DGFND-ES: Evidence-Enhanced Dual-Graph Fake News Detection under Sociological Constraints

Anonymous ACL submission

Abstract

The coordinated dissemination of multimodal content on social media has become the norm, rendering fake news increasingly covert and complex. Existing methods generally lack event-level information modeling, which limits their ability to effectively handle breaking events, and they also fail to account for group fairness and disparities in information harm under a globalized context. To address these challenges, we propose **DGFND-ES**, a dual-graph collaborative fake news detection framework that integrates evidence enhancement with sociological constraints. This framework adopts a neuro-symbolic architecture consisting of a "main graph–consistent subgraph" structure, and incorporates group fairness constraints and a harm-aware loss during training to endow the model with social responsibility. In addition, we construct a high-quality SSS dataset for systematic evaluation of model performance. Experimental results demonstrate that DGFND-ES consistently outperforms existing methods on the Weibo-21, Fakeddit, and SSS datasets.

1 Introduction

The explosive growth of social media has greatly accelerated the dissemination of information, while various information platforms have simultaneously become fertile ground for fake news. Multimodal fake news often maximizes its reach and persuasiveness by carefully combining deceptive text with manipulated or misappropriated images (Yang et al., 2025; Liu et al., 2025a; Chen et al., 2023). Compared with text-only content, multimodal fake news exhibits stronger emotional impact and apparent credibility due to the incorporation of visual information, rendering its narratives more misleading and significantly increasing the difficulty of detection (Zhang et al., 2024a; Lv et al., 2025). In the real world, the harm caused by fake news has become increasingly severe; in many major social events, false content persistently disrupts public opinion and exacerbates social disorder (Murayama, 2021). Moreover, with the continuous advancement of vision–language models, the quality of generated multimodal forgeries has markedly improved, leading to pronounced performance degradation in traditional detection methods.

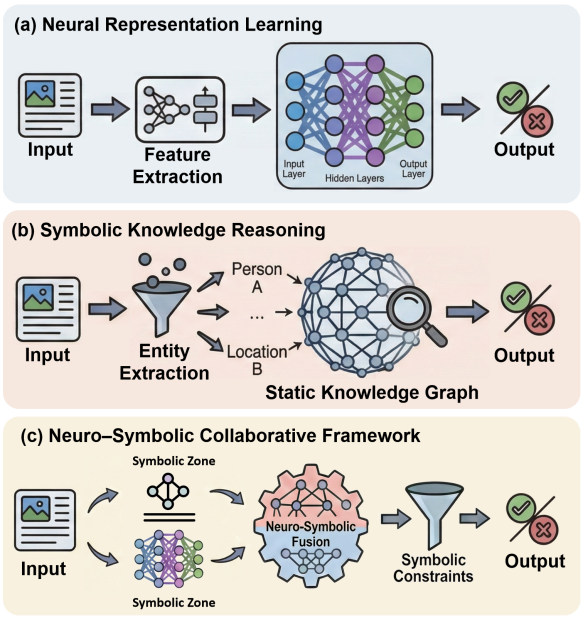

Figure 1: Comparison of fake news detection paradigms. (a) Neural representation learning with end-to-end multimodal classification; (b) Symbolic knowledge reasoning via entity extraction and static knowledge graphs; (c) Neuro-symbolic collaborative framework combines neural representations with symbolic reasoning and constraints for fake news detection.

Current multimodal fake news detection techniques can be broadly categorized into neural representation learning methods and symbolic knowledge reasoning methods. Neural representation learning approaches typically rely on pretrained vision–language models to learn deep semantic representations and cross-modal consistency through end-to-end modeling (Lu et al., 2025; Bian et al., 2020; Singhal et al., 2019; Zhang et al., 2025; Liu

et al., 2025b; Zheng et al., 2022). However, such models remain opaque "black boxes," whose predictions lack interpretable reasoning paths and explicit factual grounding. Symbolic knowledge reasoning methods construct more interpretable decision processes by explicitly incorporating structured knowledge and logical constraints. These approaches usually depend on entity linking, relation matching, and external knowledge verification mechanisms to enhance discriminative capability and reliability (Xuan et al., 2024; Yuliani et al., 2019). Nevertheless, their overall performance is largely constrained by the coverage of knowledge graphs, limiting their adaptability to emerging information and rapidly evolving events.

Although multimodal fake news detection has made notable progress, from the perspective of practical deployment and social impact, existing methods still face two fundamental challenges. First, most multimodal detection approaches rely primarily on neural representation learning for decision making. Even when external information is incorporated, the data sources are often static, fragmented, and limited in quality, which makes it difficult to effectively model the dynamic evolution of real-world events and continuous fact updates. Second, existing methods largely overlook the heterogeneous social impacts of fake news. Their training and inference processes neither adequately account for data distribution imbalances across different languages, regions, and cultural contexts, nor distinguish the varying degrees of real-world harm caused by different types of misinformation.

To address these challenges, we propose the DGFND-ES framework, which deeply integrates symbolic logic into neural representation learning and jointly optimizes group fairness and harm-aware objectives during training. The main contributions of this paper are summarized as follows:

- We propose a neuro-symbolic collaborative architecture based on a "*main graph–consistency subgraph*" design, which enables event-level relational modeling across information sources and cross-modal consistency modeling within individual news items.
- We introduce *evidence* as a dynamic event-level verification unit and leverage LLMs to quantitatively model the stance of evidence toward news claims, substantially enhancing the model's capability to verify emerging and evolving events.
- We incorporate sociological constraints into the training process by jointly optimizing group fairness and harm-aware objectives, thereby mitigating group bias while maintaining detection accuracy and prioritizing the identification of high-harm fake content.
- We construct a high-quality SSS dataset and demonstrate the overall superiority of DGFND-ES over multiple mainstream methods on this benchmark as well as several public datasets.

