EVIDENTIAL MULTI-VIEW UNCERTAINTY LEARNING FOR MULTIMODAL FAKE NEWS DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal fake news detection benefits from complementary textual and visual cues; however, existing methods often treat local text tokens and image patches as equally reliable. Consequently, intra-modal noise may propagate through cross-modal interaction and impair final decisions. To overcome this challenge, we present Evidential Multi-view Uncertainty Learning (EMUL), a unified framework that seamlessly integrates uncertainty as a continuous regulatory signal across representation learning, semantic alignment, and decision fusion. Rather than processing components in isolation, EMUL first estimates token- and patch-level intra-modal uncertainty to perform confidence-weighted aggregation for fine-grained noise suppression. These estimated confidences directly condition cross-modal semantic alignment, reducing the distribution gap between reliable local modality representations and global image-text semantics while down-weighting uncertain samples. Ultimately, at the decision stage, this uncertainty-awareness enables EMUL to jointly calibrate the text, image, and fused views using intra-modal and evidential uncertainty prior to multi-view evidence fusion. Extensive experiments on three public benchmarks demonstrate that EMUL achieves state-of-the-art performance.

1 INTRODUCTION

Social media has reshaped how news is produced, disseminated, and consumed Lazer et al. (2018). Compared with traditional news, social media news spreads faster, reaches broader audiences, and takes more diverse forms, enabling fake news to rapidly influence users' perceptions. False information can spread faster and more broadly than true information on social networks Vosoughi et al. (2018); Allcott & Gentzkow (2017). Since social media news combines text, images, and their semantic relations, a single modality cannot capture all cues required for veracity assessment. Multimodal fake news detection has thus become an important problem in social media computing, multimedia understanding, and trustworthy Artificial Intelligence Shu et al. (2017); Zhou & Zafarani (2018); Zafarani et al. (2019); Bondielli & Marcelloni (2019); Zubiaga et al. (2018).

Multimodal fake news detection has progressed from direct feature fusion to cross-modal interaction, multi-view learning, and pretrained vision-language modeling. Early studies learn event-invariant features, image-text similarity, or cross-modal ambiguity Wang et al. (2018); Zhou et al. (2020); Chen et al. (2022); Ying et al. (2023). More recent methods address event-level inconsistency and noisy external knowledge Ma et al. (2024); Zhang et al. (2024), domain variation and heterogeneous modality interactions Tong et al. (2024); Lu et al. (2025b); Liu et al. (2025), or introduce LLM-based supervision and rationale verification Hu et al. (2025); Chen et al. (2026a). Although these advances strengthen multimodal reasoning, they generally treat encoded local units as reliable inputs to subsequent interaction and fusion, leaving intra-modal token- and patch-level noise insufficiently controlled.

We observe that noise arises not only from image-text inconsistency but also within each modality, where it may be progressively amplified. Exaggerated or ambiguous expressions can weaken textual representations, while background regions, low-quality patches, and weakly related content can disrupt visual representations. As shown in Fig. 1, encoding all local units without reliability-aware weighting may mix unreliable information with key cues and disperse feature responses. This motivates estimating local reliability before cross-modal interaction and evidence fusion. We therefore

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

estimate token/patch-level intra-modal uncertainty, convert it into confidence, and weight feature aggregation to suppress uncertain information and emphasize veracity-related semantics.

Recent uncertainty-aware multi-view learning methods explicitly model view reliability and inter-view conflict to improve predictive robustness and uncertainty estimation Lu et al. (2025a); Duan et al. (2025). However, existing methods mainly estimate reliability at the modality view, leaving token- and patch-level uncertainty largely underexplored. Consequently, local noise may propagate through encoding and semantic alignment before later calibration can mitigate its influence.

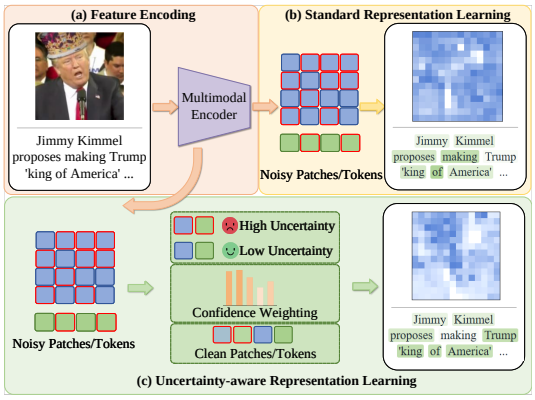

Figure 1: Motivation of intra-modal uncertainty-aware representation learning.

To address these challenges, we introduce Evidential Multi-view Uncertainty Learning (EMUL), a unified framework that seamlessly regulates representation learning, semantic alignment, and decision fusion with uncertainty. First, through temporal-context uncertainty estimation, EMUL estimates token- and patch-level intra-modal uncertainty and converts it into confidence scores for weighted modality aggregation. Guided by these confidences, uncertainty-guided semantic alignment aligns reliable local representations with global image-text semantics while down-weighting the contributions of uncertain samples. Ultimately, reliability-aware evidential fusion calibrates the text, image, and fused views using both intra-modal and evidential uncertainty prior to multi-view evidence fusion. Our main contributions are summarized as follows:

- We propose Evidential Multi-view Uncertainty Learning (EMUL), a framework that integrates uncertainty as a regulatory signal across representation learning, semantic alignment, and decision fusion, achieving state-of-the-art performance on public benchmarks.
- We develop fine-grained uncertainty estimation to quantify token- and patch-level intra-modal unreliability, enabling confidence-weighted aggregation to suppress noisy semantic units at their source.
- We introduce uncertainty-guided semantic alignment and reliability-aware evidential fusion, leveraging the estimated confidences to bridge cross-modal semantic discrepancies and calibrate multi-view evidence for robust decision-making.

2 Related Work

Multimodal Fake News Detection. Multimodal fake news detection has evolved from direct feature fusion to cross-modal interaction and multi-view reasoning. Early methods learn event-invariant representations or exploit image-text similarity, while subsequent studies model cross-modal ambiguity and complementary views Wang et al. (2018); Zhou et al. (2020); Chen et al. (2022); Ying et al. (2023). Recent work further addresses event-level inconsistency, noisy external knowledge, domain shifts, and heterogeneous modality interactions through Event-Radar, AKA-Fake, MMDFND, DAMMFND, and MIMoE-FND Ma et al. (2024); Zhang et al. (2024); Tong et al. (2024); Lu et al. (2025b); Liu et al. (2025). Reasoning-oriented approaches also incorporate LLM-generated supervision or verified rationales Hu et al. (2025); Chen et al. (2026a). Despite these advances, most methods operate on already encoded modality representations and rarely assess the reliability of individual tokens or patches before cross-modal interaction. Consequently, local noise may propagate into subsequent alignment and fusion stages.

Uncertainty-Aware Learning. Uncertainty modeling provides a principled way to handle noisy or conflicting evidence. Evidential deep learning represents class evidence using a Dirichlet distribution Sensoy et al. (2018), while trusted multi-view learning uses evidential uncertainty to calibrate view contributions Han et al. (2021); Lu et al. (2025a). In multimodal fake news detection, MMULN models intra- and inter-modal uncertainty, whereas Event-Radar and TMEF-BI estimate view credibility or perform trustworthy evidence fusion Wei et al. (2023); Ma et al. (2024); Chen et al.

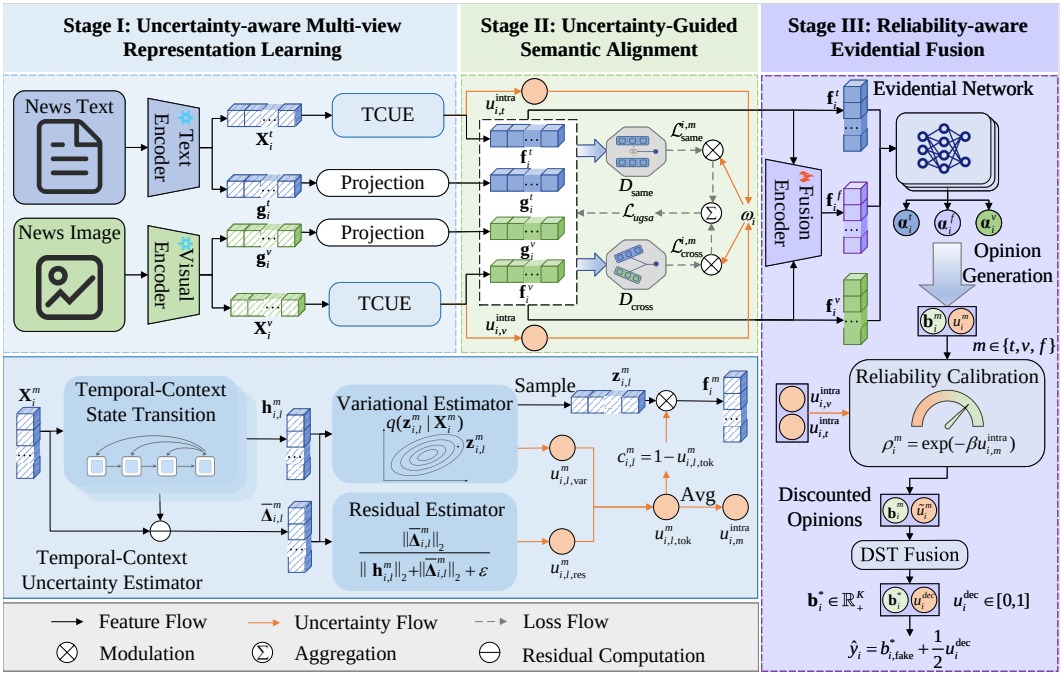

Figure 2: Overview of EMUL. Stage I estimates local uncertainty to learn reliable modality representations, Stage II performs uncertainty-guided alignment with CLIP semantics, and Stage III calibrates and fuses multi-view evidence.

