DDU: Decoupling Dual-Level Uncertainties for Reliable Multimodal Fake News Detection

Xuran Chen
Nanjing University of Science and Technology
School of Computer Science and Engineering
Nanjing, China
xuran_chen@njust.edu.cn

Xianglin Wei
National University of Defense Technology
63rd Research Institute
Nanjing, China
weixianglin@nudt.edu.cn

Yang Yang*
Nanjing University of Science and Technology
School of Computer Science and Engineering
Nanjing, China
yyang@njust.edu.cn

Abstract

The proliferation of fake news on social media motivates the development of multimodal fake news detection. While most existing methods focus on various fusion strategies, they can still yield unreliable results caused by epistemic and aleatoric uncertainties, which can arise from the restricted contextual horizon of isolated sample pairs or inherent noise in unimodal data. To address this, we propose a novel **D**ecoupling **D**ual-level **U**ncertainties (DDU) framework for reliable multimodal fake news detection. First, for epistemic uncertainty, we retrieve semantically similar instances along with their veracity labels, distilling them into dynamic prompts to anchor and calibrate the perception of semantic relationships. Second, for aleatoric uncertainty, we employ a Mixture-of-Experts architecture to quantify feature-level uncertainty. A novel uncertainty-aware routing mechanism is then introduced to dynamically down-weight features with high uncertainty, ensuring reliable multimodal fusion. 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://github.com/njustkmg/MM26-DDU.

CCS Concepts

• **Information systems** → **Data mining**; **Social networks**; • **Computing methodologies** → *Artificial intelligence*.

Keywords

Fake News Detection, Multimodal learning, Uncertainty Learning

ACM Reference Format:

Xuran Chen, Xianglin Wei, and Yang Yang. 2026. DDU: Decoupling Dual-Level Uncertainties for Reliable Multimodal Fake News Detection. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3835912

*Corresponding author.

MM '26, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3835912

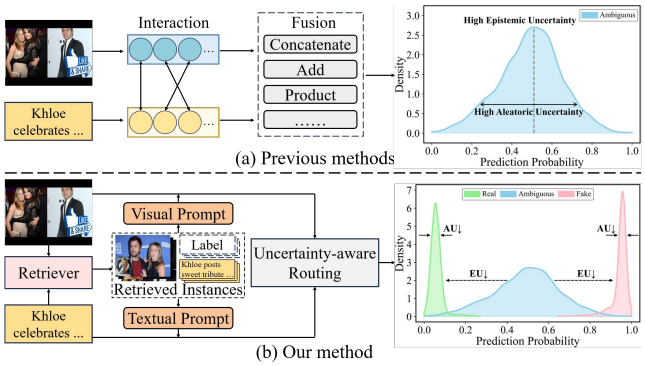


Figure 1: Comparison of the fusion and detection processes. (a) Previous methods primarily rely on the interaction of isolated sample pairs and various fusion strategies, resulting in high epistemic and aleatoric uncertainties. (b) Our method retrieves similar instances along with their veracity labels as semantic anchors to alleviate epistemic uncertainty, and employs an uncertainty-aware routing mechanism to mitigate aleatoric uncertainty.

1 Introduction

The proliferation of social media has fueled the spread of multimodal fake news, which combines text and images to appear more credible and deceptive. Such disinformation poses significant threats to societal trust and stability [19, 34]. As manually fact-checking this massive volume of content is infeasible, the development of automated Multimodal Fake News Detection (MFND) techniques has become an urgent research priority [23, 37].

Existing MFND methods primarily explore various fusion strategies to integrate multimodal signals, which can be categorized into three stages: (1) Representation-centric Basic Fusion methods [31, 36] focus on learning strong unimodal representations and fusing them with simple aggregation, but often ignore the intricate semantic interactions between modalities. (2) Correlation-aware Alignment Fusion methods [3, 46, 47] explicitly model the semantic interactions between modalities, such as similarity or ambiguity, as key detection signals. However, these methods primarily rely on global representations, which limits their effectiveness in identifying deceptive news with fine-grained inconsistencies. (3) Fine-grained Adaptive Fusion methods [25, 32, 41, 42] employ advanced architectures with multi-granularity or multi-view branches to dynamically capture complex modality interactions.

Although existing methods have achieved a certain degree of performance, they often rely on a fundamental assumption that the input multimodal data is inherently complete and reliable [43]. This deterministic perspective overlooks the potential unreliability of features, leading to two critical challenges: **(1) Epistemic uncertainty stemming from sophisticated semantic deception.** Fake news often exploits maliciously constructed cross-modal relationships to mislead detectors [46]. Relying solely on isolated image-text pairs renders it difficult for models to verify these complex deceptive cues. As shown in Figure 1(a), this limited contextual horizon leads to high epistemic uncertainty, where the model produces a fuzzy prediction around 0.5 because it cannot effectively anchor and disambiguate the underlying semantic claims. **(2) Aleatoric uncertainty arising from inherent data noise.** In real-world scenarios, visual and textual data are inevitably corrupted by inherent noise such as blurs or typos [43]. As illustrated in Figure 1(a), existing fusion mechanisms typically rely on operations like concatenation or addition to blindly aggregate these unreliable features without quantifying their quality. Consequently, the unperceived noise results in high aleatoric uncertainty, manifested as a wide and flattened probability distribution that signifies a severe lack of decision-making reliability.

To address the challenges above, we propose a novel **D**ecoupling **D**ual-level **U**ncertainties (DDU) framework for reliable multimodal fake news detection. Specifically, to alleviate epistemic uncertainty, we first design a Retrieval-Augmented Context Prompter (RACP) module. As shown in Figure 1(b), RACP bridges the limited contextual horizon of isolated samples by retrieving semantically similar instances along with their veracity labels. These retrieved instances are then distilled into dynamic prompts that act as semantic reference anchors to calibrate the model's perception of complex semantic relationships. By introducing these external anchors, the prediction mean is effectively shifted from the ambiguous 0.5 toward confident endpoints (0 or 1), thereby resolving the underlying semantic ambiguities.

