IMEX-FND: A Traceable Interaction-Aware Mixture-of-Experts Framework for Multimodal Fake News Detection

Yuchen Miao[1], Zijun Wang[1✉], Ke Liu[1], Peixuan Wang[2], and Chang Han[1]

[1] Sydney Smart Technology College, Northeastern University, China
[2] School of Computer and Communication Engineering, Northeastern University at Qinhuangdao, China
[✉]Corresponding author.

Abstract. Multimodal fake news detection (FND) increasingly demands verdicts that are not only accurate but traceable—revealing how cross-modal evidence is combined—yet two coupled difficulties remain. First, text–image relations are heterogeneous: uniqueness, redundancy, and synergy coexist and vary from post to post, so a single global fusion rule is brittle and opaque. Second, the dominant modality shifts across instances, which static encoders and a fixed fusion pathway handle poorly. We present IMEX-FND, an interaction-aware mixture-of-experts framework that couples adaptive routing with explicit interaction decomposition. A Multi-Modal Expert Gateway (MMEG) performs instance-wise, within-modality routing over specialized and shared experts and builds a CLIP-grounded cross-modal stream, yielding three refined, interaction-ready representations that adapt to the dominant modality of each post. An Interaction-aware Multi-Modal Expert Fusion (IMEF) module then decomposes the interactions among these streams into *uniqueness*, *redundancy*, and *synergy*, producing transparent, sample-wise weights via a modality-replacement training signal. The two stages form a single route-then-decompose pipeline whose routing and interaction weights are both inspectable, offering instance- and dataset-level traceability for misinformation diagnosis. Extensive experiments on the Weibo, Weibo-21, and Gossip benchmarks show that IMEX-FND achieves state-of-the-art performance, surpassing competitive baselines by 0.4–1.2% while offering superior traceability with fewer parameters.

Keywords: Fake News Detection · Multimodal Fusion · Mixture of Experts · Social Media

1 Introduction

Online social platforms have become a primary channel for producing and consuming news; while democratizing information, they also lower the cost of fabricating and amplifying deceptive content at scale. Consequently, fake news has become a persistent societal risk, as intentionally false or misleading content

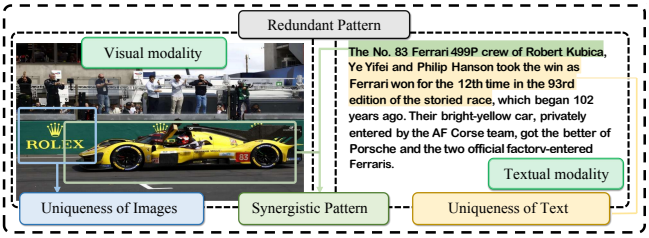

Fig. 1. An interaction-aware view of multimodal fusion showing redundant (duplicate) and modality-unique information.

can be crafted to manipulate beliefs and behavior, motivating a long line of automated detection methods.

Most deceptive posts are multimodal: a textual claim is paired with an image that may corroborate, complement, or contradict it. As Figure 1 shows, this text–image relation is heterogeneous: some evidence is *unique* to one modality, some is *redundantly shared*, and some emerges only from their combination (*synergy*). Two difficulties follow. First, the interaction that exposes a fake differs from post to post, so a single global fusion rule that compresses text and image into one vector is brittle and opaque about why a verdict is reached. Second, the dominant modality shifts across instances – one post is betrayed by its text, another by its image – and a static encoder with a fixed fusion pathway cannot adapt, capping robustness.

Existing methods tackle parts of this picture but not the whole. Late-fusion pipelines concatenate or co-attend over text–image features atop pretrained encoders [3,4]; consistency- and ambiguity-driven methods regularize cross-modal coherence via similarity, ambiguity, or contrastive objectives [24,2,17,25]; and expert-based designs use gating, domain adaptation, or mixture-of-experts (MoE) routing to pick a fusion pathway per instance [11,15,10,14,9,8]. Yet ambiguity-guided methods still collapse interaction complexity into a scalar weight, and learned experts and routes are rarely interpretable as canonical interaction types – leaving both robustness under shifting modalities and decision transparency unresolved.

We propose IMEX-FND, an interaction-aware mixture-of-experts framework that makes modality interactions first-class through a two-stage route-then-decompose design. A Multi-Modal Expert Gateway (MMEG) routes each modality instance-wise through specialized and shared experts and grounds a cross-modal stream in vision–language pretraining [25], yielding three refined representations (text, image, fusion) that adapt to whichever modality carries the decisive evidence. An Interaction-aware Multi-Modal Expert Fusion (IMEF) then decomposes their interactions into human-readable *uniqueness*, *redundancy*, and *synergy*, emitting transparent, sample-wise weights learned with a modality-replacement signal that drives each expert to specialize [7,20].

The main contributions of this paper are summarized as follows:

- We propose IMEX-FND, a unified interaction-aware MoE framework coupling instance-adaptive modality routing with explicit interaction decomposition for multimodal FND.
- We design a Multi-Modal Expert Gateway (MMEG) that refines each modality through specialized and shared experts and builds a CLIP-grounded fusion stream, improving robustness under dominant-modality shift.
- We design a traceable Interaction-aware Multi-Modal Expert Fusion (IMEF) that disentangles uniqueness, redundancy, and synergy into sample-wise weights via modality-replacement weak supervision, yielding instance- and dataset-level interpretability.
- On Weibo, Weibo-21, and Gossip, IMEX-FND attains state-of-the-art accuracy with fewer parameters while exposing interpretable interaction attributions.

