CME: Cross-Modal Consistency Enhancement for Multimodal Rumor Detection

Anonymous ACL submission

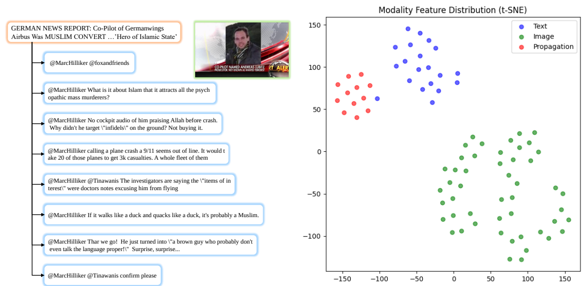
Abstract

Detecting rumors on social media has become a critical research challenge. Although existing multimodal rumor detection methods have achieved promising results, they still suffer from insufficient utilization of modality-specific information and inadequate cross-modal interaction. To address these limitations, we propose a novel Cross-Modal Consistency Enhancement (CME) model for multimodal rumor detection. It incorporates textual, visual, and propagation modalities into a unified framework and transforms each modality into a graph. The uncertainties of the three modalities are utilized to guide modality reconstruction. We design a modality alignment module, including feature alignment and structure alignment to improve the consistency of cross-modal representations. In the process of feature alignment, the aligned modality representations are used as a teacher in a graph-guided self-distillation module to supervise each unimodal student representation. Structure alignment is introduced to model structural similarities across modalities. Extensive experiments conducted on two public real-world datasets demonstrate that our CME model achieves significant improvements compared with the state-of-the-art baselines.

1 Introduction

With the proliferation of social media platforms, online information dissemination has become faster and more pervasive than ever before. The widespread of social media has greatly enhanced public communication and information access. However, it also brings significant challenges. Among them, the rapid spread of rumors stands out as a critical concern, which threatens public trust and undermines societal stability. Therefore, there is a growing need for effective approaches to detect and mitigate the dissemination of rumors.

Early approaches to rumor detection predominantly relied on manually crafted features (Castillo et al., 2011; Yang et al., 2012; Feng et al., 2012; Kwon et al., 2013). However, such methods are inherently limited by the quality of the handcrafted features. Recently, studies have leveraged deep learning techniques to automatically learn high-level feature representations (Ma et al., 2016, 2018; Liu and Wu, 2018; Li et al., 2019).



(a) Origin information (b) Distribution visualization

Figure 1: An example of multimodal encoding distribution visualization. The textual, visual, and propagation structure embeddings are projected into a two-dimensional space using t-SNE, illustrating their distributional differences in the embedding space.

With the increasing diversification of rumor propagation, textual, visual, and multimodal forms that combine both attract greater attention. A series of multimodal rumor detection approaches have been proposed to identify and analyze rumors across different modalities (Khattar et al., 2019; Chen et al., 2022; Zheng et al., 2022; Chen et al., 2025). (Qian et al., 2021) highlight the importance of the textual modality by leveraging the textual semantic representations. Several studies have emphasized the crucial role of image representations in multimodal rumor detection. (Zhou et al., 2020) convert images into textual descriptions. (Lao et al., 2024) extract frequency-domain information from images to enrich visual representations. In addition to focusing on individual modalities, some studies have also strengthened cross-modal interactions and information fusion. (Ying et al., 2023) propose multi-gate mixture-of-expert networks for feature refinement and fusion. (Liu et al., 2025) employ different fusion strategies to diverse modality interaction scenarios to achieve a more robust effect for multimodal fake news detection. However, the mentioned approaches mainly concentrate on integrating textual and visual data but neglect the social context that arises during the spread of rumors.

In real scenarios, rumors frequently circulate via user activities like reposting, commenting, and other forms of engagement on social media platforms. These social interactions reveal the fundamental propagation patterns, which are essential for enhancing the performance of multimodal rumor detection. Wu et al. (2023) incorporate comments in addition to text and images to model the dependencies among multimodal features. Zheng et al. (2022) integrate textual, visual, and social graph in a unified framework. Chen et al. (2025) utilize cross-modal and propagation network contrastive learning. The aforementioned methods leverage the modalities of the textual, visual, and propagation structure to enhance representational capacity. However, they either process each modality independently, resulting in significant distributional discrepancies that hinder effective fusion, or fail to model semantic collaboration and complementarity across modalities, thereby limiting the exploitation of cross-modal interactions. As shown in Figure 1, the representations of textual, visual, and propagation structure encodings are clearly distributed in distinct regions, which reveals a significant inconsistency across modalities. This distributional discrepancy highlights a core challenge in multimodal rumor detection.

