Cross-modal Ambiguity Learning with Heterogeneous Interaction Analysis for Rumor Detection

Zhuo Fan, Qing Zhu, Yang Xiao[†]
College of Computer Science, Beijing University of Technology
Beijing, China
fanzhuo@emails.bjut.edu.cn {ccgszq,xiaoyang}@bjut.edu.cn

Abstract

Rumor detection on social media has recently attracted significant attention. Due to the complex user group and lack of regulation, rumor-spreaders intentionally disseminate rumors to sway public opinion, severely harming the general interests. Existing approaches generally perform rumor detection by analyzing both image and text modalities, and pay less attention to the interaction behaviors in social media, which can assist in distinguishing rumors from normal information. Furthermore, the images associated with rumors are often inconsistent or manipulated, how to distinguish these different features and utilize them effectively has become crucial in preventing the widespread dissemination of rumors. To address the aforementioned issues, we propose **C**ross-modal **A**mbiguity Learning with **H**eterogeneous **I**nteraction **A**nalysis (CAHIA) for rumor detection. Specially, we design a novel heterogeneous graph feature extractor to fully utilize the different types of behavioral patterns in social interaction networks, we design an frequency inception net to extract manipulated visual features and adopt different fusing strategies to detect various types of rumors according to the ambiguity between text and image. Finally, a hierarchical cross-modal fusing mechanism is used to simulate the process users view and determine the authenticity of posts. Extensive experiments results demonstrate that CAHIA outperforms state-of-the-art models on four large-scale datasets for rumor detection in social media.

1 Introduction

People are becoming increasingly dependent on social media for daily knowledge acquisition compared to traditional news sources (Hermida, 2016). However, due to the surge in netizens and the lack of regulation on online social platforms, some users intentionally spread rumors to mislead readers, and the engaged behavior of users further contributes to the rumors quickly circulating, causing widespread panic among the public. Additionally, the deliberate spread of false information can undermine government credibility, exacerbating the situation and impacting societal stability. Therefore, timely detection and dispelling of these rumors comes at a high cost and consumes vast public resources.

As the variety of information content in rumors increases, such as text and image, these cross-modal data make rumor detection and model construction more challenging. Multimodal rumors on social media can be categorized into two scenarios: The images and content texts are consistent or inconsistent. As shown in Fig. 1, post (a) has had its image manipulated to make the text and image consistent, while post (b) uses unrelated text to exaggerate the facts. Among them, image-text correspondence is more difficult to detect, the rumor-spreaders intentionally forgery images to enhance the credibility of rumors, the image and text are so corresponding that people can easily believe they are true. In the other situation, if the exaggerated image attracts the users, they will quickly find that this post is not credible because of the inconsistency between content and image. A hypothesis drawn from this is that considering the relevance between text-image pairs and manipulated features of images is conducive to detect rumors.

To handle rumors composed of text and image, some multimodal rumor detection methods are presented, these methods explore to learn the joint representations of textual and visual features (Wang et

[†] Corresponding author.

Proceedings of the 24th China National Conference on Computational Linguistics, pages 934-945, Jinan, China, August 11-14, 2025.

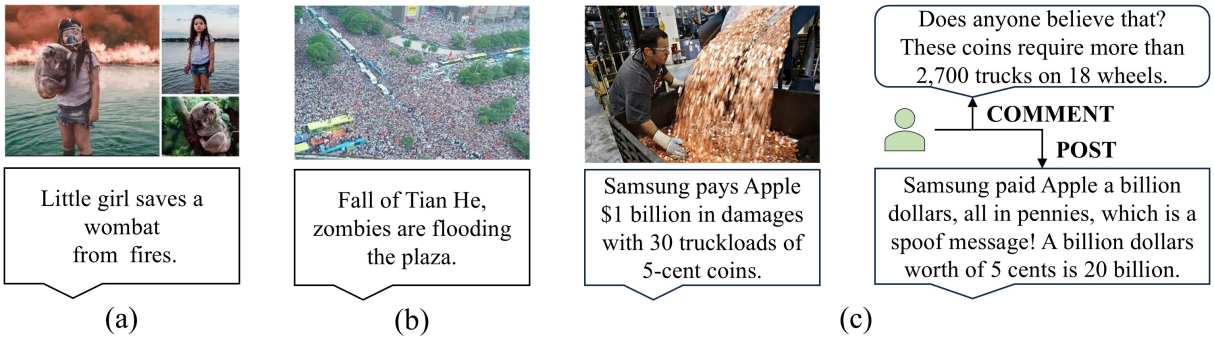

Figure 1: Different multi modal rumor types and an example of the interaction behavior.

al., 2018; Khattar et al., 2019). However, these methods ignore the intrinsic relation between the two modalities, the similarity between image and text is important for rumor detection (Zhou et al., 2020). It's inevitable that the low similarity often indicates there is a high probability that the post is a rumor. Conversely, it cannot simply judge whether the post is real when the similarity is high, because the images may be manipulated intentionally (Qi et al., 2019; Wu et al., 2021). Although these two scenarios are widespread on social media, previous methods have not considered them simultaneously.