2 Related Work

2.1 Modality Interactions

Two lines of work underpin our design: how modalities interact, and how multimodal fake news detectors exploit that interaction. Modality-interaction research studies how modalities jointly support inference, where effective reasoning often leverages both unimodal and cross-modal evidence. Factorized contrastive learning captures interaction factors from an information-theoretic view [7], while gated routing models diverse interaction patterns with separate learning pathways [11,20]. In multimodal fake news detection, models should exploit both unique and shared evidence, yet many do not explicitly model semantic-level interaction types, yielding brittle fusion under heterogeneous relations. We address this by explicitly modeling interaction patterns and exposing transparent, instance-wise fusion behavior.

2.2 Multimodal Fake News Detection

Early work combines modalities with attention and sequence models [3], while MVAE learns shared representations via variational objectives [4]. With pretrained encoders, later methods concatenate or co-attend over text–image features, though cross-modal misalignment can still harm performance.

To improve coherence, many approaches add consistency, ambiguity modeling, or contrastive objectives [2,17,24]. SAFE estimates cross-modal inconsistency by similarity [24], while CAFE quantifies ambiguity via KL-based divergence between modality distributions [2], and CMC distills cross-modal correlations implicitly [19]. COOLANT integrates cross-modal contrastive learning with ambiguity-aware fusion [17], and CLIP-guided fusion exploits vision–language pretraining for stronger semantic alignment signals [25]. Recent extensions inject high-level signals, such as intent–semantic joint learning via graph modeling [18], external knowledge augmentation with emotion guidance [26], and LLM-augmented reinforced sampling to strengthen supervision and hard-example

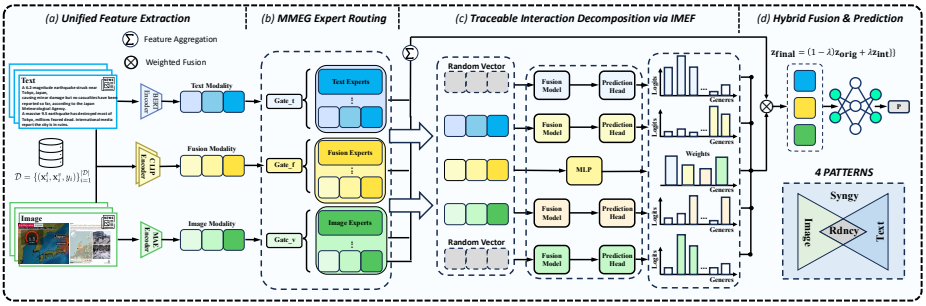

Fig. 2. Overall framework. Pretrained encoders extract token-level representations. MMEG routes each modality through a mixture of specialized and shared experts to obtain modality features. IMEF explicitly models interaction types via traceable expert weights and weakly-supervised interaction losses. Final prediction uses hybrid fusion with multi-auxiliary supervision.

coverage [16]. Beyond ambiguity-weighted aggregation, MIMoE-FND routes instances to specialized fusion experts by supervising a hierarchical MoE gate with unimodal prediction agreement and semantic alignment [8]. More recently, multimodal large language models (MLLMs) enable explicit reasoning: DIVER performs iterative visual evidence reasoning, invoking fine-grained visual tools only when text is insufficient [23], while multi-perspective rationale generation with cross-verification reconciles contradictions among MLLM-generated rationales before prediction [1]. In parallel, MViR strengthens the visual branch with multi-view visual–semantic representations via pyramid dilated convolution [6]. Retrieval-augmented detection and multimodal fact-checking resources further emphasize external evidence and retrieval difficulty [5,21].

Despite progress, ambiguity-guided methods compress interaction complexity into scalar weights, with limited insight into which interaction dominates and why, while MoE routing improves adaptivity yet rarely interprets experts as canonical interaction types. We therefore introduce an interaction-aware, traceable fusion design that disentangles heterogeneous interactions and provides transparent, instance-level attributions for robust detection.

3 Methodology

3.1 Problem Setup

We study multimodal fake news detection with a dataset $\mathcal{D} = \{(\mathbf{x}_i^t, \mathbf{x}_i^v, y_i)\}_{i=1}^{|\mathcal{D}|}$, where $\mathbf{x}^t$ is a text sequence, $\mathbf{x}^v$ is an associated image (or a representative image for multi-image posts), and $y \in \{0, 1\}$ indicates fake vs. real. Our goal is to learn a predictor $f(\mathbf{x}^t, \mathbf{x}^v) \to p(y{=}1)$.

3.2 Overview

IMEX-FND is an interaction-aware mixture-of-experts framework with four parts that turn a raw post into a traceable verdict. Pretrained encoders first provide text, image, and cross-modal signals. A Multi-Modal Expert Gateway (MMEG) then mixes experts within each modality into compact, modality-specific features, and an Interaction-aware Multi-Modal Expert Fusion (IMEF) module decomposes their interaction into *uniqueness*, *synergy*, and *redundancy* experts that emit sample-wise traceable weights. Multi-auxiliary supervision with a hybrid fusion objective then ties the pieces together, forming a single route-then-decompose pipeline whose intermediate weights stay inspectable. The pretrained backbones serve as frozen feature extractors, while the MoE modules and classifiers are trained end-to-end.

