Toward Multimodal Fake News Detection by Multi-perspective Rationale Generation and Verification

Junyang Chen[1,2], Yueqian Li[1], Ka Chung Ng[3], Huan Wang[4], Liang-Jie Zhang[*1]

[1]College of Computer Science and Software Engineering, Shenzhen University, China
[2]State Key Lab. for Novel Software Technology, Nanjing University, China
[3]Department of Management and Marketing, Faculty of Business, The Hong Kong Polytechnic University, China
[4]College of Informatics, Huazhong Agricultural University, China
junyangchen@szu.edu.cn

Abstract

The rapid proliferation of social media platforms has led to a surge in multimodal fake news, where deceptive content often combines text and images to mislead audiences. Traditional unimodal detection methods struggle to address the complexity of such content, necessitating holistic multimodal approaches. While the latest advancements in Multimodal Large Language Models (MLLMs) offer new opportunities for enhancing detection performance by analyzing multi-dimensional features, including source credibility, cross-modal contradictions, emotional bias, and manipulative writing patterns, these methods suffer from a key flaw: a susceptibility to hallucinations or erroneous reasoning, which can lead to flawed conclusions and ultimately biased detection results. To mitigate this challenge, we propose the Multimodal Fake News Detection via Multi-perspective Rationale Generation and Verification (MMRGV) model. Our method employs a cross-verification mechanism to screen and reconcile contradictions among different rationales, thereby preserving the LLM's analytical advantages while mitigating the impact of erroneous reasoning or hallucinations on the final detection. Subsequently, these optimized rationales are fused via an adaptive weighting strategy to output a robust final prediction. Extensive experiments on three benchmark datasets (Twitter, Weibo, and GossipCop) demonstrate the superiority of our method, achieving state-of-the-art accuracy of 0.9972, 0.9663, and 0.8772 respectively, and significantly outperforming existing baselines. These results validate the effectiveness of multi-perspective rationale generation and cross-verification in enhancing multimodal fake news detection, offering a resilient solution to combat misinformation in the era of generative AI.

1 Introduction

With the rapid development of social media platforms, multimodal messages—combining text, images, videos, and other media—have become ubiquitous in daily communication (Alam et al. 2022). However, this shift has also facilitated the spread of multimodal fake news, which poses severe societal risks. Compared to their text-only counterparts, multimodal fake news leverages multisensory stimuli (e.g., shocking visuals paired with exaggerated text) to enhance information virality (Khattar et al. 2019; Li et al. 2017), thereby necessitating robust automated detection systems to support effective digital governance. The key challenges in multimodal fake news detection involve cross-modal fusion and contextual reasoning (Hao et al. 2025; Cao et al. 2025). Traditional small models(e.g., BERT (Devlin et al. 2018), RoBERTa (Liu et al. 2019)) frequently lack the necessary knowledge and reasoning capabilities (Zheng, Luo, and Wang 2025; Wang et al. 2024a):

1. **Cross-modal logical contradictions.** As exemplified in Figure 1(a), a textual claim (e.g., an earthquake in Nepal) can starkly contradict the visual evidence (e.g., an undamaged location). Detecting such mismatches requires both fine-grained cross-modal alignment and commonsense validation.
2. **Factual forgery tactics.** Adversaries exploit both visual and textual channels. This includes **image manipulation** (Fig. 1(b)), where spliced visuals like sharks in a subway create sensationalism, and **textual forgery** (Fig. 1(c)), which uses biologically implausible claims (e.g., "14 children from 14 different fathers") to mislead.

Recent advancements in multimodal large language models (MLLMs), such as Qwen2-VL (Wang et al. 2024b) and LLaVA (Liu et al. 2024), can potentially address these challenges by leveraging extensive pretrained knowledge bases and sophisticated cross-modal alignment techniques. For instance, Qwen2-VL can assess whether an image's background aligns with textual claims about its location, while LLaVA can identify logical fallacies in exaggerated headlines by cross-referencing commonsense knowledge. However, LLM-based detection approaches face critical limitations: they are prone to hallucinating plausible-sounding yet fabricated details (Ji et al. 2023; Lin, Hilton, and Evans 2021), making overconfident assertions about unverified claims (Kadavath et al. 2022), and amplifying misinformation due to their autoregressive nature (Lin, Hilton, and Evans 2021). These issues can ultimately erode trust in automated verification systems and complicate the assessment of truthfulness. Thus, deploying MLLMs directly for fake news detection remains unreliable without systematic cross-verification mechanisms to validate their inferences against authenticity signals.

*Corresponding author

(a) Two and a half year old sister protected by four year old brother in # NepalEarthquake! (Inconsistency in location)

(b) New Gang moves into New York and takes over the subway...(image manipulation)

(c) In Michigan:Woman enters Guinness book of record after 14 children from 14 different fathers...(illogical)

Figure 1: Three common multimodal fake news tactics include: (a) cross-modal inconsistency between image and text, (b) manipulated or forged images, and (c) exaggerated or fabricated textual narratives.

To ensure the reliability of LLM-generated judgments, this study proposes the **M**ultimodal Fake News Detection via **M**ulti-perspective **R**ationale **G**eneration and **V**erification (**MMRGV**) framework. Inspired by the Chain of Thought (CoT) prompting strategy (Kojima et al. 2022), MMRGV directs the MLLM to generate explicit and detailed reasoning paths from multiple perspectives, which we refer to as *rationales*. To handle potential hallucinations and reasoning errors from the MLLM, a collaborative verification mechanism is then introduced. In this mechanism, multiple task-specific, fine-tuned models assess the generated rationales, effectively filtering out erroneous information while retaining high-quality evidence. The verified, multi-perspective rationales are ultimately synthesized for an accurate final prediction.

