Big Data and Information Analytics

BDIA, 10: 158–176.
DOI: 10.3934/bdia.2026008
Received: 26 May 2026
Revised: 24 June 2026
Accepted: 24 July 2026
Published: 03August 2026

https://www.aimspress.com/journal/BDIA

Research article

A multistage multimodal fusion network for cross-domain fake news detection

Qingfan Zhang* and Tenghui Wang

School of Artificial Intelligence and Big Data, Henan University of Technology, Zhengzhou 450001, China

* **Correspondence:** Email: 231210100523@stu.haut.edu.cn.

Abstract: With the rapid advancement of digital technologies and social media, news dissemination has become increasingly complex, often spanning multiple domains and encompassing diverse modalities (e.g., images and text). At the same time, the generation and spread of fake news have grown more covert and sophisticated, posing significant threats to social stability and public trust. In this context, traditional detection approaches that rely on a single modality or target a specific domain are no longer sufficient to handle the diversity and dynamic nature of real-world misinformation. To address this challenge, we propose an efficient multimodal, cross-domain fake news detection framework. Our framework incorporates a lightweight linear attention mechanism inspired by the Mamba model, along with a novel multimodal attention mechanism for efficient attention design and robust multimodal fusion. In addition, it integrates a masked linear attention module and a single-modality feature distiller to further capture modality-specific characteristics. Extensive experiments on two real-world datasets, together with comprehensive ablation studies, demonstrate that our model consistently outperforms existing methods and validates the necessity of each core component. Our source code is publicly available at: https://github.com/biubiu-cpp/zqf_Fake-News-Detection.

Keywords: fake news detection; cross-domain; cross-modal alignment; attention mechanism; multimodal fusion

1. Introduction

In recent years, the proliferation of fake news on the Internet has posed a serious threat to the health

of digital information environments. Malicious actors often employ sophisticated tactics, such as fabricating sensitive topics (e.g., social incidents or international political events) by piecing together disparate content or imitating legitimate news broadcasts, to deceive the public. The rapid dissemination and extensive reach of modern social media platforms further exacerbate this problem. Notably, fake news frequently contains images, making such multimodal content a significant vehicle for rumor propagation [1]. Consequently, developing efficient and accurate fake news detection systems is crucial to countering these threats, purifying cyberspace, and mitigating adverse effects on the public.

Early approaches to fake news detection primarily relied on textual content, often employing traditional machine learning [2] or deep learning methods such as convolutional neural networks (CNNs) [3]. However, content on modern social media is inherently multimodal, comprising rich images and text. Relying solely on text makes it difficult to accurately determine the veracity of a news piece. Consequently, recent research has increasingly focused on integrating textual and visual features to form a more comprehensive basis for judgment. Therefore, leveraging multimodal content for fake news detection represents a necessary and promising direction for future research.

Table 1. Data statistics.

Domain	Science	Military	Edu.	Disasters	Politics
Real	143	121	243	185	306
Fake	93	222	248	591	546
Total	236	343	491	776	852
Domain	Health	Finance	Ent.	Society	All
Real	485	959	1000	1198	4640
Fake	515	362	440	1471	4488
Total	1000	1321	1440	2669	9128

Recently, researchers have found that fake news has a strong domain dependence. News from different fields such as politics, health, and science and technology vary greatly in terms of content, motivation for fabrication, methods of fabrication, and psychological impact on readers. A general model attempts to detect all types of fake news, but the results are often not very good. Therefore, separating news by field and modeling for the characteristics of each field is a key strategy to improve detection results. However, this approach also brings a new problem: The amount of data in different fields is seriously unbalanced. For example, the Weibo21 dataset in Table 1 [4] shows that the number of news items in the "society" field is far greater than those in the fields of "science" and "military". This imbalance will lead to insufficient training of the model in fields with less data. In addition, the shared parameters in the model may be biased toward fields with more data and may even have a "negative transfer" effect on fields with less data. That is, when the shared information and the characteristics of the target field are too different, it will drag down the model's performance in that field [5].