3.3 Pretrained Encoders and Unified Representations

Each post is first turned into three signals that feed MMEG. The text encoder produces token features $\mathbf{T}_{\text{raw}} \in \mathbb{R}^{B\times L\times d_t}$ together with a routing summary obtained by masked attention,

$$\mathbf{t}_{\text{pool}} = \text{Attn}_{\text{mask}}(\mathbf{T}_{\text{raw}}) \in \mathbb{R}^{B\times d_t}, \tag{1}$$

and, symmetrically, the image encoder produces patch features $\mathbf{I}_{\text{raw}} \in \mathbb{R}^{B\times N\times d_v}$ with a token-attention summary,

$$\mathbf{v}_{\text{pool}} = \text{Attn}(\mathbf{I}_{\text{raw}}) \in \mathbb{R}^{B\times d_v}. \tag{2}$$

To capture cross-modal evidence directly, we also extract CLIP image and text features $\mathbf{c}^v, \mathbf{c}^t \in \mathbb{R}^{B\times d_c}$ and project their concatenation,

$$\mathbf{c} = \phi\big([\mathbf{c}^v; \mathbf{c}^t]\big) \in \mathbb{R}^{B\times d}, \tag{3}$$

where $\phi(\cdot)$ is an MLP projection and d is the unified feature dimension shared by all downstream experts (see Sec. 4.1).

3.4 Multi-Modal Expert Gateway (MMEG)

Following the expert-routing paradigm of multi-gate MoE and layered expert extraction [11,15], MMEG performs within-modality routing (Figure 3) that refines each representation while preserving its modality-specific bias. For text and image, each specialized expert is a multi-kernel 1D-CNN extractor over the token sequence,

$$E(\mathbf{X}) = \text{Concat}\Big(\max_l \text{Conv}_k(\mathbf{X})\Big)_{k\in\{1,2,3,5,10\}} \in \mathbb{R}^{B\times d}, \tag{4}$$

with max pooling over the sequence length; shared experts tie parameters across text and image, and fusion experts are lightweight MLP refiners in $\mathbb{R}^d$. Given

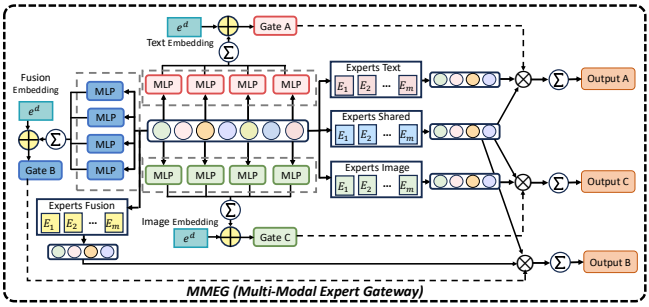

Fig. 3. Detailed architecture of the Multi-Modal Expert Gateway (MMEG). Each modality is routed by its own gate (Gate A/B/C for the text/fusion/image streams) over a pool of specialized experts together with a set of shared experts; the gated, summed combination of expert outputs yields the three refined streams (Output A/B/C, i.e. $\mathbf{t}$, $\mathbf{m}$, $\mathbf{v}$). The fusion stream is constructed from the CLIP-grounded embedding and the shared evidence of the text and image streams.

a pooled routing input $\mathbf{u}^m$ for each modality $m \in \{T, I, F\}$, a gate produces instance-wise expert weights,

$$\mathbf{w}^m = \mathrm{Softmax}\Big(\mathrm{MLP}^m_{\mathrm{gate}}(\mathbf{u}^m)\Big) \in \mathbb{R}^{K+L}, \tag{5}$$

where K and L count the specialized and shared experts (see Sec. 4.1); text and image take $\mathbf{u}^T{=}\mathbf{t}_{\mathrm{pool}}$ and $\mathbf{u}^I{=}\mathbf{v}_{\mathrm{pool}}$ and operate on the token sequences $\mathbf{X}^T{=}\mathbf{T}_{\mathrm{raw}}$ and $\mathbf{X}^I{=}\mathbf{I}_{\mathrm{raw}}$. Each modality's routed feature is then the gated combination of its experts,

$$\mathbf{H}^m = \sum_{k=1}^{K} w^m_k \, E^m_k(\mathbf{X}^m) + \sum_{l=1}^{L} w^m_{K+l} \, S_l(\mathbf{X}^m) \in \mathbb{R}^{B\times d}. \tag{6}$$

The fusion modality is built by combining the CLIP cross-modal feature with the shared evidence of the two unimodal streams,

$$\mathbf{z}_{\mathrm{fus}} = \psi\big([\mathbf{c}; \mathbf{H}^T_{\mathrm{sh}}; \mathbf{H}^I_{\mathrm{sh}}]\big) \in \mathbb{R}^{B\times d}, \tag{7}$$

where $\mathbf{H}^T_{\mathrm{sh}}, \mathbf{H}^I_{\mathrm{sh}}$ are shared-expert outputs and $\psi(\cdot)$ is an MLP. A fusion gate finally yields $\mathbf{H}^F$, giving the three MMEG outputs consumed by the next stage,

$$\mathbf{t} = \mathbf{H}^T, \qquad \mathbf{v} = \mathbf{H}^I, \qquad \mathbf{m} = \mathbf{H}^F. \tag{8}$$

3.5 Interaction-aware Multi-Modal Expert Fusion (IMEF)

Given the three MMEG streams, IMEF decomposes their interaction into a taxonomy inspired by partial information decomposition [7,20]: *uniqueness* (text, image, fusion), *synergy*, and *redundancy*. Stacking the streams into $\mathbf{z} = [\mathbf{t}; \mathbf{v}; \mathbf{m}] \in$

$\mathbb{R}^{B\times 3d}$, five interaction experts $\{\mathcal{E}_T, \mathcal{E}_I, \mathcal{E}_F, \mathcal{E}_{\text{syn}}, \mathcal{E}_{\text{red}}\}$, each a 2-layer MLP, read this joint vector,

$$\mathcal{E}_i(\mathbf{z}) = \text{MLP}_i(\mathbf{z}) \in \mathbb{R}^{B\times d}, \tag{9}$$

and we write $\mathbf{o}_i^{(0)} = \mathcal{E}_i(\mathbf{z})$ for the output on the unperturbed input. A reweighting network maps this vector to per-sample interaction weights,

$$\boldsymbol{\alpha} = \text{Softmax}\Big(\text{MLP}_{\text{rw}}(\mathbf{z})/\tau\Big) \in \mathbb{R}^{B\times 5}, \tag{10}$$

with temperature τ controlling sharpness, and the fused interaction feature is their weighted sum,

$$\mathbf{z}_{\text{int}} = \sum_{i=1}^{5} \alpha_i \, \mathbf{o}_i^{(0)} \in \mathbb{R}^{B\times d}. \tag{11}$$

Because $\boldsymbol{\alpha}$ is emitted per sample, it doubles as an attribution that can be aggregated for global interaction profiling.