(2025). However, existing methods mainly estimate uncertainty over global modality representations, view weights, or final predictions. Token- and patch-level reliability remains underexplored, and uncertainty is rarely used jointly to regulate representation learning, semantic alignment, and evidential fusion. EMUL addresses this gap by propagating intra-modal uncertainty through all three stages. Further discussions are provided in Appendix E.

3 METHODS

As shown in Fig. 2, EMUL integrates uncertainty regulation across seamlessly connected stages.

Stage I: Uncertainty-Aware Multi-view Representation Learning. The framework first estimates token- and patch-level reliability from contextual inconsistency and variational uncertainty, yielding reliable modality representations alongside sample-level intra-modal uncertainty.

Stage II: Uncertainty-Guided Semantic Alignment. Guided by these estimated confidences, same-modal and cross-modal discriminators align modality representations, while intra-modal uncertainty down-weights the contributions of unreliable samples to prevent distorted cross-modal mapping.

Stage III: Reliability-Aware Evidential Fusion. A fusion encoder constructs the fused view, and the framework calibrates the text, image, and fused-view opinions using both evidential and intra-modal uncertainty prior to Dempster-Shafer theory fusion, which produces the final prediction uncertainty.

Thus, uncertainty regulates representation learning, semantic alignment, and evidential fusion as a continuous guiding signal rather than serving merely as a post-hoc diagnostic.

3.1 UNCERTAINTY-AWARE MULTI-VIEW REPRESENTATION LEARNING

Let i index news instances and $m \in \{t, v\}$ denote text and vision, respectively. For modality m, L_m is the sequence length and $l \in \{1, \ldots, L_m\}$ indexes a token or patch. The encoders produce:

$$\mathbf{X}_i^t = [\mathbf{x}_{i,1}^t, \ldots, \mathbf{x}_{i,L_t}^t], \qquad \mathbf{X}_i^v = [\mathbf{x}_{i,1}^v, \ldots, \mathbf{x}_{i,L_v}^v], \tag{1}$$

where $\mathbf{x}_{i,l}^m$ is the local encoder output and $\mathbf{X}_i^m$ the complete sequence.

Ambiguous or irrelevant text and low-quality visual content may introduce noise into alignment and evidential fusion. We estimate local reliability to suppress low-confidence contributions. To capture both individual features and intra-modal context, the Temporal-Context Uncertainty Estimator (TCUE) applies branch-specific temporal-context transitions to the l-th local unit:

$$\tilde{\mathbf{x}}_{i,l}^{m,a} = \mathcal{T}_a\left(\mathbf{X}_{i,1:l}^m\right), \qquad \mathbf{h}_{i,l}^m = \mathrm{Agg}\left(\tilde{\mathbf{x}}_{i,l}^{m,Q}, \tilde{\mathbf{x}}_{i,l}^{m,K}, \tilde{\mathbf{x}}_{i,l}^{m,V}\right), \tag{2}$$

Here, $\mathbf{X}_{i,1:l}^m = [\mathbf{x}_{i,1}^m, \ldots, \mathbf{x}_{i,l}^m]$ denotes the feature sequence up to position l. The index $a \in \{Q, K, V\}$ identifies the query, key, and value branches. The branch-specific operator $\mathcal{T}_a(\cdot)$ produces the contextual state $\tilde{\mathbf{x}}_{i,l}^{m,a}$, and $\mathrm{Agg}(\cdot)$ combines the three branch outputs into the context-enhanced representation $\mathbf{h}_{i,l}^m$.

We compute the average dynamic residual between the original token representation and the branch-specific contextual states:

$$\bar{\mathbf{\Delta}}_{i,l}^m = \frac{1}{3} \sum_{a \in \{Q,K,V\}} \left[\mathbf{x}_{i,l}^m - \tilde{\mathbf{x}}_{i,l}^{m,a}\right]. \tag{3}$$

Here, $\mathcal{T}_a(\cdot)$ denotes the temporal-context state transition on branch a. A larger $\bar{\mathbf{\Delta}}_{i,l}^m$ indicates greater inconsistency with intra-modal context, which may reflect local noise, semantic shifts, or weakly relevant content.

The context-enhanced representation $\mathbf{h}_{i,l}^m$ and dynamic residual $\bar{\mathbf{\Delta}}_{i,l}^m$ are concatenated and passed to a variational estimation network, which predicts the mean $\boldsymbol{\mu}_{i,l}^m$ and standard deviation $\boldsymbol{\sigma}_{i,l}^m$ of a diagonal Gaussian distribution. Following variational autoencoding (Kingma & Welling, 2013), we model the latent local representation as:

$$q(\mathbf{z}_{i,l}^m | \mathbf{X}_i^m) = \mathcal{N}\left(\boldsymbol{\mu}_{i,l}^m, \mathrm{diag}\left((\boldsymbol{\sigma}_{i,l}^m)^2\right)\right). \tag{4}$$

During training, we sample $\mathbf{z}_{i,l}^m$ using the reparameterization trick, namely $\mathbf{z}_{i,l}^m = \boldsymbol{\mu}_{i,l}^m + \boldsymbol{\sigma}_{i,l}^m \odot \boldsymbol{\xi}_{i,l}^m$, where $\boldsymbol{\xi}_{i,l}^m \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\odot$ denotes element-wise multiplication. During inference, we use the mean $\boldsymbol{\mu}_{i,l}^m$ as the local representation.

We estimate token-level uncertainty from variational variance and contextual residual. The former measures representation instability:

$$u_{i,l,\mathrm{var}}^m = 1 - \exp\left(-\frac{1}{D}\sum_{d=1}^{D}(\sigma_{i,l,d}^m)^2\right), \tag{5}$$

where D is the latent dimension and d indexes its components. The latter measures inconsistency with intra-modal context:

$$u_{i,l,\mathrm{res}}^m = \frac{\|\bar{\mathbf{\Delta}}_{i,l}^m\|_2}{\|\mathbf{h}_{i,l}^m\|_2 + \|\bar{\mathbf{\Delta}}_{i,l}^m\|_2 + \varepsilon}. \tag{6}$$

Here, $\varepsilon > 0$ ensures numerical stability. Token-level intra-modal uncertainty is defined as:

$$u_{i,l,\mathrm{tok}}^m = \frac{u_{i,l,\mathrm{var}}^m + \lambda_r u_{i,l,\mathrm{res}}^m}{1 + \lambda_r}, \tag{7}$$

where λ_r controls the contribution of residual-based uncertainty. This combines uncertainty from representation variance and temporal-context residuals. We use $c_{i,l}^m = 1 - u_{i,l,\mathrm{tok}}^m$ as token confidence to aggregate a reliable modality representation:

$$\mathbf{f}_i^m = \frac{\sum_{l=1}^{L_m} c_{i,l}^m \mathbf{z}_{i,l}^m}{\sum_{l=1}^{L_m} c_{i,l}^m + \varepsilon}. \tag{8}$$

The sample-level intra-modal uncertainty $u_{i,m}^{\mathrm{intra}}$ is the sequence-wise average of token uncertainties and is used in subsequent alignment and decision stages. We further regularize the latent distributions in both modality branches toward a Gaussian prior:

$$\mathcal{L}_{\mathrm{var}} = \frac{1}{2N} \sum_{m \in \{t,v\}} \frac{1}{L_m} \sum_{i=1}^{N} \sum_{l=1}^{L_m} \mathrm{KL}\left[q(\mathbf{z}_{i,l}^m | \mathbf{X}_i^m) \| \mathcal{N}(\mathbf{0}, \mathbf{I})\right]. \tag{9}$$

Here, N denotes the number of instances over which the loss is evaluated. The regularizer averages the KL divergence across the two modalities, instances, and local units, and is incorporated into the overall training objective with weight λ_{var}.

3.2 Uncertainty-Guided Semantic Alignment

We introduce UGSA to reduce the gap between local representations and global text-image semantics. Following adversarial distribution alignment Goodfellow et al. (2020); Ganin et al. (2016), UGSA projects reliable modality representations into a shared space and aligns them with CLIP semantics through same-modal and cross-modal discriminators.