Second, to mitigate aleatoric uncertainty, we develop the Uncertainty-Aware Routing Experts (UARE) module. Inspired by [2], the module quantifies feature-level aleatoric uncertainty via probabilistic modeling within a Mixture-of-Experts architecture. Specifically, each expert maps its features onto a Gaussian distribution, where the resulting variance serves as a direct measure of uncertainty. A novel uncertainty-aware routing mechanism, as illustrated in Figure 1(b), is then introduced to achieve reliable multimodal fusion by dynamically down-weighting experts with high uncertainty. Consequently, the final probability distribution becomes significantly concentrated and sharp, yielding reliable fusion that effectively suppresses the negative impact of inherent noise.

In summary, our main contributions are:

- To the best of our knowledge, this is the first work to decouple dual-level uncertainties for multimodal fake news detection. We reveal that the prevalent deterministic assumption in existing methods results in high epistemic and aleatoric uncertainties.
- To address these issues, we propose DDU, which first retrieves similar instances along with their veracity labels as semantic anchors to alleviate epistemic uncertainty. Furthermore, DDU quantifies aleatoric uncertainty and dynamically down-weights experts with high uncertainty for reliable multimodal fusion.
- We conduct experiments on three real-world datasets spanning two languages, demonstrating that our method 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 stages: **(1) Representation-centric Basic Fusion methods** [31, 36] focus on learning strong unimodal representations and fusing them with simple aggregation, but often ignore the intricate semantic interactions between modalities. **(2) Correlation-aware Alignment Fusion methods** [3, 46, 47] explicitly model the semantic relationship between modalities, such as similarity or ambiguity, as key detection signals. However, these methods primarily rely on global representations, which limits their effectiveness in identifying deceptive news with fine-grained inconsistencies. **(3) Fine-grained Adaptive Fusion methods** [26, 32, 41, 42] employ advanced architectures with multi-granularity or multi-view branches to dynamically capture complex modality interactions.

However, these methods largely rely on a fundamental assumption that the input multimodal data is inherently complete and reliable [12, 39, 40, 43]. This deterministic perspective overlooks the potential unreliability of features, leading to high epistemic uncertainty caused by isolated semantic horizons and aleatoric uncertainty triggered by data noise. Recently, some efforts have attempted to introduce external knowledge to enhance detection. For instance, KEN [47] leverages retrieved evidence and large vision-language models to break information silos. While such knowledge-augmented methods provide supplementary context to alleviate semantic deficiency, they typically do not account for the aleatoric uncertainty arising from inherent data noise. In contrast, our proposed DDU framework decouples epistemic and aleatoric uncertainties, providing a more reliable fusion mechanism for multimodal fake news detection.

2.2 Uncertainty in Deep Learning

In general, uncertainty in deep learning can be categorized into epistemic uncertainty and aleatoric uncertainty [16]. The former represents the uncertainty inherent in the neural network itself, indicating the confidence level in its predictions, which typically originates from a lack of knowledge or limited training data. The latter refers to the inherent noise present in the data, such as image blurriness or text ambiguity. Both types of uncertainty pose significant challenges to the model's predictions, potentially leading to unreliable outputs in complex scenarios. In recent years, some efforts have been made to address uncertainty in various fields, such as computer vision [2, 15], information retrieval [21, 22], and multimedia understanding [8, 13]. It is worth noting that such uncertainty is also prevalent in multimodal fake news detection, where sophisticated semantic deception and inherent data noise