Specifically, building upon the three typical fake news patterns illustrated in Figure 1, we devise three corresponding verification perspectives. These perspectives are grounded in established detection paradigms and systematically cover all core modality combinations (text-only, text-image, and image-only) to ensure a comprehensive analysis:

1. **Textual Description (Text-only):** MLLMs deconstruct news content for textual cues (e.g., source, summary, style, tone), mirroring established linguistic-based detection (e.g., (Xiao et al. 2024)).
2. **Image-Text Consistency (Text-image):** MLLMs examine key event elements (e.g., time, location, people) for visual-textual alignment, addressing the core task of cross-modal verification (e.g., (Chen et al. 2022; Lao et al. 2024)).
3. **Image Description (Image-only):** MLLMs analyze the image for origin, forgery/tampering, and commonsense validation, drawing from principles of image forensics (e.g., (Sharma et al. 2023)).

To fully integrate these multi-perspective rationales, we introduce a method for multimodal, multi-view Rationale verification and fusion. This approach first filters erroneous information from the rationales to obtain refined multimodal cues, and subsequently fuses these cues from all perspectives to make the final prediction.

In summary, our main contributions are threefold:

1. We propose a novel detection framework, **MMRGV**, which introduces a collaborative verification mechanism to filter and refine rationales generated by a Chain-of-Thought (CoT) prompted LLM. This allows us to harness the advanced reasoning of LLMs while ensuring high reliability.
2. We establish a systematic analysis framework that deconstructs complex multimodal fake news into three core verification perspectives: textual description, image-text consistency, and image description. This provides a structured and actionable approach to the detection task.
3. We conduct extensive experiments on three real-world datasets, demonstrating that our proposed MMRGV model achieves state-of-the-art performance and significantly outperforms existing baselines.

2 Related Work

2.1 Multimodal Fake News Detection

Multimodal fake news detection leverages both advanced unimodal features (e.g., syntax-aware textual analysis (Xiao et al. 2024)) and cross-modal alignment to identify inconsistencies across different modalities. Early approaches primarily relied on naive fusion strategies, such as direct concatenation of image and text features (Singhal et al. 2019). However, these methods struggled to capture meaningful cross-modal relationships due to inherent semantic gaps between modalities and the lack of temporal or spatial alignment. To address these limitations, subsequent studies (Chen et al. 2021; Wu et al. 2021) introduced self-attention mechanisms to align and fuse modality-specific features, thereby facilitating improved semantic integration. Variational autoencoder-based methods (Chen et al. 2022) further advanced this process by modeling the latent spatial distributions of different modalities. These methods defined cross-modal ambiguity in terms of temporal or spatial misalignment, using symmetric Kullback–Leibler divergence between latent distributions as a trade-off between cross-modal coherence and unimodal distinctiveness, ultimately improving prediction accuracy. More recently, the emergence of Transformer-based pre-trained models, such as CLIP (Radford et al. 2021), BERT (Devlin et al. 2018), ViT (Dosovitskiy et al. 2020), and Swin-T (Liu et al. 2021), has established a new paradigm for multimodal fake news detection. For example, Zhou et al. (2023) utilized CLIP,

BERT, and Swin-T to perform multi-granularity, multimodal fusion, incorporating similarity-weighted unimodal branches to adaptively regulate each modality's contribution and mitigate cross-modal ambiguity. Beyond time-domain attention mechanisms, FSRU (Lao et al. 2024) proposed a frequency-domain approach, leveraging the Fast Fourier Transform to fuse multimodal features by transforming them into the frequency domain. While existing multimodal fake news detection methods have achieved significant progress, the recent emergence of LLMs highlights the need to integrate these models into traditional frameworks to further improve performance.

2.2 Application of LLMs in Fake News Detection

With the rise of LLMs, studies have increasingly explored their innovative applications in fake news detection. In the context of model collaboration, Hu et al. (2024) systematically reveals the complementary characteristics of LLMs and Smaller Language Models (SLMs). Specifically, while the zero-shot performance of unfine-tuned GPT-3.5 is limited, its reasoning capabilities improve significantly with CoT prompting. In contrast, fine-tuned BERT consistently demonstrates superior performance in task-specific scenarios. Building on these insights, the RAG collaboration framework (Lyu et al. 2025) effectively integrates the clue-discovery strengths of LLMs with the multi-dimensional feature synthesis abilities of SLMs. To address the limitations of LLMs in fact-checking, Zhang and Gao (2023) proposed a hierarchical, step-by-step prompting method. By decomposing complex news claims into sub-statements and performing multi-step verification through search engines, this approach effectively mitigates information omissions and factual hallucinations in LLM-based fact-checking. Notably, innovative counterfactual data augmentation techniques have also emerged. For example, Nan et al. (2024) employed LLMs to generate diverse user comments, especially those from typically silent users, and leverages multi-subgroup feedback analysis to enhance fake news detection. Experimental results show that these generated comments are often more effective and comprehensive than real ones. Despite these advancements, the persistent issue of reasoning hallucinations in LLMs continues to hinder their direct application in fake news detection, thus motivating a growing body of research focused on further improving LLM reliability in this domain.

3 Methodology

The proposed MMRGV framework, as illustrated in Figure 2, is architecturally divided into two primary stages: (1) Multi-perspective Rationale Generation, driven by a MLLM, and (2) Multi-perspective Cross-Verification, which assesses the reliability of the generated rationales. The latter stage forms the core of our verification process and is implemented through three key modules: feature extraction, rationale content gate fusion, and multi-view aggregation.

3.1 Multi-perspective Rationale Generation

In this stage, we leverage the inherent reasoning capabilities of a Multimodal Large Language Model (MLLM) to produce rationales from the following three perspectives:

(1) **Textual Description (TD)**: The MLLM analyzes the news text to extract various text-related cues, including event dates, news sources, logical coherence, and emotional sentiment.

