CARD: Adaptive Causal Subgraph Learning Model for Rumor Detection

Anonymous ACL submission

Abstract

Rumor detection on social media has become increasingly important due to the rapid spread of misinformation and its detrimental effects on society. Existing graph-based approaches either rely on global representations for rumor detection or employ fixed-ratio causal subgraphs. However, these methods fail to effectively disentangle causal patterns from environment-specific ones, leading to interference from environmental factors in the causal components. To address these challenges, we propose CARD, a novel adaptive CAusal subgraph learning model for Rumor Detection. CARD introduces adaptive prior constraints and an environment-invariance regularization to generate causal subgraphs that are valid to environmental perturbations. The adaptive prior constraints consist of feature-level and structure-level priors, implemented through a masked attention GATv2 encoder. Specifically, the masks act as feature priors on nodes and structure priors on edges, guiding the model to identify the causal subgraph most relevant to rumor veracity. Furthermore, the environment-invariance regularization encourages the learned causal subgraph to remain stable across noisy or confounding environments. Experimental results on three public real-world datasets demonstrate the effectiveness of our model compared with the state-of-the-art baselines.

1 Introduction

Social media has become a primary medium for communication and information dissemination in daily life. With its extensive usage, a vast amount of user-generated content is continuously produced and shared online. However, such openness also accelerates the spread of rumors, which can undermine public trust and distort social discourse. Consequently, effective rumor detection on social media has become essential to curb their negative societal impact.

Conventional rumor detection methods primarily rely on machine learning techniques that depend heavily on handcrafted features for rumor identification (Castillo et al., 2011; Yang et al., 2012; Feng et al., 2012; Kwon et al., 2013). With the rapid advancement of deep learning, a variety of neural architectures, such as Recurrent Neural Networks (RNNs) (Ma et al., 2016, 2018), Convolutional Neural Networks (CNNs) (Liu and Wu, 2018), and Graph Neural Networks (GNNs) (Nguyen et al., 2020; Bian et al., 2020), have been developed to facilitate automatic rumor detection.

GNNs have shown remarkable effectiveness in modeling rumor propagation, enabling a deeper understanding of the underlying diffusion patterns. Recent studies (Bian et al., 2020; Wei et al., 2021; Sun et al., 2022; Zhu et al., 2024; Peng et al., 2025; Cui and Jia, 2025) have explored the role of propagation structures in rumor dissemination, utilizing global diffusion information for rumor detection. Building on this line of work, more recent approaches (Yin et al., 2024; Tao et al., 2024) have incorporated masking strategies on global propagation graphs to enable encoder–decoder mechanisms. Despite their promising results, these methods are limited to global propagation graphs and do not explicitly capture causal subgraph structures, which is crucial for understanding the underlying diffusion process.

Recent studies have explored various strategies for subgraph learning in the context of rumor detection. He et al. (2021); Liu et al. (2023) propose randomly dropping edges and nodes to generate subgraphs, concluding that such randomly constructed subgraphs do not capture causal relationships. In contrast, Cui and Jia (2024) employ a fixed-ratio masking approach to generate subgraphs, ensuring a predetermined proportion of nodes and edges is preserved. Another line of work (Zhang et al., 2023) focuses on sampling the top-k most important nodes to form subgraphs, thereby prioritizing

nodes with higher relevance scores. While these approaches have provided insights into different subgraph learning techniques, they exhibit two notable limitations. First, subgraphs obtained via fixed proportions or heuristic sampling may limit the expressive capacity of graph semantics. Second, existing methods do not explicitly examine whether causal subgraphs remain robustness and stability under environmental or structural perturbations.

To address these challenges, we propose CARD, a novel adaptive CAusal subgraph learning model for Rumor Detection, which leverages adaptive prior constraints and environment-invariance regularization to extract causal subgraphs for rumor detection task. Specifically, the adaptive prior constraints enable the model to adaptively learn the importance of posts and their interrelations within the propagation graph, guiding it to identify the causal subgraph most relevant to rumor veracity. These priors are defined at both the feature-level and structure-level: feature-level priors facilitate the identification of key nodes, while structure-level priors guide the learning of critical edges. The learned information is subsequently integrated via a masked-attention GATv2 encoder to obtain causal and environmental representations. In addition, the environment-invariance regularization is designed to capture causal subgraphs that are robust to environmental variations, enabling the model to learn subgraphs that remain stable under noisy or confounded conditions.

- We introduce adaptive prior constraint that allows the model to dynamically learn the importance of nodes and edges in the propagation graph, facilitating the identification of the causal subgraph most relevant to rumor veracity.
- We propose environment-invariance regularization to enhance the model's robustness against environmental variations.
- We conduct extensive experiments on three real-world datasets to demonstrate the effectiveness of our model for rumor detection.

