Multimodal Inverse Attention Network with Intrinsic Discriminant Feature Exploitation for Fake News Detection

Tianlin Zhang[1] , **En Yu**[2] , **Yi Shao**[1] , **Jiande Sun**[1]

[1]School of Information Science and Engineering, Shandong Normal University
[2]Faculty of Engineering & Information Technology, University of Technology Sydney
{taylorzhang19951019, isenn.yu, yi.shao.mail}@gmail.com, jiandesun@hotmail.com

Abstract

Multimodal fake news detection has garnered significant attention due to its profound implications for social security. While existing approaches have contributed to understanding cross-modal consistency, they often fail to leverage modal-specific representations and explicit discrepant features. To address these limitations, we propose a Multimodal Inverse Attention Network (MIAN), a novel framework that explores intrinsic discriminative features based on news content to advance fake news detection. Specifically, MIAN introduces a hierarchical learning module that captures diverse intra-modal relationships through local-to-global and local-to-local interactions, thereby generating enhanced unimodal representations to improve the identification of fake news at the intra-modal level. Additionally, a cross-modal interaction module employs a co-attention mechanism to establish and model dependencies between the refined unimodal representations, facilitating seamless semantic integration across modalities. To explicitly extract inconsistency features, we propose an inverse attention mechanism that effectively highlights the conflicting patterns and semantic deviations introduced by fake news in both intra- and inter-modality. Extensive experiments on benchmark datasets demonstrate that MIAN significantly outperforms state-of-the-art methods, underscoring its pivotal contribution to advancing social security through enhanced multimodal fake news detection.

1 Introduction

The rapid evolution of information technology and social networks has shifted news consumption from traditional media to online platforms, enabling individuals to act as publishers and accelerating the delivery of fast, diverse, and personalized information. However, it has also enabled the spread of misinformation, frequently driven by misinterpretations, exaggerations, or deliberate falsifications, which pose significant risks to social security and stability [Shu *et al.*, 2017]. Moreover, advancements in GenAI models [Ouyang *et al.*, 2022; Li *et al.*, 2022] have further lowered the barriers to creating and distributing fake news, exceeding the capacity of human judgment to identify disinformation effectively [Wang *et al.*, 2024b]. While relevant security departments have established verification mechanisms to combat fake news, the vast amount of content has made manual verification increasingly unfeasible [Zhou and Zafarani, 2020]. Consequently, the development of automated fake news detection techniques has emerged as a critical research focus to safeguard the reliability and integrity of public information.

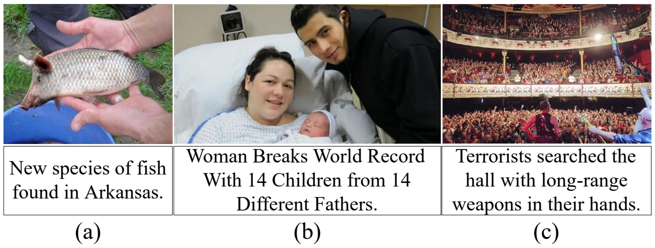
New species of fish found in Arkansas.
Woman Breaks World Record With 14 Children from 14 Different Fathers.
Terrorists searched the hall with long-range weapons in their hands.
(a)
(b)
(c)

Figure 1: Typical types of fake news. (a)/(b) fabricated fake news where either the text or image is authentic while the other is manipulated to create fake content. Specifically, in (a), the subject in the image features a fish's body alongside a pig's ears and snout, highlighting categorical inconsistencies between regions; In (b), the overall semantics of the news text conflict with specific phrases when placed in the realistic context. (c) mismatched text and images from unrelated real news sources, creating inconsistent narratives.

Research on fake news detection has evolved from relying exclusively on text-based approaches [Yu *et al.*, 2017; Wu *et al.*, 2024] to incorporating multimodal content, driven by the increasing prevalence of multimedia content on social media platforms [Yu *et al.*, 2022]. However, the lack of a systematic categorization of news content poses a substantial barrier to advancing effective multimodal fake news detection methods. To address this limitation, this study introduces a novel perspective that classifies multimodal fake news into two distinct categories: a) **fabricated fake news** encompasses cases where textual and visual content appear consistent but involve deliberate manipulations or synthetic generation in one or more modalities. It presents a more formidable detection challenge, particularly given the rapid advancements in generative AI technologies [Zhao *et al.*, 2023; Li *et al.*, 2023], as demonstrated in Figure 1 (a) and (b);

b) **mismatched fake news** refers to instances where textual and visual content are semantically or logically inconsistent. These inconsistencies constitute a form of misinformation capable of distorting public perception and influencing behavior, as illustrated in Figure 1 (c).

On the one hand, detecting fabricated fake news is highly challenging, yet research in this area remains limited. While some methods leverage external knowledge to improve content understanding, they often rely heavily on the timeliness and relevance of such knowledge, introducing potential biases that may undermine performance [Wu *et al.*, 2023; Zhang *et al.*, 2024b]. Therefore, it derives the main research question of our study, ***RQ 1: How to exploit the intrinsic discriminative information within each modality to identify fabricated fake news without relying on external knowledge?*** On the other hand, despite recent advancements in detecting mismatched fake news, current methods still face notable limitations. These methods can be broadly categorized into auxiliary task-based approaches [Zhou *et al.*, 2020; Chen *et al.*, 2022] and attention mechanism-based approaches [Qian *et al.*, 2021; Wei *et al.*, 2022]. The former quantifies the similarity between modality representations in a shared subspace or fuses multimodal representations, while the latter captures cross-modal feature similarities to identify inconsistencies. However, both approaches often fail to extract explicit inconsistency or distinguish the inter-modal relationships into consistency and inconsistency features, leading to confusion in the model. Thus, our second research question can be summarized as, ***RQ 2: How to effectively detect inconsistencies in mismatched fake news while avoiding the overemphasis on cross-modal similarities and implicit pseudo-consistency?***