In summary, current cross-domain, multimodal fake news detection faces three major challenges. First, there exists a semantic gap between different modalities (e.g., text and images). Second, intra-modality noise interference poses a significant issue, as multimodal data contains substantial high-dimensional noise that is irrelevant to veracity, necessitating careful filtering. Third, cross-domain data imbalance and negative transfer arise due to significant disparities in sample sizes across domains, which bias models toward data-rich domains and lead to poor performance in data-scarce ones. To address these challenges, we propose a comprehensive multistage fusion framework that provides a

unified approach to processing multimodal data. Our approach proceeds as follows. First, we employ pretrained models as feature extractors in the first stage to obtain high-quality, generalizable textual and visual features, thereby avoiding the substantial computational overhead of training from scratch. Second, we adopt a lightweight Mamba-like linear attention (MLLA) [6] model to perform efficient linear attention computation on visual features, reducing complexity while preserving spatial structure. Simultaneously, this model employs a masked linear attention mechanism to rapidly process critical textual segments. Third, we utilize unimodal feature distillers to individually optimize and enhance the features of each modality, ensuring that these modality-specific characteristics are not diluted during subsequent fusion. Finally, a multimodal fusion mechanism adaptively assigns weights to textual, visual, and their fused features, enabling the model to dynamically focus on the most informative type of information. This deep network architecture facilitates learning cross-domain generalizable feature representations, thereby mitigating the issues caused by data imbalance. The main contributions of this paper are summarized as follows:

1) To bridge the discrepancy between text and images, we map them into a unified semantic space using heterogeneous backbone networks (bidirectional encoder representations from transformers (BERT) and masked autoencoder (MAE)). Furthermore, we leverage the cross-modal alignment capability of contrastive language–image pretraining (CLIP) to obtain cross-modal features, thereby preliminarily addressing the semantic gap problem across different modalities.

2) To mitigate intramodality noise, we design two lightweight modules: (i) a masked linear attention mechanism to filter out invalid tokens in text sequences and (ii) an MLLA-Lite module to suppress visual noise while preserving spatial structure.

3) We introduce a squeeze-and-excitation multimodal fusion (SEMF) module, which learns to adapt modality weights across different domains. This mechanism prevents the model from over-relying on any single type of information and alleviates negative transfer to data-sparse domains. Concurrently, we employ a unimodal feature distiller (UFD) to preserve the unique characteristics of each modality prior to fusion.

4) To validate the effectiveness of our approach, we conduct extensive experiments on two real-world datasets and evaluate the contribution of each component through ablation studies.

2. Related works

2.1. Cross-domain fake news detection

To address the inherent generalization limitations of conventional single-domain approaches, multidomain fake news detection has increasingly become a pivotal area of inquiry. Initial investigations in this field concentrated on cross-domain knowledge transfer. For instance, the multidomain fake news detection model (MDFEND) [4] developed a domain gating network that dynamically modulates the contribution of textual features from disparate domains, thus effectively capturing cross-domain divergences. Subsequently, researchers have devised more refined mechanisms to simultaneously acquire domain-agnostic commonalities and domain-specific characteristics. A notable example is M3DFEND [7], which leveraged a memory network to facilitate the model's capacity to discern and represent multidomain information.

More recently, the research focus in this field has gravitated toward the fine-grained characterization of domain-specific attributes, alongside efforts to counteract the adverse effects of

negative transfer induced by data imbalance. Wu et al. [8] put forward a domain- and category-aware clustering paradigm that refines stylistic representation learning via multilevel contrastive loss, thereby bolstering both generalization and discriminative power in cross-domain settings. To tackle the persistent issue of negative transfer, Lu et al. [9] devised a domain disentanglement module that factorizes news representations into domain-invariant and domain-specific partitions. This design enables cross-domain knowledge propagation while safeguarding the distinctiveness of each individual domain. From a hierarchical transfer perspective, macro- and micro-hierarchical transfer learning framework (MMHTLF) [10] proposes a macro- and micro-hierarchical transfer framework that jointly captures domain-level commonalities (macro) and instance-level discriminative cues (micro). By aligning shared semantic patterns at the macro level while preserving domain-specific, fine-grained features at the micro level, this work demonstrates that hierarchical transfer effectively mitigates negative transfer in cross-domain scenarios.

In summary, these studies have advanced the field of multidomain fake news detection. The current trend emphasizes the use of more refined methods for disentangling and leveraging domain information to improve adaptability and accuracy.