2 Related Work

2.1 Rumor Detection

Traditional approaches to rumor detection primarily utilize machine learning techniques that rely heavily on manually designed features (Castillo et al., 2011; Yang et al., 2012; Kwon et al., 2013). With the rapid progress of deep learning, numerous neural network-based methods have been introduced for rumor detection, including content-based methods (Ma et al., 2019; Dun et al., 2021; Xu et al., 2022; Nguyen et al., 2020; Dou et al., 2021; Tian et al., 2022; Min et al., 2022) and propagation structure-based methods (Bian et al., 2020; He et al., 2021; Wei et al., 2021; Sun et al., 2022; Ma et al., 2022; Liu et al., 2023).

Propagation-based methods often employ Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs), and Graph Neural Networks (GNNs) (Ma et al., 2016, 2018; Liu and Wu, 2018; Bian et al., 2020). Recent studies have shown that GNNs are particularly effective at modeling rumor propagation (Bian et al., 2020; Wei et al., 2021; Sun et al., 2022; Zhu et al., 2024; Cui et al., 2025). Bian et al. (2020); Wei et al. (2021) explore bidirectional propagation structures, both top-down and bottom-up, and perform rumor detection using global representations of the propagation graph. Tao et al. (2024) model global and local semantic learning, where local semantic learning is conducted on parent–child pairs for feature reconstruction, without involving subgraph learning.

Recent studies have explored various strategies for subgraph learning in the rumor detection task. He et al. (2021) propose learning subgraphs by randomly dropping edges and nodes. Such randomness can lead to subgraphs that lack strong causal relationships, thereby failing to effectively capture the semantics of information propagation for rumor detection. Cui and Jia (2024) employ a fixed-ratio masking approach to generate subgraphs. While fixed-ratio masking guarantees a predetermined fraction of nodes and edges, it does not consider the relative importance of different nodes or edges. Crucial structural or semantic information may be lost during subgraph generation, which could degrade downstream task performance. Zhang et al. (2023) focuses on sampling the top-k most important nodes to construct subgraphs. This approach neglects the importance of edges, failing to ensure semantic consistency and coherence between nodes and edges, and potentially overlooking informative structures that cannot be captured by node-level relevance scores alone.

While prior work has investigated various subgraph learning strategies, it largely overlooks the coherence of nodes and edges in causal subgraphs and the impact of such subgraphs on environment-

invariant modeling. Addressing these gaps through causal subgraph learning is critical for enhancing rumor detection performance.

2.2 Causal Representation Learning on Graphs

Causal representation learning aims to uncover structured representations that disentangle true underlying causes from spurious correlations (Schölkopf et al., 2021). In graph domains, this corresponds to identifying causal subgraphs whose topological structure and node features directly influence the target prediction, while filtering out environment-specific confounding factors. (Chen et al., 2022) uses causal-inspired subgraph extraction to capture invariant label-related information, achieving robustness to structural and attribute distribution shifts without requiring domain partitions. (Wu et al., 2022) proposes discovering invariant rationales for GNNs by using interventional distributions to filter spurious patterns, improving interpretability and out-of-distribution generalization. (Sui et al., 2022) separate causal and shortcut features, and enforces stable causal-prediction relationships, improving interpretability and generalization.

However, most existing methods have limitations. They often rely on fixed sparsity constraints, which predefine the proportion of nodes or edges to keep, limiting adaptability to diverse graph structures. Few approaches integrate both feature-level and structure-level priors in a unified, adaptive manner. Distinct from the aforementioned studies, CARD bridges these gaps by introducing adaptive prior constraints, which incorporate node masks and edge masks as feature and structural priors, respectively, together with an unsupervised environment-invariance loss. This design enables the model to identify causal subgraphs within realistic rumor propagation graphs without relying on explicit environment annotations, thereby enhancing both predictive performance and interpretability.

3 Methodology

3.1 Problem Definition

Rumor detection is defined as a classification task. Formally, we represent the rumor detection dataset as $\mathcal{C} = \{C_1, C_2, \ldots, C_{|\mathcal{C}|}\}$, where C_i is the i-th event and $|\mathcal{C}|$ is the number of events. For each event $C = \{r_1, r_2, \ldots, r_n, G\}$, r_1 is the source post, other r_j represents the j-th responsive post, and n is the number of posts in the event C. $G = (V, E)$ is the propagation graph with the root node r_1, where V is the set of nodes corresponding to posts and E is the set of edges capturing the relationships between posts. We denote the adjacency matrix as $A \in \{0,1\}^{n \times n}$, where if there is a response relationship between r_s and r_t, $A_{st} = A_{ts} = 1$, otherwise $A_{st} = A_{ts} = 0$. Denote the feature matrix as $X \in \mathbb{R}^{n \times d}$, where d is the node embedding dimension. Rumor detection aims to learn a classifier $f : \mathcal{C} \rightarrow \mathcal{Y}$ that maps each event into one of the categories $\mathcal{Y} \in \{F, T\}$, where F and T denote Rumor and Non-Rumor, respectively.