To address these research questions, we propose the Multimodal Inverse Attention Network (MIAN), which leverages unimodal and multimodal intrinsic discriminative information through hierarchical interaction and explicit inconsistency without relying on external knowledge. Specifically, to capture the intrinsic discriminant features of each modality, we build a hierarchical learning module consisting of two aspects: 1) Local-to-Local, which focuses on the contextual associations within the local features; and 2) Local-to-Global, which models the relationships between global and local features. This module fully explores unimodal information from news text and images and produces refined modality-specific representations to aid MIAN in detecting fabricated fake news, i.e., targeting for ***RQ 1***. To fuse multimodal features, we propose a cross-modal interaction module that facilitates interaction between the enhanced local features from different modalities, enabling effective detection of mismatched fake news, i.e., targeting for ***RQ 2***. Additionally, both the hierarchical learning module and the cross-modal interaction module equip our designed inverse attention mechanism, allowing MIAN to extract explicit inconsistent features that are crucial for fake news detection. Overall, the contributions of this research can be summarized as:

- We propose MIAN, a multimodal framework that leverages intrinsic discriminant information to improve news content-based fake news detection while offering a novel perspective for capturing discriminative relationships.
- We design a hierarchical learning module that enhances unimodal features through hierarchical relationships and a cross-modal interaction module that enables deep integration of inter-modal dependencies, improving both intra- and inter-feature learning.
- We present an inverse attention mechanism to detect inconsistencies across multiple levels, enabling the model to effectively identify diverse fake news. Extensive experiments demonstrate that MIAN outperforms state-of-the-art methods in accuracy.

2 Related Work

2.1 News Content-based Methods

The lack of correlation between textual and visual content in news is a key characteristic of certain types of multimodal fake news. Consequently, several studies have focused on measuring multimodal consistency to verify the credibility of news. Zhou et al. [Zhou *et al.*, 2020] utilized a pre-trained image captioning model to convert images into textual descriptions which achieved cross-modal semantic space alignment, then assessed the multimodal consistency between the original text and the generated captions. Similarly, Xue et al. [Xue *et al.*, 2021] employed shared weights to enforce the alignment of textual and visual representations within a unified semantic space, calculating the similarity of the transformed multimodal representations. To mitigate the semantic gap between modalities, Jiang et al. [Jiang *et al.*, 2023] leveraged CLIP [Radford *et al.*, 2021] to extract features from news text and images, subsequently using cosine similarity to guide multimodal fusion. Chen et al. [Chen *et al.*, 2022] introduced cross-modal contrastive learning as an auxiliary task, and then used the Kullback-Leibler (KL) divergence to measure the ambiguity score between the latent distributions of text and image sampled from the autoencoder. Qi et al. [Qi *et al.*, 2021] measured multimodal entity inconsistency by calculating the similarity between entities in the news text and the corresponding visual content.

Under limited available information, some methods refine unimodal features and employ hierarchical cross-attention mechanisms [Lu *et al.*, 2019] to model inter-modal relationships. Qian et al. [Qian *et al.*, 2021] extracted text features at different hierarchical levels and fused them with visual features using a contextual transformer to learn cross-modal contextual features and supplementary information. Wu et al. [Wei *et al.*, 2022] extracted spatial and frequency domain features from images and progressively fused them with textual features through stacked cross-attention mechanisms. However, their approaches primarily focus on exploring inter-modal relationships. Even though a few methods attempt to enhance unimodal feature representations, they overlook intra-modal interactions, which are crucial for detecting fake news, particularly fabricated news. Moreover, most existing methods model high-level semantic interactions through implicit pseudo-consistency, making it challenging to explicitly capture cross-modal inconsistencies.