(2) **Image-Text Consistency (ITC)**: The MLLM examines the alignment between the news image and its corresponding text, identifying discrepancies such as temporal inconsistencies, location mismatches, character incongruities, and conflicting event details.

(3) **Image Description (ID)**: The MLLM independently analyzes the news image to detect cues related to image provenance and manipulation, such as watermarks, splicing, and tampering.

We adopt the following notation: For each news instance evaluated by MLLM, the view set is defined as $\mathcal{V} = \{\mathrm{TD}, \mathrm{ITC}, \mathrm{ID}\}$, where the original news label is denoted as $y \in \{0, 1\}$ (0 indicates "real", 1 indicates "fake"). The MLLM predicted label for each view $v \in \mathcal{V}$ is $y_J^{(v)} \in \{0, 1, 2\}$ (0 indicates "real", 1 indicates "fake", and 2 indicates "undetermined"). The MLLM judgment correctness label for each view v, $y_A^{(v)} \in \{0, 1\}$, is 1 if $y_J^{(v)} = y$, and 0 if $y_J^{(v)} \neq y$ or $y_J^{(v)} = 2$. This label is crucial as it serves as the ground-truth supervisory signal for training the models in our subsequent verification stage.

3.2 Feature Extraction

We extract feature representations using RoBERTa (Liu et al. 2019) for textual content and Swin-T (Liu et al. 2021) for visuals. RoBERTa encodes the original news text and each rationale into their respective representations, X_{TD} and $R_v, v \in \mathcal{V}$, while Swin-T yields the image features X_{ID}. To capture cross-modal consistency, we first prompt the MLLM to generate an image caption. The news text, concatenated with this caption, is then encoded by RoBERTa into the consistency-focused feature vector, X_{ITC}.

To bridge the modality gap between the visual features (X_{ID}) and their corresponding textual ID Rationales (R_{ID}), we employ a weighted contrastive loss for alignment. First, we project the pooled outputs of the image features $X_{\mathrm{ID}}^{(i)}$ and the ID Rationale $R_{\mathrm{ID}}^{(i)}$ into their respective latent vectors, $z_{\mathrm{img}}^{(i)}$ and $z_{\mathrm{text}}^{(i)}$, using Multi-Layer Perceptron(MLP)[1]. The contrastive loss $\mathcal{L}_C$ is then computed as the negative log-likelihood of matching correct pairs. To mitigate noise from flawed ID Rationales, this loss incorporates the correctness label $y_A^{(\mathrm{ID})} \in \{0, 1\}$, ensuring only correct rationales contribute:

$$\mathcal{L}_C = \frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} -y_A^{(i,\mathrm{ID})} \log p(i, i), \tag{1}$$

where $p(i, i)$ is the softmax probability of pairing the i-th

[1]For conciseness, here and subsequently, we use the general notation 'MLP'. Note that while sharing the same notation, each MLP has independent parameters and a distinct architecture. Detailed architectures are provided in the supplementary materials.

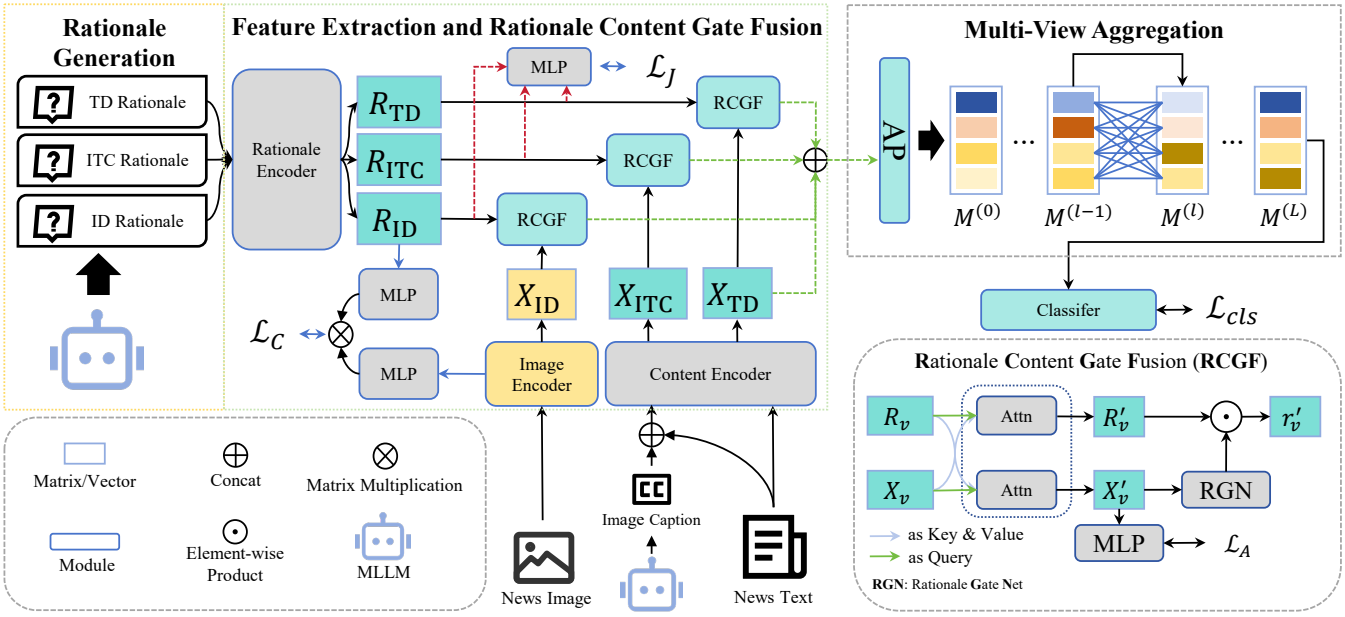

Figure 2: Overview of the MMRGV framework. The process includes two main stages: (1) An MLLM generates multi-perspective rationales. (2) A verification and fusion stage that comprises: (i) feature extraction from multimodal content and rationales; (ii) Rationale Content Gate Fusion (RCGF), which filters and re-weights rationales based on their predicted correctness; and (iii) Multi-View Aggregation for final classification.