To address these challenges, we propose a novel Cross-Modal Consistency Enhancement (CME) model which integrates textual, visual, and propagation graph modalities into a unified framework for multimodal rumor detection. First, each modality is transformed into a graph, and the corresponding graph representations are obtained via graph encoders. The uncertainties of the three modalities are then estimated to guide the reconstruction of the modality, and the reconstruction of unreliable modality features is considered. Furthermore, we designed a modality alignment module, including feature alignment and structure alignment to enhance the cross-modal representations. To achieve feature alignment, the completed modality representations are fused into a unified global representation, which is then used as a teacher in a graph-guided self-distillation module to supervise each unimodal student representation. Then we introduce a structure alignment mechanism to model graph-level structural similarities across modalities. Finally, we utilize the fused student representations to enhance the effectiveness of the proposed model. The main contributions of this paper are as follows:

- We propose a modality reconstruction strategy guided by uncertainty estimation, which can improve the model's ability to handle unreliable modality features.
- We propose a modality alignment module, in which a graph-guided self-distillation component achieves feature alignment and incorporates a structure alignment mechanism to enhance cross-modal consistency.
- We propose a unified multimodal framework based on graphs, which represents text, images, and propagation structures within a structurally consistent space.
- We conduct extensive experiments on two real-world datasets to demonstrate the effectiveness of the proposed model in rumor detection.

2 Related Work

2.1 Unimodal Rumor Detection

Many unimodal approaches to rumor detection have been proposed, including methods based on text (Ma et al., 2019; Nguyen et al., 2020b; Dou et al., 2021; Dun et al., 2021; Yang et al., 2022), images (Jin et al., 2016; Qi et al., 2019), and propagation structures (He et al., 2021; Wei et al., 2021; Ma et al., 2022; Min et al., 2022; Liu et al., 2023). Text-based approaches focus on the content of the text and detect rumors by leveraging features such as linguistic patterns and semantic information. Xu et al. (2022) explored textual semantics by modeling text as graph-structured data to capture long-range semantic dependencies. Liao et al. (2023) proposed evidence-enhanced method, which models the human process of reading news and assessing its veracity through multi-step retrieval. Image-based methods typically rely on neural networks to extract visual features and learn image representations. Such purely visual approaches often suffer from limited model performance. Propagation structures-based methods model the spread of event to simulate the dissemination of information within social networks. Bian et al. (2020) modeled the bidirectional propagation patterns in rumor spread by capturing both top-down and bottom-up structural patterns. Sun et al. (2022) proposed a graph adversarial contrastive learning method to learn the robust representations. Tao et al. (2024) developed fine-grained semantic learning by constructing global semantic information from entire graph and local semantic representations from parent-child nodes. Each modality has its own advantages. However, relying on a single modality often limits the model's capacity. This limitation has driven researchers to develop rumor detection methods that integrate multiple modalities.

2.2 Multimodal Rumor Detection

In recent years, multimodal information has been extensively explored to enhance rumor detection (Khattar et al., 2019; Singhal et al., 2019, 2020; Chen et al., 2022; Zheng et al., 2022; Chen et al., 2025). These methods primarily focus on the textual and visual content of the information. Several studies (Zhou et al., 2020; Qian et al., 2021) modeled the textual content by using the semantic representations and integrating them into multimodal frameworks. Some works have conducted in-depth image research by converting images into textual descriptions for visual information processing (Zhou et al., 2020). Works have further investigated the impact of multi-scale image inputs on model performance (Wang et al., 2024) and explored the role of frequency-based visual features in multimodal rumor detection (Wu et al., 2021; Lao et al., 2024). In addition to enhancing features within single modalities, some works (Ying et al., 2023; Liu et al., 2025) have also strengthened cross-modal interactions and fusion, underscoring the importance of complementary information across modalities. Wang et al. (2018) have introduced event discrimination as an auxiliary task to support detection. However, the aforementioned methods primarily focus on the fusion of textual and visual information, while overlooking the social contextual information generated during the propagation of rumors. Wu et al. (2023) incorporated user comments alongside text and images based on human reading habits, and proposed a cognition-aware fusion method to model the dependencies among multimodal features. Zheng et al. (2022) integrated textual, visual, and social graph features in a unified framework to achieve better complement and alignment relationships between different modalities. Chen et al. (2025) utilized intrinsic features from text, images, and propagation networks, capturing intermodal relationships for accurate fake news detection. Compared with prior multimodal approaches, the key difference in our work lies in the emphasis on modality reconstruction and alignment, which enhances the model's cross-modal representation capabilities.