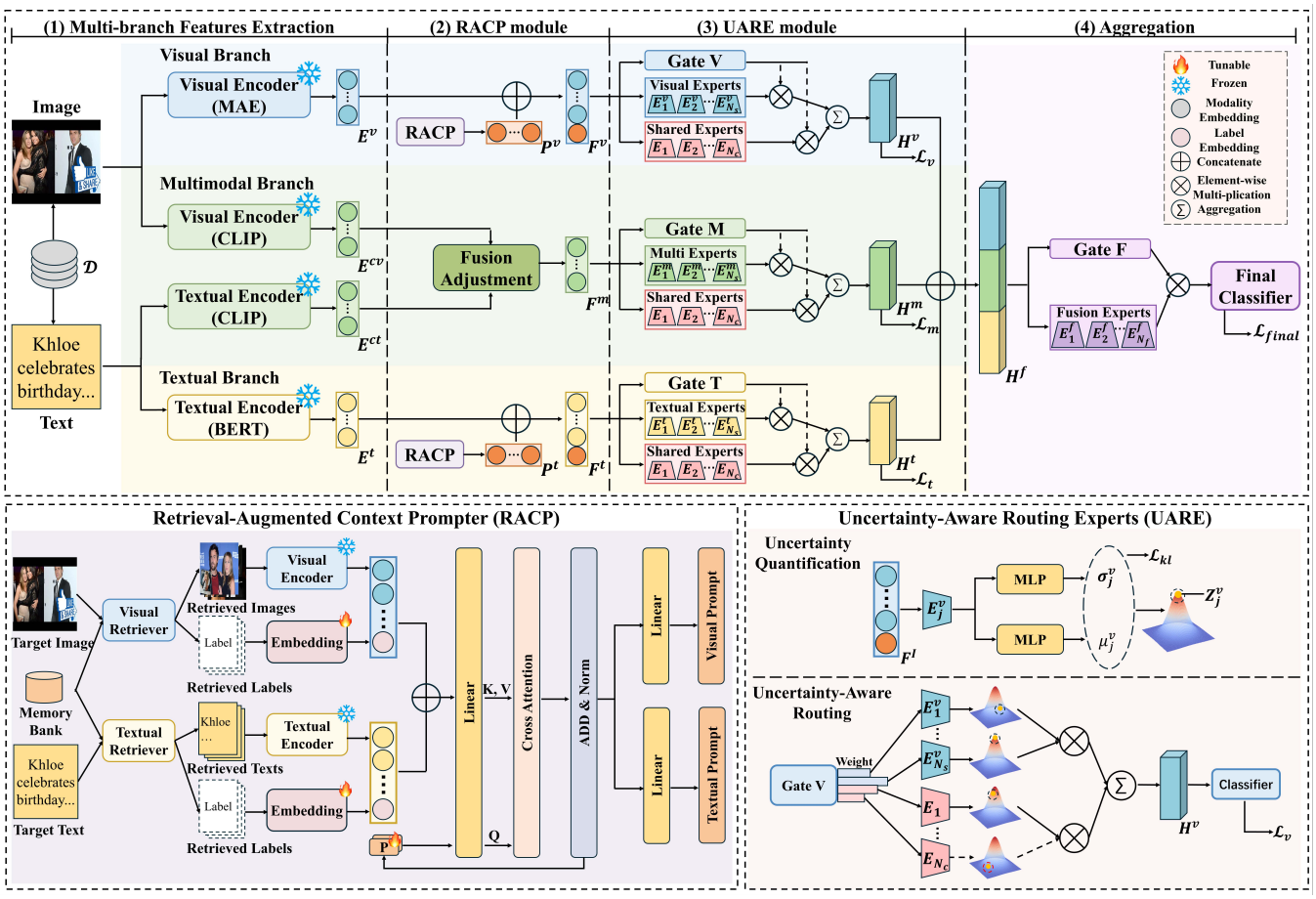


Figure 2: The overall framework of DDU consists of four components: (1) Multi-branch Feature Extraction: extracting unimodal and cross-modal representations from raw inputs; (2) Retrieval-Augmented Context Prompter (RACP): retrieving semantically similar instances as anchors to alleviate epistemic uncertainty; (3) Uncertainty-Aware Routing Experts (UARE): quantifying aleatoric uncertainty via probabilistic modeling and performing adaptive routing to mitigate aleatoric uncertainty; (4) Aggregation: integrating uncertainty-rectified features through fusion experts for final fake news detection.

lead to significant epistemic and aleatoric uncertainties, respectively, which can severely mislead the final decisions. Recently, some research [7, 44] has focused on addressing uncertainty in multimodal learning. However, these efforts often focus on a single uncertainty source or provide a unified quantification without distinguishing their distinct origins. Therefore, this work aims to fill this gap by decoupling epistemic and aleatoric uncertainties, aiming to achieve more reliable multimodal fake news detection.

3 Methodology

In this section, we introduce the proposed DDU, designed to alleviate epistemic and aleatoric uncertainties. As illustrated in Figure 2, it consists of four components: (1) multi-branch feature extraction (§3.1), (2) Retrieval-Augmented Context Prompter (RACP) module (§3.2), (3) Uncertainty-Aware Routing Experts (UARE) module (§3.3), and (4) final aggregation and loss function (§3.4).

3.1 Multi-branch Feature 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 [4] 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 [11] 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 [28] to acquire aligned embeddings E^{ct} and E^{cv}. For the multimodal branch, following [33], we generate a fused feature $F^m \in \mathbb{R}^{d_m}$ via fusion adjustment, which weights the concatenated CLIP embeddings by their cosine similarity to bridge the cross-modal semantic gap:

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

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

3.2 Retrieval-Augmented Context Prompter

To mitigate epistemic uncertainty, we propose the RACP module. RACP retrieves semantically similar instances along with their veracity labels as reference anchors, thereby bridging the limited contextual horizon of isolated samples. By distilling these external signals into dynamic prompts, the module calibrates the model's perception of complex semantic relationships, effectively resolving underlying ambiguities.

Multi-branch Retriever To retrieve relevant evidence, we first construct a memory bank $\mathcal{M} = \{(t_m, v_m, y_m)\}_{m=1}^{|\mathcal{M}|}$, where t_m, v_m, and y_m denote the raw text, image, and associated ground-truth veracity label of the m-th instance, respectively. To prevent a high-quality query in one modality from being contaminated by potential unreliability 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:

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

The visual context set $\mathcal{R}^v$ is obtained analogously. This multi-branch retrieval yields two modality-specific sets of reference instances, ensuring that each modality is enhanced by its most semantically relevant evidence without cross-modal interference.

Dynamic Prompt Generation and Enhancement To transform raw retrieved instances into actionable semantic signals, we design a dynamic prompt generation mechanism. Our key insight is to enrich the context with explicit veracity signals (i.e., ground-truth labels), empowering a cross-attention mechanism to selectively amplify salient information that helps disambiguate deceptive claims.

Specifically, we first construct the context for each modality. For the retrieved textual set $\mathcal{R}^t$, we form a rich context representation 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 $E^{l,t} \in \mathbb{R}^{2 \times d_e}$. This label embedding provides crucial veracity semantics, serving as a benchmark for the model to distinguish between authentic and fabricated patterns. After generating the visual context C^v analogously, both are concatenated into a unified context representation C to capture potential synergistic signals across modalities.

To adaptively distill the most effective evidence 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 (acting as keys and values) 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:

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

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

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

where $f^Q(\cdot)$, $f^K(\cdot)$, and $f^V(\cdot)$ are linear projections.

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

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

By incorporating these dynamic prompts, the RACP module effectively provides essential semantic anchors to guide the model's judgment, thereby significantly mitigating epistemic uncertainty.

