CURE: Context-aware Uncertainty-adaptive Rebalancing Experts for Multimodal Fake News Detection

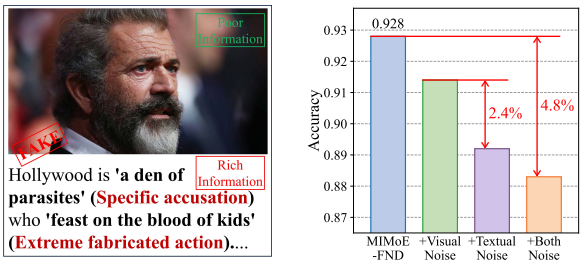
Anonymous ACL submission

Abstract

The proliferation of fake news on social media motivates the development of automatic Multimodal Fake News Detection. While most existing methods focus on various fusion strategies, they largely overlook the heterogeneous modality imbalance issue (inter-modal information disparities and noise interference) in real-world scenarios, which hinders effective fusion through biased feature representations. To address this, we propose a novel **C**ontext-aware **U**ncertainty-adaptive **R**ebalancing **E**xperts (CURE) framework for multimodal fake news detection. First, to bridge inter-modal information disparities, a Retrieval-Augmented Context Prompter (RACP) module retrieves similar instances and distills them into dynamic prompts, enhancing the features of each modality by providing supplementary context. Second, to mitigate noise interference, an Uncertainty-Adaptive Rebalancing Experts (UARE) module first quantifies feature-level noise via uncertainty modeling. An uncertainty-adaptive routing mechanism then achieves robust modality rebalancing by adaptively down-weighting features with high uncertainty. Extensive experiments are conducted on three real-world datasets spanning two languages, demonstrating significant performance improvements of our method. The code is available at https://anonymous.4open.science/r/CURE1.

1 Introduction

The proliferation of social media has fueled the spread of multimodal fake news, which combines text and images to be more credible and deceptive. Such disinformation poses significant threats to societal trust and stability (Vosoughi et al., 2018; Lazer et al., 2018). As manually fact-checking this massive volume of content is infeasible, developing automated Multimodal Fake News Detection (MFND) techniques has become an urgent research priority.



(a) An example of inter-modal information disparities: an image with poor information vs. a text with rich information.

(b) A pilot study of noise interference on the MIMoE-FND (Liu et al., 2025) baseline with injected Gaussian noise on the Weibo dataset.

Figure 1: Illustration of the heterogeneous modality imbalance issue, which manifests as (a) inter-modal information disparities and (b) noise interference.

Existing MFND methods primarily focus on various fusion strategies, which can be broadly divided into three categories: (1) Representation-centric Early Fusion methods (Wang et al., 2018; Singhal et al., 2019), (2) Cross-Modal Consistency and Alignment methods (Chen et al., 2022; Zhou et al., 2023; Wang et al., 2024b), and (3) Adaptive Strategy-based Fusion methods (Ying et al., 2023; Yu et al., 2025; Liu et al., 2025). Although these methods have achieved notable performance, they largely overlook the heterogeneous modality imbalance issue in real-world scenarios that ultimately hinders effective multimodal fusion (Zhang et al., 2024b). This issue manifests primarily through two critical challenges:

(1) Inter-modal information disparities. Raw multimodal inputs in MFND often exhibit a significant information disparity across modalities. For instance, as shown in Figure 1(a), the text may deliver an information-rich, extreme accusation that serves as primary evidence for a fake verdict. In contrast, the accompanying image is merely a portrait of the accuser, offering little contextual information. This imbalance naturally causes the model to over-rely

on the information-dense text modality while neglecting potentially complementary signals from the image, which results in a strong modality bias. Therefore, bridging this information gap is crucial for alleviating the overall modality imbalance.

(2) Noise Interference. In real-world scenarios, both visual and textual data are inevitably corrupted by noise, such as blurry images or textual typos (Luvembe et al., 2024). To investigate this, we conducted a pilot study by injecting Gaussian noise into the strong MIMoE-FND baseline (Liu et al., 2025) on the Weibo dataset, testing four settings: (1) a no-noise baseline, (2) visual noise only, (3) textual noise only, and (4) noise in both modalities. As shown in Figure 1(b), the results reveal two critical phenomena. First, introducing any noise causes a significant performance drop. Second, the model proves more sensitive to textual noise than visual noise, revealing a modality-dependent impact. Therefore, developing a mechanism to dynamically identify and down-weight high-noise features is critical for achieving a robust fusion and mitigating this aspect of modality imbalance.

To address the challenges above, we propose a novel **C**ontext-aware **U**ncertainty-adaptive **R**ebalancing **E**xperts (CURE) framework. Specifically, to tackle inter-modal information disparities, we first design a Retrieval-Augmented Context Prompter (RACP) module. Drawing inspiration from the prompt learning paradigm and its recent extension from natural language processing to multimodal learning (Zhou et al., 2022), the module retrieves semantically similar instances and distills them into dynamic prompts. As shown in Figure 2, these prompts then enhance the features of each modality by providing supplementary context, effectively bridging the information gap between them.

Second, to mitigate noise interference, we introduce the Uncertainty-Adaptive Rebalancing Experts (UARE) module. This module employs a Mixture-of-Experts architecture with both modality-specific and shared experts to learn intra-modal and cross-modal feature interactions, respectively. Inspired by Chang et al. (2020), the module quantifies feature-level noise via uncertainty modeling. Specifically, each expert maps its features onto a Gaussian distribution, where the resulting variance serves as a direct measure of uncertainty. An uncertainty-adaptive routing mechanism then achieves robust modality rebalancing by adaptively down-weighting features with high uncertainty.