In addition to text and image, the social connections can also help distinguish between these two scenarios. The most prominent of social media is that it contains a large amount of graph structures. By utilizing graph convolutional networks (Kipf and Welling, 2016) and graph attention networks (Veličković et al., 2017) to capture these intrinsic social relations, the rumor detection performance can be improved effectively (Yuan et al., 2019; Yang et al., 2021; Zheng et al., 2022), which reveals the important role of social networks in rumor detection.

Moreover, there are many social interaction behaviors present on social media. Fig. 1 (c) shows that one user publishes exaggerated information to sway user opinion, but the other points out the illogical part of the post and posts the actual situation to curb the rumor's propagation. From this, it becomes evident that interaction behaviors of users offer a remarkable influence in social media, but existing multimodal approaches ignored these important signals. These interaction behaviors involve different nodes of the interaction network, and the nodes form interaction paths with varying semantic information through different types of relations. Effectively extracting this hierarchical heterogeneous information is key to leveraging interaction networks to their fullest potential.

These interactions can be combined with multimodal information to prove the efficiency of rumor detection. In fact, the aforementioned two scenarios, as well as social interaction networks, are important bases for users to judge the credibility of posts. Simulating users' attentional shifts in this process is beneficial to rumor detection. Specifically, the way users determine the credibility of a post can be formalized into two stages. In the first stage, they make a preliminary judgment based on the content of the post, like the similarity between text and image, and tend to judge this post as a rumor if the similarity is low or the image is manipulated obviously. In the second stage, they will combine the social network information to make further judgments, in which the comments in opposition and the poster's bad posting history indicate the post is untrustworthy.

To leverage multimodal information on social media to its fullest potential, we proposed a novel multimodal rumor detection approach, which is named **C**ross-modal **A**mbiguity Learning with **H**eterogeneous **I**nteraction **A**nalysis (CAHIA). Given a post, follow how users determine whether it is a rumor or not, we first extract textual features and visual features in both spatial and frequency domains, and adopt different fusing strategies according to ambiguity values that calculate with text-image pair, enhance the model's ability to detect different types of rumors. Then we construct heterogeneous social interaction behavior paths to simulate the way people use social network information to make judgments. In summary, our contributions are threefold:

- To distinguish the rumors with manipulated images and irrelevant images, we design an frequency inception net to extract the manipulated image features from the frequency domain, calculate the ambiguity value between text-image pairs, and adopt different fusion strategies to obtain enhanced cross-modal features according to different ambiguity values.
- To capture and utilize the abundant social interaction information, we construct a heterogeneous interaction network with different types of user interaction behavior paths, then learn this semantic information from various perspectives by heterogeneous graph feature extractor.
- We use a hierarchical cross-modal co-attention mechanism to simulate the way users view a post and determine whether it's a rumor. Extensive experiments results demonstrate that CAHIA outperforms the state-of-the-art baselines on four large-scale real-world datasets.

2 Related work

According to the different types of modal information involved, existing multimodal rumor detection methods can be divided into two categories: those relying on text-image modalities and those relying on text-image-social modalities.

Text-Image Modalities. SpotFake (Singhal et al., 2019) exploits an article's textual and visual features without considering the rich inter-modal relationships between different modalities. However, it simply concatenate features from different modalities and overlook the relationships between the modalities. To address this deficiency, HMCAN (Qian et al., 2021) jointly modeling the multimodal context information and the hierarchical semantics of text in a unified deep model. MCAN (Wu et al., 2021) first uses a co-attention mechanism to fuse textual and visual features better. To further explore the intrinsic relationship between text and image, BMR (Ying et al., 2023) exploit milti-view of text and image, then use MMoE for feature refinement and fusion, obtain the final multimodal representations. Game-on (Dhawan et al., 2024) extracts entities from different modalities separately and constructs a graph based on the relationships between these entities, aiming to explore and leverage the correlations between modalities. MKV (Li et al., 2024) use a mapping mechanism learns low-dimensional mapping scheme and key semantics with discrimination from the different modal features respectively. FSRU (Lao et al., 2024) makes the first attempt at efficiently transforms textual and visual features into the frequency spectrum and obtains highly discriminative spectrum features for multimodal representation.

Text-Image-Social Modalities. Social interaction behaviors on social media carry abundant features. Modeling the social modality allows for the full utilization of the rich information available on social platforms. MFAN (Zheng et al., 2022) makes the first attempt to integrate textual, visual, and social graph features in one unified framework. CLFFRD (Xu et al., 2024) update the node features using a graph attention network followed by a mean pooling operation to obtain the feature of social graph. However, they construct social interaction information as a homogeneous graph, resulting in the loss of significant semantic interaction details. Additionally, due to the absence of a node sampling process, the interaction information between nodes contains noise, especially in scenarios where millions of nodes exist on social media.

Our uniqueness lies in leveraging heterogeneous information between nodes on social networks through sampling and defining interaction paths, and addressing the common issues of tampered and low-resolution visual information on social media through cross-modal ambiguity learning.

3 Methodology

Let multi-modal post $x = \{x^t, x^v, x^u, x^r\}$, which denote the text, image, user, and comments respectively. We construct the heterogeneous interaction graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, which consists of an object set $\mathcal{V}$ and a link set $\mathcal{E}$, where $\mathcal{V}$ and $\mathcal{E}$ represent the set of nodes (Post, User and Comment) and edges (User-Post, User-Comment, Post-Comment), respectively. Rumor detection can be formulated as a binary classification task, $y \in \{0, 1\}$ is the label corresponding to the post x, $y = 1$ indicates that the post is not a rumor and vice versa. Our work aims to incorporate text, image, and social network features to predict the label $\hat{y} \in \{0, 1\}$ of a given post.