3.2 Overview

In this section, we propose a novel adaptive CAusal subgraph learning model for Rumor Detection, named as CARD. As illustrated in Figure 1, CARD consists of three main components—Feature Extractor, Adaptive Prior Constraints, and Environment-Invariant Regularization—which will be described in detail in the following sections.

3.3 Feature Extractor

To better capture both the node characteristics and their interrelations within the propagation graph, we employ a Graph Attention Network (GATv2) (Brody et al., 2021) as the feature extractor to encode the propagation graph and obtain informative node representations. We employ GATv2 to compute graph representations. This variant addresses the limitation of static attention in the standard GAT. Specifically, given a propagation graph G, the attention coefficient between node v_i and node v_j is denoted as α_{ij}.

$$e_{ij} = a^T LeakyReLU(W_a h_i + W_b h_j) \quad (1)$$

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k \in \mathcal{N}(i)} exp(e_{ik})} \quad (2)$$

where a denotes a learnable parameter vector. W_a and W_b are trainable weight matrices used to project nodes representation. $\mathcal{N}(i)$ denotes the set of neighboring nodes of node v_i. α_{ij} represents the normalized attention coefficient between nodes v_i and v_j.

In our proposed method, we employ a two-layer GATv2 as the feature encoder. At the $l-$th layer, the GATv2 encoder updates each node representation by aggregating it's neighbors' features, weighted by the learned attention coefficients. This

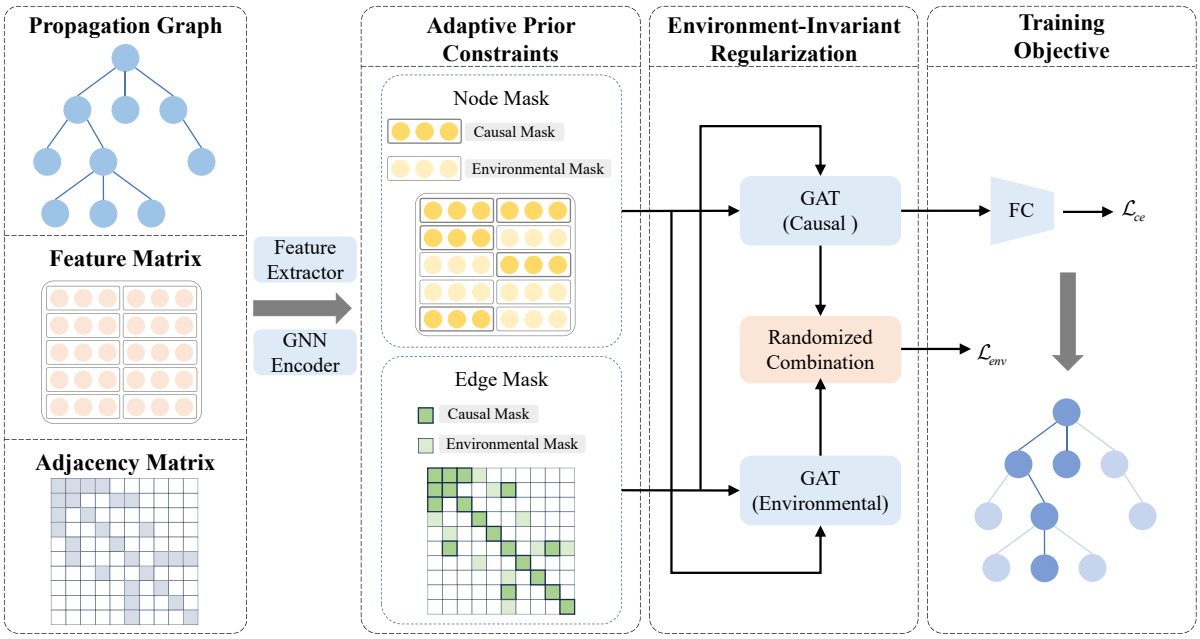


Figure 1: Overview of the proposed framework CARD. It consists of three components: Feature Extractor, Adaptive Prior Constraints, and Environment-Invariant Regularization.

operation is applied iteratively to all nodes up to the l-th layer.

$$h_i^{(l+1)} = \sigma(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l)} W_h^{(l)} h_j^{(l)}) \tag{3}$$

where σ denotes a nonlinear function, and W_h is a trainable weight matrix.

Finally, we obtain the node representations H for the graph G, where $H = [h_1, h_2, \ldots, h_n]$ and h_i denotes the representation of node v_i.

$$H = GAT(A, X) \tag{4}$$

3.4 Adaptive Prior Constraints

To obtain effective causal subgraphs, we adopt adaptive prior constraints to regulate both node features and graph structural information. Specifically, we leverage node causal masks and edge causal masks, which respectively capture the contributions of nodes and edges to the graph feature representations.