2.2. Multimodal fake news detection

With the growing prevalence of misinformation in the form of multimodal content, encompassing both text and imagery, multimodal fake news detection has become a pivotal research domain. The fundamental objective of this line of work lies in assessing the semantic coherence between visual and textual modalities. Groundbreaking efforts, such as event adversarial neural networks (EANNs) [11], incorporated an event-adversarial discriminator to acquire event-agnostic representations, thereby establishing a foundational framework for subsequent multimodal investigations. Advancing along this direction, later research has increasingly focused on the integration of heterogeneous modalities. For instance, Wei et al. [12] utilized cross-modal knowledge distillation to supervise the training of modality-specific networks, with the goal of reinforcing representational alignment across modalities. Wu et al. [13] introduced an alternative perspective that synthesizes textual and visual cues by modeling users' heterogeneous browsing behaviors, subsequently inferring cross-modal semantic incongruities. Meanwhile, Ying et al. [14] put forward a multiview representation guidance mechanism designed to bolster the model's discriminative power from complementary angles. More recently, vision-language pre-training has been leveraged to improve cross-modal alignment. FACT-CLIP [15] pioneers CLIP-guided multimodal fusion, demonstrating that cross-modal similarity scores serve as a strong prior for detecting text–image inconsistency. From the explainability perspective, Sniffer [16] proposes a two-stage instruction tuning framework on multimodal large language models (LLMs) to generate natural language explanations alongside veracity predictions.

Given that multidomain and multimodal attributes naturally co-occur in practical settings, the integration of these two dimensions has emerged as a cutting-edge research direction. The knowledge augmented transformer for adversarial multidomain multiclassification multimodal fake news detection framework (KATMF) [17] pioneered this line of inquiry by utilizing a knowledge-augmented transformer in conjunction with adversarial multi-task learning to capture modality-specific feature distributions across varying domains. Advancing upon this groundwork, MMDFEND [18], proposed by Tong et al., achieved performance gains through a refined progressive hierarchical extraction module designed to model both shared inter-domain characteristics and domain-exclusive peculiarities.

The latest advancements, exemplified by domain-aware multimodal multiview fake news detection (DAMMFND) [9], place explicit emphasis on the dynamic role that domain information plays in shaping multimodal decision-making processes. This framework innovatively introduces a domain-aware multiview discriminator alongside a dedicated decision layer, which adaptively calibrates the contribution weights of textual, visual, and fused representations in the final verdict according to the domain affiliation of each news item. Such a mechanism facilitates more fine-grained and context-sensitive detection outcomes.

3. Methodology

In this section, we detail our proposed model (see Figure 1), which refines unimodal and cross-modal features across three stages, followed by an efficient weighted fusion for cross-domain multimodal fake news detection.