To make each expert specialize in its interaction type, we add a weakly-supervised modality-replacement signal, applied only during training. For $r \in \{T, I, F\}$ we overwrite one stream with a noise vector $\mathbf{n}_r$ while keeping the others intact,

$$\tilde{\mathbf{z}}^{(r)} = \big[\tilde{\mathbf{t}}; \tilde{\mathbf{v}}; \tilde{\mathbf{m}}\big], \quad \tilde{\mathbf{h}} = \begin{cases} \mathbf{n}_r, & \text{if } \mathbf{h} \text{ corresponds to modality } r, \\ \mathbf{h}, & \text{otherwise,} \end{cases} \tag{12}$$

with $(\mathbf{h}, \tilde{\mathbf{h}}) \in \{(\mathbf{t}, \tilde{\mathbf{t}}), (\mathbf{v}, \tilde{\mathbf{v}}), (\mathbf{m}, \tilde{\mathbf{m}})\}$, and pass each perturbed input through the experts to obtain $\mathbf{o}_i^{(r)} = \mathcal{E}_i(\tilde{\mathbf{z}}^{(r)})$. Each interaction type then receives a loss matching its meaning. The *uniqueness* expert of modality r should react only when r is removed, so we use a triplet margin with the unperturbed output as anchor, the r-replaced output as negative, and the remaining replacements as positives,

$$\mathcal{L}_{\text{uniq}}^{(r)} = \frac{1}{2} \sum_{r'\in\{T,I,F\}\setminus\{r\}} \ell_{\text{tri}}\Big(\mathbf{o}_r^{(0)},\ \mathbf{o}_r^{(r')},\ \mathbf{o}_r^{(r)}\Big), \tag{13}$$

where $\ell_{\text{tri}}(\mathbf{a}, \mathbf{p}, \mathbf{n}) = \max\big(0, \|\mathbf{a} - \mathbf{p}\|_2 - \|\mathbf{a} - \mathbf{n}\|_2 + \mu\big)$ and μ is the margin. The *synergy* expert, in contrast, should collapse whenever any modality is missing, so we penalize its similarity to each perturbed output,

$$\mathcal{L}_{\text{syn}} = \frac{1}{3} \sum_{r\in\{T,I,F\}} \cos\big(\mathbf{o}_{\text{syn}}^{(0)}, \mathbf{o}_{\text{syn}}^{(r)}\big), \tag{14}$$

whereas the *redundancy* expert should stay invariant to single-modality replacement, which we encourage with the complementary objective

$$\mathcal{L}_{\text{red}} = \frac{1}{3} \sum_{r\in\{T,I,F\}} \Big(1 - \cos\big(\mathbf{o}_{\text{red}}^{(0)}, \mathbf{o}_{\text{red}}^{(r)}\big)\Big). \tag{15}$$

Summing the five terms gives the interaction regularizer

$$\mathcal{L}_{\text{int}} = \mathcal{L}_{\text{uniq}}^{(T)} + \mathcal{L}_{\text{uniq}}^{(I)} + \mathcal{L}_{\text{uniq}}^{(F)} + \mathcal{L}_{\text{syn}} + \mathcal{L}_{\text{red}}. \tag{16}$$

Replacement is confined to training; at inference IMEF uses only $\mathbf{z}_{\text{int}}$ from the unperturbed input.

3.6 Prediction Heads and Hybrid Fusion Objective

Three lightweight MLP classifiers attach auxiliary predictions to the per-modality streams,

$$p_T = \sigma(C_T(\mathbf{t})), \quad p_I = \sigma(C_I(\mathbf{v})), \quad p_F = \sigma(C_F(\mathbf{m})), \tag{17}$$

while the final decision blends the summed representation $\mathbf{z}_{\text{orig}} = \mathbf{t} + \mathbf{v} + \mathbf{m}$ with the interaction feature,

$$\mathbf{z}_{\text{final}} = (1 - \lambda)\mathbf{z}_{\text{orig}} + \lambda \mathbf{z}_{\text{int}}, \tag{18}$$

where $\lambda \in [0, 1]$ controls the contribution of interaction-aware fusion (see Sec. 4.1). The blended feature is classified by a final head,

$$p_{\text{fake}} = \sigma(C(\mathbf{z}_{\text{final}})). \tag{19}$$

The full model is trained end-to-end by minimizing the total objective

$$\begin{aligned} \mathcal{L}_{\text{total}} = & \underbrace{\mathcal{L}_{\text{main}}}_{\text{BCE}(p_{\text{fake}}, y)} + \frac{1}{3}\Big(\underbrace{\mathcal{L}_T}_{\text{BCE}(p_T, y)} + \underbrace{\mathcal{L}_I}_{\text{BCE}(p_I, y)} + \underbrace{\mathcal{L}_F}_{\text{BCE}(p_F, y)} \Big) \\ & + \sum_{b=1}^{5} w_b \, \mathcal{L}_{\text{int}}^{(b)} + \delta \|\Theta\|_2^2. \end{aligned} \tag{20}$$

where $\{\mathcal{L}_{\text{int}}^{(b)}\}_{b=1}^{5}$ correspond to $\mathcal{L}_{\text{uniq}}^{(T)}$, $\mathcal{L}_{\text{uniq}}^{(I)}$, $\mathcal{L}_{\text{uniq}}^{(F)}$, $\mathcal{L}_{\text{syn}}$, and $\mathcal{L}_{\text{red}}$. We use fixed interaction weights $\{w_b\}_{b=1}^{5}$ and weight decay δ; hyperparameters are given in Sec. 4.1.