Node Causal Mask. Given the node v_i, the node causal mask can be generated as follows.

$$m_i^v = \Phi(\phi_v(h_i)) \tag{5}$$

where ϕ_v are an MLP networks. m_i^v denotes the entries of the node mask matrix $\mathbf{M}^v$.

Edge Causal Mask. To ensure that the edge causal mask generation process is intrinsically dependent on the selection of their corresponding endpoint node causal masks, we construct causal subgraph that preserves both semantic coherence and structural consistency. This design prevents the formation of disconnected or isolated edges and encourages the extraction of meaningful relational structures that accurately reflect the underlying graph topology and node-level semantics. Given the edge $(v_i, v_j) \in E$, the edge causal mask can be generated as follows.

$$m_{ij}^e = \Phi(\phi_e([h_i || h_j])) \cdot m_i^v \cdot m_j^v \tag{6}$$

where ϕ_e are an MLP networks. m_{ij}^e denotes the entries of edge mask matrix $\mathbf{M}^e$. This process indicates that the generation of the edge causal mask is influenced not only by the edge features but also by the causal masks of its endpoint nodes.

In order to use a learnable mask to select key nodes and edges, we employ the function $\Phi(\cdot)$ (i.e., Gumbel-Sigmoid) to approximate the binarization.

$$g = -log(-log(U + \epsilon) + \epsilon) \tag{7}$$

$$\Phi(\cdot) = \sigma(\phi_*(\cdot) + g) \tag{8}$$

where U denotes uniform random noise. ϵ is a small constant introduced for numerical stability. σ denotes the sigmoid function. ϕ_* represents the MLP employed to compute causal masks at the node level or edge level.

Environmental Mask. We introduce environmental masks to explicitly capture contextual information that is complementary to causal features.

While the causal mask focuses on task-relevant dependencies, the environmental mask allows the model to aggregate environmental features. This separation facilitates more expressive feature representations of causal and environmental information, enhancing both performance and interpretation. Specifically, we generate node environment mask $\overline{\mathbf{M}}^v$ and edges environment mask $\overline{\mathbf{M}}^e$.

$$\overline{\mathbf{M}}^v = \mathbb{1} - \mathbf{M}^v \tag{9}$$

$$\overline{\mathbf{M}}^e = \mathbb{1} - \mathbf{M}^e \tag{10}$$

where $\mathbb{1}$ is the all-one matrix.

Feature-level Priors Constraint. We employ these masks to capture both causal and environmental feature of the graph G.

At the node level, we apply the mask to the feature matrix to strengthen the representation of both causal and contextual features.

$$\hat{X} = X + X \odot \mathbf{M}^v \tag{11}$$

where $\odot$ is element-wise multiplication.

Structure-level Priors Constraint. For the edge-level mask, we apply it to the corresponding edges to enable edge-wise weighting.

$$\hat{e}_{ij} = m^e_{ij} e_{ij} \tag{12}$$

We then normalize all incoming edges of a node to derive the final attention weights.

$$\hat{\alpha}_{ij} = \frac{exp(\hat{e}_{ij}/\tau)}{\sum_{k \in \mathcal{N}(i)} exp(\hat{e}_{ik}/\tau)} \tag{13}$$

where τ denotes a temperature parameter.

The masked attention mechanism can be integrated into GATv2 as follows:

$$\mathcal{H}_{cau} = GAT(\hat{X}, A, \mathbf{M}^v) \tag{14}$$

The causal subgraph representation is obtained as follows:

$$\mathcal{G}_{cau} = MEAN(\mathcal{H}_{cau}) \tag{15}$$

Similarly, we can obtain the environment subgraph node representation $\mathcal{H}_{env}$ and graph representation $\mathcal{G}_{env}$.

3.5 Environment-Invariant Regularization

The propagation graph is factorized into a causal subgraph and an environmental subgraph. The causal subgraph encodes invariant, outcome-relevant dependencies that remain stable across environments, whereas the environmental subgraph captures environment-specific patterns that should not influence the prediction.

To achieve environment-invariant prediction, we introduce a randomized feature combination layer to simulate diverse environments, thereby alleviating the impact of environmental biases on causal subgraph prediction. Specifically, given the causal subgraph representation $\mathcal{G}_{cau}$ and environmental subgraph representation $\mathcal{G}_{env}$, we define the combined feature by incorporating additive Gaussian noise:

$$\mathcal{G} = \mathcal{G}_{cau} + \mathcal{G}_{env} + \varepsilon \tag{16}$$

where $\varepsilon \in \mathcal{N}(0, \sigma^2)$.

To promote invariance to environmental variations, we introduce a KL-divergence loss between the predicted logits and a uniform class distribution, encouraging the model to produce uniform predictions across classes.

$$\mathcal{L}_{env} = KL(Softmax(W_g \mathcal{G}) || \mathbf{u}) \tag{17}$$

where $\mathbf{u} = \frac{1}{N}\mathbb{1}$ is a uniform distribution over N classes.