image with the i-th ID Rationale:

$$p(i,i) = \frac{\exp\left((z_{\text{img}}^{(i)} \cdot z_{\text{text}}^{(i)})/\tau\right)}{\sum_{k=1}^{|\mathcal{B}|} \exp\left((z_{\text{img}}^{(i)} \cdot z_{\text{text}}^{(k)})/\tau\right)}, \tag{2}$$

where τ is the temperature hyperparameter used to scale the similarity scores. Next, to ensure the rationale encoder better captures the semantic information of the generated rationales, we introduce an auxiliary task of predicting the MLLM's original judgment $y_J^{(v)}$. This auxiliary task can distill the MLLM's reasoning into the rationale encoder by optimizing the following objective:

$$\mathcal{L}_J = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \text{CE}\left(\text{MLP}\left(R_v\right), y_J^{(v)}\right), \tag{3}$$

where $|\mathcal{V}|$ represents the number of views (fixed to 3 in this work) and CE stands for cross-entropy Loss. In general, we use $\mathcal{L}_J$ and $\mathcal{L}_C$ to enhance features extracted from images and LLM-generated rationales.

3.3 Rationale Content Gate Fusion for Cross-Verification

We design a rationale content gate fusion mechanism to perform cross-verification for each view $v \in \mathcal{V}$. We use scaled dot-product attention to fuse R_v and X_v (for $v \in \mathcal{V}$), defined as follows:

$$\text{Attn}(Q,K,V) = \text{softmax}\left(Q'(K')^\top/\sqrt{d}\right)V', \tag{4}$$

where $Q' = QW_q, K' = KW_k, V' = VW_v$ (with $W_q, W_k, W_v \in \mathbb{R}^{d\times d}$ being learnable projection matrices), and d is the feature dimension. Then, the updated representations R'_v and X'_v are computed for each view $v \in \mathcal{V}$:

$$R'_v = \text{Attn}(R_v, X_v, X_v), \tag{5}$$

$$X'_v = \text{Attn}(X_v, R_v, R_v). \tag{6}$$

To distinguish beneficial from detrimental rationales, we predict the rationale correctness $y_A^{(v)}$ for each view $v \in \mathcal{V}$. As most rationales are correct, the resulting class imbalance motivates our use of Focal Loss (Lin et al. 2017). The objective $\mathcal{L}_A$ is the average Focal Loss over all views:

$$\mathcal{L}_A = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \text{FL}(y_A^{(v)}, \hat{y}_A^{(v)}), \tag{7}$$

where the predicted probability is $\hat{y}_A^{(v)} = \sigma(\text{MLP}(X'_v))$. The Focal Loss function $\text{FL}(y, \hat{y})$ is defined as:

$$\begin{aligned}\text{FL}(y,\hat{y}) = &-\alpha(1-\hat{y})^\gamma y \log(\hat{y}) \\ &-(1-\alpha)\hat{y}^\gamma(1-y)\log(1-\hat{y}),\end{aligned} \tag{8}$$

with α and γ as its standard hyperparameters. To further leverage the learned rationale representation R_v, we propose the Rationale Gate Net, which utilizes X'_v to determine gate value g_v for each view $v \in \mathcal{V}$:

$$r'_v = g_v \odot \text{Avg}(R'_v), \quad v \in \mathcal{V}, \tag{9}$$

$$g_v = \sigma(\text{MLP}(X'_v)), \quad g_v \in \mathbb{R}, \tag{10}$$

where $\odot$ is the element-wise product, $\text{Avg}(\cdot)$ averages over the sequence dimension, and $r'_v \in \mathbb{R}^d$ is the scaled, pooled rationale feature for perspective v. Note that while the gate value g_v and the correctness probability $\hat{y}_A^{(v)}$ share an identical functional form ($\sigma(\text{MLP}(X'_v))$), their underlying MLP modules use independent parameters.

3.4 Multi-View Aggregation Module for Classification

To preserve the original news text information, we apply attentive pooling (AP) (Hu et al. 2024) to the news's textual features X_{TD}, yielding a vector $x_{\text{TD}} \in \mathbb{R}^d$. We then concatenate the scaled rationales $r'_v, v \in \mathcal{V}$ with x_{TD} along the first dimension, resulting in the multi-view feature matrix $M^{(0)}$:

$$x_{\text{TD}} = \text{AP}(X_{\text{TD}}) = \text{softmax}(w \cdot X_{\text{TD}}^{\top}) X_{\text{TD}}, \tag{11}$$

$$M^{(0)} = [x_{\text{TD}}; r'_{\text{TD}}; r'_{\text{ITC}}; r'_{\text{ID}}], \quad M^{(0)} \in \mathbb{R}^{4\times d}, \tag{12}$$

where $w \in \mathbb{R}^{1\times d}$ is a learnable vector.

The resulting matrix $M^{(0)}$, combining features from the original text and the three rationales, serves as input to our multi-view aggregation module. This module iteratively refines these representations over L layers via a view-specific co-attention mechanism. At each layer l, for each view i, we compute a unique attention vector $\tau_i^{(l)}$ that represents the importance of all views relative to view i. This vector is then used to perform a weighted sum over all features in $M^{(l-1)}$ to create a view-specific context vector, which updates the prior state $M_i^{(l-1)}$ via a residual connection:

$$M_i^{(l)} = (\tau_i^{(l)})^{\top} \cdot M^{(l-1)} + M_i^{(l-1)}, \quad 0 < l \leq L, \tag{13}$$

$$\tau_i^{(l)} = \text{softmax}(M^{(l-1)} \cdot w_i^{(l)}), \quad \tau_i^{(l)} \in \mathbb{R}^{4\times 1}, \tag{14}$$

where $w_i^{(l)} \in \mathbb{R}^d$ is a learnable, view-specific parameter vector.