3 Methodology

3.1 Problem Definition

The rumor detection task can be defined as a binary classification problem. Formally, let $\mathcal{D} = \{D_1, D_2, \dots, D_n\}$ be a rumor detection dataset, where D_i is the i-th event and n is the number of events. Each event $D = (t, v, p)$, where t denotes the text, and v denotes the image. $p = \{\mathrm{p}_0, \mathrm{p}_1, \dots, \mathrm{p}_{|V_p|-1}\}$ is the propagation structure. p_0 is the source post, and other p_j represents the j-th responsive post. $|V_p|$ is the number of posts in the propagation structure. The propagation graph of event D is $G_p = \langle V_p, A_p, X_p \rangle$, where V_p refers to the set of post nodes. $A_p \in \{0,1\}^{|V_p| \times |V_p|}$ represents the adjacency matrix to describe the relationships between nodes. $X_p \in \mathbb{R}^{|V_p| \times d}$ denotes the node feature matrix, where d is the node embedding dimension.

Rumor detection aims to learn a function $f : \mathcal{D} \rightarrow \mathcal{Y}$ that classifies each event into one of the categories $\mathcal{Y} \in \{F, T\}$ (i.e., Rumor or Non-Rumor).

3.2 Overview

In this section, we propose a novel Cross-Modal Consistency Enhancement (CME) model for multimodal rumor detection tasks. As illustrated in Figure 2, we will provide a detailed explanation, including Multimodal Feature Extractor, Modality Reconstruction, and Modality Alignment.

3.3 Multimodal Feature Extractor

3.3.1 Textual Feature Extractor.

The text of the event D is represented as $t = \{\mathrm{t}_1, \mathrm{t}_2, \dots, \mathrm{t}_n\}$, where n denotes the number of words in the text and t_i denotes the i-th word. In order to encode the words and their contextual semantic information, we use BERT (Nguyen et al., 2020a) as the textual feature extractor to obtain the text embedding $X_t = \{e_{\mathrm{t}_1}, e_{\mathrm{t}_2}, \dots, e_{\mathrm{t}_n}\}$, where e_{t_i} is the transformed feature of t_i.

To represent text information in a structured form, we transform the text into a text graph. We

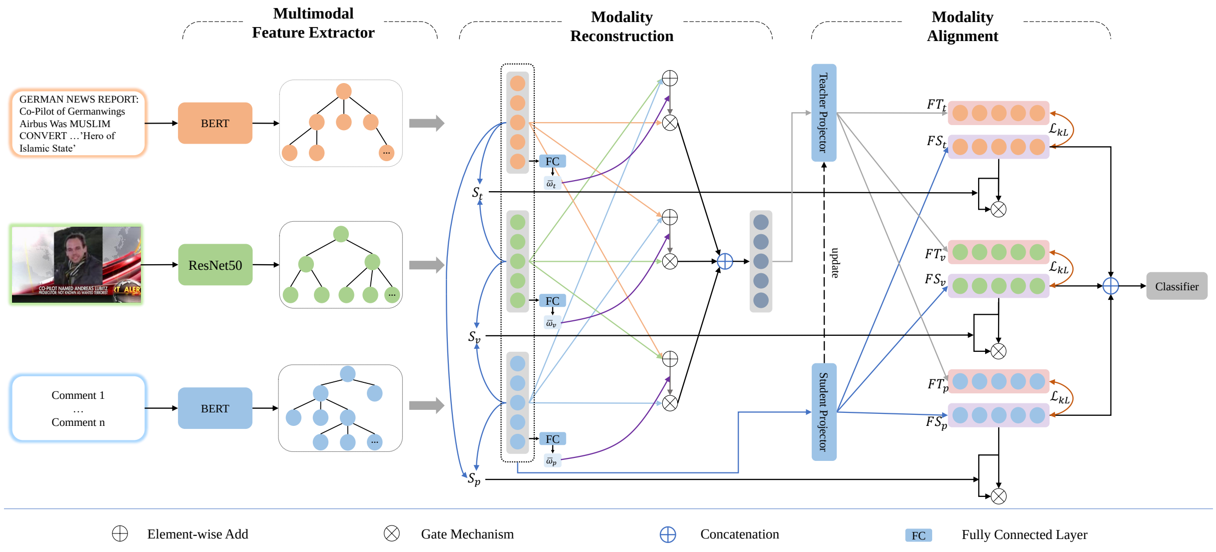


Figure 2: Overview of the proposed CME framework.