3.6 Training Objective

Classification. To calculate the labels of the rumors, we apply a fully connected layer followed by a softmax layer,

$$\hat{y} = softmax(W_f \mathcal{G}_{cau} + b_f) \tag{18}$$

where $\hat{y}$ is the predicted probability distribution. W_f and b_f are weight and bias parameters respectively.

Classifier Loss. The aim of rumor detection is trained with the cross-entropy loss $\mathcal{L}_C$ of the predictions and ground truth labels.

$$\mathcal{L}_C = -\sum_{i}^{|\mathcal{Y}|} y_i log \hat{y}_i \tag{19}$$

where y_i denotes ground-truth label and $\hat{y}_i$ denotes the predicted probability distribution for the i-th event.

Total Loss. The overall training objective is then defined as:

$$\mathcal{L} = \mathcal{L}_C + \gamma \mathcal{L}_{env} \tag{20}$$

where γ is the trade-off coefficient.

Statistics	TWITTER	DRWeibo	Weibo
# Events	1077	5998	4310
# Posts	60207	376659	816217
# Non-Rumors	564	3160	2187
# Rumors	513	2838	2123

Table 1: Statistics of the datasets.

4 Experiments

In this section, we evaluate the effectiveness of our proposed model by applying our method and other baseline methods to real-world datasets.

4.1 Datasets

We evaluate the proposed model using three publicly available real-world datasets: TWITTER (Lin et al., 2022), DRWeibo (Cui and Jia, 2024) and Weibo (Ma et al., 2016). TWITTER is collected from social media site Twitter (in English). DRWeibo and Weibo are collected from Weibo platform (in Chinese). Each event in the datasets contains both the textual content of the post and its propagation structure. All three datasets have two categories, Rumor (F) and Non-Rumor (T), used for binary classification of rumor detection. The statistics for datasets are shown in Table 1.

4.2 Implementation Details

The proposed model is implemented in PyTorch and optimized using the Adam optimizer. We leverage BERT to encode post texts, obtaining 768-dimensional vector representations that capture their textual content. The hidden dimension of node representations is set to 128. The temperature parameter τ is set to 0.6. The trade-off coefficient γ is set to 1. The model is trained for 200 epochs with a batch size of 128. Following Bian et al. (2020), we adopt the same evaluation protocol and report the final performance by averaging results over 10 runs of 5-fold cross-validation. Model performance is evaluated using four metrics: Accuracy (Acc.), Precision (Prec.), Recall (Rec.), and F1-score (F1).

4.3 Comparison Models

We compare the proposed model with the following baselines:

- **Bi-GCN** (Bian et al., 2020) utilizes bidirectional GCN propagation and root node feature to detect rumors.
- **EBGCN** (Wei et al., 2021) employs an edge-enhancement strategy to model uncertainty during the propagation process.
- **GACL** (Sun et al., 2022) uses adversarial training and contrastive learning for rumor detection.
- **RDEA** (He et al., 2021) leverages data augmentation strategies and self-supervised learning to classify rumor.
- **TrustRD** (Liu et al., 2023) incorporates self-supervised learning and Bayesian network to identify rumor.
- **DCE-RD** (Zhang et al., 2023) explores diverse counterfactual evidence to interpret rumor propagation.
- **PSGT** (Zhu et al., 2024) models a propagation-aware transformer designed to capture propagation patterns.

4.4 Experimental Results

Table 2 reports the rumor detection performance of various methods on the TWITTER, DRWeibo, and Weibo datasets. The results consistently highlight the strong competitiveness of CARD across all datasets. Across all evaluation metrics, CARD consistently outperforms the baseline models. This confirms the effectiveness of integrating adaptive prior constraints and environment-invariant regularization for modeling the propagation of rumors.

BiGCN models the complete rumor propagation graph and performs rumor detection by leveraging the features of the root node, thereby capturing global propagation patterns. EBGCN further investigates the latent relational dependencies among edges, aiming to enhance the expressiveness of the graph structure. RDEA, TrustRD, and GACL generate subgraphs by randomly dropping edges or nodes to introduce random perturbations into the training process. Meanwhile, DCE-RD adopts a top-k node sampling strategy to construct causal subgraphs and corresponding counterfactual evidence. PSGT focuses on extracting task-relevant attention subgraphs to highlight informative propagation patterns for classification. However, such subgraphs derived from random noise or heuristic top-k sampling inevitably lose crucial semantic and structural information, leading to an incomplete representation of the original propagation graph and limiting the model's reasoning capability. In contrast, our proposed CARD model introduces adaptive prior constraints and environment-