2 Related Works

2.1 Evolution of Multimodal Fake News Detection

The development of multimodal fake news detection has closely followed advances in feature extraction techniques. In the early stage, researchers primarily employed CNNs (Yang et al., 2018) and RNNs (Jin et al., 2017) to extract visual and textual features, respectively, and fused them via simple concatenation or attention mechanisms. EANN (Wang et al., 2018) introduced adversarial learning to disentangle event-specific features, while MVAE (Khattar et al., 2019) leveraged variational autoencoders to reconstruct latent representations. With breakthroughs in pretraining, the field entered an era of fine-grained cross-modal interaction. Representative methods include CAFE (Chen et al., 2022), which employs fuzzy reasoning to quantify cross-modal ambiguity, and InfoSurgeon (Fung et al., 2021), which explicitly extracts textual entities and visual regions to construct fine-grained consistency graphs. More recently, end-to-end alignment approaches (Cui et al., 2025; Wang et al., 2023) attempt to directly capture inconsistencies in a contrastive learning space.

2.2 Knowledge-Enhanced Detection Methods

To overcome the limitations of internal information, incorporating external knowledge has become a critical pathway. The first category is based on static knowledge graphs. (Qian et al., 2021a; Lin et al., 2024) A representative work, CompareNet (Hu et al., 2021), aligns news text with a pre-constructed knowledge graph and determines veracity by comparing relationships among entities. More recently, AKA-Fake (Zhang et al., 2024a) introduces a reinforcement learning mechanism to adaptively retrieve the most relevant knowledge subgraphs from the KG according to the news

context. The second category is based on open retrieval, where methods leverage search engines to obtain real-time context. A typical example is GET (Xu et al., 2022), which constructs a heterogeneous graph neural network consisting of news, retrieved evidence, and their sources, and performs prediction by aggregating evidence semantics and source credibility. RAMA (Yang et al., 2025) further adopts a retrieval-augmented multi-agent framework to conduct debate-based reasoning.

2.3 LLM-Based Detection Approaches

In recent years, large language models (LLMs) have been reshaping the detection paradigm (Allen et al., 2025; Jin et al., 2024). FKA-Owl (Liu et al., 2024) leverages the knowledge capabilities of LLMs to enhance multimodal semantic understanding, enabling the capture of implicit facts and commonsense cues across modalities. GLPN-LLM (Hu et al., 2025) generates pseudo-labels via LLMs and integrates global graph propagation to alleviate annotation scarcity and insufficient supervision. LLM-GAN (Wang et al., 2025) incorporates LLMs into a generative adversarial framework, improving detection performance while providing explainable decision rationales. FactAgent (Li et al., 2024) organizes LLMs in an agent-based manner to incrementally integrate external evidence, achieving structured and interpretable news verification.

3 Methodology

Our goal is to construct a neuro–symbolic dual-graph collaborative framework for fake news detection that integrates external evidence augmentation with sociological constraints. The framework takes multimodal news as input and outputs the final authenticity prediction. During the decision-making process, the model jointly leverages event-related external evidence and the internal cross-modal consistency of news content: on the one hand, stance-aware evidence modeling is employed to characterize the supportive or refutational effects of evidence on news authenticity; on the other hand, consistency modeling explicitly captures potential conflicts between text and images at the modal, stylistic, and logical levels. These two types of information are respectively modeled through a main graph and a consistency subgraph, and are adaptively fused within a unified framework to achieve comprehensive authenticity assessment of news.

3.1 Evidence Retrieval Framework

Existing methods often rely on static knowledge graphs when incorporating external information; however, such decontextualized knowledge struggles to adapt to the highly dynamic, event-driven decision requirements of real-world news scenarios. In contrast, external evidence oriented toward specific news events exhibits stronger discriminative power in terms of event specificity and stance divergence. Motivated by this observation, we design and implement a coarse-to-fine evidence retrieval and filtering framework, whose workflow is summarized as follows:

1. **Internet-scale retrieval.** Given an input news text, we perform online retrieval to collect the top N relevant webpages ($N = 10$ in our study) and extract their abstracts as candidate evidence texts.

2. **LLM-driven stance quantification.** We leverage a large language model to quantitatively assess the relationship between each candidate evidence item and the original post, producing a continuous stance score $s_{ij} \in [0, 1]$ that characterizes the strength of support or refutation.

3. **Polarity-aware filtering and normalization.** Based on the polarity distribution of stance scores, we retain only the Top-K evidence items with the most extreme stances ($K = 5$ in our study), thereby filtering out neutral or ambiguous content.

3.2 Feature Encoding and Representation

3.2.1 Feature Encoding of News and Evidence

For a news post p_i, we employ feature encoders to obtain representations of the news text and image, respectively. Let h_i^{text} and h_i^{img} denote the textual and visual features extracted by the model. We fuse them via an attention mechanism to obtain the unified semantic representation of the post z_i^{post}:

$$z_i^{\text{post}} = \alpha_t \left(W_t h_i^{\text{text}}\right) + \alpha_v \left(W_v h_i^{\text{img}}\right) \quad (1)$$

where W_t and W_v are linear projection matrices, and α_t and α_v are attention weights adaptively computed by a learnable query vector. Similarly, for each retrieved evidence item e_{ij}, we apply an encoder to obtain its semantic representation z_{ij}^{evid}.