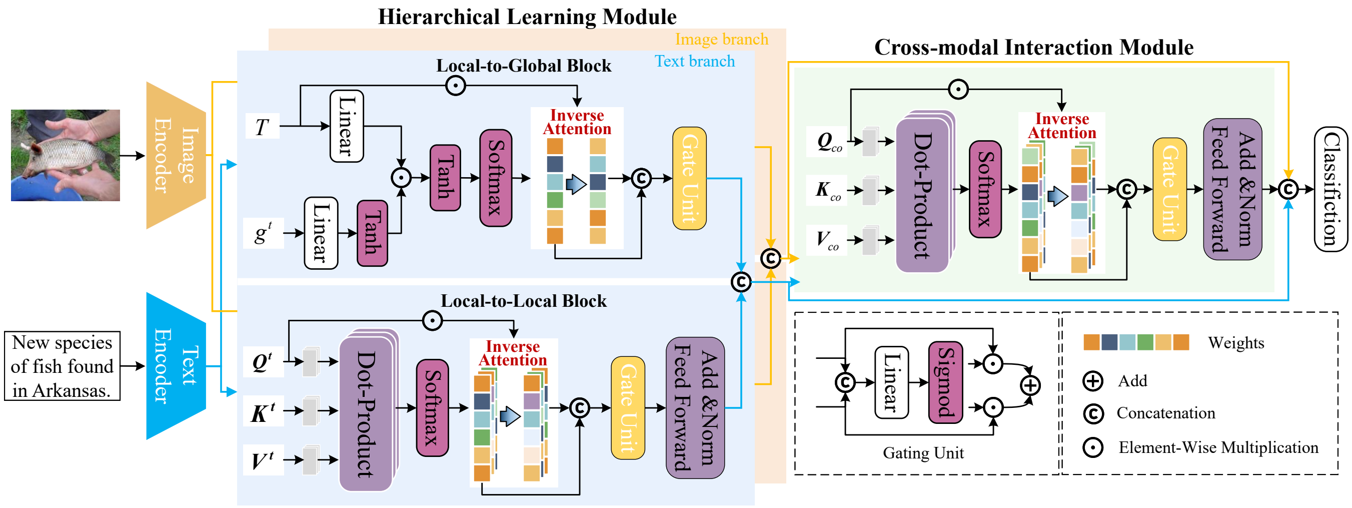

Figure 2: The proposed framework, MIAN, aims to detect fake news by fully leveraging both textual and visual content in news articles. Given a news piece, the model first utilizes modality-specific encoders to extract unimodal embeddings. Next, we apply a hierarchical learning module with different attention mechanisms in the Local-to-Global and Local-to-Local Blocks to capture and enhance hierarchical feature interactions. The enhanced unimodal features are then input into a co-attention mechanism to generate multimodal fused features. Throughout the various attention mechanisms, we incorporate an inverse attention module to explicitly extract inconsistencies between different targets. Finally, all enhanced unimodal and multimodal features are fused for fake news detection.

2.2 External Knowledge-based Methods

Due to limitations in text length or the emergence of novel terms, discrepancies in understanding news content may arise. As a result, several studies have incorporated external knowledge extracted from social media [Zhang *et al.*, 2021], knowledge graphs [Zhang *et al.*, 2019; Dun *et al.*, 2021] or internet retrieval [Zhang *et al.*, 2024a; Qi *et al.*, 2024; Yu *et al.*, 2020] to enhance the performance of fake news detection. Zheng et al. [Zheng *et al.*, 2022] employed an attention mechanism to combine social graph features, composed of user and comment data, with textual and visual features of news, generating discriminative features for fake news detection. Wu et al. [Wu *et al.*, 2023] performed the deep fusion of text and image features based on three distinct reading patterns while leveraging the relationship between news comments and content to capture semantic inconsistencies. Zhang et al. [Zhang *et al.*, 2023] leveraged news text entities as prompts to guide a large-scale vision-language model in generating image entities that enhance the semantic knowledge of the content. Zhang et al. [Zhang *et al.*, 2024b] converted news images and externally relevant knowledge into pure textual representations, which were then combined with the original text and fed into a prompt-based large language model to predict the authenticity of the news. While some methods enhance the representation of news content by incorporating external knowledge, this simultaneously introduces the risk of noise, which is crucial for the accuracy of fake news detection. These challenges significantly hinder the progress of multimodal fake news detection.

3 Methodology

Fake news detection based on multimodal content aims to leverage both textual and corresponding visual information to evaluate the authenticity of news. Given a dataset of fake news detection $D = \{X, Y\}$, where $X = [X^T, X^V]$ consists of the textual component X^T and the visual component X^V, and $Y \in \{0, 1\}$ is the corresponding label indicating whether the news is fake (0) or real (1). The objective of the model is to learn a mapping function $F : (X^T, X^V) \to Y$ from the training set of D, which maps the input space (X^T, X^V) to the labels Y. The learned function F is then used to predict labels for the test instances.

Our MIAN aims to fully learn intrinsic discriminant features within and across modalities. The detailed architecture is illustrated in Figure 2. We first extract unimodal embeddings from the news text and images using modality-specific encoders. For each multimodal news instance, the text is represented as a sequence of m tokens $\boldsymbol{T} = \{t_1, t_2, ..., t_m\}$, where $\boldsymbol{T} \in \mathbb{R}^{m\times d}$ and each token $t_i \in \mathbb{R}^d$ is a d-dimensional vector obtained from a pre-trained BERT model [Devlin *et al.*, 2018]. Meanwhile, we employ a pre-trained ViT model [Dosovitskiy *et al.*, 2020] to encode the image content into u visual tokens $\boldsymbol{O} = \{o_1, o_2, ..., o_u\}$, where $\boldsymbol{O} \in \mathbb{R}^{u\times d}$. We then introduce a hierarchical learning module consisting of two components, the Local-to-Global Block and the Local-to-Local Block. This module enables multi-granularity interactions to enhance unimodal representations for both text and image content. A detailed description is provided in Section 3.1. Subsequently, MIAN fuses the enhanced multimodal representations using a co-attention mechanism in the cross-modal interaction module. A more detailed description is given in Section 3.2. Additionally, to enable the model to explicitly learn inconsistency features, we design an inverse attention mechanism and integrate it into both the hierarchical learning module and the cross-modal interaction module. The specifics are discussed in Section 3.3. Finally, the enhanced modality-specific representations from the hierarchical learning modules in both the text and image branches, along with the fused representation from the cross-modal interaction module, are concatenated to form the final news representation, which is then fed into the classifier to assess the news authenticity.

3.1 Hierarchical Learning Module (HLM)

To comprehensively explore the internal relationships within news text and image representations and to produce enhanced unimodal features, we propose the Hierarchical Learning Module (HLM). After obtaining modality-specific embeddings from pre-trained encoders for text and images, HLM further models intra-modal associations at different levels from two perspectives: the Local-to-Local Block, which captures contextual interactions among local features, and the Local-to-Global Block, which establishes dependency relationships between local and global features.