4 Experiments and Results

4.1 Experimental Settings

Implementation Details We freeze all pretrained backbones: a BERT-style text encoder, an MAE-pretrained ViT-B/16 for images, and CLIP ViT-B/16 for cross-modal features. Text is truncated to at most 197 tokens and images are resized to 224×224 with patch size 16 (196 patches plus a [CLS] token), and every stream is projected to a unified dimension d=320. Within MMEG, each modality uses K=6 specialized and L=12 shared experts, the multi-kernel CNN has kernel sizes $\{1, 2, 3, 5, 10\}$ with 64 filters each, and the routing gate is a 2-layer SiLU MLP with dropout 0.1; IMEF's five interaction experts are

2-layer MLPs (hidden 320, dropout 0.1) and its reweighting network is a 2-layer MLP (hidden 256, τ=1.0), while all classifier heads are 2-layer MLPs (hidden 384, dropout 0.2). We optimize with Adam (learning rate 1×10^{-4}, weight decay 5×10^{-5}) at batch size 64 for up to 50 epochs with early stopping, using BCE for the main and auxiliary heads (auxiliary losses averaged), a triplet margin μ=1.0 for uniqueness and cosine penalties for synergy and redundancy, with all interaction losses sharing weight w_b=0.01 and a hybrid-fusion weight λ=0.3.

Baselines We compare our method against baselines from three paradigm families: (i) Representation and Robustness Learning, including the domain-robust model DAMMFND [10], cross-modal distillation frameworks CMC [19] and BMR [22], and the recent multi-view visual-representation method MViR [6]; (ii) Cross-Modal Consistency and Alignment, covering SAFE [24], CAFE [2], foundation-model-based FND-CLIP [25], and expert-based MIMoE-FND [8]; and (iii) Advanced Semantics and Reasoning, representing the latest state-of-the-art methods: GSFND [16], KEN [26], and InSide [18].

4.2 Overall Performance Comparison

Table 1. Performance on *Weibo*, *Weibo-21*, and *Gossip*. Results are averaged over 10 runs; shaded cells indicate the best (red) and second-best (blue) values in each column. For compactness, leading zeros are omitted for values below one.

Method	*Weibo*							*Weibo-21*							*Gossip*						
	Acc	Fake			Real			Acc	Fake			Real			Acc	Fake			Real		
		P	R	F1	P	R	F1		P	R	F1	P	R	F1		P	R	F1	P	R	F1
EANN	.825	.845	.810	.827	.805	.841	.823	.868	.900	.823	.860	.839	.910	.873	.862	.700	.515	.593	.885	.954	.918
SAFE	.760	.829	.722	.772	.693	.809	.747	.903	.891	.906	.898	.914	.899	.906	.836	.756	.556	.641	.855	.935	.893
SpotFake	.890	.900	**.964**	.931	.845	.654	.737	.850	.951	.731	.827	.784	**.964**	.865	.856	.730	.370	.491	.864	.961	.910
CAFE	.838	.853	.828	.840	.823	.849	.836	.880	.855	.913	.883	.905	.842	.872	.865	.730	.488	.585	.885	.955	.919
CMC	.871	.880	.865	.872	.843	.858	.850	.893	.903	.888	.895	.881	.896	.888	.881	.750	.558	.640	.898	**.961**	.928
BMR	.916	.880	**.948**	.913	**.942**	.877	.908	.927	.906	.947	.926	.944	.904	.924	.893	.750	.637	.689	.919	**.965**	.934
FND-CLIP	.905	.912	.899	.905	.912	.899	.905	**.943**	.935	.945	.940	**.950**	.942	**.946**	.878	.759	.547	.636	.897	.957	.926
DAMMFND	.907	.921	.906	.913	.891	.906	.898	.931	.948	.933	.940	.919	.934	.926	.891	.756	.648	.700	.914	.960	.937
MIMoE-FND	.926	**.938**	.923	.930	.913	.928	.920	.941	**.956**	.941	**.948**	.929	.944	.936	.896	.762	.644	.698	.918	.962	.939
InSide	.881	.720	.651	.684	.925	.939	.932	.910	.900	.910	.905	.912	.918	.915	**.900**	**.785**	**.809**	**.797**	**.930**	.938	**.934**
KEN	**.935**	.931	.939	**.935**	**.930**	.938	**.934**	**.948**	.930	.935	.932	.935	.940	.937	.885	.730	.660	.693	.905	.945	.925
GSFND	.915	.915	.921	.918	.910	.914	.912	.899	.880	.888	.884	.870	.880	.875	.892	.750	.780	.765	.922	.940	.931
MViR	.913	.918	.904	.911	.907	.919	.913	.917	.926	.902	.914	.912	.923	.917	.884	.753	.629	.685	.913	.951	.932
Ours	**.939**	**.953**	.938	**.945**	.929	**.944**	**.936**	**.952**	**.965**	**.950**	**.957**	**.939**	**.954**	**.946**	**.908**	**.800**	**.810**	**.805**	**.935**	.943	**.939**