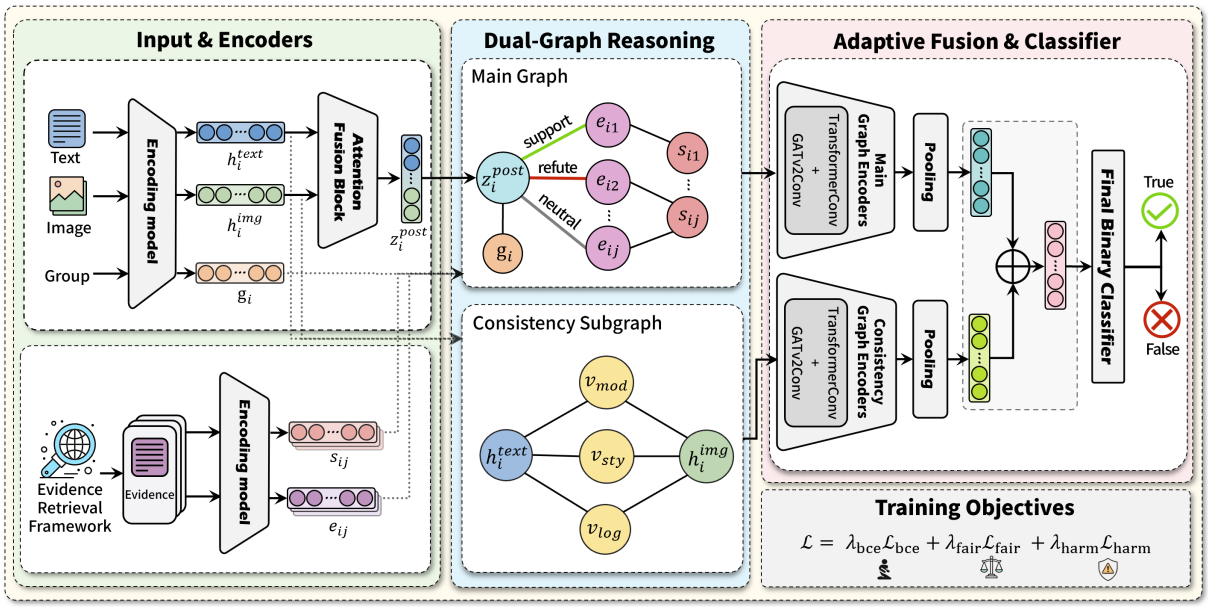

Figure 2: Overview of the DGFND-ES framework.

3.2.2 Multidimensional Consistency Feature Representation

Distinct from traditional approaches relying on entity extraction, this work directly models image-text consistency within the latent representation space across three orthogonal dimensions: modality, style, and logic. Specifically, modality consistency x_{modal} integrates feature cosine similarity with the prediction divergence of unimodal classifiers to simultaneously capture semantic alignment and discrepancies in falsity indicators; style consistency x_{style} measures coherence by projecting features into a subspace via specific style projection heads; and logic consistency x_{logic} quantifies the image-text entailment relationship by directly leveraging the zero-shot prior knowledge of CLIP (Radford et al., 2021).

3.3 Dual-Graph Neural Network Architecture

3.3.1 Main Graph Construction

The main graph G_{main} is designed to model the structured interactions between a news post and its associated external information. Specifically, the graph contains four types of nodes. The post node is initialized as $v_P = z_i^{\text{post}}$, representing the multimodal semantics of the news post. The evidence node is initialized as $v_E = z_{ij}^{\text{evid}}$, corresponding to external evidence related to the post. In addition, a group node v_G and a source node v_S are introduced to model social group attributes and information source characteristics, respectively.

The core of the main graph lies in the construction of stance-weighted post–evidence edges. Based on the stance score s_{ij} quantified in Section 3.2, we establish stance-aware edges between posts and evidence, where the edge type and weight w_{ij} are defined according to the polarity of s_{ij}:

- **Support Edge**: Established when $s_{ij} \geq \tau_{\text{pos}}$. The edge weight depends not only on semantic similarity but is also positively amplified by the stance strength:

$$w_{ij}^{+} = s_{ij} \cdot \text{sim}(v_P, v_E) \tag{2}$$

- **Refute Edge**: Established when $s_{ij} \leq \tau_{\text{neg}}$. The edge weight is determined by the inverse stance strength and the semantic similarity:

$$w_{ij}^{-} = (1 - s_{ij}) \cdot \text{sim}(v_P, v_E) \tag{3}$$

3.3.2 Consistency Subgraph Construction

The consistency subgraph G_{coms} is designed to capture cross-modal conflicts within a post. Based on the features extracted in Section 3.2, we construct a bridging graph consisting of five types of nodes. The graph includes two endpoint nodes, namely textual semantics $v_T = h_i^{\text{text}}$ and visual semantics $v_I = h_i^{\text{img}}$, as well as three intermediate bridging nodes, including a modality node $v_{\text{mod}} = x_{\text{modal}}$, a style node $v_{\text{sty}} = x_{\text{style}}$, and a logic node $v_{\text{log}} = x_{\text{logic}}$. In terms of topology, we establish three parallel semantic channels $v_T \leftrightarrow v_{\text{bridge}} \leftrightarrow v_I$, where $v_{\text{bridge}} \in \{v_{\text{mod}}, v_{\text{sty}}, v_{\text{log}}\}$.

3.3.3 Two-Layer Hybrid Encoding Architecture

First Layer: TransformerConv. In the first layer, we adopt a multi-head TransformerConv to model attribute-aware interactions between nodes and their neighbors. By explicitly incorporating edge features into the message-passing process, the model jointly captures node semantics and relational attributes, enabling attribute-aware neighborhood aggregation. The node representation is updated as

$$h_i^{(1)} = \Phi^{(1)} \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(1)} m(z_i^{\text{post}}, z_{ij}^{\text{evid}}) \right) \quad (4)$$

where $\alpha_{ij}^{(1)}$ denotes the attention weight jointly determined by node and edge attributes, $m(\cdot)$ is a message function that fuses node and edge features, and $\Phi^{(1)}(\cdot)$ represents the update operator including normalization and nonlinear transformations.