For sample i and modality $m \in \{t, v\}$, $\mathbf{f}_i^m$ denotes the reliable modality representation, $\mathbf{g}_i^m$ the CLIP global representation, and $\bar{m}$ the complementary modality. CLIP provides training-time alignment references, not direct inputs to the fused-view encoder. D_{same} distinguishes $\mathbf{f}_i^m$ from same-modal global semantics, while D_{cross} distinguishes it from complementary-modal semantics. Both estimate an input's probability of belonging to the corresponding global distribution. Appendix A.1 provides the discriminator losses and aggregate objective $\mathcal{L}_D$, with $\ell(\cdot, \cdot)$ denoting binary cross-entropy. Labels 1 and 0 denote global and local distributions, respectively. Conversely, the encoder aligns modality representations with global distributions at both levels:

$$\mathcal{L}_G^i = \sum_{m \in \{t,v\}} \left[\ell(D_{\text{same}}(\mathbf{f}_i^m), 1) + \ell(D_{\text{cross}}(\mathbf{f}_i^m), 1) \right]. \tag{10}$$

High-uncertainty samples may contain local noise or modality-inconsistent information, making strong alignment potentially detrimental to semantic-space optimization. We therefore define the alignment weight as

$$\begin{aligned} \omega_i &= \omega_0 + (1 - \omega_0) \prod_{q \in \{t,v\}} \left(1 - u_{i,q}^{\text{intra}}\right), \qquad \omega_i \in [\omega_0, 1], \\ \frac{\partial \omega_i}{\partial u_{i,q}^{\text{intra}}} &= -(1 - \omega_0) \prod_{\substack{r \in \{t,v\} \\ r \neq q}} \left(1 - u_{i,r}^{\text{intra}}\right) \leq 0. \end{aligned} \tag{11}$$

Here, ω_0 is the minimum alignment coefficient, preventing alignment supervision from vanishing for highly uncertain samples. The non-positive derivative shows that increasing the uncertainty of either modality cannot increase the corresponding sample's alignment contribution. Finally, the uncertainty-weighted UGSA objective is $\mathcal{L}_{\text{ugsa}} = \frac{1}{N} \sum_{i=1}^{N} \omega_i \mathcal{L}_G^i$.

3.3 Reliability-Aware Evidential Fusion

TCUE produces reliable text and image representations, $\mathbf{f}_i^t$ and $\mathbf{f}_i^v$, whose encoders are regularized by UGSA during training. A fused-view encoder constructs $\mathbf{f}_i^f = F_f(\mathbf{f}_i^t, \mathbf{f}_i^v)$ to capture complementary information. The resulting views provide complementary cues, but their evidence quality may vary across samples. We therefore introduce the Reliability-Aware Evidential Fusion (RAEF) module to calibrate each view's contribution using evidential and intra-modal uncertainty.

For each $m \in \{t, v, f\}$, an evidential network produces a subjective opinion $(\mathbf{b}_i^m, u_i^m)$, where $\mathbf{b}_i^m \in \mathbb{R}^K$ is the belief mass vector over K classes and u_i^m is the evidential uncertainty. The Dirichlet construction and EDL objective are detailed in Appendix A.2.

Evidential uncertainty alone cannot capture local reliability during representation learning. RAEF therefore calibrates each subjective opinion with intra-modal uncertainty. The text and image reliability coefficients are

$$\rho_i^t = \exp(-\beta u_{i,t}^{\text{intra}}), \qquad \rho_i^v = \exp(-\beta u_{i,v}^{\text{intra}}), \tag{12}$$

where β is a decay coefficient. Since the fused view is built from both modalities, its base reliability uses their average intra-modal uncertainty:

$$\rho_i^f = \exp\left(-\beta \frac{u_{i,t}^{\text{intra}} + u_{i,v}^{\text{intra}}}{2}\right). \tag{13}$$

Given these coefficients, RAEF discounts each subjective opinion to obtain calibrated belief and uncertainty masses:

$$\tilde{\mathbf{b}}_i^m = \rho_i^m \mathbf{b}_i^m, \qquad \tilde{u}_i^m = 1 - \rho_i^m + \rho_i^m u_i^m. \tag{14}$$

The discounting preserves normalization by transferring discounted belief mass to residual uncertainty, as shown in Appendix A.3. RAEF then sequentially applies the Dempster-Shafer Theory (DST) rule Dempster (2008); Shafer (2020); Yager (1987) to fuse the calibrated text, image, and fused-view opinions, yielding $(\mathbf{b}_i^*, u_i^{\mathrm{dec}})$. The pignistic transformation gives the probability:

$$\hat{y}_i = b_{i,\mathrm{fake}}^* + \frac{1}{K} u_i^{\mathrm{dec}}, \tag{15}$$

where u_i^{dec} measures evidential sufficiency and consistency after multi-view fusion.

3.4 Training Objective

The model is optimized end-to-end via a joint objective combining classification, auxiliary supervision, evidential learning, variational regularization, and semantic alignment. Specifically, $\mathcal{L}_{\mathrm{cls}}$ and $\mathcal{L}_{\mathrm{aux}}$ supervise fused and single-view predictions, while $\mathcal{L}_{\mathrm{edl}}$ optimizes Dirichlet evidence distributions and $\mathcal{L}_{\mathrm{var}}$ defined in Eq. 9 regularizes the token- and patch-level latent distributions. To align local modality representations with global semantic distributions, $\mathcal{L}_{\mathrm{ugsa}}$ is optimized alternately with the discriminator loss $\mathcal{L}_D$ (see Appendix A.4). The overall objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{aux}}\mathcal{L}_{\mathrm{aux}} + \lambda_{\mathrm{edl}}\mathcal{L}_{\mathrm{edl}} + \lambda_{\mathrm{var}}\mathcal{L}_{\mathrm{var}} + \lambda_{\mathrm{ugsa}}\mathcal{L}_{\mathrm{ugsa}}, \tag{16}$$

where $\mathcal{L}_{\mathrm{cls}} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{BCE}(\hat{y}_i, y_i)$ is the final classification loss, and $\mathcal{L}_{\mathrm{aux}}$ is the average auxiliary loss over text, image, and fused-view predictions. λ_{aux}, λ_{edl}, λ_{var}, and λ_{ugsa} weight their corresponding loss terms. The training procedure is summarized in Algorithm 1 in Appendix A.5.

4 Experiments

4.1 Experimental Settings

Datasets. This paper evaluates the proposed method on three public multimodal fake news detection datasets, namely WeiboJin et al. (2016), Weibo21Nan et al. (2021), and GossipCopShu et al. (2020). In the experiments, this paper retains only the samples that contain both text and images and constructs the training and test sets according to the data splits used in prior work. Dataset details are provided in Appendix B.1.

Baselines. We compare EMUL with representative multimodal fake news detection methods, including SAFE Zhou et al. (2020), CAFE Chen et al. (2022), BMR Ying et al. (2023), FND-CLIP Zhou et al. (2023), MMDFND Tong et al. (2024), MIMoE-FND Liu et al. (2025), and MaLSF Chen et al. (2026b). For methods with publicly available implementations or pretrained models, we follow their released codebases or reported settings as closely as possible. Further baseline descriptions are provided in Appendix B.2.

Implementation Details. For text encoding, we use chinese-roberta-wwm-ext on the Chinese datasets Cui et al. (2020) and bert-base-uncased on GossipCop Devlin et al. (2019). For image encoding, we adopt the pretrained MAE ViT-Base Patch16 He et al. (2022); Dosovitskiy et al. (2020) as the visual encoder to extract patch-level image sequence features. For global cross-modal semantic modeling, we use CLIP to extract global text-image representations, which serve as semantic references for the uncertainty-guided semantic alignment module. We report Accuracy, Precision, Recall, and F1 as the main evaluation metrics. More details are provided in Appendix B.3.

4.2 Overall Performance

Table 1 compares EMUL with representative multimodal fake news detection methods on three public datasets. EMUL achieves the best results on Weibo and Weibo-21, and performs competitively on GossipCop. The results demonstrate that EMUL maintains effective detection performance across datasets with different languages and data characteristics.

Table 1: Performance metrics for different classification methods on three datasets.

Dataset	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F1-score	Precision	Recall	F1-score
Weibo	SAFE	0.762	0.831	0.724	0.774	0.695	0.811	0.748
	CAFE	0.840	0.855	0.830	0.842	0.825	0.851	0.837
	BMR	0.918	0.882	**0.948**	0.914	**0.942**	0.879	0.904
	FND-CLIP	0.907	0.914	0.901	0.908	0.914	0.901	0.907
	MMDFND	0.916	0.899	0.920	0.910	0.930	0.912	0.921
	MIMoE-FND	0.928	0.942	0.913	0.928	0.913	0.942	0.927
	MaLSF	0.935	0.948	0.917	0.932	0.924	0.953	0.938
	Ours	**0.950**	**0.960**	0.943	**0.951**	0.940	**0.958**	**0.949**
GossipCop	SAFE	0.838	0.758	0.558	0.643	0.857	0.937	0.895
	CAFE	0.867	0.732	0.490	0.587	0.887	0.957	0.921
	BMR	**0.895**	0.752	0.639	0.691	**0.920**	0.965	0.936
	FND-CLIP	0.880	0.761	0.549	0.638	0.899	0.959	0.928
	MMDFND	0.867	0.657	**0.646**	0.651	0.916	0.919	0.918
	MIMoE-FND	**0.895**	0.762	0.644	**0.698**	**0.920**	0.953	0.936
	Ours	**0.895**	**0.825**	0.578	0.680	0.906	**0.971**	**0.937**
Weibo-21	SAFE	0.905	0.893	0.908	0.901	0.916	0.901	0.890
	CAFE	0.882	0.857	0.915	0.885	0.907	0.844	0.876
	BMR	0.929	0.908	0.947	0.927	0.946	0.906	0.925
	FND-CLIP	0.943	0.935	0.945	0.940	0.950	0.942	0.946
	MMDFND	0.950	0.948	0.951	0.949	0.951	0.948	0.950
	MIMoE-FND	0.956	0.953	0.957	0.955	0.959	0.956	0.957
	MaLSF	0.955	0.952	0.958	0.955	0.957	0.951	0.954
	Ours	**0.963**	**0.964**	**0.961**	**0.963**	**0.961**	**0.964**	**0.962**

EMUL shows particularly clear advantages on the two Chinese datasets. Compared with existing strong baselines, it consistently improves accuracy and the F1 scores of both real and fake news classes. On Weibo-21, where most baselines already achieve high performance, EMUL further improves all reported metrics, indicating that its gains are retained in a relatively saturated setting.