Classification Task Finally, the aggregated representation $M^{(L)}$ is passed to a classifier to obtain the final prediction probability $\hat{y} = \sigma(\text{MLP}(M^{(L)}))$. To address potential class imbalance in datasets like GossipCop (Shu et al. 2020), we use the FL function (Eq. (8)) as our final classification loss:

$$\mathcal{L}_{\text{cls}} = \text{FL}(y, \hat{y}; \alpha_{\text{cls}}, \gamma_{\text{cls}}), \tag{15}$$

where $\alpha_{\text{cls}}, \gamma_{\text{cls}}$ are task-specific hyperparameters for this loss. For balanced datasets, we use the standard CE loss instead.

Optimization Objective Training is carried out in two stages to enhance optimization stability. The first stage optimizes the contrastive loss $\mathcal{L}_C$ to establish a reliable cross-modal alignment, which serves as initialization for the main multi-task learning. The second stage optimizes a unified objective $\mathcal{L}_{S2}$ that combines classification and auxiliary losses:

$$\mathcal{L}_{S2} = \lambda_1 \mathcal{L}_A + \lambda_2 \mathcal{L}_J + \mathcal{L}_{\text{cls}}, \tag{16}$$

where λ_1 and λ_2 are hyperparameter weights that balance the influence of the auxiliary tasks.

4 Experiments

We evaluate our method on three benchmark datasets to demonstrate its effectiveness across different platforms and languages: the Chinese Weibo dataset (Jin et al. 2017) (7,853 samples, 53.6% fake), the English Twitter dataset (Boididou et al. 2018) (14,195 samples, 54.6% fake), and the highly imbalanced English GossipCop dataset (Shu et al. 2020) (12,331 samples, 22.0% fake). We partition all datasets into training, validation, and test sets using a 6:2:2 stratified split.

4.1 Comparison Methods

We have included several state-of-the-art methods, including DeepSeek-R1 (Guo et al. 2025), Qwen2-VL (Wang et al. 2024b), RoBERTa (Liu et al. 2019), ARG (Hu et al. 2024), FSRU (Lao et al. 2024), and CSFND (Peng et al. 2024). For the ARG model (Hu et al. 2024), which employs GPT-3-generated rationales to assist a lightweight RoBERTa classifier, we re-implemented the method using Qwen2-VL to generate rationales for fairness. This is necessary because the original ARG results rely on a rationale dataset created with a different LLM and preprocessing pipeline, which is not directly applicable to our setting. As a result, the reproduced performance may differ from that reported in the original paper. A detailed description of each baseline is available in the supplementary material.

4.2 Implementation Details

All experiments are conducted on a single NVIDIA Tesla A800 GPU. We use Qwen2-VL-72B (Wang et al. 2024b) as the MLLM for rationale and caption generation, with inference accelerated by the vLLM framework (Kwon et al. 2023). A few-shot Chain-of-Thought (CoT) prompting strategy is employed, where each query includes four example samples.

We use two separate RoBERTa-based encoders with independent parameters: one dedicated to encoding the original news content (i.e., the news text and image caption), and the other to encoding all LLM-generated rationales. Training follows a two-stage strategy: in each stage, only the final layer of each encoder is fine-tuned, while all other layers remain frozen. The initial learning rate is set to 2e-5 for Weibo and 5e-5 for both Twitter and GossipCop.

Further implementation details, including model sizes, rationale lengths, and loss weights, are provided in the supplementary material.

4.3 Performance Comparison

Table 1 reports the performance of our model against state-of-the-art baselines.

On the relatively balanced Weibo and Twitter datasets, MMRGV achieves strong results, with a 0.9662 Macro-F1 on Weibo and 0.9972 on Twitter. While the Twitter dataset appears saturated—reflected in near-perfect scores by top models such as CSFND (0.9979)—our model still remains competitive. Notably, MMRGV achieves a fake news F1 score of 0.9681 on Weibo, highlighting its strength in Chinese-language scenarios.

On the highly imbalanced GossipCop dataset, most models struggle to detect fake news due to its low prevalence (22%). For example, Qwen2-VL and CSFND only achieve fake-news F1 scores of 0.2230 and 0.4585, respectively. In contrast, MMRGV achieves a much higher F1 score of 0.6886, demonstrating strong robustness. This gain is attributed to our multi-view aggregation and cross-verification mechanisms, which enhance the model's ability to identify minority-class samples. We further validate this in the ablation study.