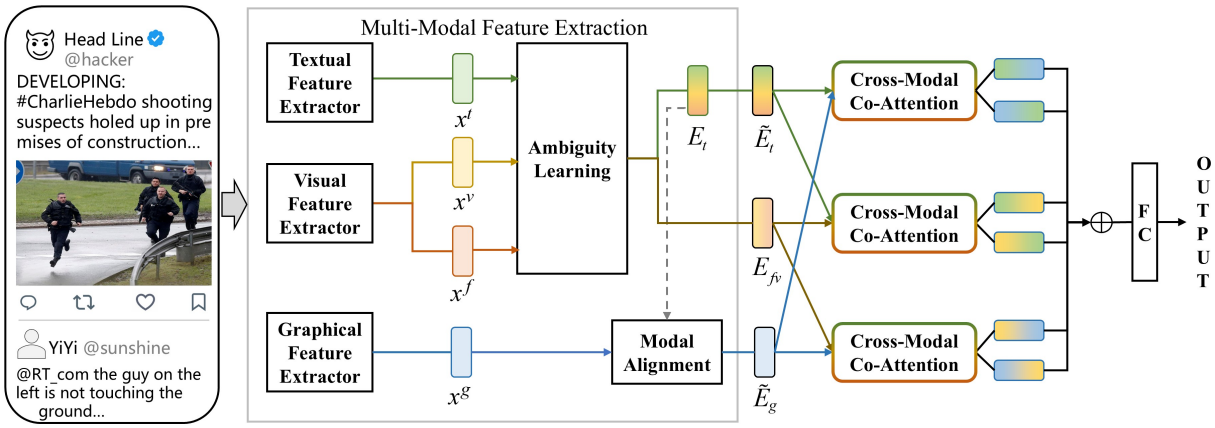

Figure 2: The overall architecture of our proposed framework.

As illustrated in Fig. 2, The whole process of CAHIA is summarized as follows: We first obtain the three modal features for a given post through three feature extractors. We then calculate the ambiguity value between text-image pairs, and adopt different fusing strategies to enhance the model's ability to detect different types of rumors. We use a modal alignment mechanism to refine the representation learned in textual and graph modalities, and finally obtain the enhanced features between every two modalities through multiple cross-modal co-attention mechanisms, which were integrated for rumor detection.

3.1 Text and Visual Feature Extractor

To exploit underlying semantic information of each posts x, we utilize the Text-CNN (Zhang and Wallace, 2015) and RestNet50 (He et al., 2016) to extract textual and spatial domain visual features $\mathbf{x}^t$ and $\mathbf{x}^v$ in x respectively.

Most images associated with rumors exhibit stronger high-frequency characteristics in the frequency domain compared to real images, we extract their frequency domain features using frequency Inception net, with its detailed architecture illustrated in Fig. 3.

First, the input image x^v is transformed using DCT (Discrete Cosine Transform) to obtain its frequency components, which produce a matrix $\mathcal{F} \in R^{(64\times250)}$. This matrix is then processed through the frequency inception net to yield the final frequency visual features $\mathbf{x}^f \in \mathbb{R}^{1\times d}$.

3.2 Cross-modal Ambiguity Learning

Cross-modal ambiguity learning is used not only to enhance the model's ability to distinguish between different types of rumors but also to improve the model's utilization of frequency domain features when spatial features of visual information are insufficient.

As shown in Fig.4 (a), given N text-image pairs, image representation $\mathcal{I}_N^f$ and text representation $\mathcal{T}_N^f$ are obtained by the image encoder and text encoder of CLIP (Radford et al., 2021) or Chinese-CLIP (Yang et al., 2023) respectively. Then we normalized these representations and calculated their cosine similarity to obtain ambiguity value $\mathcal{S}$ as follows:

$$\mathcal{S} = tr\left(\mathcal{P}\right) = tr\left(\mathcal{I}_N^f \cdot \mathcal{T}_N^{f^T}\right) \tag{1}$$

where $\mathcal{P} \in \mathbb{R}^{N\times N}$ is the cosine similarity matrix, $tr(\cdot)$ is the trace of martix.

We use thresholds λ to determine whether the fusing strategy will be adopted. If $\mathcal{S} \geq \lambda$, it means that the similarity between image and text is high, in addition to visual features in the spatial domain, the model should also consider manipulated features in the frequency domain, thus we use enhanced visual features E_{fv} and textual features $\mathbf{x}^t$ for cross-modal fusion to obtain enhanced textual features E_t. If $\mathcal{S} < \lambda$, it means that the similarity between image and text is low, the impact of visual features

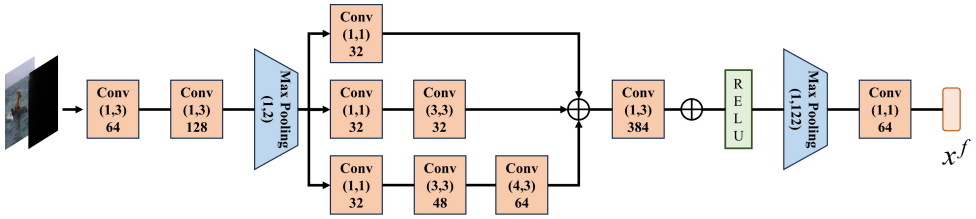

Figure 3: Frequency Inception Net.

in the spatial domain should be decreased, the model will use frequency visual and textual features for cross-modal fusion.