On GossipCop, EMUL achieves results comparable to the strongest baseline and obtains the best values on several metrics. This outcome further supports the effectiveness of the proposed framework on English news data. The smaller margin relative to the Chinese datasets also suggests that GossipCop remains a challenging benchmark for multimodal fake news detection.

Overall, the results show that EMUL consistently matches or surpasses existing methods across the three datasets. The individual contributions of uncertainty-aware representation learning, semantic alignment, and reliability-aware evidential fusion are further examined in the ablation study.

4.3 Ablation Study

To evaluate the contributions of the proposed designs, we remove or replace one component from EMUL at a time. Table 2 reports the results. Overall, the full model achieves the best performance across all three datasets, showing that TCUE, UGSA, and RAEF provide complementary benefits.

Table 2: Ablation results of the three core components.

Dataset	Method	Accuracy	F1(Fake)	F1(Real)
Weibo	w/o TCUE	0.948	0.950	0.946
	w/o UGSA	0.941	0.942	0.939
	w/o RAEF	0.948	0.950	0.947
	Ours	**0.950**	**0.951**	**0.949**
GossipCop	w/o TCUE	0.883	0.619	0.931
	w/o UGSA	0.878	0.636	0.927
	w/o RAEF	0.884	0.655	0.930
	Ours	**0.895**	**0.680**	**0.937**
Weibo-21	w/o TCUE	0.948	0.948	0.948
	w/o UGSA	0.945	0.944	0.945
	w/o RAEF	0.950	0.949	0.950
	Ours	**0.963**	**0.963**	**0.962**

Effect of TCUE. Removing TCUE consistently degrades performance, with the largest drop on GossipCop, where Fake F1 decreases from 0.680 to

0.619. This supports the value of token- and patch-level intra-modal uncertainty modeling for learning reliable modality representations.

Effect of UGSA. Removing UGSA leads to performance drops, with accuracy decreasing to 0.941 on Weibo and 0.945 on Weibo-21. This confirms the benefit of uncertainty-guided semantic alignment for multimodal detection.

Effect of RAEF. We replace the evidential fusion component in RAEF with ordinary decision fusion. This replacement reduces performance on all datasets, demonstrating the effectiveness of evidence-level multi-view fusion.

Effects of Auxiliary Designs. To further examine the two auxiliary designs, Figure 3 reports performance drops in Accuracy, F1(Fake), and F1(Real) on Weibo and Weibo-21 relative to the full model. Removing either reliability discounting or variational regularization degrades all three metrics. On Weibo, removing variational regularization causes drops of approximately 0.3 percentage points across all metrics, compared with approximately 0.2 points after removing reliability discounting. On Weibo-21, reliability discounting has a larger effect: its removal reduces Accuracy, F1(Fake), and F1(Real) by 0.3, 0.4, and 0.3 points, respectively, whereas removing variational regularization results in drops of approximately 0.2 points. These results indicate that the two designs provide complementary, dataset-dependent benefits. The corresponding results on GossipCop are provided in Appendix D.1.

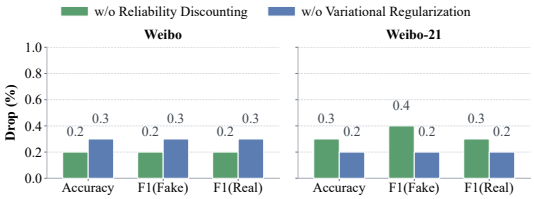

Figure 3: Additional ablations on Weibo and Weibo-21. Bars show drops in Accuracy, F1(Fake), and F1(Real) from the full model without reliability discounting or variational regularization. Larger values indicate greater degradation.

Overall, the ablation results verify that TCUE improves modality representation reliability, UGSA facilitates semantic alignment, and RAEF enhances reliable multi-view evidence fusion.

4.4 Robustness to Local Corruption

We examine robustness to local corruption on Weibo at inference. For text, we randomly replace 10%, 20%, or 30% of non-special tokens with ordinary tokens sampled from the vocabulary. For images, we add Gaussian noise with standard deviation 0.2 to the same proportions of 16×16 patches. Perturbations are synchronized between the base encoders and CLIP, and Full and w/o TCUE use identical corruption masks.

Table 3: Accuracy under local corruption on Weibo.

Type	Ratio	w/o TCUE	Full
Clean	–	0.948	**0.950**
Text	10%	0.926	**0.928**
	20%	0.914	**0.920**
	30%	0.885	**0.898**
Image	10%	0.938	**0.948**
	20%	0.935	**0.946**
	30%	0.933	**0.943**

As shown in Table 3, Full maintains higher accuracy than w/o TCUE across all corruption settings. At 30% text corruption, accuracy improves from 0.885 to 0.898; under 30% image corruption, it improves from 0.933 to 0.943. These results support the benefit of TCUE when local textual or visual evidence is corrupted. The gains are consistent with confidence-weighted aggregation, which reduces the contribution of unreliable tokens and patches, limiting their influence on subsequent alignment and fusion. Complete results on all three datasets are provided in Appendix C.2.

4.5 Visualization Analysis

To examine the learned representation space, we use t-SNE Van der Maaten & Hinton (2008) to visualize features on Weibo under three variants: w/o TCUE, w/o UGSA, and the full model. As shown in Figure 4, the full model forms coherent class-dependent regions with a clear separation between real and fake news. Removing TCUE introduces more locally mixed samples, suggesting that unreliable tokens or patches affect the aggregated representations. Without UGSA, the feature

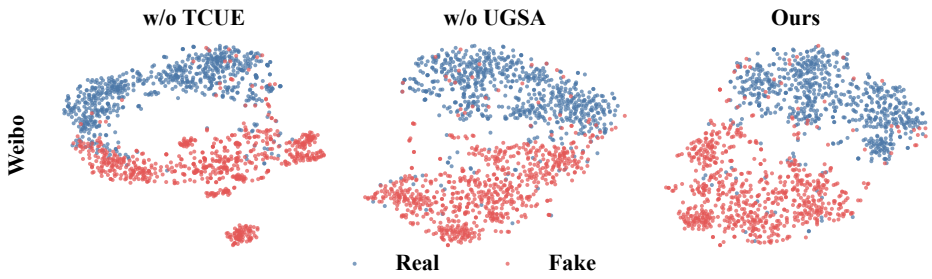

Figure 4: t-SNE visualization on Weibo for w/o TCUE, w/o UGSA, and the full model. Blue and red points denote real and fake news samples, respectively.

distribution becomes more diffuse and the class boundary weakens, indicating the benefit of aligning reliable modality representations with global semantics. These observations illustrate the complementary roles of TCUE and UGSA in representation learning. TCUE regulates local feature contributions, while UGSA provides global semantic guidance for the resulting representations. Their combination can help preserve class-relevant information while limiting the influence of unreliable local evidence. Visualizations on GossipCop and Weibo-21 are provided in Appendix D.2.

4.6 Uncertainty Analysis

To assess whether the estimated decision uncertainty reflects prediction risk, we conduct selective prediction Geifman & El-Yaniv (2017). Test samples are ranked by uncertainty, and those with the highest uncertainty are progressively rejected. We report accuracy on the remaining samples at different rejection ratios.

As shown in Figure 5, accuracy exhibits an overall upward trend as high-uncertainty samples are removed. On Weibo-21, it increases from 0.963 to a peak of approximately 0.981, followed by a slight decrease at the highest rejection ratio, while remaining substantially above the original accuracy. Accuracy also increases from 0.950 to 0.987 on Weibo and from 0.895 to 0.935 on GossipCop. These results indicate that the uncertainty estimated by EMUL generally assigns higher risk to error-prone samples and therefore provides a meaningful signal for selective prediction rather than merely serving as an auxiliary output.

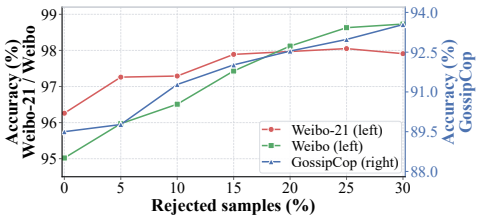

Figure 5: Selective prediction results on three datasets. Accuracy is evaluated after rejecting increasing proportions of high-uncertainty samples.

Additional local-corruption results are provided in Appendix C.2. Further analyses of uncertainty-weighted semantic alignment and decision calibration are provided in Appendix C. Parameter sensitivity and CLIP backbone comparisons are reported in Appendix D.