3.3 Uncertainty-Aware Routing Experts

To mitigate the aleatoric uncertainty arising from inherent data noise, we propose the UARE module. Unlike traditional fusion methods that blindly aggregate features regardless of their reliability, UARE first quantifies feature-level aleatoric uncertainty via probabilistic modeling and then employs an uncertainty-aware routing mechanism to achieve reliable multimodal fusion.

Uncertainty Quantification Given that noise patterns across modalities are inherently diverse, we employ a Mixture-of-Experts (MoE) architecture to model these complex distributions. 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:

$$E^k = [E_{s,1}^k, \ldots, E_{s,N_s}^k, E_{c,1}, \ldots, E_{c,N_c}] \tag{8}$$

where $E_{s,j}^k$ 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 aleatoric uncertainty by modeling each expert's output as a diagonal multivariate Gaussian distribution, inspired by Chang et al. [2]:

$$p(Z_j^k | E_j^k(F^k)) \sim \mathcal{N}(\mu_j^k, (\sigma_j^k)^2 I) \tag{9}$$

$$\mu_j^k = f_\mu(E_j^k(F^k)), \sigma_j^k = f_\sigma(E_j^k(F^k)) \tag{10}$$

where the mean μ_j^k represents the refined feature and the variance $(\sigma_j^k)^2$ serves as a direct quantification of its uncertainty level. These parameters are generated by two distinct fully-connected layers, f_μ and f_σ. To ensure differentiability for back-propagation, we apply the re-parameterization trick:

$$Z_j^k = \mu_j^k + \epsilon \cdot \sigma_j^k, \epsilon \sim \mathcal{N}(0, I) \tag{11}$$

To ensure the variance term meaningfully captures the inherent uncertainty rather than collapsing to zero, we introduce a KL divergence regularization term:

$$\mathcal{L}_{kl}^k = \frac{1}{N_e} \sum_{j=1}^{N_e} \text{KL}[\mathcal{N}(\mu_j^k, (\sigma_j^k)^2 I) || \mathcal{N}(0, I)] \tag{12}$$

Uncertainty-Aware Routing Having quantified the uncertainty level of each expert, we employ an uncertainty-aware routing mechanism to prioritize reliable information for fusion. A gating network G^k processes the input feature F^k to generate routing weights:

$$W^k = \text{Softmax}(G^k(F^k)) \tag{13}$$

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

$$H^k = \sum_{j=1}^{N_e} W_j^k \cdot \mu_j^k \tag{14}$$

To explicitly guide the gating network to down-weight unreliable outputs, we introduce an uncertainty-aware routing loss $\mathcal{L}_u^k$:

$$\mathcal{L}_u^k = \frac{1}{N_e} \sum_{j=1}^{N_e} W_j^k \cdot (\sigma_j^k)^2 \tag{15}$$

This loss penalizes assigning high weights to experts with large variance, thereby prioritizing features with low uncertainty for the subsequent fusion stage. Consequently, this mechanism ensures that reliable features dominate the decision-making process, effectively suppressing the negative impact of inherent data noise and leading to a more concentrated probability distribution.

3.4 Aggregation and Loss Function

Finally, the uncertainty-rectified 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}:

$$W^f = \text{Softmax}(G^f(H^f)) \tag{16}$$

$$H_{\text{final}} = \sum_{i=1}^{N_f} W_i^f \cdot E_i^f(H^f) \tag{17}$$

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 $\mathcal{L}_b^k$. Both losses are formulated using the Binary Cross-Entropy (BCE). The overall training objective is to minimize a total loss function $\mathcal{L}$:

$$\mathcal{L} = \mathcal{L}_{\text{final}} + \sum_k \mathcal{L}_b^k + \alpha \sum_k \mathcal{L}_{kl}^k + \beta \sum_k \mathcal{L}_u^k \tag{18}$$

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-aware routing constraint, respectively.

4 Experiments

4.1 Experimental Settings

• **Datasets.** To evaluate the effectiveness of DDU, we evaluate our method on three real-world datasets: Weibo [14], Weibo-21 [27], and GossipCop [30], 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 [41]. Dataset statistics are detailed in Table 1.

• **Baselines.** We compare DDU with several competitive baselines divided into three groups: **(1) Representation-centric Basic Fusion methods**, including: EANN [36] and SpotFake [31]. **(2) Correlation-aware Alignment Fusion methods**, including: SAFE [45], CAFE [3], MSACA [35], FND-CLIP [46] and KEN [47]. **(3) Fine-grained Adaptive Fusion methods**, including: BMR [41], RaCMC [42], MIMoE-FND [24] and DAAD [32].

• **Implementation Details.** For the RACP module, the memory bank $\mathcal{M}$ is constructed from the training 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 [17] as the textual expert, CNN [20] 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. For the multimodal branch, we employ Chinese-CLIP [38] for Chinese datasets (Weibo and Weibo-21) and OpenCLIP [29] for the English dataset (GossipCop) to ensure language-specific alignment. All models are trained for 50 epochs with a batch size of 64, using the Adam optimizer [18] with an initial learning rate of 0.0001. All experiments are conducted on an NVIDIA 4090 GPU.

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 1: Statistics of three fake news datasets.

4.2 Overall Performance

Table 2 compares DDU with several state-of-the-art methods across four evaluation metrics. It can be observed that DDU achieves the highest accuracy of 94.0%, 96.1%, and 90.2% on the three datasets, respectively. Notably, DDU consistently ranks first in both Accuracy and F1-score for fake and real news categories across all datasets. Specifically, DDU demonstrates a consistent performance lead on the Weibo dataset. On the Weibo-21 dataset, it surpasses the strongest baseline MIMoE-FND across all evaluation metrics. Notably, DDU is the only framework to exceed the 90% accuracy threshold on the challenging GossipCop dataset. These results highlight DDU's superior ability to provide reliable detection by decoupling dual-level uncertainties. Furthermore, we retrain DDU and the best-performing baselines five times. The calculated p-values for all metrics are significantly lower than 0.05, demonstrating that the improvements are statistically significant.