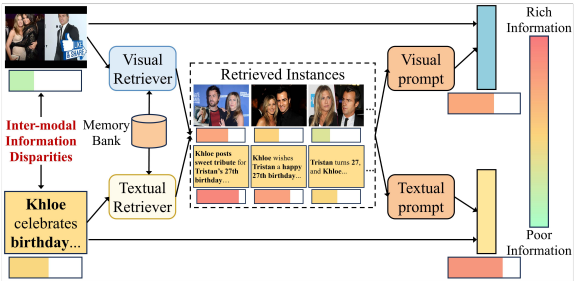


Figure 2: The proposed RACP module retrieves similar instances and distills them into dynamic prompts, bridging inter-modal information disparities by enhancing each modality with supplementary context.

In summary, our main contributions are:

- We propose CURE, a novel framework for multimodal fake news detection that aims to alleviate heterogeneous modality imbalance by tackling its two primary manifestations: inter-modal information disparities and noise interference.
- We design a Retrieval-Augmented Context Prompter module that leverages context-aware prompts to effectively bridge inter-modal information disparities.
- We develop an Uncertainty-Adaptive Rebalancing Experts module that implements an uncertainty-adaptive routing mechanism to robustly mitigate noise interference.
- We conduct extensive experiments on three real-world datasets spanning two languages, demonstrating that our method significantly outperforms state-of-the-art baselines.

2 Related Work

2.1 Multimodal Fake News Detection

Researchers address the MFND task using various fusion methods that can be broadly divided into three categories: **(1) Representation-centric Early Fusion methods** (Wang et al., 2018; Singhal et al., 2019) focus on learning strong unimodal representations and fusing them with simple aggregation, but often overlook complex inter-modal correlations. **(2) Cross-Modal Consistency and Alignment methods** (Chen et al., 2022; Zhou et al., 2023; Wang et al., 2024b; Zhu et al., 2025) explicitly model the semantic relationship between modalities, such as similarity or ambiguity, as a key detection signal. **(3) Adaptive Strategy-based Fusion methods** (Ying et al., 2023; Yu et al., 2025; Liu et al., 2025; Su et al., 2025) employ dynamic architectures to learn sample-specific fusion strate-

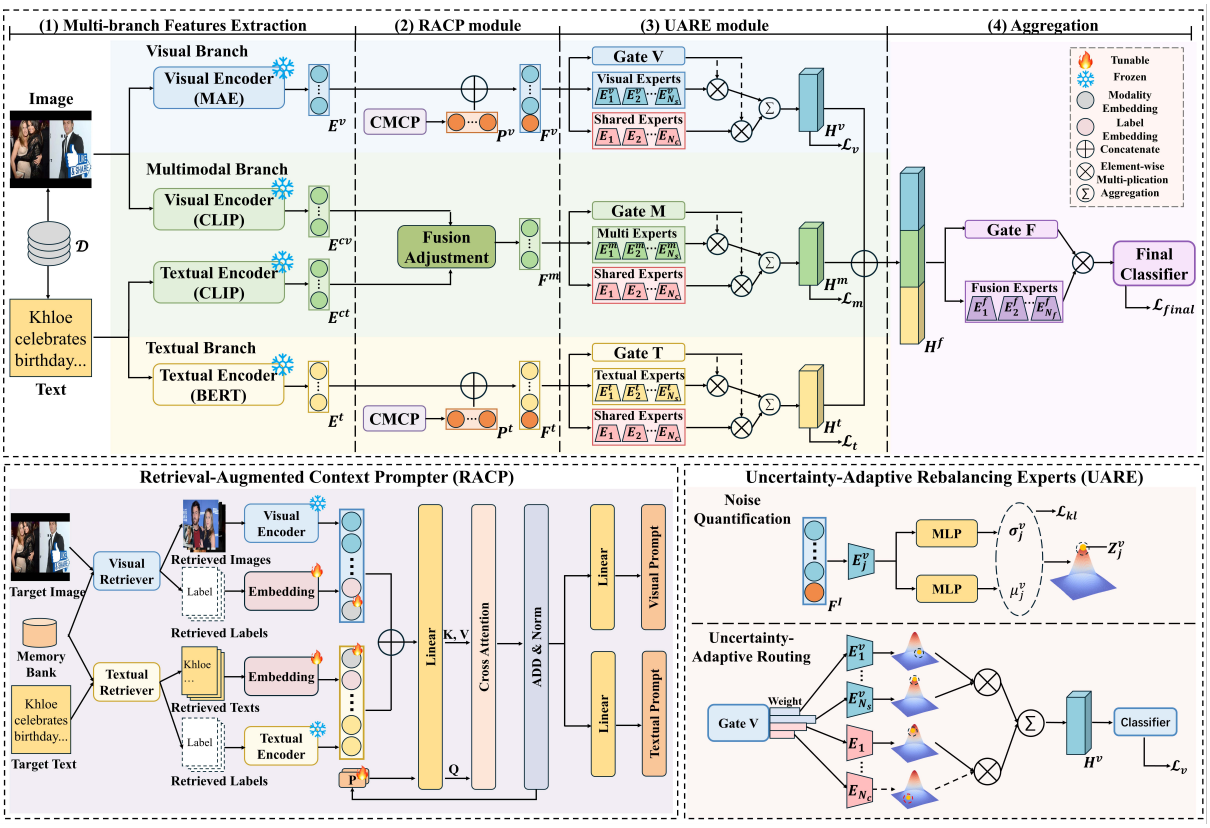


Figure 3: Overall framework of our proposed CURE.

gies. However, these methods largely overlook the heterogeneous modality imbalance issue. Recently, some work theoretically confirmed that this issue can significantly hinder effective multimodal fusion (Wang et al., 2020; Huang et al., 2022). To address this, our proposed CURE framework aims to alleviate this modality imbalance by tackling its two primary manifestations: bridging inter-modal information disparities and mitigating noise interference.

2.2 Imbalanced Multimodal Learning

Multimodal learning often fails to achieve its expected performance gains in real-world scenarios due to the presence of low-quality data. This issue primarily manifests as two critical challenges: imbalanced modality quality and imbalanced noisy modalities (Zhang et al., 2024b).