Second Layer: GATv2Conv. To further refine high-level semantic information and enhance contextual adaptivity, the second layer employs GATv2Conv with a dynamic attention mechanism. This layer adaptively adjusts the importance of neighbors based on the joint representations of the central node and its neighbors. The node update is given by

$$h_i^{(2)} = \Phi^{(2)} \left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(2)} \, h_j^{(1)} \right) \quad (5)$$

where $\alpha_{ij}^{(2)}$ is the dynamically computed attention weight, and $\Phi^{(2)}(\cdot)$ denotes the corresponding nonlinear mapping.

Finally, global mean pooling is applied to all nodes to obtain the graph-level representations $h_{\text{main}} \in \mathbb{R}^{d_{\text{out}}}$ and $h_{\text{cons}} \in \mathbb{R}^{d_{\text{cons}}}$.

3.4 Loss Functions

3.4.1 Main Classification Loss

As the fundamental objective of fake news detection, we adopt the Binary Cross-Entropy (BCE) loss to minimize the distributional discrepancy between the predicted probability $\hat{y}_i$ and the ground-truth label y_i:

$$\mathcal{L}_{\text{bce}} = -\frac{1}{N} \sum_{i=1}^{N} \Big(y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i) \Big) \quad (6)$$

3.4.2 Group Fairness Loss

To balance discriminatory bias across different groups, we introduce a group fairness regularization term. To address the non-differentiable thresholding involved in standard TPR/FPR computation, we employ a differentiable approximation based on soft predictions. Specifically, the soft decision function is defined as

$$\pi_i(\tau, T) = \sigma \left(\frac{\hat{y}_i - \tau}{T} \right) \quad (7)$$

where $\hat{y}_i$ denotes the predicted probability, τ is the decision threshold, and T is a temperature parameter controlling smoothness.

Let TPR_g and FPR_g denote the approximate true positive rate and false positive rate of group g, computed using $\pi_i(\tau_{\text{fair}}, T_{\text{fair}})$. We further denote $\overline{\text{TPR}}$ and $\overline{\text{FPR}}$ as the corresponding global averages across all groups. The group fairness loss is then defined as

$$\mathcal{L}_{\text{fair}} = \sum_{g \in \mathcal{G}} \Big[\big(\text{TPR}_g - \overline{\text{TPR}} \big)^2 + \big(\text{FPR}_g - \overline{\text{FPR}} \big)^2 \Big] \quad (8)$$

3.4.3 Harm-Aware Loss

In real-world scenarios, misinformation differs substantially in societal impact and potential risk. To prioritize the suppression of highly harmful misinformation, we assign harm-dependent penalty weights to false negatives and false positives, thereby imposing stronger optimization pressure on misclassification of high-risk samples. Based on a differentiable soft-prediction formulation, the harm-aware loss is defined as:

$$\mathcal{L}_{\text{harm}} = \frac{1}{N} \sum_{i=1}^{N} \Big[c_{\text{FN}}(h_i) \, y_i \, \big(1 - \pi_i^{(\text{harm})} \big) + c_{\text{FP}}(h_i) \, (1 - y_i) \, \pi_i^{(\text{harm})} \Big] \quad (9)$$

where h_i denotes the harm coefficient of sample i, derived from LLM-based societal risk quantification and human verification. The cost functions for false negatives and false positives are defined as $c_{\text{FN}}(h_i) = 1 + \gamma \cdot h_i$ and $c_{\text{FP}}(h_i) = 1$.

3.4.4 Overall Optimization Objective

We adopt a weighted multi-task learning strategy for end-to-end training. The final optimization objective is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{bce}} \mathcal{L}_{\text{bce}} + \lambda_{\text{fair}} \mathcal{L}_{\text{fair}} + \lambda_{\text{harm}} \mathcal{L}_{\text{harm}} \quad (10)$$

where λ denotes hyperparameters that balance the contributions of the respective loss terms.

4 Experiments

4.1 Experimental Settings

4.1.1 Datasets

This study conducts extensive experiments on two public multimodal benchmark datasets and one self-constructed dataset. The public datasets include Weibo-21 (Nan et al., 2021) and Fakeddit (Nakamura et al., 2020). Specifically, Weibo-21 is a Chinese multimodal dataset collected from Sina Weibo, while Fakeddit is a large-scale English multimodal dataset based on Reddit. To ensure the reliability of experimental evaluation, we apply a unified data cleaning and preprocessing pipeline to both datasets, systematically removing duplicate samples, invalid data, and excessively short texts.

To address the limitations of existing datasets regarding sociological annotations and data quality, we construct SSS, a high-quality dataset incorporating multidimensional sociological annotations. At the news level, SSS integrates data from six public sources (Weibo-17 (Jin et al., 2017), Weibo-21 (Nan et al., 2021), CFND (Zhang et al., 2024b), MR2 (Hu et al., 2023), Fakeddit (Nakamura et al., 2020), and FineFake (Zhou et al., 2024)), with an initial scale exceeding 1.14 million samples. Through a rigorous screening process involving automated cleaning, LLM-assisted filtering, and manual review, 7,687 data points were obtained. Additionally, we collected and supplemented 910 high-quality data points covering languages such as English, Spanish, French, and Russian. The final SSS dataset comprises 8,597 news samples. Distinct from previous works, each sample in the SSS dataset includes group labels and social harm coefficients, along with corresponding evidence. The final statistics of each dataset and their training, validation, and test splits are reported in Table 1.

Table 1: Data Composition and Partitioning of the Weibo-21, Fakeddit, and SSS Datasets. The evidence for Weibo-21 and Fakeddit originates from our evidence retrieval framework.