4.3 Ablation Studies

To evaluate each component, we conduct ablation studies and summarize the results in Table 3.

• **Effect of RACP module.** We analyze the RACP module using five variants: 1) **w/o RACP** removes the entire module; 2) **Joint Retrieval** utilizes fused multimodal embeddings for a single retrieval; 3) **Static Prompt** uses fixed non-dynamic prompts; 4) **w/o Label** excludes veracity labels; 5) **kNN Label Voting** predicts labels by voting over the retrieved instances. Results indicate that removing RACP or utilizing joint retrieval leads to substantial performance degradation, confirming that relevant context is vital for anchoring semantics to alleviate epistemic uncertainty. Notably, the performance drop of the w/o Label variant underscores veracity signals as crucial anchors for disambiguating deceptive patterns. Meanwhile, the clear gap between full DDU and kNN Label Voting shows that

Datasets	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F1	Precision	Recall	F1
Weibo	EANN [36]	0.827	0.847	0.812	0.829	0.807	0.843	0.825
	SpotFake [31]	0.892	0.902	0.964	0.932	0.847	0.656	0.739
	SAFE [45]	0.762	0.831	0.724	0.774	0.695	0.811	0.748
	CAFE [3]	0.840	0.855	0.830	0.842	0.825	0.851	0.837
	FND-CLIP [46]	0.907	0.914	0.901	0.908	0.914	0.901	0.907
	BMR [41]	0.918	0.882	**0.948**	0.914	**0.942**	0.879	0.904
	MSACA [35]	0.903	0.935	0.873	0.903	0.872	0.935	0.902
	RaCMC [42]	0.915	0.910	0.924	0.917	0.921	0.906	0.914
	MIMoE-FND [24]	0.928	0.942	0.913	0.928	0.913	0.942	0.927
	DAAD [32]	0.932	0.942	0.915	0.928	0.922	**0.947**	0.934
	KEN [47]	0.935	0.937	0.934	0.935	0.932	0.935	0.934
	DDU (Ours)	**0.940**	**0.943**	0.941	**0.942**	0.937	0.939	**0.938**
	p-value	2.43×10^{-5}	4.12×10^{-5}	1.87×10^{-6}	3.56×10^{-6}	5.21×10^{-5}	2.94×10^{-5}	4.08×10^{-6}
Weibo-21	EANN [36]	0.870	0.902	0.825	0.862	0.841	0.912	0.875
	SpotFake [31]	0.851	0.953	0.733	0.828	0.786	0.964	0.866
	SAFE [45]	0.905	0.893	0.908	0.901	0.916	0.901	0.890
	CAFE [3]	0.882	0.857	0.915	0.885	0.907	0.844	0.876
	FND-CLIP [46]	0.943	0.935	0.945	0.940	0.950	0.942	0.946
	BMR [41]	0.929	0.908	0.947	0.927	0.946	0.906	0.925
	DAAD [32]	0.942	0.951	0.925	0.938	0.931	0.955	0.943
	MIMoE-FND [24]	0.956	0.953	0.957	0.955	0.959	0.956	0.957
	DDU (Ours)	**0.961**	**0.958**	**0.965**	**0.961**	**0.964**	**0.957**	**0.961**
	p-value	1.25×10^{-5}	3.68×10^{-5}	2.14×10^{-6}	1.03×10^{-5}	4.72×10^{-6}	5.81×10^{-5}	2.33×10^{-5}
GossipCop	EANN [36]	0.864	0.702	0.518	0.594	0.887	0.956	0.920
	SpotFake [31]	0.858	0.732	0.372	0.494	0.866	0.962	0.914
	SAFE [45]	0.838	0.758	0.558	0.643	0.857	0.937	0.895
	CAFE [3]	0.867	0.732	0.490	0.587	0.887	0.957	0.921
	FND-CLIP [46]	0.880	0.761	0.549	0.638	0.899	0.959	0.928
	BMR [41]	0.895	0.752	0.639	0.691	0.920	0.965	0.936
	MSACA [35]	0.887	0.816	0.538	0.648	0.897	0.971	0.933
	RaCMC [42]	0.879	0.745	0.563	0.641	0.902	0.954	0.927
	MIMoE-FND [24]	0.895	0.762	**0.644**	0.698	**0.920**	0.953	0.936
	DDU (Ours)	**0.902**	**0.830**	0.614	**0.706**	0.911	**0.973**	**0.941**
	p-value	6.71×10^{-5}	2.84×10^{-5}	5.12×10^{-6}	4.29×10^{-5}	1.96×10^{-5}	3.77×10^{-6}	8.15×10^{-5}

Table 2: Comparison between DDU and state-of-the-art multimodal fake news detection methods on Weibo, Weibo-21 and GossipCop.

Module	Variant	Accuracy		
		Weibo	Weibo-21	GossipCop
DDU	**All**	**0.940**	**0.961**	**0.902**
RACP	w/o RACP	0.930	0.950	0.891
	Joint Retrieval	0.931	0.952	0.892
	Static Prompt	0.936	0.957	0.898
	w/o Label	0.933	0.954	0.896
	kNN Label Voting	0.918	0.941	0.876
UARE	w/o UARE	0.929	0.950	0.890
	w/o U-Aware	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 DDU on three datasets.

RACP does not simply transfer retrieved labels. The improvement over the static prompt further validates the effectiveness of our dynamic generation mechanism.