3.3 Graph Feature Extractor

Graph feature extractor is an improved heterogeneous graph attention network (Wang et al., 2019) to obtain the comprehensive social interaction behaviors features. The overall process of the extractor is shown in Fig. 4 (b).

3.3.1 Interaction Behavior Path Construction

Meta-path Φ (Sun et al., 2011) in the heterogeneous social graph $\mathcal{G}$ represents the interaction pathways of users in social media platforms. We constructed two behavior-paths, $\Phi_1\,(P-C-U-P)$ and $\Phi_2\,(P-U-P)$. The former indicates user posting or commenting on posts, including the rumor-spreaders post rumors, or make false comments under normal posts, and debunkers posting the truth and commenting under rumor posts to debunk the rumor. The latter refers to different posts posted by the same user, including the user post a rumor on one topic and post the truth on another topic. The embedding h represents the heterogeneous features of nodes. We initial it in different ways, for post node P and comment node C, we use their content to initialization, for user node U, initialization through its posts and comments.

3.3.2 Interaction Behavior Path Intra Attention

After obtain the initial embedding of each node, we enhance the representations of nodes in the same behavior path through path intra attention. For each source node on the path, intra path attention aggregates features based on the importance of the remaining nodes to the source node.

Take behavior path Φ_1 as an example, i denotes one post node, and $\mathcal{N}_i^{\Phi_1}$ denotes the remaining nodes(comments and users nodes) connected with i, comments made by different users vary in their importance to post node i, we measure this differ by $e_{ij}^{\Phi_1}$ of each node pair and normallize them to get the weight coefficient α^{Φ_1}. Then we adopts multi-head mechanism and calculates path intra attention K times repeatedly to obtain enhanced features $z_i^{\Phi_1}$ of node i, the calculation process is as follows:

$$e_{ij}^{\Phi_1} = att_{intra}\left(h_i, h_j; \Phi_1\right) \tag{2}$$

$$\alpha_{ij}^{\Phi_1} = \frac{\exp\left(\sigma\left(a_{\Phi}^{\mathrm{T}} \cdot [h_i \parallel h_j]\right)\right)}{\sum_{k \in \mathcal{N}_i^{\Phi}} \exp\left(\sigma\left(a_{\Phi}^{\mathrm{T}} \cdot [h_i \parallel h_k]\right)\right)} \tag{3}$$

$$z_i^{\Phi_1} = \Big|\Big|_{k=1}^{K} \sigma\left(\sum \alpha_{ij}^{\Phi_1} \cdot h_j\right), j \in \mathcal{N}_i^{\Phi} \tag{4}$$

where att_{intra} denote the deep nerual network that perform the behavior path intra attention, a_{Φ} is the node-level attention vector for meta-path Φ.

For each node on the behavior path Φ_1, we apply the same operation to obtain the path-specific embeddings $Z_{\Phi_1} = z_1^{\Phi_1}, z_2^{\Phi_1}, \ldots, z_n^{\Phi_1}$. Similarly, we obtain Z_{Φ_2}.

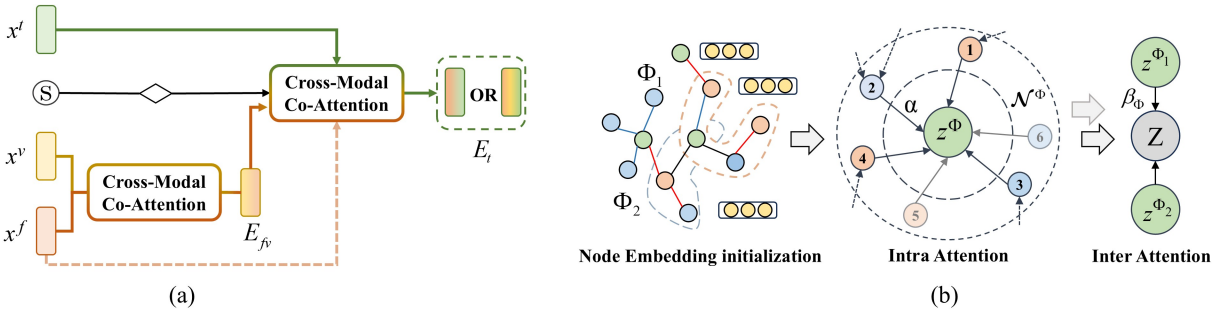

Figure 4: Ambiguity learning mechanism (a) and graph feature extractor (b).

3.3.3 Interaction Behavior Path Inter Attention

The node in the social interaction graph has multiple semantic information, and the path-specific Z_{Φ_i} reflects its features from one aspect. By using behavior path inter attention, we can calibrate and obtain comprehensive node features from various behavior aspects.