Dataset	Method	Macro-F1	Acc	Fake News			Real News		
				Precision	Recall	F1	Precision	Recall	F1
Weibo	Qwen2-VL (Wang et al. 2024b)	0.8709	0.8710	0.9199	0.8319	0.8731	0.8250	0.9163	0.8683
	DeepSeek-R1 (Guo et al. 2025)	0.8329	0.8382	0.7924	0.9461	0.8625	0.9196	0.7133	0.8035
	RoBERTa (Liu et al. 2019)	0.9177	0.9179	0.9370	0.9068	0.9215	0.8975	0.9305	0.9137
	FSRU (Lao et al. 2024)	0.9315	0.9317	0.9515	0.9197	0.9353	0.9107	0.9456	0.9278
	CSFND (Peng et al. 2024)	0.8974	0.8982	0.8913	0.9211	0.9060	0.9065	0.8719	0.8888
	ARG (Hu et al. 2024)	0.8703	0.8704	0.9462	0.7987	0.8662	0.8100	0.9497	0.8743
	MMRGV (ours)	**0.9662**	**0.9663**	**0.9769**	**0.9594**	**0.9681**	**0.9546**	**0.9741**	**0.9644**
Twitter	Qwen2-VL (Wang et al. 2024b)	0.7363	0.7399	0.9272	0.5691	0.7053	0.6453	0.9461	0.7673
	DeepSeek-R1 (Guo et al. 2025)	0.6477	0.6480	0.7281	0.5687	0.6386	0.5882	0.7436	0.6568
	RoBERTa (Liu et al. 2019)	0.8842	0.8896	0.8673	0.9555	0.9093	0.9289	0.7992	0.8593
	FSRU (Lao et al. 2024)	0.9871	0.9874	0.9898	0.9885	0.9891	0.9842	0.9860	0.9851
	CSFND (Peng et al. 2024)	**0.9979**	**0.9979**	0.9974	**0.9987**	**0.9980**	**0.9984**	0.9969	**0.9976**
	ARG (Hu et al. 2024)	0.9077	0.9085	0.9185	0.9138	0.9162	0.8964	0.9020	0.8993
	MMRGV (ours)	0.9972	0.9972	**0.9987**	0.9961	0.9973	0.9953	**0.9984**	0.9968
GossipCop	Qwen2-VL (Wang et al. 2024b)	0.5548	0.8018	0.8163	0.1292	0.2230	0.8013	**0.9918**	0.8864
	DeepSeek-R1 (Guo et al. 2025)	0.7109	0.8303	0.6842	0.4260	0.5250	0.8535	0.9445	0.8966
	RoBERTa (Liu et al. 2019)	0.7920	0.8658	0.7335	0.6133	0.6679	0.8957	0.9371	0.9158
	FSRU (Lao et al. 2024)	0.7639	0.8579	0.7625	0.5177	0.6167	0.8752	0.9539	0.9129
	CSFND (Peng et al. 2024)	0.6808	0.8354	**0.8309**	0.3168	0.4585	0.8358	0.9818	0.9030
	ARG (Hu et al. 2024)	0.7957	0.8753	0.8115	0.5678	0.6681	0.8870	0.9626	0.9233
	MMRGV (ours)	**0.8060**	**0.8772**	0.7791	**0.6169**	**0.6886**	**0.8979**	0.9506	**0.9235**

Table 1: Comparison of MMRGV with the baselines. The highest and second-highest values of each metric are highlighted in bold and underlined, respectively.

	Weibo				Twitter				GossipCop			
	Mac-F1	Acc	$F1_{Real}$	$F1_{Fake}$	Mac-F1	Acc	$F1_{Real}$	$F1_{Fake}$	Mac-F1	Acc	$F1_{Real}$	$F1_{Fake}$
w/o TD	0.9527	0.9529	0.9500	0.9555	0.9968	0.9968	0.9965	0.9971	0.7932	0.8666	0.9164	0.6700
w/o ITC	0.9560	0.9561	0.9536	0.9583	0.9943	0.9944	0.9938	0.9948	0.8034	0.8703	0.9181	0.6887
w/o ID	0.9578	0.9580	0.9554	0.9603	0.9897	0.9898	0.9887	0.9907	**0.8064**	0.8731	0.9201	**0.6928**
Full model	**0.9662**	**0.9663**	**0.9644**	**0.9681**	**0.9972**	**0.9972**	**0.9968**	**0.9973**	0.8060	**0.8772**	**0.9235**	0.6886

Table 2: The ablation results of different rationale perspectives removed from the full model.

4.4 Ablation Study

Ablation study for rationale perspectives Table 2 presents the results of removing each rationale component (TD, ITC, and ID) from our model. On the Weibo and Twitter datasets, removing any module leads to noticeable performance degradation across all metrics. Among them, removing the TD module (w/o TD) results in the largest drop—reducing Macro-F1 from 0.9662 to 0.9527 on Weibo, and from 0.9972 to 0.9968 on Twitter—underscoring the pivotal role of text-based rationales in these linguistically rich platforms. In contrast, removing the ID module (w/o ID) yields the smallest performance drop, suggesting that image-based rationales have a more limited effect in relatively text-centric datasets.

For the highly imbalanced GossipCop dataset, the impact of rationale removal is more nuanced. Excluding the TD module still leads to the most severe decline (Macro-F1 drops from 0.8060 to 0.7932; fake-news F1 from 0.6886 to 0.6700), reaffirming the centrality of textual reasoning in identifying subtle misinformation patterns. Interestingly, removing the ID module slightly improves performance (Macro-F1 rises to 0.8064), implying that visual cues may introduce noise or modality misalignment in low-resource settings. Meanwhile, excluding the ITC module results in negligible performance variations, reflecting its auxiliary role in cross-modal verification.

Overall, these results highlight the complementary nature of all three rationale perspectives, with text-based reasoning playing a dominant role. Their integration proves essential, especially for handling class imbalance and enhancing model robustness.