Method	Class	TWITTER				DRWeibo				Weibo			
		Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
Bi-GCN	F	0.7591	0.7595	0.7419	0.7506	0.8118	0.8033	0.8031	0.8032	0.9103	0.9099	0.9110	0.9104
	T		0.7771	0.7732	0.7751		0.8256	0.8187	0.8221		0.9147	0.9109	0.9128
EBGCN	F	0.7539	0.7449	0.7428	0.7438	0.8212	0.8080	0.8116	0.8098	0.9166	0.9151	0.9157	0.9154
	T		0.7657	0.7640	0.7648		0.8333	0.8303	0.8318		0.9176	0.9173	0.9175
GACL	F	0.7609	0.7987	0.6781	0.7335	0.8541	0.8454	0.8463	0.8458	0.9367	0.9352	0.9366	0.9359
	T		0.7454	0.8377	0.7889		0.8628	0.8609	0.8618		0.9386	0.9369	0.9378
RDEA	F	0.7855	0.7942	0.7496	0.7713	0.8610	0.8564	0.8482	0.8523	0.9340	0.9294	0.9378	0.9336
	T		0.7857	0.8181	0.8016		0.8656	0.8723	0.8689		0.9393	0.9303	0.9348
TrustRD	F	0.7695	0.7751	0.7359	0.7550	0.8500	0.8528	0.8278	0.8401	0.9312	0.9258	0.9355	0.9306
	T		0.7749	0.7974	0.7860		0.8494	0.8698	0.8595		0.9375	0.9266	0.9320
DCE-RD	F	0.7796	0.7914	0.7579	0.7743	0.8527	0.8421	0.8485	0.8453	0.9303	0.9252	0.9339	0.9296
	T		0.7954	0.8010	0.7982		0.8623	0.8551	0.8587		0.9380	0.9270	0.9325
PSGT	F	0.8089	0.8148	0.7814	0.7977	0.8427	0.8342	0.8336	0.8339	0.9235	0.9171	0.9295	0.9233
	T		0.8141	0.8332	0.8235		0.8529	0.8510	0.8520		0.9315	0.9175	0.9244
CARD	F	**0.8243**	**0.8259**	**0.8073**	**0.8165**	**0.8754**	**0.8807**	**0.8533**	**0.8668**	**0.9477**	**0.9409**	**0.9538**	**0.9473**
	T		**0.8319**	**0.8390**	**0.8355**		**0.8717**	**0.8952**	**0.8833**		**0.9549**	**0.9415**	**0.9482**

Table 2: Rumor detection results on TWITTER, DRWeibo and Weibo. Abbrev.: Rumor (F), Non-Rumor (T).

Method	TWITTER		DRWeibo		Weibo	
	Acc.	F1	Acc.	F1	Acc.	F1
CARD	**0.8243**	**0.8165**	**0.8754**	**0.8668**	**0.9477**	**0.9473**
w/o F-priors	0.8076	0.7926	0.8266	0.8200	0.9164	0.9161
w/o S-priors	0.8125	0.8047	0.8707	0.8618	0.9412	0.9406
w/o E-regularization	0.8191	0.8034	0.8683	0.8611	0.9388	0.9397

Table 3: Results of ablation study.

invariant regularization to explicitly capture the intrinsic consistency. Rather than relying on heuristic sampling or fixed attention mechanisms, CARD adaptively guides subgraph construction based on learned priors, ensuring that the extracted structural representations are both causally meaningful and robust to environmental shifts. This design enables CARD to achieve superior effectiveness and generalization compared with existing graph-based rumor detection models.

4.5 Ablation Study

We validate the effectiveness of the CARD framework from two perspectives: adaptive prior constraints and environment-invariant regularization, with the adaptive prior constraints comprising both feature-level and structure-level priors. The variants are defined as follows: (1) w/o F-priors: removing feature-level priors; (2) w/o S-priors: removing structure-level prior; (3) w/o E-regularization: removing environment-invariant regularization.

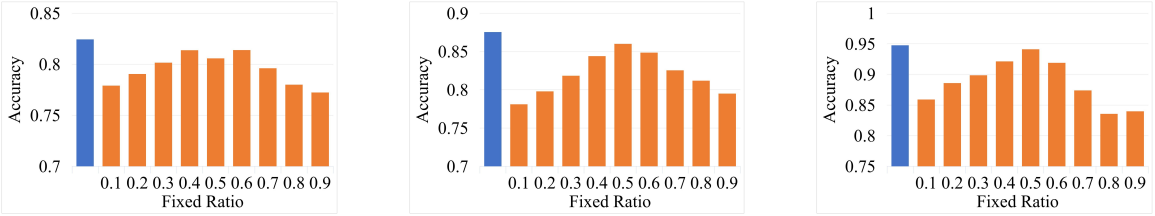
As shown in Table 3, removing any individual component of the CARD framework results in a significant decline in overall performance, highlighting the importance of each component in the overall model architecture. The feature-level priors guide the model to focus on semantically informative representations, enabling it to better capture the intrinsic characteristics of textual inputs. The structure-level priors further enhance the model's capability to reason over complex relational dependencies, allowing it to preserve critical topological information and maintain structural coherence during graph-based learning. Meanwhile, the environment-invariant regularization serves as a stabilizing mechanism that constrains the model to learn environment-agnostic representations, effectively mitigating overfitting to spurious correlations. These components jointly contribute to a more robust and generalizable framework.