Learning with Imbalanced Modality Quality. Imbalanced modality quality often causes over-reliance on the information-dense modality. To address this, some efforts suppress the dominant modality via on-the-fly gradient modulation (Peng et al., 2022) or adaptively mask its network components (Yang et al., 2025). Recently, some methods in the MFND task attempt to equalize modal contributions by dynamically re-weighting branch losses (Wu et al., 2025b) or regulating modal contribution scores (Wu et al., 2025c). However, they pay less attention to the intrinsic inter-modal information disparities. In contrast, we design the RACP module that leverages context-aware prompts to effectively bridge this disparity.

Learning with Noisy Modalities. Noise interference in various modalities, such as vision and text, is a common issue that degrades model robustness. In recent years, some efforts have been made to address this in fields like multimedia recommendation (Xv et al., 2024; He et al., 2025) and multimodal sentiment analysis (Zhang et al., 2024a; Li and Li, 2025). It is worth noting that noise is also prevalent in multimodal fake news detection, imposing a negative impact on final decisions. However, dedicated research in this area remains limited. To fill this gap, we develop the UARE module, which employs a robust uncertainty-adaptive routing mechanism to mitigate this noise interference.

3 Methodology

In this section, we introduce the proposed CURE, designed to alleviate the heterogeneous modality imbalance issue. As illustrated in Figure 3, it con-

sists of four components: (1) a multi-branch feature extraction (§3.1), (2) the Retrieval-Augmented Context Prompter (RACP) module (§3.2), (3) the Uncertainty-Adaptive Rebalancing Experts (UARE) module (§3.3), and (4) the final aggregation and loss function (§3.4).

3.1 Multi-branch Features Extraction

We employ a multi-branch architecture to extract features from both unimodal and cross-modal perspectives. For the textual branch, we use a pre-trained BERT (Devlin et al., 2019) to obtain token-level representations $E^t \in \mathbb{R}^{n \times d_e}$, where n denotes the text sequence length and d_e is the feature dimension. For the visual branch, we use a pre-trained MAE (He et al., 2022) to obtain patch-level representations $E^v \in \mathbb{R}^{m \times d_e}$, where m denotes the number of image patches. Additionally, we use CLIP's encoders (Radford et al., 2021) to acquire aligned embeddings E^{ct} and E^{cv}. For the multimodal branch, we adopt the strategy from Tong et al. (2024), generating a fused feature $F^m \in \mathbb{R}^{d_m}$ by weighting the concatenated CLIP embeddings with their cosine similarity:

$$\text{sim} = \frac{E^{ct} \cdot (E^{cv})^T}{\|E^{ct}\| \|E^{cv}\|} \tag{1}$$

$$F^m = \text{sim} \cdot \text{MLP}(E^{ct} \oplus E^{cv}) \tag{2}$$

3.2 Retrieval-Augmented Context Prompter

To tackle inter-modal information disparities, we propose the RACP module. Inspired by recent advancements in injecting prompts directly into hidden feature representations (Zhou et al., 2022), it first retrieves semantically similar instances and distills them into dynamic prompts. These prompts then enhance the features of each modality by providing supplementary context, effectively bridging the information gap between them.

Multi-branch Retriever To retrieve relevant instances, we first construct a memory bank $\mathcal{M} = \{(t_m, v_m, y_m)\}_{m=1}^{|\mathcal{M}|}$, where $|\mathcal{M}|$ represents the total number of instances. To prevent a high-quality query in one modality from being contaminated by a low-quality one in another, we design a multi-branch retriever. It leverages the pre-aligned embeddings E^{ct} and E^{cv} as independent query vectors to retrieve the Top-K relevant instances for each modality based on cosine similarity. The retrieval for the textual branch is defined as:

$$\mathcal{R}^t = \underset{m \in \mathcal{M}}{\text{Top-K}} \left(\frac{E^{ct} \cdot (E^{ct}_m)^T}{\|E^{ct}\| \|E^{ct}_m\|} \right) \tag{3}$$

The visual context set $\mathcal{R}^v$ is obtained analogously. This multi-branch retrieval yields two modality-specific sets of (text, label) and (image, label) pairs, preventing cross-modal contamination.

Dynamic Prompt Generation and Enhancement To best leverage the retrieved context, we design a dynamic prompt generation mechanism. Our key insight is to enrich this context with explicit signals (i.e., ground-truth labels and modality types), empowering a cross-attention mechanism to selectively amplify salient information.

Specifically, we first construct the context for each modality. For the retrieved textual set $\mathcal{R}^t$, we form a rich context C^t by concatenating its token-level features $E^{r,t} \in \mathbb{R}^{K \times n \times d_e}$ with a learnable label embedding $E^{l,t} \in \mathbb{R}^{2 \times d_e}$ and a modality type embedding $E^{type} \in \mathbb{R}^{2 \times d_e}$. These embeddings provide crucial label semantics and source-distinguishing information, respectively. After generating the visual context C^v analogously, both are concatenated into a unified context representation C with potential synergistic signals across both modalities.

To adaptively learn the most effective information from this unified context, we introduce a set of learnable query tokens $P_q \in \mathbb{R}^{L_p \times d_e}$, where L_p is the prompt length. These tokens attend to the context C, which serves as the key and value, via a cross-attention block to generate a shared prompt P. This prompt is subsequently specialized into modality-specific prompts, P^t and P^v, through two independent Feed-Forward Networks. The process is defined as:

$$P = \text{Attn}(f^Q(P_q), f^K(C), f^V(C)) \tag{4}$$

$$\text{Attn}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V \tag{5}$$

$$P^t = \text{FFN}_t(P), \quad P^v = \text{FFN}_v(P) \tag{6}$$

where $f^Q(\cdot), f^K(\cdot)$, and $f^V(\cdot)$ are linear projections, and d is the dimension of the key vectors.