5 Conclusion

This paper presented EMUL for multimodal fake news detection, where uncertainty serves as a continuous regulatory signal across representation learning, semantic alignment, and decision fusion. Specifically, EMUL first estimates token- and patch-level intra-modal uncertainty to suppress unreliable local information during feature aggregation. Guided by these estimated confidences, it aligns reliable local modality representations with global image-text semantics while down-weighting uncertain samples. Ultimately, at the decision stage, EMUL leverages both intra-modal and evidential uncertainty to jointly calibrate the text, image, and fused views prior to multi-view evidence fusion. Extensive experiments demonstrate that EMUL achieves state-of-the-art performance on public benchmarks.

AI USE STATEMENT

We used generative AI tools to improve the grammar, wording, and readability of this manuscript. We also used these tools for a pre-submission check of references, formatting, and consistency among the text, tables, and figures. Generative AI was not used to generate or alter data, conduct experiments, or make the final scientific judgments reported in this paper. The authors reviewed all AI-assisted suggestions incorporated into the submitted manuscript and take full responsibility for its methods, results, claims, citations, and text.

REFERENCES

Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236, 2017.

Alessandro Bondielli and Francesco Marcelloni. A survey on fake news and rumour detection techniques. *Information sciences*, 497:38–55, 2019.

Junyang Chen, Yueqian Li, Ka Chung Ng, Huan Wang, and Liang-Jie Zhang. Toward multimodal fake news detection by multi-perspective rationale generation and verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 66–74, 2026a.

Xuelong Chen, Yiping Chen, Yang Liu, and Jinchao Pan. Tmef-bi: Trusted multimodal evidential fusion considering behavior information for fake news detection. *IEEE Transactions on Computational Social Systems*, 12(5):2901–2914, 2025.

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

Zizhao Chen, Ping Wei, Ziyang Ren, Huan Li, and Xiangru Yin. Bridging pixels and words: Mask-aware local semantic fusion for multimodal media verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26561–26571, 2026b.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 657–668, 2020.

Arthur P Dempster. Upper and lower probabilities induced by a multivalued mapping. In *Classic works of the Dempster-Shafer theory of belief functions*, pp. 57–72. Springer, 2008.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pp. 4171–4186, 2019.

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

Siyuan Duan, Yuan Sun, Dezhong Peng, Guiduo Duan, Xi Peng, and Peng Hu. Deep fuzzy multi-view learning for reliable classification. In *Forty-second International Conference on Machine Learning*, 2025.

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

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. *arXiv preprint arXiv:2102.02051*, 2021.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Shuguo Hu, Jun Hu, and Huaiwen Zhang. Synergizing llms with global label propagation for multimodal fake news detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1426–1440, 2025.

Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia*, 19(3):598–608, 2016.

Audun Jsang. *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated, 2018.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, pp. 2915–2921, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

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

Jueqing Lu, Wray Buntine, Yuanyuan Qi, Joanna Dipnall, Belinda Gabbe, and Lan Du. Navigating conflicting views: Harnessing trust for learning. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*. PMLR, 2025a.

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

Zihan Ma, Minnan Luo, Hao Guo, Zhi Zeng, Yiran Hao, and Xiang Zhao. Event-radar: Event-driven multi-view learning for multimodal fake news detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5809–5821, 2024.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

Glenn Shafer. A mathematical theory of evidence. 2020.

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

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 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, 2020.

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

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

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

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

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pp. 849–857, 2018.

Lingwei Wei, Dou Hu, Wei Zhou, and Songlin Hu. Modeling both intra-and inter-modality uncertainty for multimodal fake news detection. *IEEE Transactions on Multimedia*, 25:7906–7916, 2023.

Ronald R Yager. On the dempster-shafer framework and new combination rules. *Information sciences*, 41(2):93–137, 1987.

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

Reza Zafarani, Xinyi Zhou, Kai Shu, and Huan Liu. Fake news research: Theories, detection strategies, and open problems. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 3207–3208, 2019.

Litian Zhang, Xiaoming Zhang, Ziyi Zhou, Feiran Huang, and Chaozhuo Li. Reinforced adaptive knowledge learning for multimodal fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 16777–16785, 2024.

Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *arXiv preprint arXiv:1812.00315*, 2018.

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

Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. Multimodal fake news detection via clip-guided learning. In *2023 IEEE international conference on multimedia and expo (ICME)*, pp. 2825–2830. IEEE, 2023.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. *Acm Computing Surveys (Csur)*, 51(2):1–36, 2018.

A Method and Optimization Details

A.1 Discriminator Objectives

For modality m, the same-modal and cross-modal alignment losses are

$$\mathcal{L}^{i,m}_{\text{same}} = \ell(D_{\text{same}}(\mathbf{g}^m_i), 1) + \ell(D_{\text{same}}(\mathbf{f}^m_i), 0), \tag{17}$$

and

$$\mathcal{L}^{i,m}_{\text{cross}} = \ell(D_{\text{cross}}(\mathbf{g}^{\bar{m}}_i), 1) + \ell(D_{\text{cross}}(\mathbf{f}^m_i), 0), \tag{18}$$

where $\ell(\cdot,\cdot)$ is the binary cross-entropy loss. Label 1 denotes the global semantic distribution, and label 0 denotes the local modality representation distribution. The discriminator objective is

$$\mathcal{L}_D = \frac{1}{N}\sum_{i=1}^{N}\sum_{m\in\{t,v\}}\left(\mathcal{L}^{i,m}_{\text{same}} + \mathcal{L}^{i,m}_{\text{cross}}\right). \tag{19}$$

A.2 Evidential Opinion Construction

For each view $m \in \{t, v, f\}$, an evidential network maps $\mathbf{f}^m_i$ to non-negative evidence and converts it into Dirichlet parameters:

$$\boldsymbol{\alpha}^m_i = E_m(\mathbf{f}^m_i) + \mathbf{1}. \tag{20}$$

Let K denote the number of classes and $S^m_i = \sum_{k=1}^{K} \alpha^m_{i,k}$ the total evidence strength. Under subjective logic Jsang (2018), the belief and uncertainty masses are

$$b^m_{i,k} = \frac{\alpha^m_{i,k} - 1}{S^m_i}, \qquad u^m_i = \frac{K}{S^m_i}. \tag{21}$$

We denote the belief mass vector as $\mathbf{b}^m_i = [b^m_{i,1}, \ldots, b^m_{i,K}]$. A smaller S^m_i yields a larger u^m_i, indicating insufficient discriminative evidence. We optimize each evidential distribution with the standard evidential deep learning (EDL) loss Sensoy et al. (2018), comprising a Bayes risk term, a Dirichlet variance term, and an annealed KL regularizer. This loss promotes evidence for correct classes while suppressing overconfident incorrect predictions.

A.3 Opinion Discounting Property

The subjective opinion in Eq. 21 satisfies $\sum_{k=1}^{K} b^m_{i,k} + u^m_i = 1$. Applying the RAEF discounting rule in Eq. 14 gives

$$\sum_{k=1}^{K} \tilde{b}^m_{i,k} + \tilde{u}^m_i = 1, \qquad \tilde{u}^m_i - u^m_i = (1-\rho^m_i)(1-u^m_i) \geq 0. \tag{22}$$

Thus, reliability discounting preserves total opinion mass and reallocates the discounted belief mass to residual uncertainty.

A.4 Alternating Optimization

For UGSA, we alternately optimize the discriminator objective $\mathcal{L}_D$ and encoder-side weighted alignment objective $\mathcal{L}_{\text{ugsa}}$. $\mathcal{L}_D$ updates only the same-modal and cross-modal discriminators, whereas $\mathcal{L}_{\text{ugsa}}$ is included in the overall objective to update the encoder and its upstream representation modules.

A.5 End-to-End Training Procedure

B Experimental Settings

B.1 Datasets

This paper evaluates the proposed method on three public multimodal fake news detection datasets, namely WeiboJin et al. (2016), Weibo21Nan et al. (2021), and GossipCopShu et al. (2020). Weibo

Algorithm 1 End-to-end training procedure of EMUL.

Require: Training set $\mathcal{D}$; modality encoders; CLIP; TCUE; discriminators D_{same} and D_{cross}; fusion encoder F_f; evidential networks $\{E_m\}_{m\in\{t,v,f\}}$
Ensure: Optimized model parameters