• **Effect of UARE module.** To assess the UARE module, we evaluate four key variants: 1) **w/o UARE** removes the entire module; 2) **w/o U-Aware** omits uncertainty quantification; 3) **w/o U-Loss** and 4) **w/o KL-Loss** ablate the uncertainty-aware routing loss and the KL divergence loss respectively. The decline observed when removing UARE or its quantification mechanism shows that modeling feature-level uncertainty is important for suppressing aleatoric uncertainty. Additionally, the necessity of both loss functions confirms that uncertainty-aware routing is vital for reliable fusion by effectively prioritizing low-noise features.

4.4 Probability Distribution Analysis

To further investigate the learning dynamics of epistemic and aleatoric uncertainties, we visualize the prediction probability distributions of DDU and the strong MIMoE-FND baseline across different training stages in Figure 3.

• **Epistemic Uncertainty and Mean Calibration.** As illustrated in Figure 3(a), both models initially exhibit high uncertainty, with overlapping distributions centered around 0.5. However, as training

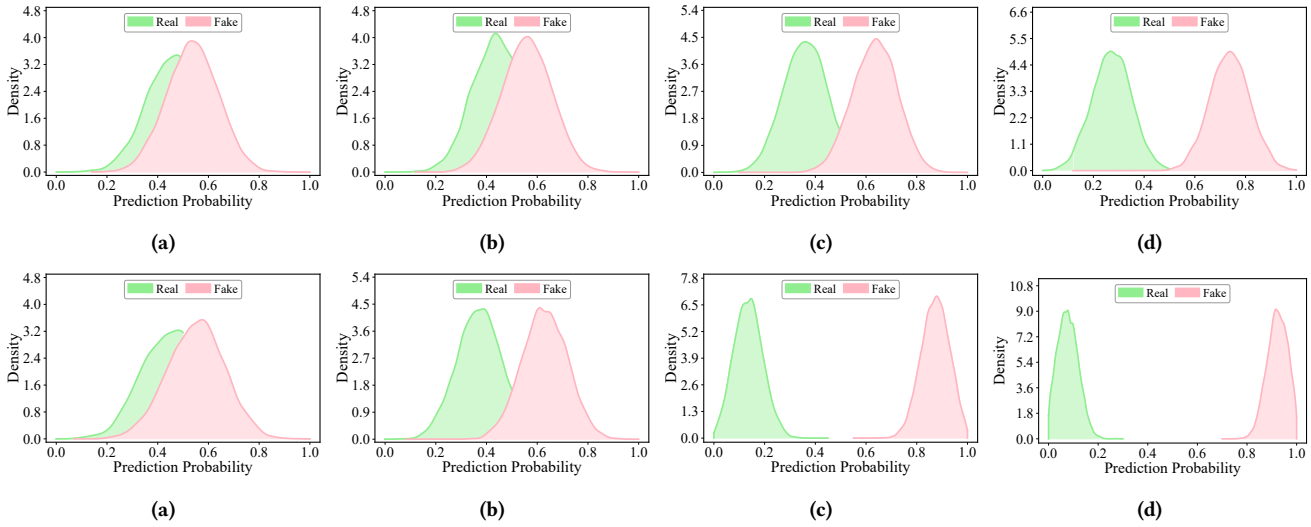


Figure 3: Visualization of the prediction probability distributions at different training stages. The top row displays the MIMoE-FND baseline, and the bottom row shows our DDU. (a) Initial Distribution (b) Epoch 5 (c) Epoch 15 (d) Epoch 30.

Dataset	Method	ECE	Brier	NLL
Weibo	MIMoE-FND	0.071	0.094	0.318
	+ MC Dropout	0.058	0.086	0.291
	+ Temp. Scaling	0.049	0.081	0.274
	DDU	**0.039**	**0.066**	**0.236**
Weibo-21	MIMoE-FND	0.053	0.058	0.206
	+ MC Dropout	0.046	0.053	0.193
	+ Temp. Scaling	0.039	0.049	0.181
	DDU	**0.030**	**0.041**	**0.158**
GossipCop	MIMoE-FND	0.086	0.124	0.392
	+ MC Dropout	0.074	0.116	0.365
	+ Temp. Scaling	0.065	0.109	0.342
	DDU	**0.052**	**0.096**	**0.308**

Table 4: Uncertainty calibration comparison on three datasets.

Figure 4: Examples of Top-2 Retrieved Instances along with their veracity labels. Red texts highlight similar content with the query news.

progresses to epoch 5 shown in Figure 3(b), DDU demonstrates a significantly faster polarization, with its probability means rapidly shifting towards 0 for real news and 1 for fake news. This rapid mean calibration indicates that the RACP module effectively mitigates epistemic uncertainty by providing external semantic anchors, enabling the model to disambiguate deceptive patterns even in early training phases. In contrast, the baseline's means remain closer to the ambiguous center for a much longer duration.

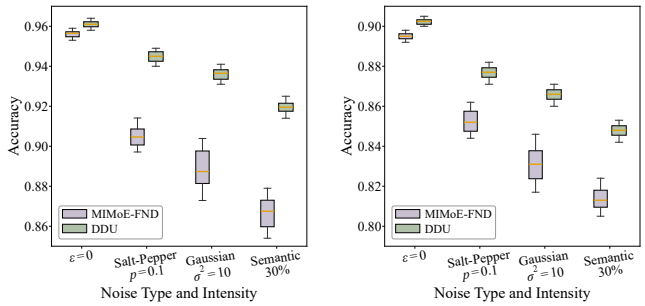
• **Aleatoric Uncertainty and Variance Compression.** Additionally, a significant contrast emerges in the concentration of the probability mass between the two models. At the convergence stage of epoch 30 in Figure 3(d), while the baseline produces relatively wide and flattened peaks, the distributions of DDU are remarkably narrow and concentrated. This significant variance compression signifies the effective suppression of aleatoric uncertainty through our UARE module. By dynamically down-weighting experts with high uncertainty, DDU filters out inherent data noise, resulting in a more reliable fusion.