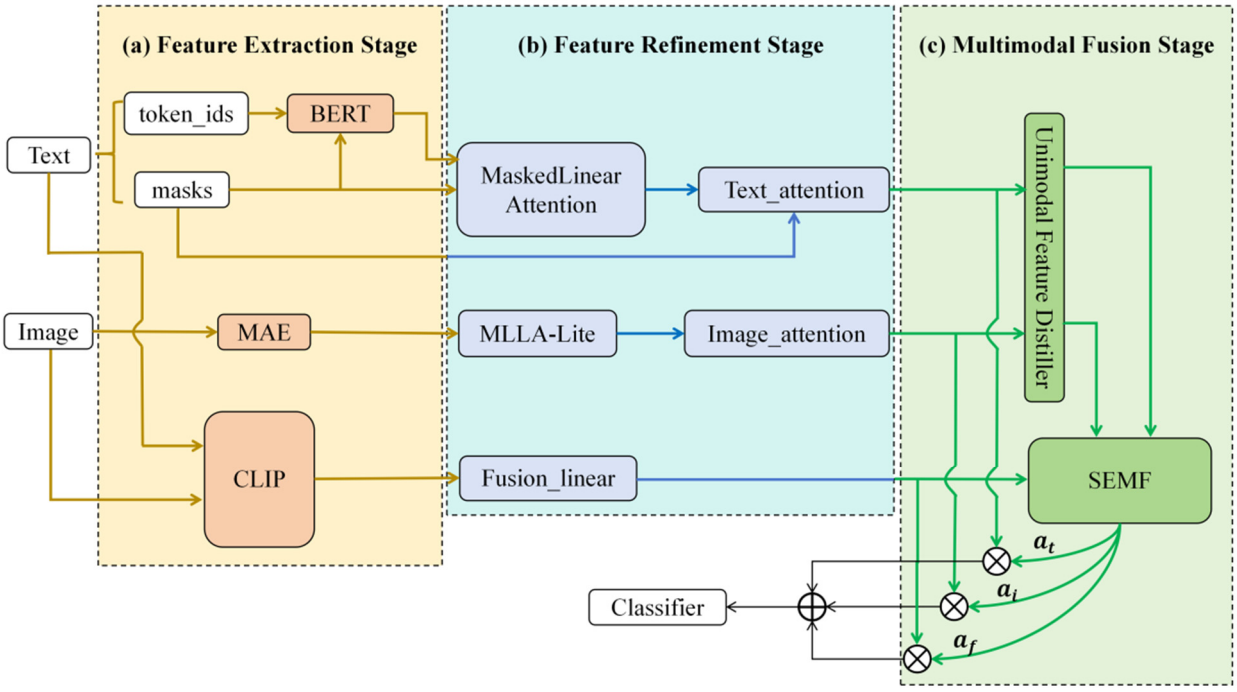

Figure 1. Our network structure.

3.1. Feature extraction stage

To overcome semantic discrepancies between unimodal features, we first employ CLIP to align textual and visual features. Given a text–image pair (T, I), for the text, we use BertTokenizer [22] to obtain token IDs T_i, and the corresponding attention mask M_t. These are fed into a pretrained BERT model [22] to obtain the textual representation e_t. For the image, we use a pretrained masked autoencoder [20] to extract deep visual features e_i. For cross-modal alignment, a significant semantic gap typically exists between the features derived from BERT and MAE, making direct fusion challenging. Therefore, we leverage the CLIP model [21], which is pretrained on a large corpus of

image-text pairs. We encode both the image and the text using CLIP's encoders to obtain aligned embedding representations f_{clip_t} and f_{clip_i}. These embeddings are passed through a multilayer perceptron (MLP) network, and their cosine similarity is computed. The final aligned multimodal features e_f are obtained via the following weighted sum:

$$\cos\_sim = \frac{f_{clip\_t} \cdot (f_{clip\_i})^T}{\| f_{clip\_t} \| f_{clip\_i} \|} \tag{1}$$

$$e_f = \cos\_sim \cdot MLP(f_{clip\_t} \oplus f_{clip\_i}) \tag{2}$$

3.2. Feature refinement stage

3.2.1. Text feature refinement

To suppress noise and focus on informative tokens in textual data, we design a masked linear attention mechanism. Given textual features e_t and a mask M_t, the masked linear attention mechanism first applies a linear transformation to the textual features to obtain attention scores A (where $A \in \mathbb{R}^{B\times L\times d}$). Then, the token-level attention weights $\alpha \in \mathbb{R}^{B\times L\times d}$ are computed via Softmax as follows:

$$A = e_t \cdot W^T + b \tag{3}$$

$$\alpha = \frac{exp(A_{ijl})}{\sum_{k=1}^{L} exp(A_{ijk})} \tag{4}$$

Here, (i, j, l) denote the sequence length, feature dimension, and batch size, respectively. Unlike traditional self-attention, which computes QK^T via linear projections, we directly transform the features obtained from BERT and apply Softmax. Our rationale is that BERT already encodes rich contextual information, making direct optimization of these features more efficient. Finally, we incorporate the mask to obtain the final output:

$$f'_{text} = A \odot (\alpha \odot M_t) \tag{5}$$

The resulting output f'_{text} has the same dimensions as the input e_t, yet critical information is enhanced, and noise is suppressed. We posit that different tokens contribute unequally to the information content. This module focuses on the distribution of token importance within the sequence while preventing invalid tokens from interfering with the optimization process, and it is also computationally lightweight. However, the detection task additionally requires a global semantic feature. Therefore, we next employ a Text_attention module to compress the refined features into a global vector. This module maps the refined features f'_{text} to compute a scalar score $s \in \mathbb{R}^{B\times L}$ for each token. Then, based on the mask M_t, the original scores of valid tokens are retained, and the scores of padding positions are set to negative infinity, ensuring that their weights approach zero after Softmax. Finally, the attention weights β are obtained via Softmax normalization, and a weighted sum is performed using these weights to obtain the final textual feature $f_{text} \in \mathbb{R}^{B\times d}$:

$$s = f'_{text} \cdot W_a + b_a \tag{6}$$

$$\beta = softmax(s, dim = -1) \in \mathbb{R}^{B \times L} \tag{7}$$

$$f_{text} = \sum_{j=1}^{L} \beta_j f_j \in \mathbb{R}^{B \times d} \tag{8}$$

3.2.2. Image feature refinement

Similarly, to optimize visual features while maintaining computational efficiency, we propose MLLA-Lite. Prior to this, we note that Mamba [19] has been effective for high-resolution visual tasks. However, although the standard Mamba excels at handling long sequences via a data-dependent state-space model (SSM), its effectiveness diminishes when processing low-resolution image patches commonly found in news images, as the strict ordering of pixels is far less important than local spatial consistency. This motivates us to adopt MLLA (Mamba-like linear attention) [6], which integrates forget gates and a modular design into the linear attention framework. MLLA has demonstrated superior performance over Mamba on visual tasks while maintaining parallelizability and faster inference speed. Nevertheless, we observe that the original implementation of MLLA is relatively complex and includes certain task-specific components, which hinders modular analysis and efficient deployment.

To address this issue, we propose MLLA-Lite, a lightweight and decoupled variant of the linear attention mechanism. Compared with the original MLLA, MLLA-Lite incorporates three key design differences to better suit cross-domain fake news detection.

First, linearity and efficiency: By adopting a simple single-projection mechanism for q and k and reusing the input v, MLLA-Lite reduces the number of parameters by 22% compared to standard Mamba. Specifically, whereas the original MLLA employs a task-coupled multi-stage projection, our simplified version is formulated as

$$Attention(q, k, x) = softmax\left(\frac{qk^T}{\sqrt{d}}\right)x \tag{9}$$

This is equivalent to a data-dependent smoothing operation.

Second, robustness of the activation function: We replace the original Gaussian error linear unit (GELU) with ELU+1.0. This modification provides a nonzero gradient for non-negative values, thereby stabilizing the training of multimodal features while reducing computation time by 15%.

Third, dynamic resolution support: Unlike fixed-size rotary position embeddings, our improved diffusion space rotatory positional embedding (D-RoPE) enables the model to dynamically compute the rotation matrix during training, thereby directly accommodating the various image aspect ratios commonly encountered on social media.

Furthermore, we simplify the local feature pathway by retaining only LayerNorm and switching to grouped depthwise convolution. The computational complexity of MLLA-Lite is $O(L \cdot d \cdot (d / h))$, where L is the number of tokens, h is the number of attention heads, and d is the dimension per head. In summary, compared with standard Mamba and the original MLLA, MLLA-Lite achieves a better trade-off between efficiency and robustness in the fake news detection context, where images are often low-resolution and spatially noisy, rather than merely differing in structure.

Similar to the text pipeline, the output of MLLA-Lite, denoted as $f'_{image} = MLLA\_Lite(e_i)$, presents the refined local features. We design an Image_attention module to compress this token sequence into a global feature vector. This module computes un-normalized importance scores using a two-layer linear transformation with sigmoid linear unit (SiLU) activation, which are then used for

the weighted sum as follows:

$$scores = W_2 \cdot SiLU(W_1 \cdot f'_{image} + b_1) + b_2 \tag{10}$$

$$f_{image} = scores^T \cdot f'_{image} \tag{11}$$

3.2.3. Fusion feature projection

To enable the multimodal features e_f to be better processed by the subsequent SEMF module, we use a standard two-layer nonlinear projection network to reduce its dimensionality:

$$f_{fusion} = RELU\left(W_2 \cdot RELU(W_1 \cdot e_f + b_1) + b_2\right) \tag{12}$$

3.3. Multimodal fusion stage

To address domain shift and data imbalance, we introduce two complementary mechanisms: a UFD, which preserves the unique information of each modality prior to fusion, and an SEMF module, which learns to adapt modality weights across different domains. Not all modalities are equally important for every decision, particularly when a news piece spans multiple domains with varying cross-modal characteristics. To tackle domain shift and data imbalance, we employ a domain adaptation strategy to reweight these features.