Local-to-Local Block. We employ a multi-head attention mechanism [Vaswani *et al.*, 2017] to learn consistent relationships between tokens and capture intrinsic structural features, thereby generating context-aware, refined local representations within each unimodal branch. Specifically, for the text modality, given the text embedding $\boldsymbol{T}$, we first perform different multi-head linear transformations to derive the query, key, and value matrices: $\boldsymbol{Q}_t^i = \boldsymbol{T}\boldsymbol{W}_{ll-q}^i$, $\boldsymbol{K}_t^i = \boldsymbol{T}\boldsymbol{W}_{ll-k}^i$ and $\boldsymbol{V}_t^i = \boldsymbol{T}\boldsymbol{W}_{ll-v}^i$, and then concatenate the outputs from the attention mechanism.

$$\boldsymbol{Att}_t^i = softmax\left(\frac{\boldsymbol{Q}_t^i \boldsymbol{K}_t^{i\top}}{\sqrt{d_k}}\right), \tag{1}$$

$$\boldsymbol{h}_t^i = \boldsymbol{Att}_t^i \boldsymbol{V}_t^i, \tag{2}$$

$$\bar{\boldsymbol{R}}_t = \left[h_t^1, ..., h_t^n\right] \boldsymbol{W}_{cat}, \tag{3}$$

where all $\boldsymbol{W}$ are learnable parameters, and d_k denotes the dimensionality. The n weighted value vectors are concatenated using $[\cdot]$. Finally, a series of residual connections and feedforward layers are applied to produce the context-aware, local-to-local enhanced features.

$$\hat{\boldsymbol{R}}_t^{ll} = LN(\boldsymbol{T} + \boldsymbol{W}_t^{fc1} \bar{\boldsymbol{R}}_t), \tag{4}$$

$$\boldsymbol{R}_t^{ll} = LN\left(\hat{\boldsymbol{R}}_t^{ll} + ReLU\left(\boldsymbol{W}_t^{fc2} \hat{\boldsymbol{R}}_t^{ll} + b\right)\right), \tag{5}$$

where $\boldsymbol{W}_{fc1}^t$, $\boldsymbol{W}_{fc2}^t$ and b are learnable parameters. $LN(\cdot)$ and $ReLU(\cdot)$ refer to layer normalization and activation function.

Similarly, the ViT encoder transforms the segmented image into a sequential representation of local features, enabling the application of the self-attention mechanism. Following the same procedure as described above, we derive $\boldsymbol{R}_o^{ll}$.

Local-to-Global Block. This block refines local features by integrating global context, using global features to bridge fine-grained local information with overarching semantics. Specifically, for the visual modality, given the image embedding O, we construct the global feature to encapsulate the overall semantics of the image from two complementary perspectives. On the one hand, the mean-pooled vector $g^{mean} = 1/n \sum_{i=1}^{n} o_i$ summarizes the spatial information of the entire image. On the other hand, we use the $[cls]$ token g^{cls} output from the modality-specific encoder to represent the semantic information of the salient objects in the image. Finally, these two representations are concatenated to form the global feature:

$$g_o = \left[g^{mean}, g^{cls}\right]. \tag{6}$$

In this block, the global features serve as the query guidance vector, while the local features function as the matched vectors. Their inner product is computed to measure the similarity between the global and local features. The resulting similarity scores are normalized via a softmax function, ensuring that the attention weights across all regions sum to 1. Specifically, the attention weights reflecting the consistency between the global high-level semantic features and the word-level semantic features of the text are computed as follows:

$$h_o = tanh(\boldsymbol{W}_o^1 \boldsymbol{O} \odot tanh(\boldsymbol{W}_o^2 g_o)), \tag{7}$$

$$att_o^{lg} = softmax\left(\boldsymbol{W}_o^3 h_o\right), \tag{8}$$

$$\boldsymbol{R}_o^{lg} = att_o^{lg} \boldsymbol{O}, \tag{9}$$

where all $\boldsymbol{W}$ are learnable parameters, h_o denotes the hidden state of the text attention function, and $\odot$ represents the element-wise multiplication. The process of calculating the weighted relationship $\boldsymbol{R}_t^{\mathrm{lg}}$ between global and local information in the news text branch follows the same procedure as in the image branch.

Finally, we concatenate the enhanced modality-specific features produced by the Local-to-Local Block and Local-to-Global Block in each modality branch to obtain the hierarchically enhanced representations. Specifically, for the text modality, the representation is $\boldsymbol{R}_t = [\boldsymbol{R}_t^{ll}, \boldsymbol{R}_t^{\mathrm{lg}}]$, and for the visual modality, it is $\boldsymbol{R}_o = [\boldsymbol{R}_o^{ll}, \boldsymbol{R}_o^{\mathrm{lg}}]$.

3.2 Cross-modal Interaction Module (CIM)

To fuse multimodal representations, we propose the Cross-modal Interaction Module (CIM), which leverages a co-attention mechanism to capture inter-modal dependencies and effectively integrate the hierarchically enhanced representations of news text and images. Specifically, to obtain the text features enriched with image information, given the hierarchical enhanced representations $\boldsymbol{R}_t$ and $\boldsymbol{R}_o$ from the HLM, we apply distinct multi-head linear projections $\boldsymbol{Q}_t^{co} = \boldsymbol{R}_t \boldsymbol{W}_t^{co-q}$, $\boldsymbol{K}_t^{co} = \boldsymbol{R}_t \boldsymbol{W}_t^{co-k}$ and $\boldsymbol{V}_t^{co} = \boldsymbol{R}_t \boldsymbol{W}_t^{co-v}$ to derive the query, key, and value matrices, respectively. Due to the similarity of operations to the Local-to-Local Block and space constraints, the specific details of the multi-head attention mechanism are omitted here.

$$\boldsymbol{Att}_{t \to o}^{co} = softmax\left(\frac{\boldsymbol{Q}_t^{co} \boldsymbol{K}_t^{co\top}}{\sqrt{d_k}}\right), \tag{10}$$

$$\bar{\boldsymbol{R}}_{t \to o}^{co} = \boldsymbol{Att}_{t \to o}^{co} \boldsymbol{V}_t^{co}, \tag{11}$$

$$\hat{\boldsymbol{R}}_t^{co} = LN(\boldsymbol{R}_t + \boldsymbol{W}_{co}^{fc1} \bar{\boldsymbol{R}}_{t \to o}^{co}), \tag{12}$$

$$\boldsymbol{R}_t^{co} = LN\left(\hat{\boldsymbol{R}}_t^{co} + ReLU\left(\boldsymbol{W}_{co}^{fc2} \hat{\boldsymbol{R}}_t^{co} + b\right)\right). \tag{13}$$

Similarly, after applying the above process, the fused features $\boldsymbol{R}_o^c o$ are obtained, which capture the influence of the image on the text.