define the word set as node set V_t, and X_t serves as the feature matrix. The edges are constructed based on the adjacency relationships of words. Let (i, j) represent an adjacent pair of words. The edge set $\mathcal{E}_t$ in the text graph is defined as:

$$\mathcal{E}_t = \{(i, j) | |i - j| = 1, i, j \in V_t\} \tag{1}$$

Based on this, we construct the adjacency matrix A_t as follows:

$$[A_t]_{ij} = \begin{cases} 1 & if \quad (i, j) \in \mathcal{E}_t \\ 0 & otherwise \end{cases} \tag{2}$$

Finally, we construct the text graph $G_t = \langle V_t, A_t, X_t \rangle$.

3.3.2 Visual Feature Extractor.

The image of the event D is represented as v. We adopt a pretrained ResNet50 (He et al., 2016) as the visual encoder and remove its final global average pooling and fully connected layers to preserve the spatial structure of intermediate features. The resulting output is a convolutional feature map, which is denoted as I.

$$I = ResNet50(v) \tag{3}$$

where $I \in \mathbb{R}^{c \times h \times w}$, c is the number of channels, h and w are the spatial dimensions of the feature output. Then the feature output I is reshaped into a set of $m = h \times w$ region-level feature vectors, each with dimension c, resulting in a node feature matrix $X_v \in \mathbb{R}^{m \times c}$.

To represent visual information in a structured form, we construct a region-level image graph for each text-associated image. We define the region-level feature set as node set V_v, and X_v serves as the feature matrix. To construct the graph structure, we use the K-Nearest Neighbors (KNN) algorithm in the feature space. Each node connects to its k nearest neighbors based on cosine similarity in the feature space.

$$\mathcal{N}(i) = KNN(x_i, X_v, k) \tag{4}$$

where x_i is the feature vector of the i-th node. $\mathcal{N}(i)$ is the set of the k nearest neighbors for x_i. Then edges are constructed based on each neighbor pair between node i and its nearest neighbors $\mathcal{N}(i)$. The edge set $\mathcal{E}_v$ in the image graph is defined as:

$$\mathcal{E}_v = \{(i, j) | i \neq j, j \in \mathcal{N}(i)\} \tag{5}$$

Based on this, we construct the adjacency matrix A_v as follows:

$$[A_v]_{ij} = \begin{cases} 1 & if \quad (i, j) \in \mathcal{E}_v \\ 0 & otherwise \end{cases} \tag{6}$$

Finally, we construct the image graph $G_v = \langle V_v, A_v, X_v \rangle$.

3.3.3 Graph Representation.

To effectively model and exploit the structural characteristics of each modality, we extract modality-specific graph representations.

Specifically, the text graph is fed into a Graph Attention Network (GAT) (Veličković et al., 2017) to capture structure-aware representations.

$$\mathcal{H}_t = GAT(G_t) \tag{7}$$

where $\mathcal{H}_t$ represents the node representations. Then we use mean-pooling operators (MEAN) to aggregate the information of $\mathcal{H}_t$ and pass it through a fully connected layer.

$$h_t = W(MEAN(\mathcal{H}_t)) + b \tag{8}$$

where W and b are learned parameters. h_t is the graph representation of textual modality.

Similarly, for the visual modality and propagation graph modality, we obtain h_v and h_p respectively.

3.4 Modality Reconstruction

To address the incomplete or unreliable representation of certain modalities, we propose a Modality Reconstruction (MR) module. MR adaptively reconstructs weak modalities using complementary information from other modalities.

Specifically, we first model the uncertainty of each modality using a fully connected Layer. Given the modality graph representation h_m, we estimate the standard deviation s_m of the modality as:

$$s_m = \exp(W_s h_m + b_s) \tag{9}$$

where $m \in \{t, v, p\}$. W_s and b_s are learned parameters. The exponential function ensures positivity of the standard deviation.

Then we transform these uncertainties s_m into confidence weights.

$$\omega_m = \exp(-s_m) \tag{10}$$

$$\bar{\omega}_m = \frac{\omega_m}{\sum_i \omega_i} \tag{11}$$

where the exponential function ensures that the predicted uncertainty is strictly positive.