We first transform z_i^{Φ} through a nonlinear transformation and measure the importance w_{Φ_i} between them at the inter path level. Then we normalize the importance of each behavior path via softmax function to obtain the weight β_{Φ_i}. After that, we obtain the final social interaction graph embedding Z, the calculation process is as follows:

$$w_{\Phi_i} = \frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}} q^{\mathrm{T}} \cdot tanh\left(W \cdot z_i^{\Phi} + \mathrm{b}\right) \tag{5}$$

$$\beta_{\Phi_i} = \exp\left(w_{\Phi_i}\right) / \sum_{j=1}^{2} \exp\left(w_{\Phi_j}\right) \tag{6}$$

$$Z = \sum_{i=1}^{2} \beta_{\Phi_i} \cdot Z_{\Phi_i} \tag{7}$$

where q is the path inter level attention vector, W is the weight martix, b is the bias, $|\mathcal{V}|$ is the total node of social interaction network. The graph feature $\mathbf{x}^g \in \mathbb{R}^{1\times d}$ of the given post S_i corresponding to the i-th column of Z.

3.4 Hierarchical Cross-modal Feature Fusing

After capturing these single model features, Co-Attention block (Lu et al., 2019) is used to interact with information from different modalities, which reflects the combined use of different modal information by the user. We first perform multi-head self-attention to improve the intra modal feature representation for each modal to obtain E_t, E_v, E_f, E_g. Then we use the Co-Attention block to obtain the enhanced visual features E_{fv} as follows:

$$E_{fv} = \left(||_{h=1}^{H} \, softmax\left(Q_f K_v^T / \sqrt{d}\right) V_v\right) W_{fv}^O \tag{8}$$

Where $Q_f = E_f W_f^Q$, $Q_v = E_v W_v^Q$, $K_v = E_v W_f^K$ is the query, key and value matrix, ${W_{fv}}^O \in \mathbb{R}^{d\times d}$ is the output linear transformations. E_{fv} reflect the comprehensive visual information capture by user, then according to the ambiguity value $\mathcal{S}$, we obtain enhanced textual feature E_{fvt} based on the same operation, which reflect the user's initial judgment of the post's authenticity.

The modal alignment is used to refine the representation learned in text and graph modality. Specifically, for E_g and E_{fvt}, they are transformed into the same modal feature space, that is:

$$E_t{}' = W_t{}' E_{fvt}, E_g{}' = W_g{}' E_g \tag{9}$$

where $W_g{}'$ and $W_t{}'$ are learnable parameters. Then we use the distance between them with the MSE loss for modal alignment to get refined textual features $\tilde{E}_t$ and social graph features $\tilde{E}_g$.

$$\mathcal{L}_{align} = \frac{1}{n}\sum_{i=1}^{n}\left(E_g{}' - E_t{}'\right)^2 \tag{10}$$

Table 1: The statistics of four datasets.

Dataset	Rumors	Reals	Images	Users	Comments	Edges
PHEME (Zubiaga et al., 2017)	590	1428	2018	894	7388	16819
WeiboCED (Song et al., 2019)	590	877	1467	985	4534	58628
Weibo (Jin et al., 2017)	4749	4779	9528	-	-	-
WeiboLate	886	548	1434	996	4015	24386

We perform a co-attention fusing again among each pair of the three modal features to get $\tilde{E}_{tv}$, $\tilde{E}_{vt}$, $\tilde{E}_{gt}$, $\tilde{E}_{tg}$, $\tilde{E}_{gv}$, $\tilde{E}_{vg}$, then concatenate them as the final multi-modal feature $\widehat{E}$, which represent the user's comprehensive judgment on the authenticity of the post. Then the final multi-modal feature is used to predict whether it is a rumor or not.

$$\hat{y}_i = softmax\left(W_c\widehat{E} + b\right) \tag{11}$$

where $\hat{y}_i$ denotes the predicted probability of post being a rumor. Then we use the cross-entropy loss function as

$$\mathcal{L}_{classify} = -y\log\left(\hat{y}_i\right) - (1-y)\log\left(1-\hat{y}_i\right) \tag{12}$$

the final loss can be written as follows:

$$\mathcal{L}_{all} = \mathcal{L}_{classify} + \mathcal{L}_{align} \tag{13}$$

4 Experiment

4.1 Experiment Setup

Dataset. We use PHEME (Zubiaga et al., 2017), WeiboCED (Song et al., 2019), Weibo (Jin et al., 2017) and our newly construct WeiboLate to evaluate our model [2]. WeiboLate is primarily used to differentiate from the older WeiboCED dataset, as it contains more recent posts. We collected rumor data from the Sina Weibo misinformation reporting platform and completed the missing image modality information. We split the datasets for train, dev, and test with a ratio of 7:1:2, and retain the original split of the Weibo dataset.The statistics of these datasets are shown in Table 1.

Model Settings. Our model is implemented with PyTorch 2.2.1 and DGL 2.0, conducted all experiments on a NVIDIA RTX 4090 GPU. The times K in node-level attention and number of heads H are set to 8. The images are transformed into 224*224, the dimension of feature embedding is set to 300, and the thresholds λ are set to 0.7, 0.8, 0.3 and 0.8 for the PHEME, WeiboCED, WeiboLate and Weibo dataset. The learning rate used in the training process is 2e-3. The batch size was set to 64, and the optimization algorithm used was Adam. We adopt the accuracy, precision, recall, and F1 score as the evaluation metrics and report the average results. We use Faknow (Zhu et al., 2024), a unified library for rumor detection to help with reproduction several baselines models.