Finally, the generated dynamic prompts are concatenated to the initial features to yield the enhanced features F^t and F^v:

$$F^t = P^t \oplus E^t, \quad F^v = P^v \oplus E^v \tag{7}$$

This process enriches the original features with distilled supplementary context, thus effectively bridging the inter-modal information gap.

3.3 Uncertainty-Adaptive Rebalancing Experts

To mitigate noise interference in the enhanced features, which can hinder effective fusion, we propose the UARE module. It first quantifies feature-level noise via uncertainty modeling. An uncertainty-adaptive routing then achieves robust modality rebalancing by adaptively down-weighting features with high uncertainty.

Noise Quantification Motivated by our pilot study revealing the modality-dependent impact of noise, we employ a Mixture-of-Experts architecture to handle both modality-specific and cross-modal noise patterns. For each branch $k \in \{t, v, m\}$, we define an expert ensemble E^k with a total of N_e experts, composed of N_s modality-specific experts and N_c shared experts:

$$E^k = [E^k_{s,1}, \ldots, E^k_{s,N_s}, E_{c,1}, \ldots, E_{c,N_c}] \quad (8)$$

where $E^k_{s,j}$ and $E_{c,j}$ denote the j-th modality-specific and modality-shared expert, respectively.

With the expert architecture established, we quantify the feature-level noise by modeling the uncertainty, inspired by Chang et al. (2020). Specifically, each expert E^k_j maps its output to a diagonal multivariate Gaussian distribution:

$$p(Z^k_j | E^k_j(F^k)) \sim \mathcal{N}(\mu^k_j, (\sigma^k_j)^2 I) \quad (9)$$

$$\mu^k_j = f_\mu(E^k_j(F^k)), \sigma^k_j = f_\sigma(E^k_j(F^k)) \quad (10)$$

where the mean μ^k_j represents the refined feature and the variance $(\sigma^k_j)^2$ serves as a direct quantification of its noise level. They are generated by two distinct fully-connected layers, f_μ and f_σ. To enable differentiability for back-propagation, we apply the re-parameterization trick:

$$Z^k_j = \mu^k_j + \epsilon \cdot \sigma^k_j, \epsilon \sim \mathcal{N}(0, I) \quad (11)$$

To ensure the variance term meaningfully captures noise, rather than collapsing to zero under the pressure of the main task loss (Chang et al., 2020), we introduce a KL divergence regularization term:

$$\mathcal{L}^k_{kl} = \frac{1}{N_e} \sum_{j=1}^{N_e} \mathrm{KL}[\mathcal{N}(\mu^k_j, (\sigma^k_j)^2 I) || \mathcal{N}(0, I)] \quad (12)$$

Uncertainty-Adaptive Routing Having quantified the noise level of each expert's output, we employ an uncertainty-adaptive routing mechanism to achieve robust modality rebalancing. A gating network G^k processes the input feature F^k to generate routing weights:

$$W^k = \mathrm{Softmax}(G^k(F^k)) \quad (13)$$

These weights then aggregate the refined features from all experts, yielding the final denoised feature H^k:

$$H^k = \sum_{j=1}^{N_e} W^k_j \cdot \mu^k_j \quad (14)$$

To explicitly guide the gating network to down-weight outputs with high uncertainty, we introduce an uncertainty-adaptive routing loss $\mathcal{L}^k_u$:

$$\mathcal{L}^k_u = \frac{1}{N_e} \sum_{j=1}^{N_e} W^k_j \cdot (\sigma^k_j)^2 \quad (15)$$

This loss penalizes assigning high weights to outputs with large variance, thereby prioritizing features with low noise for fusion. Consequently, this mechanism ensures that high-quality modalities contribute more to the final representation, effectively promoting robust modality rebalancing.

3.4 Aggregation and Loss Function

Finally, the denoised features from all branches are concatenated into a unified representation H^f. Without loss of generality, we employ a classic expert gating mechanism, consisting of N_f fusion experts and a gating network G^f, to produce the final representation H_{final}:

$$W^f = \mathrm{Softmax}(G^f(H^f)) \quad (16)$$

$$H_{\text{final}} = \sum_{i=1}^{N_f} W^f_i \cdot E^f_i(H^f) \quad (17)$$

The resulting H_{final} is then fed into a classifier to yield the prediction $\hat{y}_{\text{final}}$.

The main task loss is denoted as $\mathcal{L}_{\text{final}}$, and an auxiliary loss for each branch is denoted as $\mathcal{L}^k_b$. Both losses are formulated using the Binary Cross-Entropy (BCE). The overall training objective is to minimize a total loss function $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_{\text{final}} + \sum_k \mathcal{L}^k_b + \alpha \sum_k \mathcal{L}^k_{kl} + \beta \sum_k \mathcal{L}^k_u \quad (18)$$

where $k \in \{t, v, m\}$ indicates the modality branch. α and β serve as trade-off hyperparameters to balance the contribution of the KL divergence regularization and the uncertainty-adaptive constraint, respectively.