First, we employ UFD. This step independently optimizes the textual and visual features, filters out domain-specific noise, and employs batch normalization to normalize the numerical range of the features. Taking the textual feature f_{text} as an example (the resulting image feature z_{image} is processed similarly):

$$h_{text} = RELU\left(BN(W_1 \cdot f_{text})\right) \tag{13}$$

$$z_{text} = RELU\left(BN(W_2 \cdot h_{text})\right) \tag{14}$$

Subsequently, inspired by SE-Net [23], we designed the SEMF module to learn adaptive modality weights (see Figure 2). We name this module SEMF because it inherits the "squeeze-and-excitation" concept from SE-Net, that is, recalibrating the importance of features at the channel level, and extends it to the multimodal fusion scenario. Specifically, for each text–image pair, SEMF dynamically assigns higher weights to the modality (text, image, or their fused features) that is more indicative of fake news within the current domain, while suppressing less informative modalities. Compared with the standard SE-Net, which is designed for unimodal channelwise attention, our SEMF is constructed for cross-modal importance assignments. Theoretically, SEMF differs from SE-Net in that we integrate a 24-layer deep residual network into it. Whereas standard SE-Net typically recalibrates channels within a single layer, our deep stacking enables the network to learn higher-order interactions across modalities. This stands in contrast to existing multimodal fusion approaches such as bootstrapping multiview representations (BMR) [14] and multireading habits fusion reasoning networks (MRHFR) [13], which rely on single-layer attention or shallow gating to compute modality weights. A single-step dot-product or linear scoring, as employed in these methods, can only capture first-order cross-modal correlations; in our 24-layer architecture, each residual block incrementally refines channelwise importance upon the output of the previous layer, thereby composing elementary modality interactions into domain-

level discriminative patterns, akin to how deep vision networks hierarchically learn from edges to textures to object parts. The SEMF architecture is specifically designed to mitigate the negative transfer problem by distilling domain-invariant features. It first concatenates features from three modalities as follows:

$$F = Concat(z_{text}, z_{image}, f_{fusion}) \tag{15}$$

A one-dimensional convolution projects the concatenated features F into a unified space. This is followed by a deep network composed of 24 stacked residual blocks (ResBlocks) integrated with SE mechanisms. From a cross-domain perspective, the deep residual structure facilitates the extraction of higher-level, abstract features that are more robust to stylistic variations across different news domains.

The SE mechanism acts as a dynamic "gating" modulator, identifying domain-agnostic and generalizable feature channels, thereby enabling the network to focus on the most reliable modalities for the given sample. Furthermore, we introduce a learnable channel bias b, which allows the model to retain a baseline level for each modality even when the excitation signal is weak. Shallow gating mechanisms (e.g., in BMR) lack such a fallback: When domain shift renders one modality temporarily unreliable, the gate may suppress it entirely, causing what we term "modality collapse". The bias b in Eq (16) provides a trainable baseline that prevents any modality from being zeroed out, which is critical for data-sparse domains where the model must retain access to all available signals. This adaptive weighting facilitates knowledge transfer through shared channel-level feature importance, thereby mitigating the model's bias toward data-rich domains and ensuring robust modality calibration even for data-sparse domains. The detailed formulation of the above process is as follows:

$$SEBlock(H) = \sigma\left(W_2 \cdot RELU(W_1 \cdot avgpool(H))\right) \odot H + b \tag{16}$$

$$H_{n+1} = RELU\left(H_n + SEBlock\left(Conv1D\left(RELU(Conv1D(H_n))\right)\right)\right) \tag{17}$$

$$\{a_t, a_i, a_f\} = log\text{-}softmax\left(W_{fc} \cdot avgpool\left(Conv1D(H_N)\right)\right) \tag{18}$$