Then we utilize the normalized weights to guide cross-modal reconstruction.

$$M_t = f_t(h_t, \bar{\omega}_t \frac{h_v + h_p}{2}) \tag{12}$$

$$M_v = f_v(h_v, \bar{\omega}_v \frac{h_t + h_p}{2}) \tag{13}$$

$$M_p = f_p(h_p, \bar{\omega}_p \frac{h_t + h_v}{2}) \tag{14}$$

where $f_m(h_{target}, h_{source})$ is a modality-specific gating network. For each target modality h_{target}, the reconstruction is computed as a combination with the source modalities h_{source}.

$$g_{gate} = \sigma(W_{gate} h_{target} + b_{gate}) \tag{15}$$

$$M_m = g_{gate} \cdot h_{target} + (1 - g_{gate}) \cdot h_{source} \tag{16}$$

where W_{gate} and b_{gate} are learned parameters.

Finally, the reconstructed features are concatenated and projected via a linear layer to obtain a modality-enhanced representation.

$$M = concat(M_t, M_v, M_p) \tag{17}$$

$$\bar{M} = WM + b \tag{18}$$

where M and b are learned parameters.

3.5 Modality Alignment

3.5.1 Feature Alignment

To enhance cross-modal feature representations in multimodal rumor detection, we propose a Feature Alignment (FA) module. Specifically, we employ a graph-guided self-distillation method tailored for graph representations. The self-distillation method utilized a global fused graph representation as the teacher, while each modality-specific graph (i.e., text, image, propagation graph) serves as each student. This design ensures that each unimodal branch benefits from the semantic guidance of the global multimodal context.

Specifically, we implement two projectors as teacher projector and student projector. The projections are defined as follows:

$$FS_m = f_{student}(h_m) \tag{19}$$

$$FT_m = f_{teacher}(\bar{M}) \tag{20}$$

where $f_{student}$ and $f_{teacher}$ are the student projector and teacher projector, here we use linear layer as both the projectors. The student projector updated via backpropagation. The teacher projector updated using exponential moving average (EMA) of the student parameters to stabilize the distillation targets. The process are updated iteratively as:

$$\theta_{teacher} \leftarrow \mu \cdot \theta_{teacher} + (1 - \mu) \cdot \theta_{student} \tag{21}$$

where $\theta_{teacher}$ are the teacher projector parameters, $\theta_{student}$ are the student projector parameters. μ is the momentum coefficient.

To align the feature distributions of student and teacher representations, we adopt a KL-divergence loss as self-distillation loss between the student output and the teacher output:

$$\mathcal{L}_{FA} = \sum_{m \in \{t,v,p\}} \mathcal{L}_{KL}(FS_m)||FT_m) \quad (22)$$

During inference, we employ the student projector to obtain the feature representations of each student modality.

3.5.2 Structure Alignment

To explicitly capture structural correlations across different modalities, we design a cross-modal Structure Alignment (SA) module. Given the graph representation of each modality, we compute the pairwise cross-modal structural similarity matrices.

$$S_t = \frac{h_t {h_v}^T}{||h_t||_2 ||h_v||_2} \quad (23)$$

$$S_v = \frac{h_v {h_p}^T}{||h_v||_2 ||h_p||_2} \quad (24)$$

$$S_p = \frac{h_p {h_t}^T}{||h_p||_2 ||h_t||_2} \quad (25)$$

where $|| \cdot ||_2$ denotes the L_2-norm.

Each student representation is then enhanced via structure alignment, which incorporates cross-modal structure guidance.

$$Z_t = \gamma_t \cdot FS_t + (1 - \gamma_t) \cdot (S_t \cdot FS_t) \quad (26)$$

$$Z_v = \gamma_v \cdot FS_v + (1 - \gamma_v) \cdot (S_v \cdot FS_v) \quad (27)$$

$$Z_p = \gamma_p \cdot FS_p + (1 - \gamma_p) \cdot (S_p \cdot FS_p) \quad (28)$$

where γ_t, γ_v and γ_p are learned parameters.