Table 1 reports overall results on Weibo, Weibo-21, and Gossip using the baselines from Section 4.1. Three findings stand out. First, our route-then-decompose design achieves the best accuracy on all three benchmarks—for instance, 0.952 accuracy with 0.952 macro-F1 on Weibo-21 and 0.908 accuracy with 0.872 macro-F1 on Gossip—outperforming strong baselines such as InSide and KEN by 0.4–1.2% in accuracy, and on the harder Gossip set it keeps Real

Table 2. Ablation study of core components on three datasets, reporting Macro-F1. (Higher is better)

Variant	Weibo	Weibo-21	Gossip
Full	**0.940**	**0.952**	**0.872**
w/o IMEF	0.920 (-0.020)	0.930 (-0.022)	0.849 (-0.023)
w/o MMEG	0.928 (-0.012)	0.942 (-0.010)	0.858 (-0.014)
w/o MR & L_{int}	0.932 (-0.008)	0.945 (-0.007)	0.861 (-0.011)
w/o Aux	0.936 (-0.004)	0.948 (-0.004)	0.866 (-0.006)
SF	0.908 (-0.032)	0.921 (-0.031)	0.832 (-0.040)

F1 high (0.939) while lifting Fake F1 to 0.805, so coupling adaptive routing with explicit interaction decomposition helps most on the difficult class. Second, performance tracks how finely a method treats cross-modal interaction: simple fusion (EANN, SpotFake) trails, ambiguity- and consistency-aware methods (CAFE, CMC, BMR) improve, and recent expert-routing designs (MIMoE-FND, InSide, KEN) push the upper bound; by decomposing fusion into *uniqueness*, *redundancy*, and *synergy* rather than a single fused vector or a scalar weight, IMEF advances this frontier, confirming that resolving heterogeneous interactions, not merely adding capacity, is what matters. Third, the model stays robust as the dominant modality shifts, keeping precision and recall closely matched for both classes (e.g., Weibo-21 Fake 0.965/0.950, Real 0.939/0.954) and gaining a stable 1.3 points in accuracy from Weibo to Weibo-21, because MMEG routes each post instance-wise and can lean on whichever modality is informative—a property we probe directly in Sec. 4.4.

4.3 Ablation Studies

To dissect IMEX-FND and quantify each component's contribution, we evaluate the variants below; all keep the encoders, data splits, and optimization of the main experiments and change only the target component. Table 2 reports Macro-F1 (averaged over Fake/Real F1) on Weibo, Weibo-21, and Gossip, with drops w.r.t. the full model in red.

- **w/o IMEF**: the final head uses only the summed streams **t**+**v**+**m**, with no interaction decomposition.
- **w/o MMEG**: each modality enters IMEF as its pooled encoder feature, not routed experts.
- **w/o MR & L_{int}**: the interaction experts train without modality replacement or the interaction losses.
- **w/o Aux**: only the final classifier is kept, removing the per-modality heads p_T, p_I, p_F.
- **SF**: a simple-fusion baseline without MMEG and IMEF, classifying the concatenated pooled features.

Table 3. Robustness under dominant-modality shift: Macro-F1 on the *agreement* vs *conflict* subsets (posts where the text- and image-only heads agree / disagree).

	Weibo		Weibo-21		Gossip	
Variant	Agree	Conflict	Agree	Conflict	Agree	Conflict
Full	0.962	0.862	0.969	0.895	0.911	0.761
−MMEG	0.956	0.828	0.964	0.867	0.903	0.730
Δ	−0.006	−0.034	−0.005	−0.028	−0.008	−0.031

Removing IMEF causes the largest single-component drop on all datasets (−0.020/−0.022/−0.023 Macro-F1), confirming that explicitly decomposing heterogeneous interactions is the most important design; removing MMEG is next (−0.012/−0.010/−0.014), showing that instance-wise routing is what supplies robustness under dominant-modality shift. Disabling MR and L_{int} also weakens Macro-F1, indicating that the perturbation signal drives experts to specialize rather than collapse into redundant copies, while the auxiliary heads add a small but stable gain. The ordering w/o IMEF > w/o MMEG > w/o MR & L_{int} > w/o Aux mirrors the route-then-decompose logic: decomposition contributes most, adaptive routing second, each relying on its dedicated training signal.

4.4 Robustness to Dominant-Modality Shift

To probe difficulty (ii) directly, we test whether MMEG helps when the decisive modality varies across posts. Using the auxiliary text- and image-only heads (p_T, p_I), we split each test set into an agreement subset (the two unimodal predictions concur) and a conflict subset (they disagree); the conflict subset approximates posts whose dominant modality is ambiguous or shifts. Table 3 compares the full model with the −MMEG variant on both subsets.

Removing MMEG hurts the conflict subset far more than the agreement subset (ΔMacro-F1 ≈ -0.031 for conflict vs. −0.006 for agreement on average), confirming that instance-wise routing is the component that absorbs dominant-modality shift, whereas on easy, agreement posts the routing margin is naturally small.

4.5 Traceability and Interaction Analysis

The interaction weights $\boldsymbol{\alpha}$ expose which interaction drives each decision, addressing the transparency side of difficulty (i), and can be read at two levels. At the dataset level, aggregating $\boldsymbol{\alpha}$ over a test set yields an interaction profile: Gossip is dominated by text- and image-*uniqueness* ($\bar{\alpha} = 0.291$ and 0.268), whereas Weibo leans most on *synergy* ($\bar{\alpha} = 0.308$), with Weibo-21 in between, matching their differing modality reliance. At the single-post level, $\boldsymbol{\alpha}$ attributes the verdict to one pattern—a claim with a mismatched photo to *synergy*, a spliced image to image-*uniqueness*, and a corroborating caption–image pair to *redundancy*—as detailed in Table 4, turning an opaque fusion score into a per-sample rationale.