In summary, the evolution from flat and ambiguous to sharp and polarized distributions quantitatively validates that DDU achieves more reliable detection by decoupling and suppressing dual-level uncertainties.

4.5 Uncertainty Calibration Analysis

To further assess the reliability of detection results, we examine whether predicted confidence is well aligned with empirical correctness, rather than relying only on sharper probability distributions. Therefore, we evaluate calibration using ECE [10], Brier score [1], and NLL [9], and compare DDU with MIMoE-FND and two uncertainty-enhanced variants [6, 10].

As shown in Table 4, DDU consistently obtains the lowest ECE, Brier score, and NLL across all datasets. These results indicate



(a) Weibo-21 (b) GossipCop

Figure 5: Robustness and reliability analysis of MIMoE-FND and DDU on Weibo-21 and GossipCop datasets under various noise conditions.

that DDU improves reliability not merely by producing sharper distributions, but by yielding better calibrated predictions.

4.6 Retrieval Quality Presentation

To analyze the retrieval quality of the RACP module, we visualize the Top-2 retrieved instances for two examples from the GossipCop dataset in Figure 4. The retrieved instances exhibit strong semantic correlations with the query news in both image and text modalities. Crucially, these instances provide explicit veracity labels that serve as reliable semantic anchors. Specifically, both the query and its retrieved evidence share the same veracity labels, such as real in the first example and fake in the second example. By effectively bridging the limited contextual horizon of isolated samples with these semantic anchors, the RACP module provides the necessary evidence to alleviate epistemic uncertainty.

4.7 Robustness and Reliability Analysis

To verify the stability of DDU, we conduct pressure tests under diverse noise scenarios. As illustrated in Figure 5, we compare DDU with the strong baseline MIMoE-FND across four configurations: clean data, salt and pepper noise with a density of 0.1, Gaussian noise with a standard deviation of 10, and semantic interference. Specifically, we randomly replace 30% of the images in the test set with images from other news instances to simulate sophisticated semantic deception. The experimental results reveal a consistent trend where DDU maintains higher median accuracy and significantly narrower box heights than MIMoE-FND across all noise types.

Specifically, DDU exhibits reduced sensitivity to pixel-level or feature-level perturbations. This stability is primarily because the UARE module successfully quantifies aleatoric uncertainty and adaptively prioritizes reliable features to ensure a consistent fusion process. Furthermore, we evaluate the resilience against sophisticated semantic deception. While the performance of MIMoE-FND exhibits a sharp decline and severe instability in this challenging scenario, DDU remains remarkably robust. This resilience highlights the critical role of the RACP module, which retrieves external veracity labels as semantic anchors to calibrate judgment and effectively suppress epistemic uncertainty even when faced with cross-modal mismatches. In summary, DDU achieves a superior balance between accuracy and reliability by maintaining high precision and low variance under various interference conditions.

Method	Params (M)	FLOPs (G)	Inf. (ms/batch)	Train (s/epoch)	Conv. (epochs)
MIMoE-FND [24]	284.7	48.6	115.3 ± 4.2	112.7 ± 3.1	35
DDU (Ours)	272.5	45.2	128.6 ± 3.8	118.5 ± 3.4	30

Table 5: Comparison of complexity and efficiency on Weibo.

4.8 Complexity and Efficiency Analysis

In this section, we analyze the complexity and efficiency of DDU to demonstrate its suitability for real-world deployment.

• **Theoretical Complexity.** DDU's overhead primarily arises from UARE and RACP. For UARE, the 18 experts are implemented as lightweight TextCNN, CNN, and MLPs, yielding $O(N_e d^2)$ per branch. This is lower than the Transformer-based experts in MIMoE-FND [24], which involve $O(n_e N^2 d)$ and $O(n_e N d^2)$ terms. As shown in Table 5, DDU significantly reduces total FLOPs to 45.2G while maintaining a comparable parameter size of 272.5M. Although DDU introduces moderate inference and per-epoch training overhead, it converges in fewer epochs than MIMoE-FND.

• **Retrieval Efficiency and Deployment.** To ensure efficiency, RACP utilizes offline indexing strategy where the memory bank $\mathcal{M}$'s features are pre-computed during the pre-processing phase. In real-world deployment, the indexing and search can be further optimized using the FAISS library [5], which limits the query latency to approximately 4.5ms per batch (< 3.5% of total inference time). While RACP introduces a marginal latency of ~13.3ms compared to MIMoE-FND, the overall inference speed remains well within the requirements for real-time social media monitoring.

• **Memory Bank Update Strategy.** To handle social media's temporal dynamics, $\mathcal{M}$ can be refreshed via a sliding-window strategy (e.g., every 24h) to incorporate the latest fact-checked news. As an offline process, this update does not affect online latency or accuracy, enabling DDU to adapt to evolving deceptive patterns without compromising real-time performance.

5 Conclusion

In this paper, we propose DDU, a novel framework for reliable multimodal fake news detection by decoupling dual-level uncertainties. To alleviate epistemic uncertainty stemming from the restricted contextual horizon of isolated samples, the Retrieval-Augmented Context Prompter module retrieves semantically similar instances along with their veracity labels as semantic anchors, effectively calibrating the model's perception of complex semantic relationships. To mitigate aleatoric uncertainty arising from inherent data noise, the Uncertainty-Aware Routing Experts module quantifies feature-level noise through probabilistic modeling and introduces an uncertainty-aware routing mechanism to prioritize reliable features for fusion. Extensive experiments conducted on three real-world datasets demonstrate that DDU not only significantly outperforms state-of-the-art methods in terms of accuracy but also exhibits superior reliability and robustness under various interference conditions.