3.3 Inverse Attention

Exploring intrinsic relationships between modalities remains a significant challenge in multimodal fake news detection, especially in cases involving mismatched textual and visual content. Although the widely adopted cross-attention mechanism has shown promise in modeling cross-modal interactions, it often fails to capture inconsistencies between cross-modal features effectively.

Specifically, cross-attention calculates the similarity between query and key vectors to determine which key vectors each query vector should attend to, and then applies the resulting similarity weights to compute a weighted sum of value vectors. While this mechanism effectively emphasizes similar content across modalities, it may inadvertently encourage the network to learn implicit pseudo-consistencies, leading to confusion in fake news detection. To address this limitation, we propose the inverse attention mechanism, which amplifies dissimilar features to enhance the learning of explicit inconsistency signals. The detailed explanation of the inverse attention mechanism employed in CIM is as follows:

$$\boldsymbol{Att}_{t\to o}^{co-ic} = softmax\left(\boldsymbol{A} - \boldsymbol{Att}_{t\to o}^{co}\right), \tag{14}$$

$$\bar{\boldsymbol{R}}_{t\to o}^{co-ic} = \boldsymbol{Att}_{t\to o}^{co-ic}\boldsymbol{V}_{t}^{co}, \tag{15}$$

where $\boldsymbol{A}$ is the introduced scalar matrix, which is used to subtract from the attention matrix $\boldsymbol{Att}_{t\to o}^{co}$ The matrix $\boldsymbol{Att}_{t\to o}^{co-ic}$ represents the inverse attention weights, where the values corresponding to inconsistent vectors are larger. $\bar{\boldsymbol{R}}_{t\to o}^{co-ic}$ denotes the desired explicit inconsistency features. Given the consistency feature $\bar{\boldsymbol{R}}_{t\to o}^{co}$ and the inconsistency features $\bar{\boldsymbol{R}}_{t\to o}^{co-ic}$, which provide distinct information for fake news detection, we employ a gating unit to combine them effectively:

$$g = \sigma\left(\boldsymbol{W}_g\left[\bar{\boldsymbol{R}}_{t\to o}^{co}, \bar{\boldsymbol{R}}_{t\to o}^{co-ic}\right] + b_g\right), \tag{16}$$

$$\boldsymbol{R}_{t\to o}^{co} = g \cdot \bar{\boldsymbol{R}}_{t\to o}^{co} + (1-g)\cdot \bar{\boldsymbol{R}}_{t\to o}^{co-ic}, \tag{17}$$

where, $\boldsymbol{W}_g$ and b_g represent learnable parameters, while $[\cdot]$ denotes the concatenation operation. The function σ is the sigmoid function, and g is the gating weight computed during the process.

Given that inconsistencies may also emerge within the content of fake news texts or images, the inverse attention mechanism is incorporated into both the Local-to-Global and Local-to-Local modules. Specifically, during the computation of $\boldsymbol{R}_t^{ll}$, $\boldsymbol{R}_o^{ll}$, $\boldsymbol{R}_t^{lg}$, and $\boldsymbol{R}_o^{lg}$, similar operations to those described in Equations 14 and 15 are applied. This design enables the model to capture inconsistent features at both intra- and inter-modal levels, thereby improving its effectiveness in detecting various types of fake news.

Notably, since interactions are performed at the token level, the values of individual elements directly influence the computation of inconsistency weights, particularly when processing textual embeddings. To alleviate the impact of irrelevant or missing information, we introduce a positional mask. Furthermore, due to the potential ambiguity in expression arising from the sequence order of text and the positional information of objects in images, and since the attention mechanism lacks an inherent notion of sequence order, we also incorporate positional encoding. The specific formulation is as follows:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right), \tag{18}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right), \tag{19}$$

where pos, i and d represent the position index, dimension index, and the dimension of the positional encoding, respectively.

Loss function. Finally, we concatenate the enhanced modal-specific features, enriched with intra-modal inconsistency signals, and the multimodal fused features, capturing inter-modal inconsistencies, to construct the news representation $\boldsymbol{R}_n$. This representation is subsequently input into a multi-layer perceptron-based classifier to predict confidence scores, $\hat{y} = MLP(\boldsymbol{R}_n)$. We define the loss function L using cross-entropy as follows:

$$L = -y\log(\hat{y}) - (1-y)\log(1-\hat{y}). \tag{20}$$

4 Experiments

The experiments in this study aim to address several key challenges in fake news detection by exploring the following research questions:

- **Q1:** Can MIAN effectively improve the performance of fake news detection?
- **Q2:** Does each component contribute to improving detection?
- **Q3:** Can the proposed designs improve the accuracy of detecting specific types of fake news?

4.1 Configurations

Dataset. Our experiments are evaluated on four real-world datasets: Weibo17, Weibo21, GossipCop, and PolitiFact, which cover nearly all publicly available datasets in this field. These datasets used in the experiments not only cover multiple domains but also include specific domain datasets where fake news frequently occurs. The detailed descriptions of these datasets are provided in Appendix A.