Dataset	Type	Train		Val		Test	
		Real	Fake	Real	Fake	Real	Fake
Weibo-21	Post	1330	1406	443	468	443	469
	Evidence	12118		2503		2601	
Fakeddit	Post	1603	1947	536	649	535	649
	Evidence	14646		4893		4846	
SSS	Post	2554	2609	848	867	850	868
	Evidence	22781		7549		7732	

4.1.2 Baselines

We compare our proposed method with two categories of baseline models:

- **Without External Knowledge:** MVAE (Khattar et al., 2019), CAFE (Chen et al., 2022), HMCAN (Qian et al., 2021b), EventRadar (Ma et al., 2024), and MSACA (Wang et al., 2024).
- **With External Knowledge:** EGHGAT (Guo et al., 2024), KEHGNN-FD (Xie et al., 2023), and GLPN-LLM (Hu et al., 2025).

4.1.3 Experiment Details

All experiments are implemented using the PyTorch framework. The model is trained end-to-end using the AdamW optimizer, with a learning rate of 1×10^{-4} and a weight decay of 0.01. The batch size is set to 32, and the maximum number of training epochs is 50.

4.2 Performance on Public Benchmarks

4.2.1 Benchmark Comparison

Table 2 presents the experimental results on the Weibo-21 and Fakeddit datasets. On the Chinese Weibo-21 dataset, our model achieves an accuracy improvement of approximately 2.17% over the strongest non-knowledge-based baseline, and a gain of about 3.37% compared to the strongest knowledge-augmented baseline. On the English Fakeddit dataset, our model also shows consistent superiority, outperforming the best competing baseline by approximately 3.04%.

4.2.2 Representation Analysis

Through t-SNE visualization, we observe that the complete model yields the clearest separation between real and fake news representations; in contrast, removing either the evidence augmentation or the consistency modeling module leads to substantially increased class overlap. This phenomenon indicates that evidence enhancement and the consistency subgraph play a critical role in learning discriminative representations.

4.2.3 Ablation Study

To verify the effectiveness of individual components in the neuro–symbolic integration framework, we design two ablation variants: (1) *w/o Evidence*, which removes the external evidence augmentation module; (2) *w/o Consistency*, which removes the consistency subgraph.

Table 2: Performance comparison of DGFND-ES with other methods on the Weibo-21 and Fakeddit datasets.

Dataset	Category	Method	Acc.	Fake News			Real News		
				P	R	F1	P	R	F1
Weibo-21	Without Ext.	MVAE	0.7638	0.7554	0.7896	0.7721	0.7732	0.7372	0.7548
		CAFE	0.8147	0.8521	0.7740	0.8112	0.7819	0.8578	0.8181
		HMCAN	0.8695	0.8616	0.8891	0.8751	0.8785	0.8488	0.8634
		Event-Radar	0.8806	0.8747	0.8957	0.8851	0.8871	0.8647	0.8757
		MSACA	0.8936	0.8619	0.9446	0.9013	0.9347	0.8397	0.8847
	With Ext.	EGHGAT	0.6612	0.6581	0.7100	0.6831	0.6650	0.6095	0.6360
		KEHGNN-FD	0.7643	0.7545	0.8013	0.7772	0.7759	0.7252	0.7497
		GLPN-LLM	0.8816	0.8574	0.9232	0.8891	0.9115	0.8375	0.8729
	DGFND-ES (Ours)		**0.9153**	**0.8913**	**0.9457**	**0.9177**	**0.9423**	**0.8849**	**0.9127**
Fakeddit	Without Ext.	MVAE	0.7660	0.7657	0.8259	0.7947	0.7665	0.6935	0.7282
		CAFE	0.7863	0.7727	0.8644	0.8160	0.8079	0.6916	0.7452
		HMCAN	0.8564	0.8507	0.8952	0.8724	0.8643	0.8093	0.8359
		Event-Radar	0.8385	0.8274	0.8882	0.8567	0.8542	0.7795	0.8151
		MSACA	0.8463	0.8399	0.8891	0.8638	0.8551	0.7944	0.8236
	With Ext.	EGHGAT	0.6867	0.6687	0.8490	0.7481	0.7278	0.4897	0.5855
		KEHGNN-FD	0.7646	0.7585	0.8470	0.8003	0.7746	0.6610	0.7133
		GLPN-LLM	0.8750	0.8711	0.9060	0.8882	0.8802	0.8374	0.8582
	DGFND-ES (Ours)		**0.9054**	**0.9074**	**0.9214**	**0.9144**	**0.9029**	**0.8860**	**0.8943**

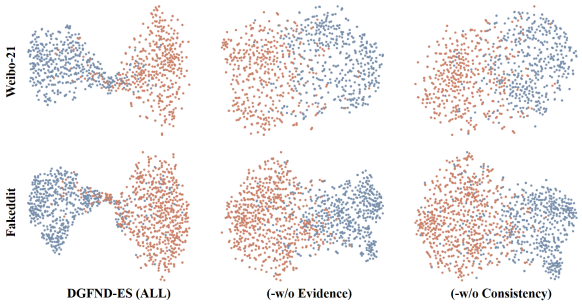

Figure 3: t-SNE visualization of representations learned by DGFND-ES and its ablated variants on Weibo-21 and Fakeddit.

Table 3: Ablation results of the DGFND-ES architecture on the Weibo-21 and Fakeddit datasets.

Dataset	Method	Accuracy	F1-score
Weibo-21	DGFND-ES (ALL)	**0.9153**	**0.9152**
	w/o Evidence	0.8904	0.8902
	w/o Consistency	0.8847	0.8847
Fakeddit	DGFND-ES (ALL)	**0.9054**	**0.9044**
	w/o Evidence	0.8834	0.8815
	w/o Consistency	0.8894	0.8881

The experimental results are reported in Table 3. When external evidence is removed (*w/o Evidence*), the model's accuracy on the Weibo-21 dataset drops by approximately 2.5%. This result indicates that external evidence provides event-level, dynamic fact-checking capabilities that are distinct from static knowledge. When the consistency subgraph is removed (w/o Consistency), the model accuracy further decreases by about 3.0%. This demonstrates that the semantic and logical consistency between textual and visual content within a news item constitutes a core basis for fake news detection, and its absence directly undermines the model's performance.