Acknowledgments

This work was supported by the NSFC (62276131), the Natural Science Foundation of Jiangsu Province of China (BK20240081), and the Special Research Project on Teaching Reform of General Artificial Intelligence Courses in Jiangsu Undergraduate Universities (No. ZNT-10).

References

[1] Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* 78, 1 (1950), 1–3.

[2] 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*. 5710–5719. doi:10.1109/CVPR42600.2020.00575

[3] 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*. 2897–2905. doi:10.1145/3485447.3511968

[4] 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)*. 4171–4186. doi:10.18653/v1/N19-1423

[5] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *IEEE Transactions on Big Data* (2025).

[6] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*. PMLR, 1050–1059.

[7] Zixian Gao, Disen Hu, Xun Jiang, Huimin Lu, Heng Tao Shen, and Xing Xu. 2024. Enhanced experts with uncertainty-aware routing for multimodal sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 9650–9659.

[8] Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. 2024. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 26876–26885. doi:10.1109/CVPR52733.2024.02538

[9] Tilmann Gneiting and Adrian E Raftery. 2007. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association* 102, 477 (2007), 359–378.

[10] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*. PMLR, 1321–1330.

[11] 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*. 16000–16009. doi:10.1109/CVPR52688.2022.01553

[12] Qing-Yuan Jiang, Longfei Huang, and Yang Yang. 2026. Rethinking multimodal learning from the perspective of mitigating classification ability disproportion. *Advances in Neural Information Processing Systems* 38 (2026), 129580–129607.

[13] Xun Jiang, Zailei Zhou, Xing Xu, Yang Yang, Guoqing Wang, and Heng Tao Shen. 2023. Faster video moment retrieval with point-level supervision. In *Proceedings of the 31st ACM International Conference on Multimedia*. 1334–1342.

[14] 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*. 795–816. doi:10.1145/3123266.3123454

[15] Lie Ju, Xin Wang, Lin Wang, Dwarikanath Mahapatra, Xin Zhao, Quan Zhou, Tongliang Liu, and Zongyuan Ge. 2022. Improving medical images classification with label noise using dual-uncertainty estimation. *IEEE transactions on medical imaging* 41, 6 (2022), 1533–1546.

[16] Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems* 30 (2017). https://proceedings.neurips.cc/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html

[17] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 1746–1751. doi:10.3115/v1/D14-1181

[18] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International conference on learning representations (ICLR)*, Vol. 5. California;.

[19] 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 (2018), 1094–1096. doi:10.1126/science.aao2998

[20] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 2002. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (2002), 2278–2324.

[21] Shenshen Li, Chen He, Xing Xu, Fumin Shen, Yang Yang, and Heng Tao Shen. 2024. Adaptive uncertainty-based learning for text-based person retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 3172–3180.

[22] Shenshen Li, Xing Xu, Yang Yang, Fumin Shen, Yijun Mo, Yujie Li, and Heng Tao Shen. 2023. DCEL: deep cross-modal evidential learning for text-based person retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*. 6292–6300.

[23] Yiheng Li, Weihai Lu, Hanyi Yu, and Yue Wang. 2026. Retrieval-Augmented Multimodal Model for Fake News Detection. *arXiv preprint arXiv:2604.18112* (2026).

[24] 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*. 5139–5150.

[25] Wayne Lu and Yiheng Li. 2026. From Blind Transfer to Wise Selection: Prototype-Driven Neighbor-Domain Adaptation for Fake News Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 40. 818–826.

[26] Weihai Lu, Yu Tong, and Zhiqiu Ye. 2025. Dammfnd: Domain-aware multimodal multi-view fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 559–567.

[27] 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*. 3343–3347.

[28] 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*, Vol. 139. PMLR, 8748–8763. http://proceedings.mlr.press/v139/radford21a.html

[29] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems* 35 (2022), 25278–25294.

[30] 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 (2020), 171–188.

[31] 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*. IEEE, 39–47.

[32] 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*. 8164–8173.

[33] 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*. 1178–1186.

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

[35] Jiandong Wang, Hongguang Zhang, Chun Liu, and Xiongjun Yang. 2024. 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*. 2406–2410.

[36] 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*. 849–857. doi:10.1145/3219819.3219903

[37] Wenyan Xu, Dawei Xiang, Tianqi Ding, and Weihai Lu. 2025. MMM-Fact: A Multimodal, Multi-Domain Fact-Checking Dataset with Multi-Level Retrieval Difficulty. *arXiv preprint arXiv:2510.25120* (2025).

[38] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335* (2022).

[39] 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* (2025).

[40] Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, and Yi Xu. 2024. Facilitating multimodal classification via dynamically learning modality gap. *Advances in Neural Information Processing Systems* 37 (2024), 62108–62122. http://papers.nips.cc/paper_files/paper/2024/hash/71b17f00017da0d73823ccf7fbce2d4f-Abstract-Conference.html

[41] 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*. 5384–5392. doi:10.1609/AAAI.V37I4.25670

[42] 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*, Vol. 39. 986–994.

[43] Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Cheng Deng, Qinghua Hu, Cai Xu, Jie Wen, Di Hu, and Changqing Zhang. 2024. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947* (2024). https://arxiv.org/abs/2404.18947

[44] Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. 2023. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*. PMLR, 41753–41769. https://proceedings.mlr.press/v202/zhang23ar.html

[45] Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-Aware Multimodal 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* (Singapore, Singapore). Springer-Verlag, 354–367. doi:10.1007/978-3-030-47436-2_27

[46] 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*. IEEE, 2825–2830.

[47] 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*. 1793–1801.