Table 4. Local interaction attribution for representative posts: per-sample IMEF weights $\boldsymbol{\alpha}$ over the five experts (dominant one in **bold**). The weight pattern matches the underlying text–image relation, and every prediction agrees with the ground-truth label.

Dataset	Text–image relation	α_T	α_I	α_F	α_{syn}	α_{red}	Pred.
Weibo	Claim–photo mismatch	0.08	0.07	0.10	**0.62**	0.13	Fake ✓
Gossip	Spliced/tampered image	0.06	**0.71**	0.08	0.09	0.06	Fake ✓
Gossip	Exaggerated text, neutral image	**0.68**	0.07	0.09	0.10	0.06	Fake ✓
Weibo-21	Caption–image corroborate	0.09	0.08	0.12	0.14	**0.57**	Real ✓

4.6 Hyperparameter Settings

We study the sensitivity of the fusion weight λ, which controls the hybrid fusion between the conventional fused representation and the IMEF interaction feature. Specifically, λ=0 reduces the model to a conventional fusion path, while larger λ places more emphasis on interaction-aware fusion. We sweep $\lambda \in \{0.0, 0.1, 0.3, 0.6, 1.0\}$ and report Accuracy together with class-wise F1 scores (Fake/Real).

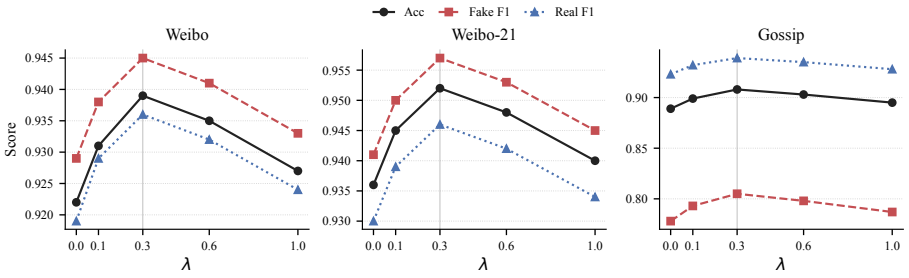

Fig. 4. Sensitivity of λ on Weibo, Weibo-21, and Gossip (left to right). Each subplot shows Accuracy and class-wise F1 scores (Fake/Real).

As Figure 4 shows, a moderate interaction weight generally performs best across datasets. Performance improves from λ=0 to λ=0.3 and then slightly saturates or decreases, indicating that interaction-aware fusion helps but should be balanced with the conventional path: routing (MMEG) and decomposition (IMEF) are complementary, not substitutes. Unless otherwise specified, we use λ=0.3 in all experiments.

4.7 Efficiency Analysis

We compare efficiency against architecturally comparable detectors—neural fusion and expert-based models trained on frozen encoders—under identical back-

Table 5. Efficiency comparison on Weibo-21 (batch size 64) against architecturally comparable baselines. Params counts trainable parameters only.

Method	Params (M)	Train (ms/iter)	Test (ms/iter)
SpotFake	20.4	88	32
EANN	22.6	94	34
CAFE	25.8	103	37
CMC	27.3	108	38
DAMMFND	29.2	115	42
FND-CLIP	31.6	117	40
IMEX-FND (ours)	33.5	121	43
BMR	36.8	127	45
MIMoE-FND	43.0	133	47

bones and input sizes (197 tokens, 224×224); we omit the LLM- and knowledge-augmented methods (GSFND, KEN, InSide), whose cost is dominated by external generation or retrieval and is thus not comparable on trainable parameters. We report trainable parameters (excluding frozen encoders) and runtime on a single NVIDIA RTX 3090 with batch size 64 (AMP). As shown in Table 5, IMEX-FND sits in the middle of the parameter range yet attains the best accuracy (Table 1), giving a favorable cost–performance trade-off; in particular, its interaction regularization adds only limited training overhead since the extra perturbed passes are confined to the lightweight IMEF module and disabled at inference, keeping its test-time latency close to the most efficient baselines.

5 Conclusion

We presented IMEX-FND, an interaction-aware framework that addresses the limited robustness and interpretability of multimodal fake news detection via a route-then-decompose design. A Multi-Modal Expert Gateway (MMEG) routes each modality through specialized and shared experts and builds a CLIP-grounded fusion stream, while an Interaction-aware Multi-Modal Expert Fusion (IMEF) disentangles cross-modal dynamics into *uniqueness*, *redundancy*, and *synergy* via traceable sample-wise weights, turning an opaque fusion score into an inspectable account of modality contributions and interactions. Experiments on Weibo, Weibo-21, and Gossip show that IMEX-FND achieves state-of-the-art performance, surpassing competitive baselines by 0.4%–1.2% while offering superior transparency with fewer parameters. Future work will explore causal reasoning [13]. Efficient refinement under limited feedback is another direction [12].

Acknowledgments. This work was supported by the National College Student Innovation and Entrepreneurship Training Program under Project No. 202619145026.