4.2 Results and Analysis

Table 2 and Table 3 shows the performance of the comparison methods, our proposed CAHIA has outperformed all the other approaches on both datasets. Besides the above verdict, more observations can be obtained as follows.

Firstly, on three datasets containing social modalities, the performance of SpotFake, HMCAN, and MCAN is inferior to that of methods incorporating social modalities, such as MFAN and CLFFRD. This is because these three datasets contain rich interaction information between users, such as comments. Judgments based solely on the text and visual information of the posts reflect only the content of the posts themselves, while ignoring the interactions between the posts and users, as well as between users. CLFFRD achieves slightly better performance than MFAN by applying curriculum learning and data

[2]Our code and newly construct dataset is available at https://github.com/2ezInCode/CAHIA-Core.

Table 2: Performance comparison on datasets containing social modalities. Bold indicates the best result, and underscore indicates the second-best result.

Method	PHEME				WeiboCED				WeiboLate			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
SpotFake (Singhal et al., 2019)	77.72	75.02	67.55	69.25	70.41	70.01	70.22	70.07	71.78	69.95	70.29	70.09
HMCAN (Qian et al., 2021)	84.15	80.93	81.51	81.21	73.52	71.98	72.82	72.26	74.56	73.10	71.00	71.65
MCAN (Wu et al., 2021)	81.43	77.86	77.17	77.49	90.81	90.57	90.88	90.69	82.63	81.38	84.61	81.85
MFAN (Zheng et al., 2022)	88.73	87.07	85.61	86.16	88.95	88.91	88.13	88.33	85.36	84.18	85.11	84.57
CLFFRD (Xu et al., 2024)	89.95	88.26	87.57	88.13	91.26	90.23	89.70	89.82	-	-	-	-
MKV (Li et al., 2024)	90.13	87.97	88.39	87.58	93.15	92.98	**93.37**	93.10	-	-	-	-
CAHIA	**91.12**	**91.04**	**91.11**	**91.07**	**93.22**	**93.24**	93.22	**93.15**	**87.50**	**85.89**	**86.57**	**86.21**

Table 3: Performance comparison on the dataset with more low-resolution and forged traces images, which does not include the social modality.

Method	Weibo			
	Acc	Prec	Rec	F1
BMR (Ying et al., 2023)	88.42	87.45	88.35	87.85
Game-on (Dhawan et al., 2024)	88.95	88.90	88.80	88.80
FSRU (Lao et al., 2024)	90.13	90.05	**90.25**	**90.05**
CAHIA	**94.04**	**90.50**	89.55	90.02

augmentation, but this comes at the cost of additional overhead. MKV achieves better results than MFAN and CLFFRD through a more effective modality fusion approach, despite not utilizing social modalities.

Secondly, methods that consider the frequency components of visual information perform better than those that rely solely on spatial representations, as evidenced by MCAN outperforming SpotFake and HMCAN. This is because some images on social media exhibit signs of forgery, while a significant portion have low resolution, making spatial features insufficient to comprehensively capture visual information. This is particularly evident on WeiboCED and WeiboLate. However, the performance gap is less pronounced on PHEME, where images are generally exhibit fewer instances of forgery.

The Weibo dataset contains more images, most of which are forged or have low resolution. Game-on builds a graph of entities to learn inter-modal relationships. BMR improves on this by using iMMoE for better entity associations. FSRU achieves good performance by integrating multimodal information in the frequency domain while reducing overhead.

Our method fully leverages all the information on social media by defining interaction paths for node sampling and hierarchical feature aggregation. In scenarios with rich social interaction information, hierarchical path attention corrects the features of the same node under different propagation modes, leading to more comprehensive node representations. When dealing with low-resolution and forged images, such as in the Weibo dataset, our method achieves better results by relying solely on cross-modal ambiguity learning.

4.3 Ablation Study

Table 4 shows the ablation study results for the full model and its sub-models: ”-w/o F” (without image frequency domain features), ”-w/o AL” (without ambiguity learning), and ”-w/o G” (without social interaction graph).

Table 4: Ablation study results.

Method	PHEME		WeiboCED		WeiboLate		Weibo	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
CAHIA - w/o F	87.24	84.70	91.55	90.15	83.37	82.15	92.37	86.32
CAHIA - w/o AL	88.01	84.80	92.54	91.55	83.65	82.60	92.56	87.05
CAHIA - w/o G	84.63	82.25	89.11	88.20	82.94	81.80	-	-
CAHIA	**91.12**	**91.07**	**93.22**	**93.15**	**87.50**	**86.21**	**94.04**	**90.02**

From the results of w/o F and w/o AL, we can observe that as forgery traces become increasingly