Ablation study for model components Our component ablation study, with results in Table 3, reveals each module's distinct contribution across the different datasets. On the Weibo dataset, the rationale content gate fusion (RCGF) module is most critical; its removal causes the largest drop in

	Weibo				Twitter				GossipCop			
	Mac-F1	Acc	$F1_{Real}$	$F1_{Fake}$	Mac-F1	Acc	$F1_{Real}$	$F1_{Fake}$	Mac-F1	Acc	$F1_{Real}$	$F1_{Fake}$
w/o RA	0.9566	0.9567	0.9545	0.9587	**0.9993**	**0.9993**	**0.9992**	**0.9994**	0.7913	0.8683	0.9180	0.6646
w/o EF	0.9566	0.9567	0.9545	0.9587	0.9957	0.9958	0.9953	0.9961	**0.8089**	**0.8788**	**0.9245**	**0.6933**
w/o RCGF	0.9483	0.9484	0.9459	0.9508	0.9936	0.9937	0.9930	0.9942	0.7891	0.8626	0.9136	0.6647
w/o MA	0.9623	0.9624	0.9596	0.9649	0.9943	0.9944	0.9938	0.9948	0.7751	0.8593	0.9127	0.6374
Full model	**0.9662**	**0.9663**	**0.9644**	**0.9681**	0.9972	0.9972	0.9968	0.9973	0.8060	0.8772	0.9235	0.6886

Table 3: Component ablation results. We ablate the contributions of: Rationale Alignment (RA, trained with $\mathcal{L}_C$), the Enhanced Feature module (EF, trained with $\mathcal{L}_J$), Rationale Content Gate Fusion (RCGF, trained with $\mathcal{L}_A$), and the Multi-View Aggregation (MA) module.

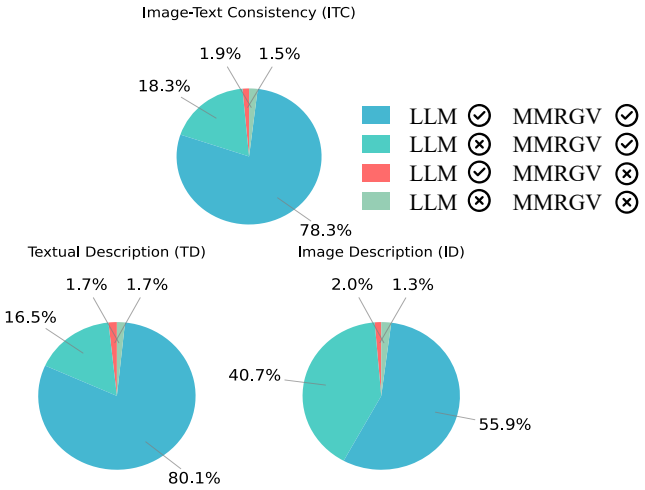

Figure 3: Comparison of raw MLLM judgments versus MMRGV's final predictions for each perspective on the Weibo test set, categorized into four outcomes (MLLM correct/MMRGV wrong, MLLM wrong/MMRGV right, both correct, and both wrong).

Macro-F1 (from 0.9662 to 0.9483), while the identical impact of removing the rationale alignment (RA) and enhanced feature (EF) modules suggests some functional overlap. A different pattern emerges on Twitter, where removing the RA module surprisingly improves the Macro-F1 to 0.9993, indicating it may introduce noise or overfit on this dataset's concise textual environment. Lastly, on the highly imbalanced GossipCop dataset, the multi-view aggregation (MA) is most vital, with its removal causing the sharpest performance decline (from 0.8060 to 0.7751). Conversely, removing the EF module slightly improves performance here, suggesting that knowledge distillation can be counterproductive on such noisy, long-tail data. Across all datasets, the RCGF module consistently proves essential, highlighting the core importance of its gating mechanism.

4.5 Robustness to Erroneous MLLM Judgments

To assess MMRGV's robustness to hallucinations and incorrect initial judgments from the MLLM, we conducted an evaluation on the Weibo test set. For each perspective, predictions were categorized into four groups based on the agreement between the MLLM's raw judgment ($y_J^{(v)}$) and MMRGV's final output ($\hat{y}$) with respect to the ground truth. The distribution is visualized in Figure 3.

The results confirm MMRGV's strong corrective capability. Across all perspectives, the model corrects more than 90% of the MLLM's erroneous predictions. Notably, for the most error-prone Image Description (ID) perspective, MMRGV achieves its highest correction rate of 95.3%, demonstrating the robustness of the proposed cross-verification and gating mechanisms **against** the impact of **hallucinated or erroneous** visual rationales.

Moreover, MMRGV rarely introduces errors when the MLLM's initial prediction is correct, with an error introduction rate of approximately 2%. These findings highlight the framework's reliability under imperfect LLM conditions. Additional analyses and case studies are provided in the supplementary material.

5 Conclusion and Limitation

We propose **MMRGV**, a multimodal fake news detection framework that integrates multi-perspective rationale generation and cross-verification to enhance reliability. By jointly analyzing textual content, image-text consistency, and image forensics, MMRGV produces diverse rationales and verifies them to reduce prediction errors arising from hallucinated reasoning. These verified signals are then fused using an adaptive weighting strategy. Experiments on three benchmark datasets demonstrate that MMRGV consistently outperforms strong baselines in both accuracy and robustness. Further analysis confirms the importance of rationale diversity and cross-verification in driving these improvements.

Despite its effectiveness, MMRGV has limitations. Its performance relies on the quality of LLM-generated rationales, which depends on manually crafted prompts that might overlook subtle misinformation cues. The multi-rationale architecture increases inference costs, limiting real-time deployment. Furthermore, its fixed, manually defined perspectives restrict adaptability across different domains. Future directions include automated prompt optimization to improve rationale quality, enhancing inference speed, and exploring the automatic selection of reasoning perspectives.

Acknowledgments

This work was supported by Stable Support Project of Shenzhen (20231120161634002), Shenzhen Science and Technology Program (JCYJ20240813141417023), Natural Science Foundation of Guangdong Province of China (2025A1515010233), Guangdong Provincial Department of Education (2024KTSCX060), Tencent "Rhinoceros Birds" - Scientific Research Foundation for Young Teachers of Shenzhen University, Open Project of State Key Lab. for Novel Software Technology of Nanjing University (KFKT2025B22).