Baselines. To evaluate the effectiveness and performance of the proposed method in multimodal fake news detection. The baseline methods include classic models that leverage deep neural networks for extraction such as EANN [Wang *et al.*, 2018], Spotfake [Singhal *et al.*, 2019], HMCAN [Qian *et al.*, 2021] and RaCMC [Yu *et al.*, 2025], as well as methods exploring cross-modal relationships such as CAFE [Chen *et al.*, 2022], CMC [Wei *et al.*, 2022], BMR [Ying *et al.*, 2023] and MSACA [Wang *et al.*, 2024a], and integrating external knowledge for enhanced detection accuracy such as MRHF [Wu *et al.*, 2023]. The detailed descriptions of these methods can be found in Appendix B.

Implementation. All experiments were implemented using the PyTorch toolkit on an NVIDIA A100 GPU. We pad the sequence length of news text to 196 for all datasets, denoted

Datasets	Methods	Acc.	Fake News			Real News		
			Pre.	Rec.	F1	Pre.	Rec.	F1
Weibo17	CAFE [Chen *et al.*, 2022]	0.840	0.855	0.830	0.842	0.825	0.851	0.837
	CMC [Wei *et al.*, 2022]	0.908	0.940	0.869	0.899	0.876	0.945	0.907
	BMR [Ying *et al.*, 2023]	0.918	0.882	**0.948**	0.914	**0.942**	0.870	0.904
	MRHF [Wu *et al.*, 2023]	0.907	0.939	0.869	0.903	0.879	0.931	0.904
	MSACA [Wang *et al.*, 2024a]	0.903	0.935	0.873	0.903	0.872	0.935	0.902
	RaCMC[Yu *et al.*, 2025]	0.915	0.910	0.924	0.917	0.921	0.906	0.914
	MIAN	**0.936**	**0.950**	0.920	**0.935**	0.923	**0.952**	**0.937**
Weibo21	EANN [Wang *et al.*, 2018]	0.870	0.902	0.825	0.862	0.841	0.912	0.875
	SpotFake [Singhal *et al.*, 2019]	0.851	**0.953**	0.733	0.828	0.786	**0.964**	0.866
	CAFE [Chen *et al.*, 2022]	0.882	0.857	0.915	0.885	0.907	0.844	0.876
	BMR [Ying *et al.*, 2023]	0.929	0.908	**0.947**	0.927	0.946	0.906	0.925
	MIAN	**0.938**	0.924	**0.947**	**0.936**	**0.950**	0.928	**0.939**
GossipCop	EANN [Wang *et al.*, 2018]	0.864	0.702	0.518	0.594	0.887	0.956	0.920
	SpotFake [Singhal *et al.*, 2019]	0.858	0.732	0.372	0.494	0.866	0.962	0.914
	CAFE [Chen *et al.*, 2022]	0.867	0.732	0.490	0.587	0.887	0.957	0.921
	CMC [Wei *et al.*, 2022]	0.893	0.826	0.657	0.692	0.920	0.963	0.935
	BMR [Ying *et al.*, 2023]	0.895	0.752	0.639	0.691	0.920	0.965	0.936
	MSACA [Wang *et al.*, 2024a]	0.887	0.816	0.538	0.646	0.897	**0.971**	0.933
	RaCMC[Yu *et al.*, 2025]	0.879	0.745	0.563	0.641	0.902	0.954	0.927
	MIAN	**0.923**	**0.834**	**0.872**	**0.853**	**0.956**	0.941	**0.948**
PolitiFact	HMCAN [Qian *et al.*, 2021]	0.864	0.738	**0.933**	0.824	**0.960**	0.828	0.889
	CAFE [Chen *et al.*, 2022]	0.864	0.724	0.778	0.750	0.895	0.919	0.907
	LII [Singhal *et al.*, 2022]	0.907	0.895	0.872	0.883	0.900	0.918	0.909
	CMC [Wei *et al.*, 2022]	0.893	0.826	0.657	0.692	0.920	0.963	0.935
	RaCMC[Yu *et al.*, 2025]	0.922	0.833	0.926	0.877	0.967	0.921	0.943
	MIAN	**0.953**	**0.971**	0.919	**0.944**	0.941	**0.980**	**0.960**

Table 1: Results of comparison among different approaches on Weibo17, Weibo21, GossipCop and PolitiFact Datasets. The best performance is highlighted in bold, and the follow-up is highlighted in underlined.

as $m = 196$, and put the processed text into the BERT. Images were resized to 224×224 pixels and subsequently divided into $u = 14 \times 14$ patches, which were then input into the vit-patch16-224. The multi-head attention used in both the Local-to-Local Block and the Cross-modal Interaction Module consists of 2 layers with 12 heads. The initial learning rate was set to $2e^{-6}$, and we utilized the StepLR decay strategy with a decay step of 20 and a decay rate of 0.5 to help the model converge better.

4.2 Overall Performance

Table 1 presents the performance results of the proposed method and the comparison methods on the public datasets. Although our method does not achieve the highest precision or recall, it outperforms the other comparison methods in terms of accuracy and F1 score, demonstrating its superior balance and overall performance, thus resolving **Q1**.

EANN and SpotFake, as early classic methods, simply concatenate the multimodal features obtained from modal-specific encoders to represent news, and do not perform well. This is likely due to their failure to further explore unimodal features and the lack of cross-modal interactions. CAFE and BMR highlight the importance of unimodal features in decision-making, with BMR achieving suboptimal results on certain datasets, suggesting that unimodal news content contributes to fake news detection. HMCAN and MSACA attempt to extract multi-level unimodal features from modal-specific encoders but do not yield strong performance, indicating that when exploring unimodal features, it is crucial to consider their specific role in the fake news detection task and fully leverage the interactions between them. Moreover, HMCAN exhibits a significantly higher F1 score for real news compared to fake news, which could be due to the use of a co-attention mechanism to model cross-modal information, where the model overemphasizes modal consistency and overlooks the stronger inconsistencies present in fake news.