for each mini-batch $\mathcal{B} \subset \mathcal{D}$ **do**
  for each $(x_i^t, x_i^v, y_i) \in \mathcal{B}$ **do**
    Encode text and image into local sequences $\mathbf{X}_i^t$ and $\mathbf{X}_i^v$
    Extract CLIP global semantics $\mathbf{g}_i^t$ and $\mathbf{g}_i^v$
    for $m \in \{t, v\}$ **do**
      $(\mathbf{f}_i^m, u_{i,m}^{\text{intra}}) \leftarrow \text{TCUE}(\mathbf{X}_i^m)$
    end for
    Project $\mathbf{f}_i^m$ and $\mathbf{g}_i^m$ into the shared semantic space
  end for
  Compute the discriminator objective $\mathcal{L}_D$ on $\mathcal{B}$
  Update only D_{same} and D_{cross} by minimizing $\mathcal{L}_D$
  for each $(x_i^t, x_i^v, y_i) \in \mathcal{B}$ **do**
    Compute the uncertainty-guided alignment weight
    $\omega_i \leftarrow \omega_0 + (1-\omega_0)(1-u_{i,t}^{\text{intra}})(1-u_{i,v}^{\text{intra}})$
    Compute the encoder-side alignment loss $\mathcal{L}_G^i$
    Construct the fused view $\mathbf{f}_i^f \leftarrow F_f(\mathbf{f}_i^t, \mathbf{f}_i^v)$
    for $m \in \{t, v, f\}$ **do**
      Generate Dirichlet parameters $\boldsymbol{\alpha}_i^m \leftarrow E_m(\mathbf{f}_i^m) + \mathbf{1}$
      Obtain subjective opinion $(\mathbf{b}_i^m, u_i^m)$
      Compute reliability ρ_i^m from intra-modal uncertainty
      Discount the opinion: $(\tilde{\mathbf{b}}_i^m, \tilde{u}_i^m) \leftarrow \textsc{Discount}(\mathbf{b}_i^m, u_i^m, \rho_i^m)$
    end for
    Fuse the calibrated opinions using DST:
    $(\mathbf{b}_i^*, u_i^{\text{dec}}) \leftarrow \text{DSTFuse}((\tilde{\mathbf{b}}_i^t, \tilde{u}_i^t), (\tilde{\mathbf{b}}_i^v, \tilde{u}_i^v), (\tilde{\mathbf{b}}_i^f, \tilde{u}_i^f))$
    Obtain prediction $\hat{y}_i \leftarrow b_{i,\text{fake}}^* + u_i^{\text{dec}}/K$
  end for
  Compute $\mathcal{L}_{\text{ugsa}} \leftarrow |\mathcal{B}|^{-1} \sum_{i\in\mathcal{B}} \omega_i \mathcal{L}_G^i$
  Compute the overall objective $\mathcal{L}$
  Update all non-discriminator parameters by minimizing $\mathcal{L}$
end for

is one of the most widely used multimodal fake news detection datasets in the Chinese social media setting. A commonly used version contains 9,528 image-text pairs, including 4,749 fake news samples and 4,779 real news samples, with a relatively balanced class distribution. Weibo21 is a multi-domain Chinese fake news detection dataset whose samples are also collected from Sina Weibo. It contains 9,128 news instances, including 4,640 real news samples and 4,488 fake news samples, and provides text, images, timestamps, comments, and domain labels. GossipCop is derived from FakeNewsNet and is collected from fact-checking websites. The original dataset contains 16,817 real news samples and 5,323 fake news samples, showing a clear class imbalance, and it includes news articles, headlines, images, and social context information. In the experiments, this paper retains only the samples that contain both text and images and constructs the training and test sets according to the data splits used in prior work.

B.2 Baselines

We compare EMUL with representative multimodal fake news detection methods, including SAFE Zhou et al. (2020), CAFE Chen et al. (2022), BMR Ying et al. (2023), FND-CLIP Zhou et al. (2023), MMDFND Tong et al. (2024), MIMoE-FND Liu et al. (2025), and MaLSF Chen et al. (2026b). These baselines cover cross-modal similarity and ambiguity modeling, multi-view representation learning, vision-language semantic guidance, multi-domain detection, expert-based modality interaction, and local semantic fusion. For methods with publicly available implemen-

Table 4: Accuracy under local corruption on GossipCop and Weibo-21. Ratios indicate the proportions of perturbed tokens or patches.

Dataset	Variant	Clean	Text corruption			Image corruption		
			10%	20%	30%	10%	20%	30%
GossipCop	w/o TCUE	0.883	0.866	0.855	0.850	0.870	0.867	0.865
	Full	**0.895**	**0.871**	**0.860**	**0.852**	**0.878**	**0.875**	**0.873**
Weibo-21	w/o TCUE	0.948	0.935	0.932	0.904	0.937	0.929	0.925
	Full	**0.963**	**0.942**	**0.935**	**0.924**	**0.951**	**0.950**	**0.943**

tations or pretrained models, we follow their released codebases or reported settings as closely as possible.

B.3 Implementation Details

For text encoding, we use chinese-roberta-wwm-ext on the Chinese datasets Cui et al. (2020) and bert-base-uncased on GossipCop Devlin et al. (2019). For image encoding, we adopt the pretrained MAE ViT-Base Patch16 He et al. (2022); Dosovitskiy et al. (2020) as the visual encoder to extract patch-level image sequence features. For global cross-modal semantic modeling, we use CLIP to extract global text-image representations, which serve as semantic references for the uncertainty-guided semantic alignment module. The maximum text length is set to 197 tokens, the batch size is set to 64, and Adam Kingma & Ba (2014) is used as the optimizer with a weight decay of 5×10^{-5}. During training, early stopping is applied to prevent overfitting. In the uncertainty-guided semantic alignment module, we use a warm-up strategy to delay the alignment process, reducing the interference from unstable representations at the early training stage. We report Accuracy, Precision, Recall, and F1 as the main evaluation metrics. The model is implemented in PyTorch Paszke et al. (2019) and trained on a single NVIDIA RTX 4090 GPU.

C Mechanism-Level Analysis of Uncertainty Regulation

C.1 Experimental Protocol

The component ablations in the main paper evaluate the contributions of TCUE, UGSA, and RAEF. We further examine their uncertainty-regulation mechanisms: robustness to local corruption, gains from uncertainty-weighted semantic alignment, and decision calibration and error awareness. The Weibo corruption results are presented in Section 4.4, with additional datasets reported below. All variants use the same data splits and evaluation protocol, and classification metrics are computed from the final prediction head.

For local-corruption experiments, we report accuracy on clean and corrupted inputs. For decision-level analysis, we report Expected Calibration Error (ECE) with 15 equal-width bins, Brier score, negative log-likelihood (NLL), error-detection AUROC and AUPR, and the area under the risk–coverage curve (AURC). Lower ECE, Brier, NLL, and AURC are better, whereas higher error-detection AUROC and AUPR are better. Ordinary Fusion uses predictive entropy as its risk score; the evidential variants use u_i^{dec}.

C.2 Additional Results on Local Corruption

We extend the Weibo analysis in Section 4.4 to GossipCop and Weibo-21. Both datasets use the same corruption procedure and comparison protocol described in the main paper. Full and w/o TCUE receive identical corruption masks within each dataset.

Table 4 shows that Full maintains higher accuracy than w/o TCUE across all tested corruption types and ratios. At 30% text corruption, the gains are 0.2 and 2.0 percentage points on GossipCop and Weibo-21, respectively. Under 30% image corruption, the corresponding gains are 0.8 and

1.8 points. Together with the Weibo results in the main paper, these findings show a consistent performance advantage for TCUE under local corruption across the three datasets.

C.3 Effect of Uncertainty-Weighted Alignment

We separate the effect of semantic alignment from that of uncertainty weighting by comparing three settings. The w/o UGSA variant disables semantic alignment, Uniform UGSA retains alignment but fixes $\omega_i = 1$, and Full UGSA uses the sample-dependent weight.

Table 5: Effect of semantic alignment and uncertainty weighting.

Dataset	Variant	Accuracy	F1(fake)	F1(real)
Weibo	w/o UGSA	0.941	0.942	0.939
	Uniform UGSA	0.943	0.945	0.942
	Full UGSA	**0.950**	**0.951**	**0.949**
GossipCop	w/o UGSA	0.878	0.636	0.927
	Uniform UGSA	0.883	0.642	0.930
	Full UGSA	**0.895**	**0.680**	**0.937**
Weibo-21	w/o UGSA	0.945	0.944	0.945
	Uniform UGSA	0.946	0.946	0.947
	Full UGSA	**0.963**	**0.963**	**0.962**

Uniform UGSA consistently improves over w/o UGSA, showing that aligning reliable modality representations with global semantics is beneficial even without uncertainty weighting. Full UGSA further improves accuracy over Uniform UGSA by 0.7, 1.2, and 1.7 percentage points on Weibo, GossipCop, and Weibo-21, respectively. The gains are also observed for both class-specific F1 scores, with the largest improvement appearing in GossipCop F1(fake), which increases from 0.642 to 0.680. These results separate the contribution of semantic alignment from that of uncertainty weighting and show that adaptive sample weighting provides additional gains across all three datasets.

C.4 Calibration and Error Awareness of RAEF

We compare Ordinary Fusion, Evidential Fusion without reliability discounting, and Full RAEF. Ordinary Fusion uses standard probabilistic outputs. Evidential w/o Discount retains evidential opinions and DST fusion but sets every view reliability coefficient to one. Full RAEF combines evidential fusion with intra-modal-uncertainty-based reliability discounting.

Table 6: Calibration and error-awareness results. Lower ECE, Brier, NLL, and AURC are better; higher error-detection AUROC and AUPR are better.