Datasets	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F1	Precision	Recall	F1
Weibo	EANN (Wang et al., 2018)	0.827	0.847	0.812	0.829	0.807	0.843	0.825
	SpotFake (Singhal et al., 2019)	0.892	0.902	0.964	0.932	0.847	0.656	0.739
	SAFE (Zhou et al., 2020)	0.762	0.831	0.724	0.774	0.695	0.811	0.748
	CAFE (Chen et al., 2022)	0.840	0.855	0.830	0.842	0.825	0.851	0.837
	FND-CLIP (Zhou et al., 2023)	0.907	0.914	0.901	0.908	0.914	0.901	0.907
	BMR (Ying et al., 2023)	0.918	0.882	**0.948**	0.914	**0.942**	0.879	0.904
	MSACA (Wang et al., 2024b)	0.903	0.935	0.873	0.903	0.872	0.935	0.902
	MINER-UVS (Wang et al., 2024a)	0.934	**0.948**	0.922	0.935	0.919	0.946	0.933
	MFUIE (Hao et al., 2024)	0.926	0.936	0.912	0.924	0.917	0.940	0.929
	RaCMC (Yu et al., 2025)	0.915	0.910	0.924	0.917	0.921	0.906	0.914
	MIMoE-FND (Liu et al., 2025)	0.928	0.942	0.913	0.928	0.913	0.942	0.927
	DAAD (Su et al., 2025)	0.932	0.942	0.915	0.928	0.922	**0.947**	0.934
	KEN (Zhu et al., 2025)	0.935	0.937	0.934	0.935	0.932	0.935	0.934
	CURE	**0.940**	0.943	0.941	**0.942**	0.937	0.939	**0.938**
Weibo-21	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
	SAFE (Zhou et al., 2020)	0.905	0.893	0.908	0.901	0.916	0.901	0.890
	CAFE (Chen et al., 2022)	0.882	0.857	0.915	0.885	0.907	0.844	0.876
	FND-CLIP (Zhou et al., 2023)	0.943	0.935	0.945	0.940	0.950	0.942	0.946
	BMR (Ying et al., 2023)	0.929	0.908	0.947	0.927	0.946	0.906	0.925
	MFUIE (Hao et al., 2024)	0.935	0.942	0.926	0.934	0.944	0.929	0.936
	DAAD (Su et al., 2025)	0.942	0.951	0.925	0.938	0.931	0.955	0.943
	MIMoE-FND (Liu et al., 2025)	0.956	0.953	0.957	0.955	0.959	0.956	0.957
	CURE	**0.961**	**0.958**	**0.965**	**0.961**	**0.964**	**0.957**	**0.961**
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
	SAFE (Zhou et al., 2020)	0.838	0.758	0.558	0.643	0.857	0.937	0.895
	CAFE (Chen et al., 2022)	0.867	0.732	0.490	0.587	0.887	0.957	0.921
	FND-CLIP (Zhou et al., 2023)	0.880	0.761	0.549	0.638	0.899	0.959	0.928
	BMR (Ying et al., 2023)	0.895	0.752	0.639	0.691	0.920	0.965	0.936
	MSACA (Wang et al., 2024b)	0.887	0.816	0.538	0.648	0.897	0.971	0.933
	MINER-UVS (Wang et al., 2024a)	0.891	0.746	0.653	0.697	0.920	0.947	0.933
	RaCMC (Yu et al., 2025)	0.879	0.745	0.563	0.641	0.902	0.954	0.927
	MIMoE-FND (Liu et al., 2025)	0.895	0.762	**0.644**	0.698	**0.920**	0.953	0.936
	CURE	**0.902**	**0.830**	0.614	**0.706**	0.911	**0.973**	**0.941**

Table 1: Comparison between CURE and state-of-the-art multimodal fake news detection methods on Weibo, Weibo-21 and GossipCop.

Datasets	Method	Accuracy	F1 score	
			Fake	Real
Weibo	PFBL (Wu et al., 2025a)	0.929	0.935	0.930
	BMLHF (Wu et al., 2025b)	0.912	0.903	0.902
	BCMPL (Wu et al., 2025c)	0.912	0.922	0.924
	DCLR (Wang et al., 2026)	0.917	0.895	0.907
	CURE	**0.940**	**0.942**	**0.938**

Table 2: Comparison between CURE and latest multimodal imbalance methods on Weibo.

4 Experiments

4.1 Experimental Settings

To evaluate the effectiveness of CURE, we benchmark it on three widely adopted datasets in MFND: Weibo (Jin et al., 2017), Weibo-21 (Nan et al., 2021), and GossipCop (Shu et al., 2020). We compare our method against several competitive baselines, as well as recent approaches specifically addressing modality imbalance. Detailed descriptions of the datasets and baseline methods are provided in Appendix A.

Implementation Details For the RACP module, the memory bank $\mathcal{M}$ is constructed from the training and validation sets to build a diverse knowledge base, with the test set strictly held out to prevent data leakage. The retrieval process is conducted as an efficient offline pre-processing step, introducing no latency during model training or inference. The number of retrieved instances K is set to 5 and the prompt length L_p is 16. For the UARE module, we utilize TextCNN (Kim, 2014) as the textual expert, CNN (LeCun et al., 2002) as the visual expert, and MLPs as multimodal experts. We configure the number of modality-specific experts $N_s = 6$ and shared experts $N_c = 12$. In the final aggregation stage, we use $N_f = 8$ MLPs as fusion experts. The trade-off hyperparameters α and β in Eq. 18 are both set to 0.001. All models are trained for 50 epochs with a batch size of 64, using the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 0.0001. All experiments are

Module	Variant	Accuracy		
		Weibo	Weibo-21	GossipCop
CURE	**All**	**0.940**	**0.961**	**0.902**
RACP	w/o RACP	0.930	0.950	0.891
	Rd Retrieval	0.931	0.952	0.892
	Static Prompt	0.936	0.957	0.898
UARE	w/o UARE	0.929	0.950	0.890
	w/o U-Adaptive	0.931	0.951	0.891
	w/o U-Loss	0.934	0.953	0.895
	w/o KL-Loss	0.935	0.956	0.897

Table 3: Ablation studies of CURE on three datasets.