Datasets	w/o models	Acc.	Fake F1	Real F1
Weibo17	**MIAN**	**0.936**	**0.935**	**0.937**
	w/o intra-lg	0.906	0.902	0.909
	w/o intra-lg-ic	0.910	0.907	0.912
	w/o intra-ll-ic	0.913	0.910	0.916
	w/o inter-ic	0.908	0.905	0.910

Table 2: The ablation experiment on the F1 score of the variant structure design on the Weibo17 dataset.

Additionally, we provide the t-SNE visualization of the features for real and fake news in Figure 3. Since the LII model also generates discriminative features through enhanced intra- and inter-modal methods, it serves as a comparison model to MIAN. It demonstrates that the features learned by the proposed model are more discriminative.

4.3 Ablation Studies

Impact of each component. In order to address **Q2**, we evaluate MIAN against its variants by systematically removing

Methods	Real	Fake		
	0	4	5	Avg.
HMCAN	0.776	0.804	0.968	0.822
CAFE	0.907	0.794	0.938	0.842
LII	0.909	0.840	0.941	0.773
MIAN	**0.920**	**0.843**	**0.972**	**0.870**

Table 3: Performance comparison to different models in terms of accuracy on the True, False Connection, and Manipulation Content classes of the Fakeddit dataset.

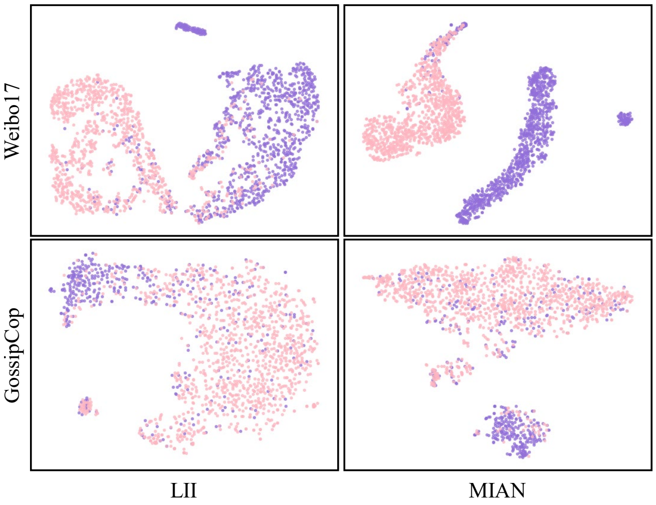

Figure 3: T-SNE visualization of the mined features on the test set. Each color indicates a different class of data points.

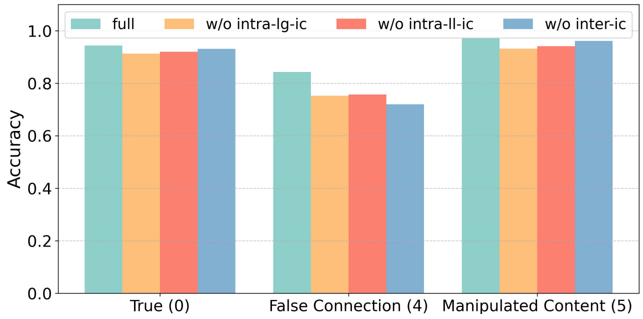

Figure 4: The performance of different variants of the proposed model in terms of accuracy on the True, False Connection, and Manipulation Content classes in the Fakeddit dataset.

specific components. Specifically, w/o intra-lg refers to eliminating the Local-to-Global Block, thereby removing the interaction between global and local features within both text and image branches. w/o intra-lg-ic and w/o intra-ll-ic refer to removing the inconsistent features generated by inverse attention at different levels within the same modality (Local-to-Global and Local-to-Local, respectively). w/o inter-ic refers to excluding the inconsistent features in the CIM.

Table 2 presents the results of the ablation experiments comparing various method variants, demonstrating the following key findings: (1) In the multimodal fake news detection task, comprehensive and in-depth unimodal information is essential. Extracting features from text or images aids the model in understanding the news content more effectively. (2) The loss of inconsistent information at any stage significantly impacts the accuracy of the detection results, particularly for fake news. Other metrics and results on the GossipCop dataset can be found in the Appendix C.1.

Impact of different types. To further validate the effectiveness of the proposed method in identifying both fabricated news and mismatched news, we introduce the Fakeddit dataset [Nakamura *et al.*, 2019], which includes multiple fake news categories. In this dataset, label 0 represents real news, label 4 corresponds to False Connection, and label 5 denotes Manipulation Content, all of which align with the focus of our study. This experiment answers **Q3**, demonstrating that our specific design significantly enhances the accuracy of detecting various types of news.

For the comparative experiments, three representative methods were selected for comparison with our approach: HMCAN, which employs a co-attention mechanism to fuse multimodal information; CAFE, which utilizes the semantic differences between modalities as auxiliary tasks; and LII, which simultaneously leverages both intra-modal and inter-modal relationships. The results, presented in Table 3, demonstrate that our method achieves superior performance.

Furthermore, the ablation study, shown in Figure 4, reveals that removing unimodal modeling or inconsistency feature extraction results in notable accuracy degradation in categories 5 (Manipulated Content) and 4 (False Connection). This indicates that the proposed modules effectively identify both fabricated news and mismatched news. The details of the comparison and ablation experiments on other categories of the Fakeddit dataset can be found in Appendix C.2.

5 Conclusion & Limitations

In this paper, we propose MIAN, a novel multimodal fake news detection method that effectively identifies various types of fake news. MIAN explores hierarchical interactions within news text and image contents to generate enhanced unimodal representations, and then utilizes a co-attention mechanism to model inter-modal dependencies for multimodal feature fusion. Additionally, we introduce an inverse attention mechanism from a new perspective to explicitly learn and extract intra- and inter-modal inconsistencies. Experimental results demonstrate that the proposed method achieves optimal accuracy and effectively detects different types of fake news.