References

1. Chen, J., Li, Y., Ng, K.C., Wang, H., Zhang, L.J.: Toward multimodal fake news detection by multi-perspective rationale generation and verification. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 40, pp. 66–74 (2026)
2. Chen, Y., Li, D., Zhang, P., Sui, J., Lv, Q., Lu, T., Shang, L.: Cross-modal ambiguity learning for multimodal fake news detection. In: Proceedings of the ACM Web Conference 2022 (WWW '22). pp. 2897–2905. Association for Computing Machinery, New York, NY, USA (2022)
3. Jin, Z., Cao, J., Guo, H., Zhang, Y., Luo, J.: Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: Proceedings of the 25th ACM International Conference on Multimedia (MM '17). pp. 795–816. Association for Computing Machinery, Mountain View, CA, USA (2017)
4. Khattar, D., Goud, J.S., Gupta, M., Varma, V.: MvaE: Multimodal variational autoencoder for fake news detection. In: Proceedings of the 2019 World Wide Web Conference (WWW '19). pp. 2915–2921. Association for Computing Machinery, New York, NY, USA (2019)
5. Li, Y., Lu, W., Yu, H., Wang, Y.: Retrieval-augmented multimodal model for fake news detection. arXiv preprint arXiv:2604.18112 (2026)
6. Liang, H., Su, X., Wang, J., Chen, C., Yu, Z.: MViR: Multi-view visual-semantic representation for fake news detection. In: ICASSP 2026 – 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 12192–12196. IEEE (2026)
7. Liang, P.P., Deng, Z., Ma, M.Q., Zou, J., Morency, L.P., Salakhutdinov, R.: Factorized contrastive learning: Going beyond multi-view redundancy. In: Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS) (2023)
8. Liu, Y., Liu, Y., Li, Z., Yao, R., Zhang, Y., Wang, D.: Modality interactive mixture-of-experts for fake news detection. In: Proceedings of the ACM on Web Conference 2025 (WWW '25). pp. 5139–5150. Association for Computing Machinery, New York, NY, USA (2025)
9. Lu, W., Li, Y.: 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, pp. 818–826 (2026)
10. Lu, W., Tong, Y., Ye, Z.: Dammfnd: Domain-aware multimodal multi-view fake news detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 559–567 (2025)
11. Ma, J., Zhao, Z., Yi, X., Chen, J., Hong, L., Chi, E.H.: Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '18). pp. 1930–1939. Association for Computing Machinery, New York, NY, USA (2018)
12. Miao, Y.: Universal refinement without interaction: Order-optimal 1-bit mean estimation (2026), arXiv preprint arXiv:2607.24358
13. Miao, Y., Cui, M., Zhu, Y., Wang, Y., Xu, S.: R3-rec: Reasoning-driven recommendation via retrieval-augmented llms over multi-granular interest signals. In: ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5951–5955. IEEE (2026)
14. Tong, Y., Lu, W., Cui, X., Mao, Y., Zhao, Z.: Dapt: Domain-aware prompt-tuning for multimodal fake news detection. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 7902–7911 (2025)

15. Tong, Y., Lu, W., Zhao, Z., Lai, S., Shi, T.: MMDFND: Multi-modal multi-domain fake news detection. In: Proceedings of the 32nd ACM International Conference on Multimedia (MM '24). pp. 1178–1186. Association for Computing Machinery, New York, NY, USA (2024)
16. Tong, Z., Gu, Y., Liu, H., Liu, Q., Wu, S., Shi, H., Zhang, X.Y.: Generate first, then sample: Enhancing fake news detection with LLM-augmented reinforced sampling. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 24276–24290. Association for Computational Linguistics, Vienna, Austria (Jul 2025)
17. Wang, L., Zhang, C., Xu, H., Xu, Y., Xu, X., Wang, S.: Cross-modal contrastive learning for multimodal fake news detection. In: Proceedings of the 31st ACM International Conference on Multimedia (MM '23). pp. 5696–5704. Association for Computing Machinery, New York, NY, USA (2023)
18. Wang, Z., Sheng, Q., Wang, D., Hu, B., Cao, J.: Bridging thoughts and words: Graph-based intent-semantic joint learning for fake news detection (2025), arXiv preprint arXiv:2509.01660
19. Wei, Z., Pan, H., Qiao, L., Niu, X., Dong, P., Li, D.: Cross-modal knowledge distillation in multi-modal fake news detection. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4733–4737. IEEE (2022)
20. Xin, J., Yun, S., Peng, J., Choi, I., Ballard, J.L., Chen, T., Long, Q.: I2moe: interpretable multimodal interaction-aware mixture-of-experts. In: Proceedings of the 42nd International Conference on Machine Learning. ICML'25, JMLR.org (2025)
21. Xu, W., Xiang, D., Ding, T., Lu, W.: Mmm-fact: A multimodal, multi-domain fact-checking dataset with multi-level retrieval difficulty. arXiv preprint arXiv:2510.25120 (2025)
22. Ying, Q., Hu, X., Zhou, Y., Qian, Z., Zeng, D., Ge, S.: Bootstrapping multi-view representations for fake news detection. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 37, pp. 5384–5392 (2023)
23. Zhou, W., Ying, Z., Meng, C., Liu, J., Zhou, H., Zou, Q., Zhang, D., Yang, D., Zhang, X.: Diver: Dynamic iterative visual evidence reasoning for multimodal fake news detection (2026), arXiv preprint arXiv:2601.07178
24. Zhou, X., Wu, J., Zafarani, R.: Safe: Similarity-aware multi-modal fake news detection. In: Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II. Lecture Notes in Computer Science, vol. 12085, pp. 354–367. Springer, Cham (2020)
25. Zhou, Y., Yang, Y., Ying, Q., Qian, Z., Zhang, X.: Multimodal fake news detection via CLIP-guided learning. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME). pp. 2825–2830. IEEE, Brisbane, Australia (2023)
26. Zhu, P., Jing, Y., Cheng, L., Tang, K., Guo, Y.: KEN: Knowledge augmentation and emotion guidance network for multimodal fake news detection. In: Proceedings of the 33rd ACM International Conference on Multimedia (MM '25). pp. 1793–1801. Association for Computing Machinery, New York, NY, USA (2025)