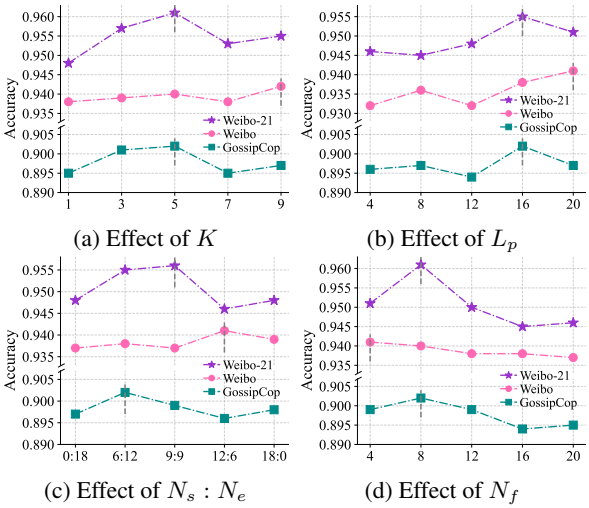


Figure 4: Sensitivity analysis of CURE on three datasets.

conducted on an NVIDIA A100 GPU, with results averaged over five independent runs.

4.2 Overall Performance

The performance of CURE is evaluated against several state-of-the-art methods, with detailed results presented in Table 1. It can be observed that CURE achieves the highest accuracy of 94.0%, 96.1%, and 90.2% on the three datasets, respectively. While some baseline methods might achieve slightly higher Precision or Recall in specific categories on Weibo or GossipCop datasets, CURE consistently secures the highest F1-scores for both fake and real news across all datasets. This highlights its superior ability to provide a robust detection by addressing the modality imbalance issue.

More importantly, we directly validate CURE's capability against modality imbalance by comparing it with specialized methods in Table 2. Our method significantly outperforms these approaches in accuracy, achieving improvements ranging from 1.1% to 2.8%. This advantage arises because while competing methods primarily adjust training dynamics, CURE tackles modality imbalance by directly addressing the issue of low-quality data, which manifests as inter-modal information disparities and noise interference. We note that as addressing this modality imbalance issue in MFND is a recent direction with few public implementations, our comparison is performed against the results reported in their respective papers on the Weibo dataset.

4.3 Ablation Studies

To evaluate the impact of each component within CURE, we conducted an extensive ablation study and summarize the results in Table 3.

• **Effect of RACP module.** We designed 3 variants to analyze the effect of the RACP module: (1) **w/o RACP** removes the entire module; (2) **Rd Retrieval** replaces semantic retrieval with random instances; (3) **Static Prompt** uses a fixed, non-dynamic prompt. Results show that removing the entire RACP module or using random retrieval causes the most significant performance drop. This confirms that leveraging retrieved, relevant context is crucial for bridging inter-modal information disparities. Additionally, the improvement over static prompt validates the effectiveness of our dynamic generation mechanism.

• **Effect of UARE module.** To assess the effect of the UARE module, we designed 4 key variants: (1) **w/o UARE** removes the entire module; (2) **w/o U-Adaptive** removes the uncertainty quantification mechanism; (3) **w/o U-Loss** and (4) **w/o KL-Loss** ablate the uncertainty-adaptive routing loss and the KL divergence loss, respectively. The significant performance degradation observed when removing the UARE module or its core uncertainty-adaptive mechanism validates our central hypothesis: explicitly modeling feature uncertainty is paramount for mitigating noise interference. The associated loss functions also prove essential for ensuring robust fusion.

4.4 Hyper-Parameter Analysis

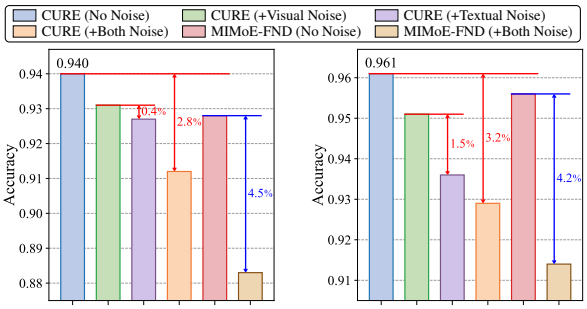
Figure 4 presents the sensitivity analysis of CURE's key parameters on three datasets.

• **Number of retrieved instances K.** As shown in Figure 4(a), the model's performance initially improves as K increases, but then declines. This suggests that a small K provides insufficient context, while a large K may introduce noise from

irrelevant instances. We thus select $K = 5$ for its

	Image 1	Text 1	Image 2	Text 2
Query		Khloe celebrates birthday for Tristan...		Judge denies Bill Cosby's request for new trial...
Retrieved Instance 1		Khloe wishes Tristan a happy 27th birthday...		Bill Cosby returns to court for his retrial...
Retrieved Instance 2		Khloe posts sweet tribute for Tristan's 27th birthday…		Bill Cosby jurors give conflicting accounts of deadlock...

Figure 5: Examples of Top-2 Retrieved Instances. Red texts highlight similar content, while purple highlights supplementary information.



(a) Weibo. (b) Weibo-21.

Figure 6: Analysis of robustness to noise interference on the Weibo and Weibo-21 datasets. We compare the accuracy of CURE and the MIMoE-FND baseline across unimodal (Visual/Textual) and multimodal noise with injected Gaussian noise ($\sigma = 1.5$).

robust balance across all datasets.

• **Prompt length L_p.** Figure 4(b) shows a similar rise-and-fall trend. A shorter prompt may fail to capture sufficient context, while a longer one can introduce noise. We thus adopt $L_p = 16$ for optimal performance.

• **Ratio of experts N_s:N_c.** Figure 4(c) confirms that using only one type of expert (specific or shared) is suboptimal. The ratio N_s:N_c = 6:12 yields robust results, suggesting this configuration is most effective for modeling both modality-specific and cross-modal noise patterns.

• **Number of fusion experts N_f.** Figure 4(d) shows that performance peaks at $N_f = 8$. We select this value as it is sufficient for effective fusion without adding redundant complexity.

4.5 Retrieval Quality Presentation

To analyze the retrieval quality of our RACP module, we visualize the Top-2 retrieved instances for two examples from the GossipCop dataset in Figure 5. The retrieved instances not only exhibit strong semantic correlation with the target news but also provide valuable supplementary context. This confirms RACP's ability to identify relevant information for bridging inter-modal disparities.