Despite the promising results, our method currently relies on a single image per news article, which could lead to information bias between the two modalities. Future work could address this by extending the approach to incorporate multiple images per news article.

Acknowledgments

This work was supported by the Joint Project for Innovation and Development of Shandong Natural Science Foundation, (No. ZR2022LZH012).

Contribution Statement

Tianlin Zhang and En Yu contributed equally to this work and should be considered as co-first authors. Corresponding Author: Jiande Sun.

References

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

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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

[Jiang *et al.*, 2023] Ye Jiang, Xiaomin Yu, Yimin Wang, Xiaoman Xu, Xingyi Song, and Diana Maynard. Similarity-aware multimodal prompt learning for fake news detection. *Information Sciences*, 647:119446, 2023.

[Li *et al.*, 2022] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[Li *et al.*, 2023] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

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

[Nakamura *et al.*, 2019] Kai Nakamura, Sharon Levy, and William Yang Wang. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854*, 2019.

[Ouyang *et al.*, 2022] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[Qi *et al.*, 2021] Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1212–1220, 2021.

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

[Qian *et al.*, 2021] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. Hierarchical multimodal 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, 2021.

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

[Shu *et al.*, 2017] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter*, 19(1):22–36, 2017.

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

[Singhal *et al.*, 2022] Shivangi Singhal, Tanisha Pandey, Saksham Mrig, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. Leveraging intra and inter modality relationship for multimodal fake news detection. In *Companion Proceedings of the Web Conference 2022*, pages 726–734, 2022.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2018] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 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, 2018.

[Wang *et al.*, 2024a] Jiandong Wang, Hongguang Zhang, Chun Liu, and Xiongjun Yang. Fake news detection via multi-scale 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, 2024.

[Wang *et al.*, 2024b] Li Wang, Xiangtao Meng, Dan Li, Xuhong Zhang, Shouling Ji, and Shanqing Guo. Deepfaker: a unified evaluation platform for facial deepfake and detection models. *ACM Transactions on Privacy and Security*, 27(1):1–34, 2024.

[Wei *et al.*, 2022] Zimian Wei, Hengyue Pan, Linbo Qiao, Xin Niu, Peijie Dong, and Dongsheng Li. Cross-modal knowledge distillation in multi-modal fake news detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4733–4737. IEEE, 2022.

[Wu *et al.*, 2023] Lianwei Wu, Pusheng Liu, and Yanning Zhang. 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, 2023.

[Wu *et al.*, 2024] Lianwei Wu, Linyong Wang, and Yongqiang Zhao. Unified evidence enhancement inference framework for fake news detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6541–6549, 2024.

[Xue *et al.*, 2021] Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management*, 58(5):102610, 2021.

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

[Yu *et al.*, 2017] Feng Yu, Qiang Liu, Shu Wu, Liang Wang, Tieniu Tan, et al. A convolutional approach for misinformation identification. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3901–3907, 2017.

[Yu *et al.*, 2020] En Yu, Jing Li, Li Wang, Jia Zhang, Wenbo Wan, and Jiande Sun. Multi-class joint subspace learning for cross-modal retrieval. *Pattern Recognition Letters*, 130:165–173, 2020.

[Yu *et al.*, 2022] En Yu, Jianhua Ma, Jiande Sun, Xiaojun Chang, Huaxiang Zhang, and Alexander G Hauptmann. Deep discrete cross-modal hashing with multiple supervision. *Neurocomputing*, 486:215–224, 2022.

[Yu *et al.*, 2025] Xinquan Yu, Ziqi Sheng, Wei Lu, Xiangyang Luo, and Jiantao Zhou. Racmc: Residual-aware compensation network with multi-granularity constraints for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 986–994, 2025.

[Zhang *et al.*, 2019] Huaiwen Zhang, Quan Fang, Shengsheng Qian, and Changsheng Xu. Multi-modal knowledge-aware event memory network for social media rumor detection. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1942–1951, 2019.

[Zhang *et al.*, 2021] Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021*, pages 3465–3476, 2021.

[Zhang *et al.*, 2023] Qiang Zhang, Jiawei Liu, Fanrui Zhang, Jingyi Xie, and Zheng-Jun Zha. Hierarchical semantic enhancement network for multimodal fake news detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3424–3433, 2023.

[Zhang *et al.*, 2024a] Fanrui Zhang, Jiawei Liu, Jingyi Xie, Qiang Zhang, Yongchao Xu, and Zheng-Jun Zha. Escnet: Entity-enhanced and stance checking network for multimodal fact-checking. In *Proceedings of the ACM on Web Conference 2024*, pages 2429–2440, 2024.

[Zhang *et al.*, 2024b] Qiang Zhang, Jiawei Liu, Fanrui Zhang, Jingyi Xie, and Zheng-Jun Zha. Natural language-centered inference network for multi-modal fake news detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, volume 2024, pages 2542–2550, 2024.

[Zhao *et al.*, 2023] Liang Zhao, En Yu, Zheng Ge, Jinrong Yang, Haoran Wei, Hongyu Zhou, Jianjian Sun, Yuang Peng, Runpei Dong, Chunrui Han, et al. Chatspot: Bootstrapping multimodal llms via precise referring instruction tuning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 1743–1752, 2023.

[Zheng *et al.*, 2022] Jiaqi Zheng, Xi Zhang, Sanchuan Guo, Quan Wang, Wenyu Zang, and Yongdong Zhang. Mfan: Multi-modal feature-enhanced attention networks for rumor detection. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, volume 2022, pages 2413–2419, 2022.

[Zhou and Zafarani, 2020] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40, 2020.

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