Finally, we concatenate the enhanced modality representations Z_t, Z_v and Z_p to obtain the structure-aware representation.

$$\hat{Z} = concat(Z_t, Z_v, Z_p) \quad (29)$$

3.6 Training Objective

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

$$\hat{y} = softmax(W_f \hat{Z} + b_f) \quad (30)$$

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

For the binary classification task, we aim to minimize the cross-entropy loss $\mathcal{L}_D$, which defined as:

$$\mathcal{L}_D = - \sum_{i}^{|\mathcal{Y}|} y_i log \hat{y}_i \quad (31)$$

Statistics	PHEME	Weibo
# Non-Rumors	1428	877
# Rumors	590	590
# Images	2018	1467
# Posts	34846	336261

Table 1: Statistics of the datasets.

where y_i denotes ground-truth label for the i-th event.

Our training object contains a cross-entropy loss and a self-distillation loss. Finally, we aim at minimizing the loss as follows.

$$\mathcal{L} = \mathcal{L}_D + \alpha \mathcal{L}_{FA} \quad (32)$$

where α is the trade-off parameters.

4 Experiments

4.1 Datasets

We evaluate the proposed model on two real-world datasets: PHEME (Zubiaga et al., 2017) and Weibo (Song et al., 2019). PHEME is an English dataset collected based on five breaking news from Twitter. Weibo is a Chinese dataset collected from the social platform Weibo. Each event in these datasets consists of text, image, and corresponding responsive posts. Both datasets are formulated as binary classification tasks, where each event is annotated as either a Rumor (F) or a Non-Rumor (T). Table 1 shows the statistics of the datasets.

4.2 Implementation Details

We use the pre-trained BERT (Nguyen et al., 2020a) to extract textual features and the pre-trained ResNet50 (He et al., 2016) to extract visual features. The proposed model is implemented using PyTorch (Ketkar et al., 2021). Adam algorithm (Kingma and Ba, 2014) is used to optimize the parameters. The model is trained for 150 epochs with a learning rate of 0.0005. The dimension of hidden layer is set to 128 and the batch size is set to 128. The trade-off parameter α is set to 0.1. We split the datasets for training and testing with a ratio of 8:2. To ensure fairness, we employ 5-fold cross-validation throughout all experiments and report the average results. The Accuracy (Acc.), Precision (Prec.), Recall (Rec.), and F1-score (F1) are adopted as evaluation metrics.

Method	Class	PHEME				Weibo			
		Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
MVAE	F	0.8061	0.6966	0.5523	0.6161	0.7947	0.7633	0.6893	0.7244
	T		0.8385	0.9066	0.8712		0.8144	0.8572	0.8352
SpotFake	F	0.7955	0.6861	0.5576	0.6152	0.8737	0.8567	0.8288	0.8425
	T		0.8308	0.8940	0.8613		0.8885	0.9040	0.8962
SpotFake+	F	0.8048	0.7252	0.5639	0.6345	0.8785	0.8560	0.8006	0.8274
	T		0.8331	0.9074	0.8687		0.8889	0.9182	0.9033
HMCAN	F	0.8209	0.6911	0.6976	0.6944	0.8937	0.9190	0.8209	0.8672
	T		0.8770	0.8701	0.8735		0.8969	0.9405	0.9182
CAFE	F	0.8206	0.7407	0.6245	0.6777	0.9004	0.8778	0.8730	0.8754
	T		0.8495	0.9066	0.8771		0.9145	0.9189	0.9167
MFAN	F	0.8229	0.7348	0.6473	0.6883	0.8962	0.8698	0.8696	0.8697
	T		0.8618	0.8927	0.8770		0.9130	0.9143	0.9137
MFCL	F	0.8170	0.7121	0.6412	0.6748	0.9214	0.8905	0.9159	0.9031
	T		0.8585	0.8880	0.8730		0.9426	0.9259	0.9342
CME	F	**0.8713**	**0.8186**	**0.7225**	**0.7675**	**0.9568**	**0.9382**	**0.9578**	**0.9479**
	T		**0.8941**	**0.9286**	**0.9110**		**0.9717**	**0.9561**	**0.9638**

Table 2: Rumor detection results on two datasets. Abbrev.: Rumor (F), Non-Rumor (T).

4.3 Baselines

We compare the proposed model with the following baselines:

(1) **MVAE** (Khattar et al., 2019) uses the bimodal variational autoencoder to classify posts for multimodal fake news detection.

(2) **SpotFake** (Singhal et al., 2019) utilizes the pretrained models to exploit both the textual and visual features for detecting fake news.

(3) **SpotFake+** (Singhal et al., 2020) leverages transfer learning to capture contextual representation for multimodal fake news detection.