4.6 Robustness to Noise Interference Analysis

Building on our pilot study, we evaluate robustness by injecting Gaussian noise ($\sigma = 1.5$) into textual, visual, and multimodal inputs. As shown in Figure 6, CURE outperforms the strong MIMoE-FND baseline in two aspects. First, CURE exhibits superior robustness to noise. Specifically, under multimodal noise, it experiences a marginal accuracy drop of 2.8% on Weibo and 3.2% on Weibo-21, compared to severe degradations of 4.5% and 4.2% for the baseline. Second, CURE demonstrates a more balanced sensitivity to noise across modalities. The performance gap between visual-only and textual-only noise is merely 0.4% on Weibo and 1.5% on Weibo-21 for CURE, significantly narrower than the baseline's 2.2% and 2.5%. This resilience is attributed to the UARE module, which effectively identifies and down-weights noisy features, rebalancing modal contributions to ensure a robust fusion.

4.7 Modality Rebalance Analysis

To thoroughly investigate how CURE addresses the modality imbalance issue, we analyze the testing accuracy curves under different modality settings, as shown in Figure 7. The results reveal that CURE fosters a synergistic rebalancing, where individual modality branches within the framework surpass their independently trained counterparts. Detailed analysis is provided in Appendix B.

5 Conclusion

In this paper, we propose CURE that aims to alleviate heterogeneous modality imbalance in multimodal fake news detection. To bridge inter-modal information disparities, the Retrieval-Augmented Context Prompter module retrieves similar instances and distills them into dynamic prompts to provide supplementary context. To mitigate noise interference, the Uncertainty-Adaptive Rebalancing Experts module quantifies feature-level noise via uncertainty modeling and leverages an uncertainty-adaptive routing mechanism for a robust fusion. Extensive experiments conducted on three real-world datasets demonstrate the superiority of our method.

Limitations

Our work has two main limitations that may affect the generalization and application of the proposed framework. First, the performance of our RACP module relies on a memory bank containing relevant instances, which limits its effectiveness on novel or niche topics. Additionally, the retrieval process introduces computational overhead, posing a concern for real-time applications. Second, while our UARE module effectively mitigates noise by quantifying its magnitude via uncertainty, it may not effectively distinguish its semantic source. For example, it may struggle to differentiate between uncertainty arising from low-quality data (e.g., a blurry image) and that from semantically anomalous content (e.g., a manipulated image). The former is pure noise to be suppressed, whereas the latter can be a strong indicator of fake news. By uniformly down-weighting features based on uncertainty magnitude, our model may overlook this crucial detection signal. We plan to address these limitations in future research.

Ethical Considerations

This paper adheres to the ACL Code of Ethics and Professional Conduct. Firstly, our research utilizes publicly available datasets, ensuring no sensitive private information is involved and no harm is posed to individuals. Secondly, proper attribution is given to all relevant prior work, data sources, and pre-trained models. Furthermore, our source code will be made publicly available and will adhere to the licenses of all artifacts it relies on. Lastly, this work aims to contribute to a healthier online information ecosystem by combating the spread of multimodal fake news.

References

Jie Chang, Zhonghao Lan, Changmao Cheng, and Yichen Wei. 2020. Data uncertainty learning in face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5710–5719.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Xiulan Hao, Wenjing Xu, Xu Huang, Zhenzhen Sheng, and Huayun Yan. 2024. Mfuie: A fake news detection model based on multimodal features and user information enhancement. *EAI Endorsed Transactions on Scalable Information Systems*, 12(1).

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

Yue He, Jingxi Xie, Fengling Li, Lei Zhu, and Jingjing Li. 2025. Flip is better than noise: Unbiased interest generation for multimedia recommendation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 6298–6306.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. 2022. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *International conference on machine learning*, pages 9226–9259. PMLR.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 795–816.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International conference on learning representations (ICLR)*, volume 5. California;.

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 2002. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Zuocheng Li and Lishuang Li. 2025. t-HNE: A text-guided hierarchical noise eliminator for multimodal sentiment analysis. In *Proceedings of the 31st International Conference on Computational Linguistics*,

pages 2834–2844, Abu Dhabi, UAE. Association for Computational Linguistics.

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

Alex Munyole Luvembe, Weimin Li, Shaohau Li, Fangfang Liu, and Xing Wu. 2024. Caf-odnn: Complementary attention fusion with optimized deep neural network for multimodal fake news detection. *Information Processing & Management*, 61(3):103653.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 3343–3347.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. 2022. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8238–8247.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8748–8763. PMLR.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 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.

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth International Conference on Multimedia Big Data*, pages 39–47. IEEE.

Xinqi Su, Zitong Yu, Yawen Cui, Ajian Liu, Xun Lin, Yuhao Wang, Haochen Liang, Wenhui Li, Li Shen, and Xiaochun Cao. 2025. Dynamic analysis and adaptive discriminator for fake news detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 8164–8173.

Yu Tong, Weihai Lu, Zhe Zhao, Song Lai, and Tong Shi. 2024. Mmdfnd: Multi-modal multi-domain fake news detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1178–1186.

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science*, 359(6380):1146–1151.

Bing Wang, Ximing Li, Changchun Li, Shengsheng Wang, and Wanfu Gao. 2024a. Escaping the neutralization effect of modality features fusion in multimodal fake news detection. *Information Fusion*, 111:102500.

Jiandong Wang, Hongguang Zhang, Chun Liu, and Xiongjun Yang. 2024b. 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.

Weiyao Wang, Du Tran, and Matt Feiszli. 2020. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12695–12705.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 849–857.

Zhenyu Wang, Chao Jiang, and Kun Wang. 2026. Dclr: mitigating modality imbalance in multimodal fake news detection via dynamic contrastive learning with re-initialization. *Multimedia Systems*, 32(1):45.