Dataset	Variant	Risk score	ECE↓	Brier↓	NLL↓	Err-AUROC↑	Err-AUPR↑	AURC↓
Weibo	Ordinary Fusion	Entropy	0.0394	0.0497	0.2431	0.8206	0.1633	0.0126
	Evidential w/o Discount	u_i^{dec}	0.0391	0.0520	**0.2066**	0.8482	**0.2543**	0.0144
	Full RAEF	u_i^{dec}	**0.0380**	**0.0480**	0.2312	**0.8547**	0.2414	**0.0110**
GossipCop	Ordinary Fusion	Entropy	**0.0903**	**0.1036**	0.6578	0.7232	0.2643	0.0586
	Evidential w/o Discount	u_i^{dec}	0.1003	0.1122	**0.5233**	0.7586	**0.2900**	0.0478
	Full RAEF	u_i^{dec}	0.1239	0.1306	0.5590	**0.7706**	0.2763	**0.0458**
Weibo-21	Ordinary Fusion	Entropy	0.0450	0.0485	0.2768	0.7750	0.2087	0.0177
	Evidential w/o Discount	u_i^{dec}	0.0451	0.0497	0.2662	**0.8193**	0.2126	0.0150
	Full RAEF	u_i^{dec}	**0.0350**	**0.0461**	**0.2129**	0.8184	**0.2322**	**0.0133**

On Weibo, Full RAEF achieves the lowest ECE, Brier score, and AURC and the highest error-detection AUROC. Relative to Evidential w/o Discount, reliability discounting reduces ECE from 0.0391 to 0.0380, Brier score from 0.0520 to 0.0480, and AURC from 0.0144 to 0.0110, although the undiscounted variant achieves lower NLL and higher error-detection AUPR.

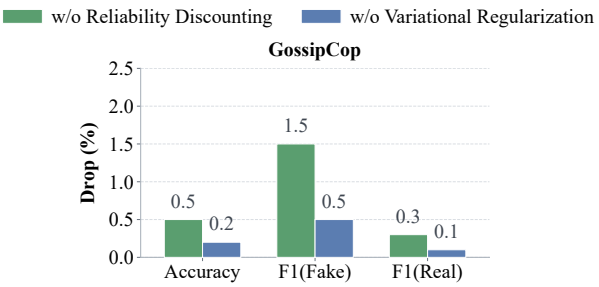

Figure 6: Additional ablation analysis on GossipCop. Bars show performance drops relative to the full model after removing reliability discounting or variational regularization.

On GossipCop, Full RAEF achieves the highest error-detection AUROC and the lowest AURC, whereas Ordinary Fusion yields the lowest ECE and Brier score and Evidential w/o Discount obtains the lowest NLL and highest error-detection AUPR. On Weibo-21, Full RAEF provides the strongest overall calibration, obtaining the lowest ECE, Brier score, NLL, and AURC and the highest error-detection AUPR; its AUROC is comparable to the best result, differing by only 0.0009. These results show that RAEF improves the ranking of risky predictions across datasets, while its effect on probability calibration depends on the dataset and metric.

D Additional Experimental Results

D.1 Additional Ablation Analysis on GossipCop

Figure 6 reports the corresponding auxiliary ablation results on GossipCop. Removing reliability discounting decreases Accuracy, F1(Fake), and F1(Real) by 0.5, 1.5, and 0.3 percentage points, respectively. Removing variational regularization produces corresponding drops of 0.2, 0.5, and 0.1 points. F1(Fake) exhibits the largest degradation under both settings, suggesting that these designs are particularly beneficial for detecting fake-news samples on GossipCop, with reliability discounting making the stronger contribution.

D.2 Additional Visualization Analysis

Figure 7 presents the corresponding t-SNE visualizations on GossipCop and Weibo-21. On GossipCop, all variants retain some class overlap, reflecting the more challenging semantic structure of this dataset. Nevertheless, the full model organizes samples into more coherent class-dominant regions. On Weibo-21, the full model shows a clearer separation between real and fake news than the variants without TCUE or UGSA. Together with the Weibo results in Figure 4, these visualizations consistently show that intra-modal uncertainty modeling and semantic alignment improve the organization of the learned representation space.

D.3 Hyperparameter Sensitivity Analysis

We analyze the sensitivity of EMUL to five hyperparameters. The reliability mapping parameter β controls the conversion from intra-modal uncertainty to view reliability, while λ_{var} and λ_{ugsa} weight the variational regularization and uncertainty-guided semantic alignment objectives, respectively. In addition, λ_{aux} controls view-level auxiliary supervision, and λ_{edl} weights the evidential learning objective. We vary one parameter at a time while keeping all others fixed and report accuracy on Weibo-21, Weibo, and GossipCop.

Module and Regularization Parameters. Figure 8 presents the results for β, λ_{var}, and λ_{ugsa}. For β, performance generally improves before slightly decreasing or fluctuating, indicating that moderate reliability discounting suppresses uncertain views without discarding useful modality information. A small positive λ_{var} generally performs well by regularizing token-level latent distributions, whereas excessive regularization may interfere with discriminative representation learning. Simi-

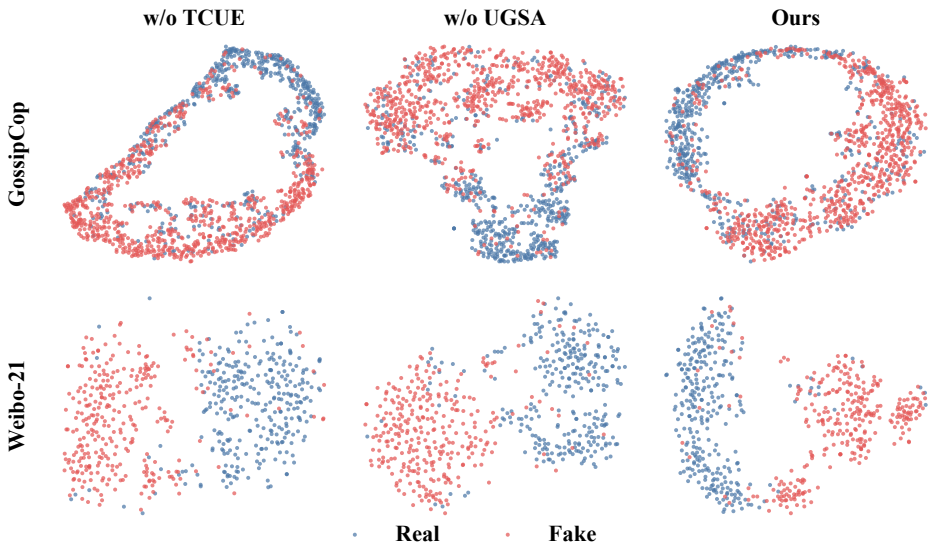

Figure 7: Supplementary t-SNE visualizations on GossipCop and Weibo-21 for w/o TCUE, w/o UGSA, and the full model. Blue and red points denote real and fake news samples, respectively.

larly, a moderate λ_{ugsa} improves performance by reducing the gap between local modality representations and global semantics, while overly strong alignment may weaken task-specific discriminative cues.

Auxiliary Objective Weights. Figure 9 examines λ_{aux} and λ_{edl}. Both objectives remain effective across a range of small-to-moderate weights, although their preferred settings vary across datasets. Intermediate values of λ_{aux} maintain competitive performance on the two Chinese datasets, while GossipCop exhibits smaller variations. For λ_{edl}, Weibo-21 benefits from moderate evidential supervision, whereas Weibo and GossipCop achieve comparable accuracy under several small-to-moderate settings. These results support the use of nonzero auxiliary and evidential supervision while indicating the need for dataset-specific weight selection.

Overall, EMUL remains stable within reasonable parameter ranges. The results also show that uncertainty estimation, reliability calibration, semantic alignment, and task supervision require balanced weighting to achieve their complementary effects.

D.4 Robustness to CLIP Backbone Scale

To assess robustness to vision-language backbone scale, we compare CLIP backbones on all three datasets, varying only the CLIP model used by UGSA to extract global text-image semantics. We use Chinese-CLIP for Weibo and Weibo-21, and OpenAI CLIP for GossipCop. Small, Base, and Large correspond to RN50, ViT-B/16, and ViT-L/14, respectively, with all backbones using 224×224 inputs.

To accommodate backbone-specific feature dimensions, modality-specific adapters project CLIP image and text features to 512 dimensions for use as semantic references in UGSA. Apart from the CLIP backbone and these adapters, all model settings and training protocols remain unchanged.

Table 7 shows that ViT-B/16 achieves the best results on all datasets. It outperforms RN50, suggesting that it provides more discriminative global semantics for UGSA, and also surpasses ViT-L/14, indicating that larger vision-language backbones do not necessarily improve fake news detection. Excessive model capacity may introduce semantic details less aligned with news veracity prediction and complicate cross-modal alignment. ViT-B/16 instead provides a suitable balance among semantic capacity, model complexity, and dataset scale. Overall, EMUL remains stable across CLIP backbones, with ViT-B/16 adopted as the default configuration.