4.3 Evaluation on the SSS Dataset

4.3.1 Overall Performance

The experimental results show that DGFND-ES achieves an accuracy of 0.8050 on the SSS dataset, outperforming all baseline models in terms of overall performance. Further analysis reveals that methods relying on generic external knowledge generally underperform most approaches without external knowledge in terms of accuracy. This is mainly because static, general-purpose knowledge graphs lack event specificity in breaking news scenarios, which easily introduces semantic noise and weakens discriminative capability. In contrast, the event-driven evidence modeling proposed in this work maintains higher semantic relevance and effectively alleviates these issues.

Table 4: Performance comparison of DGFND-ES with other methods on the SSS dataset.

Dataset	Category	Method	Acc.	Fake News			Real News		
				P	R	F1	P	R	F1
SSS	Without Ext.	MVAE	0.6318	0.6365	0.6630	0.6495	0.6264	0.5987	0.6122
		CAFE	0.7090	0.6676	0.8445	0.7457	0.7823	0.5706	0.6599
		HMCAN	0.7509	0.7607	0.7396	0.7500	0.7414	0.7624	0.7517
		Event-Radar	0.7362	0.7131	0.8019	0.7549	0.7668	0.6687	0.7144
		MSACA	0.7637	0.7590	0.7800	0.7693	0.7688	0.7471	0.7578
	With Ext.	EGHGAT	0.6170	0.6190	0.6290	0.6240	0.6148	0.6047	0.6097
		KEHGNN-FD	0.6645	0.6558	0.6827	0.6690	0.6739	0.6467	0.6600
		GLPN-LLM	0.7014	0.7735	0.5783	0.6618	0.6576	0.8271	0.7327
	DGFND-ES (Ours)		**0.8050**	**0.7820**	**0.8514**	**0.8152**	**0.8331**	**0.7576**	**0.7936**

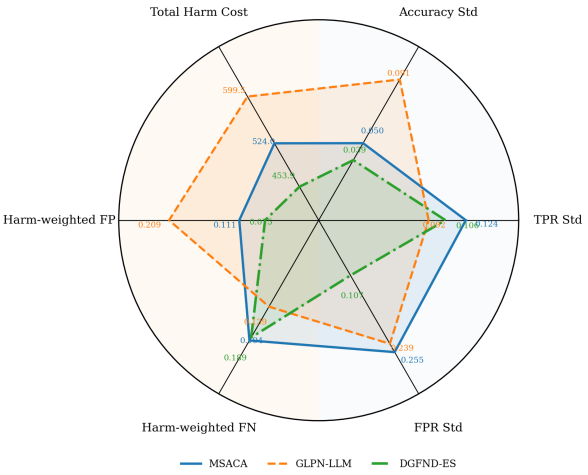

Figure 4: Comparison of different models in terms of group fairness disparity and harm-aware metrics. Lower values indicate better performance.

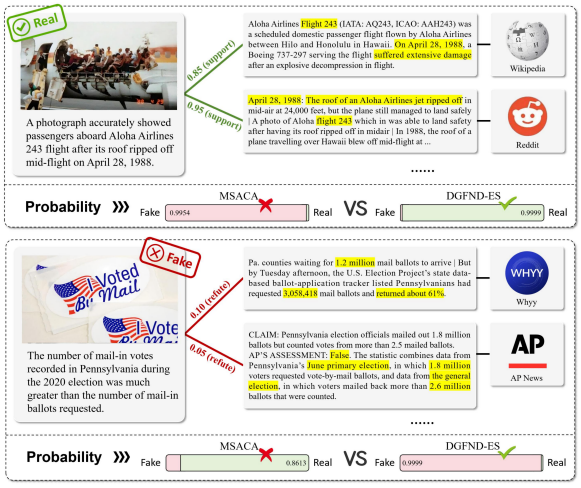

Figure 5: Case studies on challenging real and fake multimodal news comparing MSACA and DGFND-ES.

4.3.2 Fairness and Harm-Aware Evaluation

As shown in Figure 4, DGFND-ES exhibits the smallest overall area in the radar chart. It achieves lower values across multiple key metrics related to group fairness disparities and harm-aware costs, indicating that the model can more effectively balance group fairness and harm awareness while maintaining strong detection performance.

4.3.3 Case Study

As shown in Figure 5, the upper example is a piece of real news about Aloha Airlines Flight 243, where the news and the retrieved evidence are highly consistent in stance. DGFND-ES correctly classifies it as real, whereas MSACA makes an incorrect prediction. The lower example concerns a false claim about the 2020 U.S. mail-in voting. Although it appears superficially credible, evidence from authoritative sources provides a clear refutation. DGFND-ES is able to make the correct judgment, while MSACA is misled. These cases demonstrate that by integrating external evidence with internal consistency modeling, DGFND-ES achieves more reliable fake news detection.

5 Conclusion

In this study, we propose a dual-graph collaborative framework for fake news detection that integrates evidence enhancement with sociological constraints. By jointly modeling news content and external evidence within a neuro-symbolic dual-graph architecture, the framework enables collaborative reasoning, achieving improved detection performance while effectively accounting for group fairness and harm sensitivity, and providing a systematic solution for responsible fake news detection in real-world scenarios.