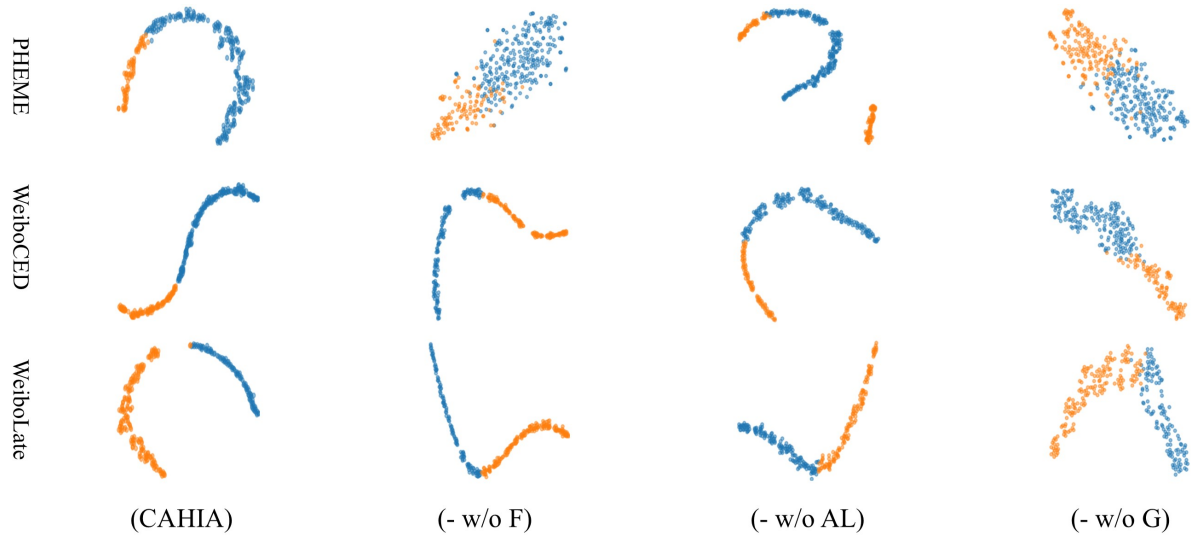

Figure 5: Visualization of learned representations of the test set.

subtle, rumor detection becomes more challenging if spatial domain information is used indiscriminately for judgment. The significant decline in w/o G highlights the indispensability of social interaction information in social media rumor detection. This phenomenon is also reflected in the features learned by the model. As shown in Fig. 5, the t-SNE visualizations of the multimodal rumor features learned by CAHIA and it's sub-models. We can observe that when the social modality is considered (e.g., w/o F and w/o AL), the rumor and non-rumor samples are more clustered, whereas w/o G shows the least discriminative ability among the samples. When text and visuals are meticulously forged, CAHIA's heterogeneous interaction information effectively exposes forgeries, enabling accurate rumor identification.

4.4 Effects of Multimodal Ambiguity Threshold λ

Fig. 6 shows the effect of threshold λ (from 0.1 to 0.9) on Accuracy and F1 socres. In PHEME, WeiboCED and WeiboLate datasets, CAHIA reaches its peak performance when set $\lambda = 0.7, 0.8, 0.3$ respectively. Threshold λ represents how the manipulated images feature used in the model. For a dataset that has more forgery images, the λ should set higher to let the model put attention on this manipulated features. We can observe that the threshold λ used in the WeiboCED is higher than the PHEME dataset, the explanation is that the PHEME dataset is based on five breaking news, which has less forgery iamges than WeiboCED. For WeiboLate dataset, it threshold is much less than WeiboCED and PHEME. The main reason is that WeiboLate contains more rumors and mismatched text-image cases, with low correlation between textual and visual modalities, requiring greater reliance on frequency domain information.

4.5 Case Study

We have shown some rumor cases in the Fig. 7, all with artificially manipulated images. In the images we present, regions with higher high-frequency energy are shown in warm colors, while low-frequency regions are displayed in cool colors to illustrate the distribution of high-frequency components.

As shown in the left of Fig.7, this image of example (a) is a Photoshop composition created using a gas mask, a wombat, and a little girl. The high-frequency information obtained from the DCT transform is reflected at the spliced regions. With ambiguity Learning, our CAHIA make greater use of the frequency-domain information in the visual modality to obtain more discriminative feature E_g, alleviates the insufficiency of spatial information in cases of rumor with image-text consistency.

As shown in example (b) and example (c), most rumor images exhibit stronger high-frequency features in the frequency domain compared to normal images. By extracting frequency-domain components from

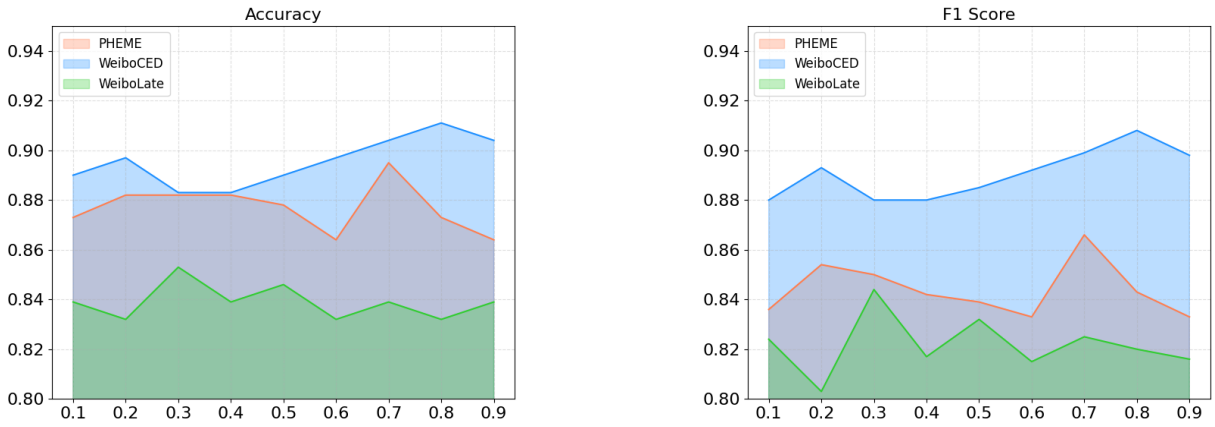

Figure 6: Effect of the λ with Accuracy(left) and F1(right).