(4) **HMCAN** (Qian et al., 2021) jointly models the multimodal context and the hierarchical semantics in a unified framework.

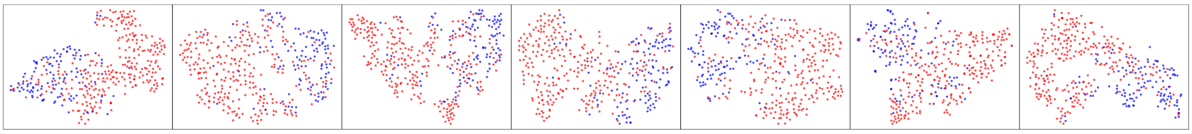
(5) **CAFE** (Chen et al., 2022) propose an ambiguity-aware fake news detection method to capture the cross-modal correlations.

(6) **MFAN** (Zheng et al., 2022) integrates textual, visual, and social graph features in a unified framework to detect multimodal rumors.

(7) **MFCL** (Chen et al., 2025) utilizes pretrained strategy and text, images, and propagation networks for multimodal fake news detection.

4.4 Experimental Results

Table 2 shows the results of rumor detection on two public real-world datasets. The experimental results demonstrate that the proposed CME model outperforms other baselines. MVAE lacks the capacity to model deep semantic relationships between textual and visual features, which makes it ineffective in capturing complex cross-modal semantic interactions. SpotFake heavily relies on pretrained models to extract textual and visual features but lacks effective interaction and alignment between the two modalities, thereby limiting the model's representational capacity. SpotFake+ does not explicitly model or align deep cross-modal information, which results in insufficient multimodal fusion and ultimately hinders model performance. The hierarchical encoding network of HMCAN provides layered semantics for text, but the insufficient exploration of features in image results in inadequate interaction of modality information. CAFE relies on cross-modal ambiguity learning, which may fail to accurately capture complex modality conflicts. However, the aforementioned models only utilize text and image modalities and lack the ability to model propagation paths and structural information, which limits their detection capabilities. MFAN models the text, image, and social graph modalities. In particular, insufficient modeling of the social propagation path can negatively impact the model's performance. MFCL relies on the pretrained propagation network and image-text matching augmentation, while the contrastive learning strategy also plays a crucial role in determining the model's performance. The proposed model CME leverages graph-based representations



(a) CME (b) w/o T (c) w/o V (d) w/o P (e) w/o MR (f) w/o FA (g) w/o SA

Figure 3: T-SNE visualization of the extracted features on the PHEME test set. Dots of the same color correspond to the same class label.

Model	PHEME		Weibo	
	Acc.	F1	Acc.	F1
CME	**0.8713**	**0.7675**	**0.9568**	**0.9479**
w/o T	0.8476	0.7251	0.9422	0.9276
w/o V	0.8499	0.7365	0.9503	0.9381
w/o P	0.8629	0.7641	0.9240	0.9090
w/o MR	0.8653	0.7627	0.9492	0.9375
w/o FA	0.8524	0.7325	0.9311	0.9156
w/o SA	0.8606	0.7453	0.9490	0.9392

Table 3: Results of ablation study on two datasets.

for each modality, enabling the capture of rich structural relationships through graph encoders. By incorporating an uncertainty-based modality reconstruction strategy, the model handles unreliable modality features, ensuring effective information reconstruction. The graph-guided self-distillation approach allows each unimodal student model to be supervised by a unified global teacher representation, improving the transfer of modality information. The structure alignment mechanism across modalities fosters the cross-modal consistency.

4.5 Ablation Study

To analyze the contribution of different components in our proposed CME model, we compare it with the variant models: (1) w/o T (without text), (2) w/o V (without image), (3) w/o P (without propagation graph), (4) w/o MR (without modality reconstruction), (5) w/o FA (without feature alignment) and (6) w/o SA (without structure alignment). The experimental results are shown in Table 3. Acc. refers to the overall results, and F1 refers to the results for the Rumor (F) category.

The experimental results demonstrate that the removal of any component results in a performance decline, highlighting the essential role of each component in the proposed model. The results indicate that: (1) Textual, visual, and propagation graph features each play a critical role in multimodal rumor detection. (2) The contribution of each component differs across datasets. On the PHEME dataset, textual information exhibits the most significant influence on model performance, whereas on the Weibo dataset, the propagation graph contributes most negatively. This discrepancy arises from the structural differences between the two datasets. The propagation graphs in PHEME are relatively shallow and sparse, and textual contents play a more central role. In contrast, the Weibo dataset features deeper and broader propagation structures, making the propagation graph a more dominant factor in determining model performance. (3) The modality reconstruction and modality alignment can facilitate the cross-modality fusion and significantly improve the multimodal feature representations.