Fei Wu, Shu Chen, Zhe-Ying Deng, Di Wu, Chao Lan, Yimu Ji, and Xiao-Yuan Jing. 2025a. Pfbl: Prototype-based fully balanced learning for multimodal fake news detection. *IEEE Transactions on Computational Social Systems*.

Fei Wu, Shu Chen, Guangwei Gao, Yimu Ji, and Xiao-Yuan Jing. 2025b. Balanced multi-modal learning with hierarchical fusion for fake news detection. *Pattern Recognition*, 164:111485.

Fei Wu, Hao Jin, Feng Chen, Yimu Ji, Xiao-Yuan Jing, and Guo-Ping Jiang. 2025c. Balanced cross-modal prompt learning and fusion network for multi-modal fake news detection. *Information Fusion*, page 103196.

Guipeng Xv, Xinyu Li, Ruobing Xie, Chen Lin, Chong Liu, Feng Xia, Zhanhui Kang, and Leyu Lin. 2024. Improving multi-modal recommender systems by denoising and aligning multi-modal content and user feedback. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3645–3656.

Yang Yang, Hongpeng Pan, Qing-Yuan Jiang, Yi Xu, and Jinhui Tang. 2025. Learning to rebalance multi-modal optimization by adaptively masking subnetworks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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

Xinquan Yu, Ziqi Sheng, Wei Lu, Xiangyang Luo, and Jiantao Zhou. 2025. 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.

Haoyu Zhang, Wenbin Wang, and Tianshu Yu. 2024a. Towards robust multimodal sentiment analysis with incomplete data. *Advances in Neural Information Processing Systems*, 37:55943–55974.

Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Cheng Deng, Qinghua Hu, Cai Xu, Jie Wen, Di Hu, and Changqing Zhang. 2024b. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947*.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. Safe: Similarity-aware multi-modal fake news detection. In *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II*, page 354–367. Springer-Verlag.

Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. 2023. Multimodal fake news detection via clip-guided learning. In *2023 IEEE International Conference on Multimedia and Expo*, pages 2825–2830. IEEE.

Peican Zhu, Yubo Jing, Le Cheng, Keke Tang, and Yangming Guo. 2025. Ken: Knowledge augmentation and emotion guidance network for multimodal fake news detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 1793–1801.

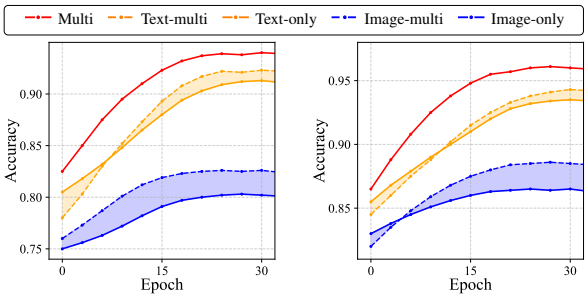
A Experimental Settings

A.1 Datasets

Datasets	Train			Test		
	Real	Fake	Total	Real	Fake	Total
Weibo	3749	3783	7532	996	1000	1996
Weibo-21	3712	3589	7301	928	898	1826
GossipCop	7974	2036	10010	2285	545	2830

Table 4: Statistics of three fake news datasets.

We evaluate our method on three real-world datasets: Weibo (Jin et al., 2017), Weibo-21 (Nan et al., 2021) and GossipCop (Shu et al., 2020), which are all widely adopted datasets in MFND. We use the original train-test splits for Weibo and GossipCop, and adopt a 9:1 ratio for Weibo-21, following BMR (Ying et al., 2023). Dataset statistics are detailed in Table 4.



(a) Weibo. (b) Weibo-21.

Figure 7: Accuracy curves of CURE across different modality settings on the Weibo and Weibo-21 datasets. The curves illustrate the performance comparison between the full multimodal model, independent single-modality models, and the individual branches within the joint framework.

A.2 Baselines

We compare CURE with several competitive baselines divided into three groups: **(1) Representation-centric Early Fusion methods**, including: EANN (Wang et al., 2018), SpotFake (Singhal et al., 2019) and MFUIE (Hao et al., 2024). **(2) Cross-Modal Consistency and Alignment methods**, including: SAFE (Zhou et al., 2020), CAFE (Chen et al., 2022), MSACA (Wang et al., 2024b), FND-CLIP (Zhou et al., 2023) and KEN (Zhu et al., 2025). **(3) Adaptive Strategy-based Fusion methods**, including: BMR (Ying et al., 2023), MINER-UVS (Wang et al., 2024a), RaCMC (Yu et al., 2025), MIMoE-FND (Liu et al., 2025) and DAAD (Su et al., 2025). We also compare our method with recent approaches that specifically address modality imbalance in MFND, namely PFBL (Wu et al., 2025a), BMLHF (Wu et al., 2025b), BCMPL (Wu et al., 2025c) and DCLR (Wang et al., 2026).

B Detailed Modality Rebalance Analysis

To thoroughly investigate how CURE addresses the modality imbalance issue, we analyze the testing accuracy curves under different modality settings, as shown in Figure 7. The results reveal two significant insights. First, the full multimodal CURE consistently achieves the highest accuracy com-

pared to any single-modality variant on both Weibo and Weibo-21 datasets, confirming the effectiveness of our fusion strategy. Second, and more importantly, the individual modality branches trained within CURE exhibit a clear performance boost. Specifically, both the textual and visual branches within the joint framework (i.e., Text-multi and Image-multi) surpass their independently trained counterparts (i.e., Text-only and Image-only) on both datasets. This demonstrates that instead of the dominant modality suppressing the weaker one, CURE effectively bridges information disparities and optimizes contribution weights. This leads to a synergistic effect where visual and textual representations are mutually enhanced, thereby achieving synergistic modality rebalancing.