparameter Settings for our Experiments.

Hyperparameter	English Dataset	Chinese Dataset
optimizer	Adam	Adam
causal learning rate	5e-4	5e-5
bias learning rate	8e-4	8e-5
α	0.4	0.4
β	0.2	0.2

D Handcrafted Dataset

To validate DICE's causal learning capability in more challenging scenarios, we randomly sampled existing data according to the proportions shown in Table.5 and artificially constructed a handcrafted dataset by introducing bias information with opposite logic into the training and test sets.

Table 5: Sampling details from the English dataset.

Dataset	Train	Valid	Test
Fakeddit	10000	1100	1400
Twitter	5000	550	800
Pheme	1080	120	100

The specific handcrafted features added are detailed in the Table.6 below. We aim to demonstrate the model's causal learning capability intuitively through the introduction of high-intensity bias information.

E Supplementary ablation experiments

We presented the ablation study results of DICE on the Fakeddit dataset in the main text. Here, we provide its performance on other datasets, as shown in Table.7, where the values in the metric column are written as "Accuracy/F1 Score."

Table 6: Specific bias settings. Due to the large size of the handcrafted dataset, we will upload the specific files at a later stage. Here, we provide detailed modification methods for replication.

Bias Type	Details
Text Bias	Add context: President Biden is concerned about this. Add context: Shocked! I can't believe this is happening! Add context: Or it will affect the world landscape.
Image Bias	Color enhancement by factor 1.5 and fogging by factor 0.15. Add Gaussian random noise with intensity 4.

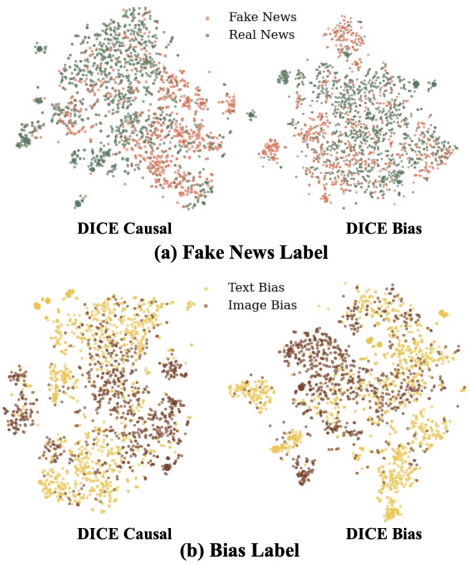


Figure 8: The t-SNE visualization results at 80% proportion of handcrafted bias features. (a) Despite 80% of training data containing bias information, DICE-Causal exhibits more distinct classification boundaries compared to DICE-Bias. (b) DICE-Bias demonstrates stronger correlation with different types of bias information used as labels.

F T-SNE Visualization of Learned Representations

To provide additional insights into DICE's ability to disentangle causal and biased representations, we employ t-SNE to visualize the latent representations learned by the model when 80% of the dataset contains handcrafted bias. As shown in Figure 8, DICE's causal learning module successfully isolates causal representations into distinct clusters, demonstrating its robustness against spurious correlations.

In contrast, the biased learning module captures patterns heavily influenced by the injected bias, as evidenced by the overlapping or fragmented clusters. These results underscore the effectiveness of DICE in leveraging causal relationships for robust and accurate predictions, even in scenarios with significant data contamination.

Table 7: Supplementary results of DICE ablation experiment.

Settings	Twitter	Weibo	Weibo21	Pheme
w/o Casual	0.846/0.832	0.875/0.875	0.853/0.850	0.849/0.821
w/o L_c	0.871/0.861	0.870/0.870	0.850/0.843	0.859/0.835
w/o L_d	0.856/0.854	0.883/0.883	0.857/0.855	0.857/0.832
GAT	0.916/0.912	0.892/0.892	0.883/0.881	0.865/0.834
Factor GCN	0.860/0.852	0.882/0.882	0.891/0.891	0.852/0.831
DICE	**0.930/0.926**	**0.897/0.897**	**0.917/0.917**	**0.867/0.846**

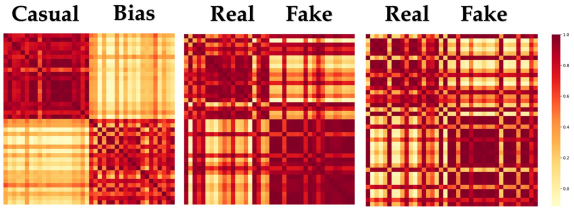


Figure 9: Heatmap visualization. Each cell in the heatmap represents the pairwise cosine similarity. The test for (a) was conducted on the Pheme dataset. The tests for (b) and (c) were conducted on the Handcrafted Bias Dataset with a 30% proportion of handcrafted bias features.

G Causal Learning Studies

To evaluate the effectiveness of disentanglement, we randomly sampled 100 data points from the dataset for disentanglement analysis. As shown in Figure 9(a), biased features and unbiased causal features exhibit clear separability, indicating good orthogonality between the two during the disentanglement process. Figures 9 (b) and (c) present the pairwise similarity of intermediate features obtained by training the DICE Causal and DICE Bias classifiers on a dataset with 30% manually introduced biased features. Despite the interference of biased features, our method maintains strong intra-class and inter-class distinctions, whereas the biased features obtained through disentanglement demonstrate suboptimal performance.