Limitations

Dependency on External Evidence. The proposed framework relies on retrieving evidence from the internet for event-level verification and reasoning. In the case of breaking news, authoritative reports may not yet be established online, making it difficult to obtain effective auxiliary information. Furthermore, if the retrieved evidence contains noise or misinformation, it may interfere with the reasoning process of the dual-graph network, thereby affecting the final detection accuracy.

High Computational Cost. This framework incorporates a Large Language Model to quantify the stance of each piece of evidence and constructs a complex dual-graph interaction network. Compared to lightweight end-to-end detection models, this architecture increases computational resource consumption and inference latency, which to some extent limits the method's deployment efficiency in large-scale real-time data streaming scenarios.

References

Bradley P Allen, Prateek Chhikara, Thomas Macaulay Ferguson, Filip Ilievski, and Paul Groth. 2025. Sound and complete neurosymbolic reasoning with llm-grounded interpretations. *arXiv preprint arXiv:2507.09751*.

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 549–556.

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

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 627–638.

ShaoDong Cui, Kaibo Duan, Wen Ma, and Hiroyuki Shinnou. 2025. Ccgn: consistency contrastive-learning graph network for multi-modal fake news detection. *Multimedia Systems*, 31(2):119.

Yi Fung, Christopher Thomas, Revanth Gangi Reddy, Sandeep Polisetty, Heng Ji, Shih-Fu Chang, Kathleen McKeown, Mohit Bansal, and Avirup Sil. 2021. Infosurgeon: Cross-media fine-grained information consistency checking for fake news detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1683–1698.

Yihao Guo, Longye Qiao, Zhixiong Yang, Jianping Xiang, Xinlong Feng, and Hongbing Ma. 2024. Fake news detection: Extendable to global heterogeneous graph attention network with external knowledge. *Tsinghua Science and Technology*, 30(3):1125–1138.

Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 2021. Compare to the knowledge: Graph neural fake news detection with external knowledge. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 754–763.

Shuguo Hu, Jun Hu, and Huaiwen Zhang. 2025. 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)*, pages 1426–1440, Vienna, Austria. Association for Computational Linguistics.

Xuming Hu, Zhijiang Guo, Junzhe Chen, Lijie Wen, and Philip S Yu. 2023. Mr2: A benchmark for multimodal retrieval-augmented rumor detection in social media. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 2901–2912.

Ruihan Jin, Ruibo Fu, Zhengqi Wen, Shuai Zhang, Yukun Liu, and Jianhua Tao. 2024. Fake news detection and manipulation reasoning via large vision-language models. *arXiv preprint arXiv:2407.02042*.

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

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

Xinyi Li, Yongfeng Zhang, and Edward C Malthouse. 2024. Large language model agent for fake news detection. *arXiv preprint arXiv:2405.01593*.

Zehang Lin, Jiayuan Xie, and Qing Li. 2024. Multimodal news event detection with external knowledge. *Information Processing & Management*, 61(3):103697.

Moyang Liu, Kaiying Yan, Yukun Liu, Ruibo Fu, Zhengqi Wen, Xuefei Liu, and Chenxing Li. 2025a. Deconfounded reasoning for multimodal fake news detection via causal intervention. *arXiv preprint arXiv:2504.09163*.

Xuannan Liu, Peipei Li, Huaibo Huang, Zekun Li, Xing Cui, Jiahao Liang, Lixiong Qin, Weihong Deng, and Zhaofeng He. 2024. Fka-owl: Advancing multimodal fake news detection through knowledge-augmented lvlms. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10154–10163.

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

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

Hongzhen Lv, Wenzhong Yang, Yabo Yin, Fuyuan Wei, Jiaren Peng, and Haokun Geng. 2025. Mdf-fnd: A dynamic fusion model for multimodal fake news detection. *Knowl. Based Syst.*, 317:113417.

Zihan Ma, Minnan Luo, Hao Guo, Zhi Zeng, Yiran Hao, and Xiang Zhao. 2024. 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)*, pages 5809–5821.

Taichi Murayama. 2021. Dataset of fake news detection and fact verification: a survey. *arXiv preprint arXiv:2111.03299*.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the twelfth language resources and evaluation conference*, pages 6149–6157.

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

Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2021a. Knowledge-aware multi-modal adaptive graph convolutional networks for fake news detection. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(3):1–23.

Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021b. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 153–162.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

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

Jiandong Wang, Hongguang Zhang, Chun Liu, and Xiongjun Yang. 2024. Fake news detection via multi-scale semantic alignment and cross-modal attention. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 2406–2410.

Longzheng Wang, Chuang Zhang, Hongbo Xu, Yongxiu Xu, Xiaohan Xu, and Siqi Wang. 2023. Cross-modal contrastive learning for multimodal fake news detection. In *Proceedings of the 31st ACM international conference on multimedia*, pages 5696–5704.

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

Yifeng Wang, Zhouhong Gu, Siwei Zhang, Suhang Zheng, Tao Wang, Tianyu Li, Hongwei Feng, and Yanghua Xiao. 2025. Llm-gan: Constructing generative adversarial network through large language models for explainable fake news detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Bingbing Xie, Xiaoxiao Ma, Jia Wu, Jian Yang, and Hao Fan. 2023. Knowledge graph enhanced heterogeneous graph neural network for fake news detection. *IEEE Transactions on Consumer Electronics*, 70(1):2826–2837.

Weizhi Xu, Junfei Wu, Qiang Liu, Shu Wu, and Liang Wang. 2022. Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM web conference 2022*, pages 2501–2510.

Keyang Xuan, Li Yi, Fan Yang, Ruochen Wu, Yi R Fung, and Heng Ji. 2024. Lemma: towards lvlm-enhanced multimodal misinformation detection with external knowledge augmentation. *arXiv preprint arXiv:2402.11943*.