4.6 Impact of Prior Constraints

To highlight the effectiveness of the adaptive prior constraints, we conduct a comparative experiment using node and edge masks generated by top-k sampling with fixed ratios as prior constraints. The comparison spans ratios ranging from 0.1 to 0.9.

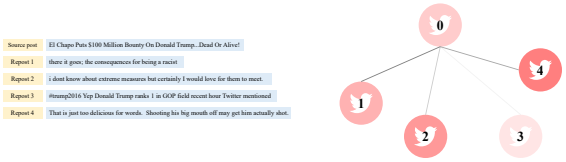
As shown in Figure 2, at lower ratios (< 0.4), the subgraph contains too few nodes and edges, missing critical causal relations and thus limiting the model's ability to capture the full causal structure. At moderate ratios (0.4–0.6), the subgraph covers most of the important causal nodes and edges while minimizing the inclusion of irrelevant ones, maximizing causal information and minimizing noise, leading to optimal performance. Beyond this range (> 0.6), the subgraph becomes excessively large, incorporating non-causal related nodes and edges that introduce noise and degrade performance.

The contribution of each node and edge to the causal structure varies across graphs and training stages. Top-k sampling with fixed ratios may ig-



(a) TWITTER dataset (b) DRWeibo dataset (c) Weibo dataset

Figure 2: Impact of Prior Constraints. In the figure, the blue bar represents the adaptive prior constraints employed by CARD, while the orange bars correspond to top-k sampling with fixed ratios as prior constraints.



(a) All post text (b) Visualization mask

Figure 3: A adaptive prior constraints example from TWITTER. We use red to indicate the mask values of nodes, and gray to indicate the mask values of edges. Darker colors correspond to higher mask values for both nodes and edges.

nore critical nodes and edges, while including non-causal ones that degrade model performance. The adaptive prior constraints dynamically adjust the subgraph based on the importance of nodes and edges, effectively capturing causal relations while avoiding noise.

4.7 Case Study

To further interpret the causal effects captured by the adaptive prior constraints in CARD, we visualize the corresponding node and edge masks in Figure 3. As a case study, we consider a rumor event from the TWITTER dataset, titled "El Chapo Puts $100 Million Bounty on Donald Trump... Dead or Alive!". The source post and its subsequent reposts are shown in Figure 3a, while Figure 3b presents the node and edge masks induced by the adaptive priors. The visualization reveals that the learned feature-level and structure-level information is well aligned with the underlying causal patterns.

The source post disseminates news content using only neutral expressions. Repost 1 conveys a schadenfreude tone, reflecting a negative stance toward the content of the source post. Repost 2 exhibits a mildly sarcastic tone, questioning the veracity of the news. Repost 3 presents objective facts, which are unrelated to the task of determining the truthfulness of the rumor. Repost 4 carries a clear sarcastic implication, strongly signaling that the original news is false. As shown in Figure 3b, nodes 1, 2, and 4, along with their corresponding edges, have the highest mask values. In particular, node 4 expresses strong sentiment and contributes significantly to the assessment of rumor veracity, indicated by the darkest color. In contrast, node 3 is irrelevant to rumor verification, reflected by the lightest color. In summary, the adaptive prior constraint effectively identifies informative nodes and edges, thereby enhancing the contribution of causally relevant subgraph information.

5 Conclusion

In this paper, we propose CARD, a novel adaptive CAusal subgraph learning model for Rumor Detection. CARD leverages adaptive prior constraints and environment-invariance regularization to extract causal subgraphs and obtain stable representations. The adaptive prior constraints are defined at both the feature-level and structure-level: feature-level priors facilitate the identification of key nodes, while structure-level priors guide the learning of critical edges. The learned information is subsequently integrated via a masked-attention GATv2 encoder to obtain causal and environmental representations. The environment-invariance regularization is designed to capture causal subgraphs that are robust to environmental variations, enabling the model to learn subgraphs that remain stable under noisy or confounded conditions. Experiments on three public datasets demonstrate that CARD outperforms the state-of-the-art baselines. In the future, we will explore more approaches for causal subgraph modeling to enhance the performance of rumor detection.

Limitations

While CARD effectively identifies causal subgraphs in rumor propagation graphs, it implicitly assumes that the causal subgraph is invariant within a single observed environment. In practice, rumor propagation can be influenced by multiple latent environments. Under such conditions, the proposed environment-invariance regularization may limit the model's ability to generalize across diverse propagation scenarios. Extending CARD to handle multiple environments explicitly is an important direction for future work.

References

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 549–556.