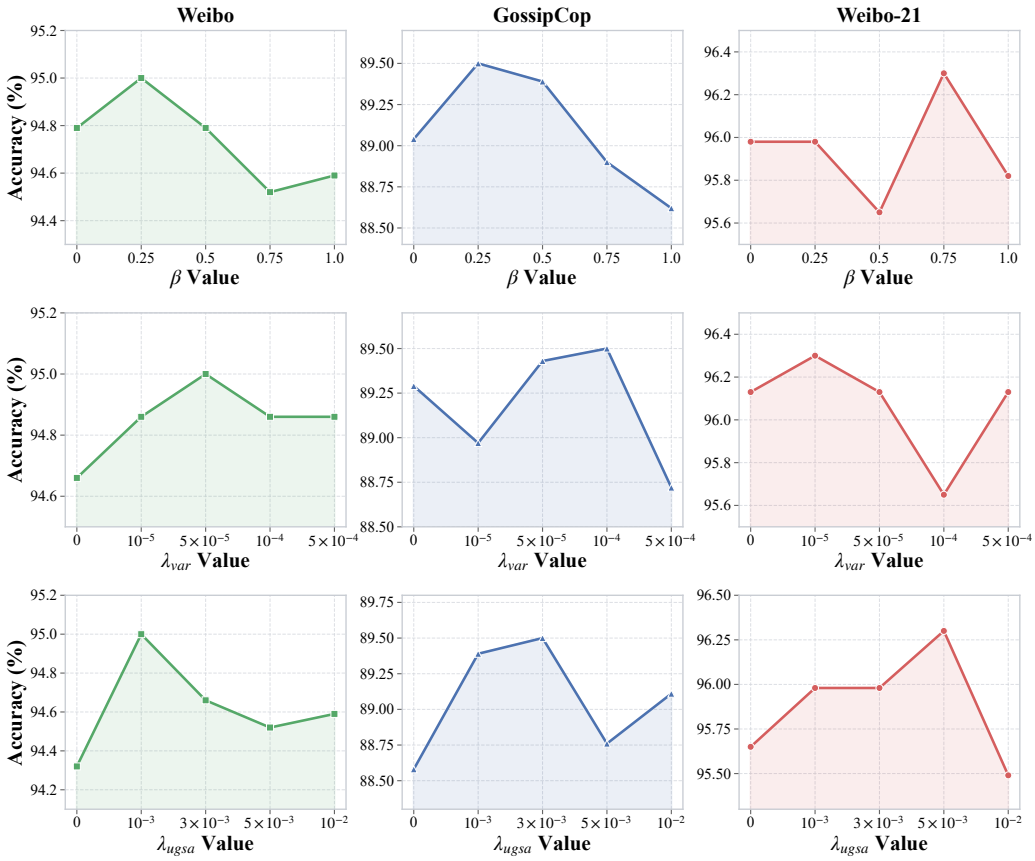

Figure 8: Sensitivity to β, λ_{var}, and λ_{ugsa} on Weibo-21, Weibo, and GossipCop. Each parameter is varied while the others remain fixed.

Table 7: Robustness under CLIP backbones with different scales.

Dataset	CLIP Backbone	Acc	F1(fake)	F1(real)
Weibo	CN-CLIP RN50	0.945	0.946	0.943
	CN-CLIP ViT-L/14	0.947	0.948	0.945
	Ours (CN-CLIP ViT-B/16)	**0.950**	**0.951**	**0.949**
GossipCop	CLIP RN50	0.887	0.665	0.932
	CLIP ViT-L/14	0.889	0.667	0.933
	Ours (CLIP ViT-B/16)	**0.895**	**0.680**	**0.937**
Weibo-21	CN-CLIP RN50	0.953	0.953	0.953
	CN-CLIP ViT-L/14	0.956	0.956	0.955
	Ours (CN-CLIP ViT-B/16)	**0.963**	**0.963**	**0.962**

E Extended Related Work

This appendix provides additional background and detailed comparisons that complement the concise discussion in Section 2.

E.1 Multimodal Fake News Detection

Multimodal fake news detection exploits complementary textual and visual cues to assess news veracity. Early methods primarily focused on learning modality-specific representations and combin-

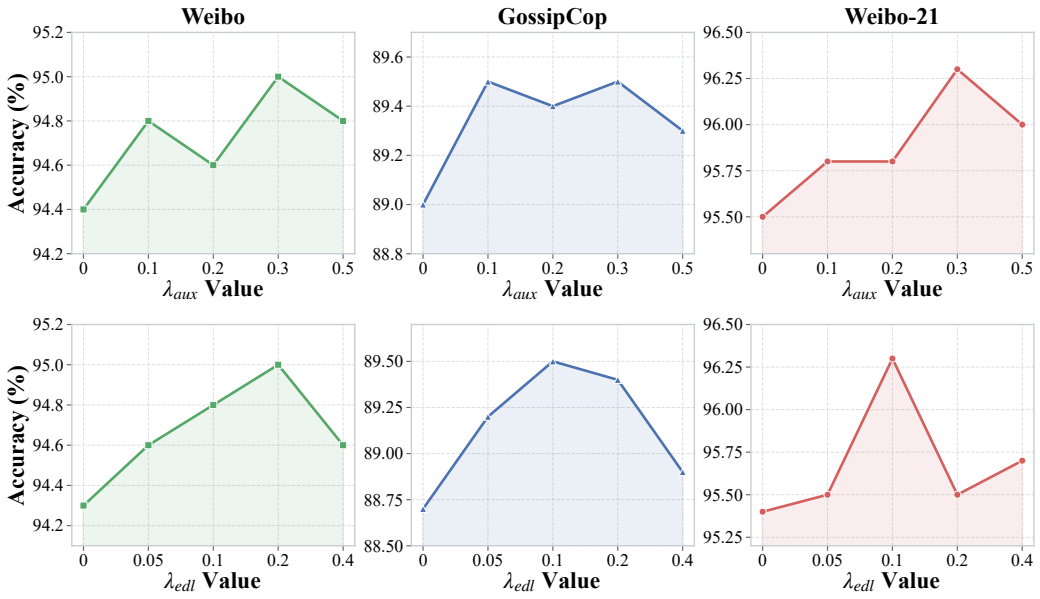

Figure 9: Sensitivity to the view-level auxiliary supervision weight λ_{aux} and evidential learning weight λ_{edl} on Weibo-21, Weibo, and GossipCop. Each parameter is varied while the others remain fixed.

ing them through direct fusion. EANN Wang et al. (2018) introduced event-adversarial learning to improve generalization across unseen events, while MVAE Khattar et al. (2019) learned shared multimodal representations through a variational autoencoder. SpotFake Singhal et al. (2019) combined pretrained textual and visual features, and SAFE Zhou et al. (2020) modeled image-text similarity as an additional cue for veracity prediction. Although these methods established the value of multimodal evidence, their relatively coarse fusion strategies provide limited modeling of fine-grained semantic relations and local noise.

Subsequent studies shifted toward cross-modal interaction, multi-view reasoning, and vision-language pretrained models. CAFE Chen et al. (2022) explicitly models cross-modal ambiguity, whereas BMR Ying et al. (2023) learns complementary representations through bootstrapped multi-view modeling. FND-CLIP Zhou et al. (2023) employs CLIP representations to introduce pretrained image-text semantic knowledge. Event-Radar Ma et al. (2024) models event-level multimodal inconsistency and estimates view credibility for low-quality news. AKA-Fake Zhang et al. (2024) selects informative external knowledge through adaptive subgraph construction. MMDFND Tong et al. (2024) and DAMMFND Lu et al. (2025b) address domain variation through multi-domain and domain-aware multi-view learning, while MIMoE-FND Liu et al. (2025) uses interaction-specific experts to accommodate heterogeneous modality relations.

Recent studies have also explored large language models and rationale-based reasoning. GLPN-LLM Hu et al. (2025) combines LLM-generated pseudo-labels with global label propagation, while MMRGV Chen et al. (2026a) generates and cross-verifies multiple rationales to reduce unreliable multimodal reasoning. MaLSF Chen et al. (2026b) further emphasizes local semantic interaction for multimodal media verification. Despite these advances, most methods treat encoded modality representations as reliable inputs to subsequent interaction. They rarely estimate whether individual tokens or patches should contribute equally before cross-modal reasoning, allowing unreliable local semantics to propagate into alignment and fusion.

E.2 Uncertainty-Aware Multimodal Learning and Trustworthy Fusion

Uncertainty estimation provides a principled means of characterizing data ambiguity, predictive risk, and model limitations. Bayesian approximations and deep ensemble methods estimate predictive

uncertainty through stochastic inference or model disagreement Gal & Ghahramani (2016); Kendall & Gal (2017); Lakshminarayanan et al. (2017). Evidential deep learning (EDL) instead predicts the parameters of a Dirichlet distribution, allowing class evidence and predictive uncertainty to be represented within a unified framework Sensoy et al. (2018). Dempster-Shafer theory (DST) and subjective logic provide complementary foundations for combining evidence from multiple sources while explicitly retaining residual uncertainty Shafer (2020).

These principles have motivated trustworthy multi-view learning. Trusted Multi-View Classification Han et al. (2021) uses evidential opinions to represent view-specific predictions and uncertainty. More recently, Lu et al. Lu et al. (2025a) introduced instance-wise trust estimation to discount unreliable or conflicting views, while Duan et al. Duan et al. (2025) jointly considered view-specific uncertainty and inter-view conflict for reliable fusion. These methods demonstrate that uncertainty is useful not only as a confidence estimate but also as a regulatory signal for multi-view decision making.

Uncertainty has also been explored in multimodal fake news detection. MM-ULN Wei et al. (2023) models intra-modal and inter-modal uncertainty over global modality representations. EventRadar Ma et al. (2024) estimates the credibility of multiple semantic views, whereas TMEF-BI Chen et al. (2025) combines evidential learning with behavioral information for trustworthy multimodal fusion. These methods improve robustness against incomplete or conflicting evidence, but uncertainty is generally estimated at the modality, view, or prediction level. In contrast, token- and patch-level reliability remains insufficiently explored, and uncertainty is rarely propagated across representation learning, semantic alignment, and evidential fusion. EMUL addresses this limitation by estimating local intra-modal uncertainty before cross-modal interaction and using it to regulate both semantic alignment and view-level evidence fusion.