References

Alam, F.; Cresci, S.; Chakraborty, T.; Silvestri, F.; Dimitrov, D.; Martino, G. D. S.; Shaar, S.; Firooz, H.; and Nakov, P. 2022. A Survey on Multimodal Disinformation Detection. In Calzolari, N.; Huang, C.-R.; Kim, H.; Pustejovsky, J.; Wanner, L.; Choi, K.-S.; Ryu, P.-M.; Chen, H.-H.; Donatelli, L.; Ji, H.; Kurohashi, S.; Paggio, P.; Xue, N.; Kim, S.; Hahm, Y.; He, Z.; Lee, T. K.; Santus, E.; Bond, F.; and Na, S.-H., eds., *Proceedings of the 29th International Conference on Computational Linguistics*, 6625–6643. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Boididou, C.; Papadopoulos, S.; Zampoglou, M.; Apostolidis, L.; Papadopoulou, O.; and Kompatsiaris, Y. 2018. Detection and visualization of misleading content on Twitter. *International Journal of Multimedia Information Retrieval*, 7(1): 71–86.

Cao, J.; Wu, J.; Shang, W.; Wang, C.; Song, K.; Yi, T.; Cai, J.; and Zhu, H. 2025. Fake News Detection Based on Cross-Modal Ambiguity Computation and Multi-Scale Feature Fusion. *Computers, Materials & Continua*, 83(2).

Chen, J.; Wu, Z.; Yang, Z.; Xie, H.; Wang, F. L.; and Liu, W. 2021. Multimodal Fusion Network with Latent Topic Memory for Rumor Detection. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6.

Chen, Y.; Li, D.; Zhang, P.; Sui, J.; Lv, Q.; Tun, L.; and Shang, L. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM web conference 2022*, 2897–2905.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints*, arXiv:1810.04805.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv e-prints*, arXiv:2010.11929.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hao, X.; Xu, W.; Huang, X.; Sheng, Z.; and Yan, H. 2025. MFUIE: A Fake News Detection Model Based on Multimodal Features and User Information Enhancement. *EAI Endorsed Transactions on Scalable Information Systems*, 12(1).

Hu, B.; Sheng, Q.; Cao, J.; Shi, Y.; Li, Y.; Wang, D.; and Qi, P. 2024. Bad Actor, Good Advisor: Exploring the Role of Large Language Models in Fake News Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20): 22105–22113.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12): 1–38.

Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; and Luo, J. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, 795–816.

Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Khattar, D.; Goud, J. S.; Gupta, M.; and Varma, V. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, 2915–2921.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J.; Zhang, H.; and Stoica, I. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, 611–626.

Lao, A.; Zhang, Q.; Shi, C.; Cao, L.; Yi, K.; Hu, L.; and Miao, D. 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, 18426–18434.

Li, C.; Wang, S.; Yang, D.; Li, Z.; Yang, Y.; Zhang, X.; and Zhou, J. 2017. PPNE: property preserving network embedding. In *Database Systems for Advanced Applications: 22nd International Conference, DASFAA 2017, Suzhou, China, March 27-30, 2017, Proceedings, Part I 22*, 163–179. Springer.

Lin, S.; Hilton, J.; and Evans, O. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2999–3007.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints*, arXiv:1907.11692.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Lyu, Y.; Li, Z.; Niu, S.; Xiong, F.; Tang, B.; Wang, W.; Wu, H.; Liu, H.; Xu, T.; and Chen, E. 2025. Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models. *ACM Transactions on Information Systems*, 43(2): 1–32.

Nan, Q.; Sheng, Q.; Cao, J.; Hu, B.; Wang, D.; and Li, J. 2024. Let Silence Speak: Enhancing Fake News Detection with Generated Comments from Large Language Models. *arXiv e-prints*, arXiv:2405.16631.

Peng, L.; Jian, S.; Kan, Z.; Qiao, L.; and Li, D. 2024. Not all fake news is semantically similar: Contextual semantic representation learning for multimodal fake news detection. *Information Processing & Management*, 61(1): 103564.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

Sharma, D. K.; Singh, B.; Agarwal, S.; Garg, L.; Kim, C.; and Jung, K.-H. 2023. A survey of detection and mitigation for fake images on social media platforms. *Applied Sciences*, 13(19): 10980.

Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3): 171–188.

Singhal, S.; Shah, R. R.; Chakraborty, T.; Kumaraguru, P.; and Satoh, S. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, 39–47. IEEE.

Wang, J.; Zhu, Z.; Liu, C.; Li, R.; and Wu, X. 2024a. LLM-Enhanced multimodal detection of fake news. *PloS one*, 19(10): e0312240.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wu, Y.; Zhan, P.; Zhang, Y.; Wang, L.; and Xu, Z. 2021. Multimodal Fusion with Co-Attention Networks for Fake News Detection. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2560–2569. Online: Association for Computational Linguistics.

Xiao, L.; Zhang, Q.; Shi, C.; Wang, S.; Naseem, U.; and Hu, L. 2024. Msynfd: Multi-hop syntax aware fake news detection. In *Proceedings of the ACM web conference 2024*, 4128–4137.

Zhang, X.; and Gao, W. 2023. Towards LLM-based Fact Verification on News Claims with a Hierarchical Step-by-Step Prompting Method. In Park, J. C.; Arase, Y.; Hu, B.; Lu, W.; Wijaya, D.; Purwarianti, A.; and Krisnadhi, A. A., eds., *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 996–1011. Nusa Dua, Bali: Association for Computational Linguistics.

Zheng, X.; Luo, M.; and Wang, X. 2025. Unveiling Fake News with Adversarial Arguments Generated by Multimodal Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics*, 7862–7869.

Zhou, Y.; Yang, Y.; Ying, Q.; Qian, Z.; and Zhang, X. 2023. Multi-modal fake news detection on social media via multi-grained information fusion. In *Proceedings of the 2023 ACM international conference on multimedia retrieval*, 343–352.