4.6 Visualization

In Figure 3, we extract feature vectors before the classification heads for CME under various ablation settings on the PHEME dataset, including without text, without image, without propagation graph, without modality reconstruction, without feature alignment, and without structure alignment. The t-SNE visualizations of these features are presented. The results demonstrate that our method achieves clear decision boundaries on PHEME datasets. Compared to other variants, CME features are more discriminative, which facilitates more accuracy predictions.

5 Conclusion

In this paper, we propose a novel Cross-Modal Consistency Enhancement (CME) model for multimodal rumor detection. We integrate textual, visual, and propagation graph modalities into a unified framework. The uncertainties of the three modalities are then estimated to guide the modality reconstruction. We also design a modality alignment module, including feature alignment and structure alignment to enhance cross-modal representation learning. Experiments on two public datasets demonstrate that the CME model outperforms state-of-the-art baselines.

Limitations

One limitation of our model is the construction method of graphs. The quality of graph-based representations for each modality can vary significantly depending on the construction method, which may affect the overall model performance. In the future, we will explore more approaches for construction method of graphs of multimodality rumor detection further.

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.

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.

Han Chen, Hairong Wang, Zhipeng Liu, Yuhua Li, Yifan Hu, Yujing Zhang, Kai Shu, Ruixuan Li, and Philip S Yu. 2025. Multi-modal robustness fake news detection with cross-modal and propagation network contrastive learning. *Knowledge-Based Systems*, 309:112800.

Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM web conference 2022*, pages 2897–2905.

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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

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.

Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2016. Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia*, 19(3):598–608.

Nikhil Ketkar, Jojo Moolayil, Nikhil Ketkar, and Jojo Moolayil. 2021. Introduction to pytorch. *Deep learning with python: learn best practices of deep learning models with PyTorch*, pages 27–91.

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*, pages 2915–2921.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18426–18434.

Quanzhi Li, Qiong Zhang, and Luo Si. 2019. Rumor detection by exploiting user credibility information, attention and multi-task learning. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1173–1179.

Hao Liao, Jiahao Peng, Zhanyi Huang, Wei Zhang, Guanghua Li, Kai Shu, and Xing Xie. 2023. Muser: A multi-step evidence retrieval enhancement framework for fake news detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4461–4472.

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.

Yifan Liu, Yaokun Liu, Zelin Li, Ruichen Yao, Yang Zhang, and Dong Wang. 2025. Modality interactive mixture-of-experts for fake news detection. In *Proceedings of the ACM on Web Conference 2025*, pages 5139–5150.

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.

Dat Quoc Nguyen, Thanh Vu, and Anh-Tuan Nguyen. 2020a. Bertweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020b. 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.

Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. 2019. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE international conference on data mining (ICDM)*, pages 518–527. IEEE.

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, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, Tanmoy Chakraborty, and Ponnurangam Kumaraguru. 2020. Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13915–13916.

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, pages 39–47. IEEE.

Changhe Song, Cheng Yang, Huimin Chen, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2019. Ced: Credible early detection of social media rumors. *IEEE Transactions on Knowledge and Data Engineering*, 33(8):3035–3047.

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.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Jiandong Wang, Hongguang Zhang, Chun Liu, and Xiongjun Yang. 2024. Fake news detection via multiscale semantic alignment and cross-modal attention. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2406–2410.

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*, pages 849–857.

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.

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*, pages 2560–2569.

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.

Zhiwei Yang, Jing Ma, Hechang Chen, Hongzhan Lin, Ziyang Luo, and Yi Chang. 2022. A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2608–2621.

Qichao Ying, Xiaoxiao Hu, Yangming Zhou, Zhenxing Qian, Dan Zeng, and Shiming Ge. 2023. Bootstrapping multi-view representations for fake news detection. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 37, pages 5384–5392.

Jiaqi Zheng, Xi Zhang, Sanchuan Guo, Quan Wang, Wenyu Zang, and Yongdong Zhang. 2022. Mfan: Multi-modal feature-enhanced attention networks for rumor detection. In *IJCAI*, volume 2022, pages 2413–2419.

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.

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 9*, pages 109–123. Springer.