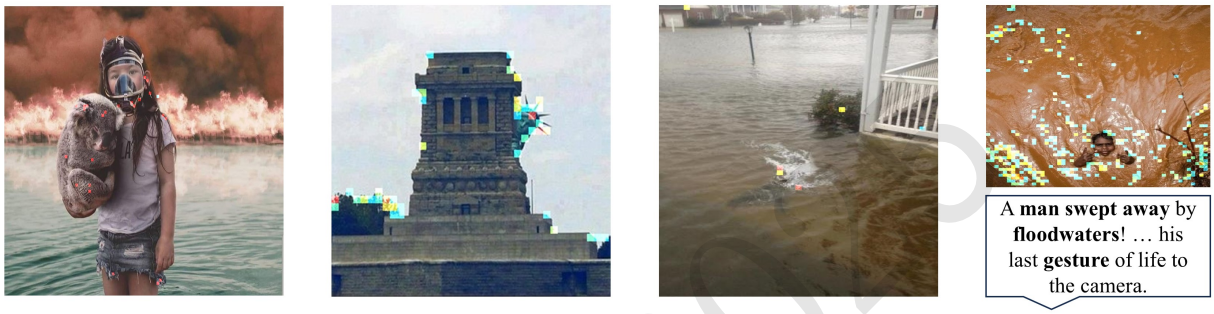

Figure 7: Some rumor cases with the visualization result of DCT.

the DCT transform and incorporating ambiguity learning, our CAHIA can better distinguish between real and false information.

Example (d) can be detected by CAHIA but missed by CAHIA-w/o G. The image of the post is consistent with the text and show more edge textures, hence the DCT can not capture the most valuable part. It can't be recognized by merely using textual and visual information. However, since the question (unsupported) comments below this post, such as 'Why can't I see fear in this man's face ?' or 'People in a death situation won't make a YES pose.' it can be identified as a rumor post by our model with social interaction network.

5 Conclusion

We proposed Cross-modal Ambiguity Learning with Heterogeneous Interaction Analysis (CAHIA) for multimodal rumor detection. CAHIA simultaneously considers scenarios where image is manipulated and the inconsistency between text and image, and it integrates heterogeneous social interaction networks to simulate the process that users determines the credibility of posts. Evaluations on four datasets demonstrate that our model can outperform state-of-the-art baselines. One potential improvement to CAHIA is that existing social networks have redundant nodes, which could introduce noise, and the graph condensation methods can extract key nodes in social networks. In this way, only a small amount of node information is needed to complete the detection of rumors.

References

Mudit Dhawan, Shakshi Sharma, Aditya Kadam, Rajesh Sharma, and Ponnurangam Kumaraguru. 2024. Game-on: Graph attention network based multimodal fusion for fake news detection. *Social Network Analysis and Mining*, 14(1):114.

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.

Alfred Hermida. 2016. Social media and the news. *The SAGE Handbook of Digital Journalism*, page 81.

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*, pages 2915–2921.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

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.

Yang Li, Liguang Liu, Jiacai Guo, Lap-Kei Lee, Fu Lee Wang, and Zhenguo Yang. 2024. Mkv: Mapping key semantics into vectors for rumor detection. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2512–2516.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

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.

Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. 2011. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment*, 4(11):992–1003.

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

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.

Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. 2019. Heterogeneous graph attention network. In *The world wide web conference*, pages 2022–2032.

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.

Fan Xu, Lei Zeng, Bowei Zou, Aiti Aw, and Huan Rong. 2024. Clffrd: Curriculum learning and fine-grained fusion for multimodal rumor detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3314–3324.

Xiaoyu Yang, Yuefei Lyu, Tian Tian, Yifei Liu, Yudong Liu, and Xi Zhang. 2021. Rumor detection on social media with graph structured adversarial learning. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 1417–1423.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2023. Chinese clip: Contrastive vision-language pretraining in chinese.

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.

Chunyuan Yuan, Qianwen Ma, Wei Zhou, Jizhong Han, and Songlin Hu. 2019. Jointly embedding the local and global relations of heterogeneous graph for rumor detection. In *2019 IEEE international conference on data mining (ICDM)*, pages 796–805. IEEE.

Ye Zhang and Byron Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.

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*, pages 2413–2419.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on knowledge discovery and data mining*, pages 354–367. Springer.

Yiyuan Zhu, Yongjun Li, Jialiang Wang, Ming Gao, and Jiali Wei. 2024. Faknow: A unified library for fake news detection. *arXiv preprint arXiv:2401.16441*.

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.