Shaked Brody, Uri Alon, and Eran Yahav. 2021. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684.

Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, MA Kaili, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. 2022. Learning causally invariant representations for out-of-distribution generalization on graphs. *Advances in Neural Information Processing Systems*, 35:22131–22148.

Chaoqun Cui and Caiyan Jia. 2024. Propagation tree is not deep: Adaptive graph contrastive learning approach for rumor detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 73–81.

Chaoqun Cui and Caiyan Jia. 2025. Towards real-world rumor detection: Anomaly detection framework with graph supervised contrastive learning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7141–7155.

Chaoqun Cui, Siyuan Li, Kunkun Ma, and Caiyan Jia. 2025. Enhancing rumor detection methods with propagation structure infused language model. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7165–7179.

Yingtong Dou, Kai Shu, Congying Xia, Philip S Yu, and Lichao Sun. 2021. User preference-aware fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2051–2055.

Yaqian Dun, Kefei Tu, Chen Chen, Chunyan Hou, and Xiaojie Yuan. 2021. Kan: Knowledge-aware attention network for fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 81–89.

Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 171–175.

Zhenyu He, Ce Li, Fan Zhou, and Yi Yang. 2021. Rumor detection on social media with event augmentations. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2020–2024.

Sejeong Kwon, Meeyoung Cha, Kyomin Jung, Wei Chen, and Yajun Wang. 2013. Prominent features of rumor propagation in online social media. In *2013 IEEE 13th international conference on data mining*, pages 1103–1108. IEEE.

Hongzhan Lin, Jing Ma, Liangliang Chen, Zhiwei Yang, Mingfei Cheng, and Chen Guang. 2022. Detect rumors in microblog posts for low-resource domains via adversarial contrastive learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2543–2556.

Leyuan Liu, Junyi Chen, Zhangtao Cheng, Wenxin Tai, and Fan Zhou. 2023. Towards trustworthy rumor detection with interpretable graph structural learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4089–4093.

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

Guanghui Ma, Chunming Hu, Ling Ge, Junfan Chen, Hong Zhang, and Richong Zhang. 2022. Towards robust false information detection on social networks with contrastive learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1441–1450.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on twitter with tree-structured recursive neural networks. Association for Computational Linguistics.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2019. Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In *The world wide web conference*, pages 3049–3055.

Erxue Min, Yu Rong, Yatao Bian, Tingyang Xu, Peilin Zhao, Junzhou Huang, and Sophia Ananiadou. 2022. Divide-and-conquer: Post-user interaction network for fake news detection on social media. In *Proceedings of the ACM web conference 2022*, pages 1148–1158.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 1165–1174.

Xingyu Peng, Junran Wu, Ruomei Liu, and Ke Xu. 2025. Rumor detection on social media with temporal propagation structure optimization. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3865–3878.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634.

Yongduo Sui, Xiang Wang, Jiancan Wu, Min Lin, Xiangnan He, and Tat-Seng Chua. 2022. Causal attention for interpretable and generalizable graph classification. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 1696–1705.

Tiening Sun, Zhong Qian, Sujun Dong, Peifeng Li, and Qiaoming Zhu. 2022. Rumor detection on social media with graph adversarial contrastive learning. In *Proceedings of the ACM Web Conference 2022*, pages 2789–2797.

Xiang Tao, Liang Wang, Qiang Liu, Shu Wu, and Liang Wang. 2024. Semantic evolvement enhanced graph autoencoder for rumor detection. In *Proceedings of the ACM Web Conference 2024*, pages 4150–4159.

Lin Tian, Xiuzhen Jenny Zhang, and Jey Han Lau. 2022. Duck: Rumour detection on social media by modelling user and comment propagation networks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4939–4949.

Lingwei Wei, Dou Hu, Wei Zhou, Zhaojuan Yue, and Songlin Hu. 2021. Towards propagation uncertainty: Edge-enhanced Bayesian graph convolutional networks for rumor detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3845–3854, Online. Association for Computational Linguistics.

Ying-Xin Wu, Xiang Wang, An Zhang, Xiangnan He, and Tat-Seng Chua. 2022. Discovering invariant rationales for graph neural networks. *arXiv preprint arXiv:2201.12872*.

Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM web conference 2022*, pages 2501–2510.

Fan Yang, Yang Liu, Xiaohui Yu, and Min Yang. 2012. Automatic detection of rumor on sina weibo. In *Proceedings of the ACM SIGKDD workshop on mining data semantics*, pages 1–7.

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.

Kaiwei Zhang, Junchi Yu, Haichao Shi, Jian Liang, and Xiao-Yu Zhang. 2023. Rumor detection with diverse counterfactual evidence. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3321–3331.

Junyou Zhu, Chao Gao, Ze Yin, Xianghua Li, and Jürgen Kurths. 2024. Propagation structure-aware graph transformer for robust and interpretable fake news detection. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4652–4663.