Shuo Yang, Zijian Yu, Zhenzhe Ying, Yuqin Dai, Guoqing Wang, Jun Lan, Jinfeng Xu, Jinze Li, and Edith CH Ngai. 2025. Rama: Retrieval-augmented multi-agent framework for misinformation detection in multimodal fact-checking. *arXiv preprint arXiv:2507.09174*.

Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, and Philip S Yu. 2018. Ti-cnn: Convolutional neural networks for fake news detection. *arXiv preprint arXiv:1806.00749*.

SY Yuliani, Mohd Faizal Bin Abdollah, Shahrin Sahib, and Yunus Supriadi Wijaya. 2019. A framework for hoax news detection and analyzer used rule-based methods. *International Journal of Advanced Computer Science and Applications*, 10(10).

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

Qiang Zhang, Jiawei Liu, Fanrui Zhang, Jingyi Xie, and Zheng-Jun Zha. 2024b. Natural language-centered inference network for multi-modal fake news detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 2542–2550.

Tianlin Zhang, En Yu, Yi Shao, and Jiande Sun. 2025. Multimodal inverse attention network with intrinsic discriminant feature exploitation for fake news detection. pages 7940–7948. Main Track.

Jiaqi Zheng, Xi Zhang, Sanchuan Guo, Quan Wang, Wenyu Zang, and Yongdong Zhang. 2022. Mfan: Multi-modal feature-enhanced attention networks for rumor detection. In *IJCAI*, volume 2022, pages 2413–2419.

Ziyi Zhou, Xiaoming Zhang, Litian Zhang, Jiacheng Liu, Senzhang Wang, Zheng Liu, Xi Zhang, Chaozhuo Li, and Philip S Yu. 2024. Fine-fake: A knowledge-enriched dataset for fine-grained multi-domain fake news detection. *arXiv preprint arXiv:2404.01336*.

A Baselines

To comprehensively evaluate the performance of DGFND-ES, we benchmark it against diverse state-of-the-art methods. These baselines are selected to cover varying levels of external information utilization and are categorized into two groups:

Methods Without External Knowledge: These approaches focus on mining features from the internal multimodal content of the posts without relying on auxiliary information.

- **MVAE** (Khattar et al., 2019): Uses a multimodal variational autoencoder to learn shared latent representations of text and images, reconstructing the input to capture cross-modal correlations.
- **CAFE** (Chen et al., 2022): Focuses on cross-modal ambiguity learning, assessing the consistency between modalities to adaptively aggregate unimodal and cross-modal features for detection.
- **HMCAN** (Qian et al., 2021b): Employs a hierarchical multi-modal contextual attention network to model deep intra-post text-image correlations and align features through multi-step reasoning.
- **Event-Radar** (Ma et al., 2024): Proposes an event-driven multi-view learning framework that aims to identify event-independent features, thereby improving the model's generalization capabilities on unseen events.
- **MSACA** (Wang et al., 2024): Utilizes a multi-scale semantic alignment mechanism to align and fuse text and image features at different semantic levels, effectively capturing fine-grained inconsistencies.

Methods With External Knowledge: These approaches enhance reasoning capabilities by incorporating external knowledge graphs or background information.

- **EGHGAT** (Guo et al., 2024): Leverages Graph Attention Networks (GAT) to process external knowledge graphs, enhancing the representation of entity relationships through neighbor information aggregation.
- **KEHGNN-FD** (Xie et al., 2023): Uses a heterogeneous graph neural network to deeply fuse external structured knowledge with multimodal content features, capturing rich semantic relations.
- **GLPN-LLM** (Hu et al., 2025): A state-of-the-art method that synergizes Large Language Models (LLMs) with global label propagation, using LLMs to generate pseudo-labels and alleviate the scarcity of supervision data.

Table 5: Prompt for High-Quality News Data Screening

Speaker	Content
User	You are a professional data screening expert for fake news research, specializing in selecting high-quality news data. Your task is to determine whether the given news content meets the standards of high-quality news. High-quality news must satisfy the following criteria: • Text completeness: The text should have a relatively complete beginning and ending. It should not be composed of isolated phrases, disordered sentences, or clearly truncated content. • Concrete elements: The content should include specific information such as time, location, people involved, or the main subject of the event, rather than vague or abstract discussions. • Narrative nature: The text should describe an event or viewpoint with a basic narrative structure, rather than purely emotional expressions or meaningless repetition. Exclusion conditions (if any of the following conditions is met, the news should be considered "unusable"): 1. The text is incomplete, missing key information, or obviously truncated. 2. There are clear grammatical errors or logical inconsistencies. 3. The text lacks explicit news elements such as time, location, or entities. 4. Fewer than 10 characters of key informational content are present. 5. The content lacks concrete details and cannot be verified. 6. The text contains a large amount of garbled characters or obvious formatting errors. Output requirement: Only return "Usable" or "Unusable". Do not provide explanations or any additional text.
User	News content: “Breaking News: Early Monday morning, a strong 7.8-magnitude earthquake struck southeastern Turkey and northern Syria. Multiple buildings collapsed, and rescue teams have been deployed. Local authorities have declared a Level-4 emergency.”
LLM	Usable
User	News content: “It rained today, and I feel very upset. I hate this kind of weather. It makes me want to stay in bed all day.”
LLM	Unusable
User	News content: “A new policy was released yesterday.”
LLM	Unusable

Table 6: Composition and Filtering Results of Multi-source Datasets

Dataset	Language	Original Size	Filtered Size
Weibo-17 (Jin et al., 2017)	Chinese	9,528	1,073
Weibo-21 (Nan et al., 2021)	Chinese	9,128	790
CFND (Zhang et al., 2024b)	Chinese	26,665	380
MR2 (Hu et al., 2023)	Chinese & English	14,700	461
Fakeddit (Nakamura et al., 2020)	English	1,063,106	2,398
FineFake (Zhou et al., 2024)	English	16,909	2,585