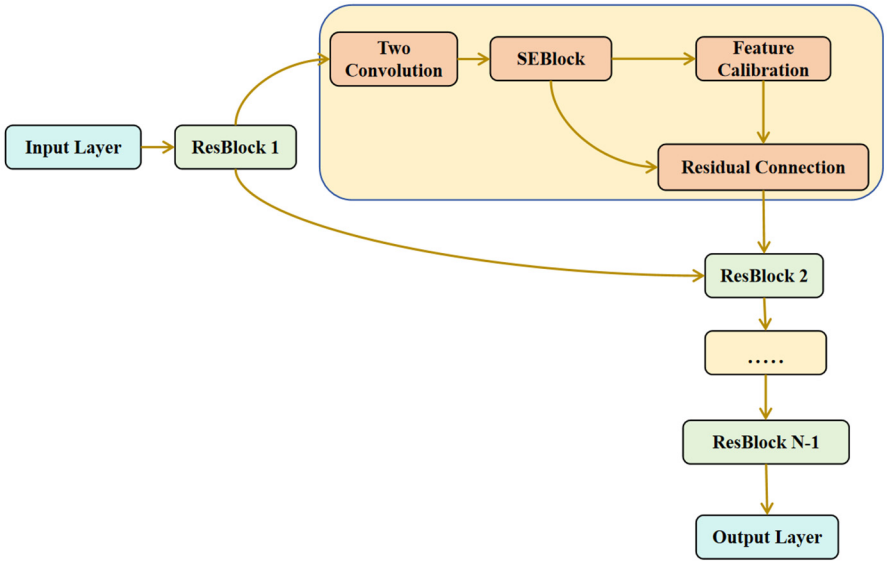

Figure 2. SEMF network structure.

3.4. Classifier and loss function

The refined features ($f_{text}, f_{image}, f_{fusion}$) and their weights (a_t, a_i, a_f) are combined in the following way:

$$f_{final} = (a_t \times f_{text}) \oplus (a_i \times f_{image}) \oplus (a_f \times f_{fusion}) \tag{19}$$

The $\oplus$ represents the concatenation operation. The combined vector will then be fed into a multilayer MLP for classification:

$$\hat{y} = softmax\left(MLP(f_{final})\right) \tag{20}$$

We use the binary cross-entropy loss function $\mathcal{L}$. For intermediate features (text, image, and fused features), we also use the binary cross-entropy loss function to calculate the loss. The total loss $\mathcal{L}_{loss}$ is a weighted sum of the final loss and the auxiliary losses from the intermediate features (text, image, and fused features), with weights as follows: $\lambda_1 = 0.4, \lambda_2 = 0.2, \lambda_3 = 0.2, \lambda_4 = 0.2$;

$$\mathcal{L}_{loss} = \lambda_1 \mathcal{L}_{final} + \lambda_2 \mathcal{L}_{text} + \lambda_3 \mathcal{L}_{image} + \lambda_4 \mathcal{L}_{fusion} \tag{21}$$

4. Experiment

4.1. Experimental setup

We evaluate our model on two real-world datasets, Weibo [11] and Weibo21 [4]. For the Weibo dataset, we adopt the original split (training set: 3749 real/3783 fake; test set: 1000 fake/996 real) and follow the domain labels used in [18]. The Weibo21 dataset consists of 4640 real news items and 4488 fake news items, partitioned according to the benchmark [4]. For each dataset, 70% of the samples in each class are used for training, 15% for testing, and 15% for validation.

All experiments are conducted using the PyTorch framework on a system equipped with an NVIDIA RTX 4090 GPU (24 GB memory), 64 GB RAM, and an Intel Core i9-14900K CPU. The detailed hyperparameter configurations are as follows:

Optimizer and learning rate: we use the Adam optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 5×10^{-5}.

Attention mechanism: for the MLLA module used for text processing, the hidden size is set to 768. For the MLLA-Lite module used for image processing, the number of attention heads is set to 8, the dimension per head is 96, and the total feature dimension is 768.

Module details: the SEMF module stacks 24 residual blocks, with the reduction ratio of its internal SE mechanism set to 16. The output dimensions of BERT and MAE are 768, and the projection alignment dimension of CLIP is also 768.

Training procedure: the batch size is set to 32. We adopt an early stopping strategy that terminates training when the model performance does not improve for more than 5 epochs, in order to prevent overfitting.

4.2. Baseline

To demonstrate the advantages of our algorithm, we compare its performance against three groups of baseline models. The first group comprises unimodal, cross-domain methods, namely multigate

mixture-of-experts (MMoE) [24], multitask mixture of sequential experts (MoSE) [25], MDFEND [4], and M3FEND [7]. The second group includes multimodal, single-domain methods, such as SpotFake [26], cross-modal ambiguity feature/fusion (CAFE) [27], cross-modal feature correlations (CMC) [12], BMR [14], and MRHFR [13]. Finally, the third group consists of multimodal, cross-domain models, specifically KATMF [17] and MMDFND [18]. The performance results for the first and third groups are presented in Table 2, and the results for the second group are shown separately in Table 3.

4.3. Results and analysis

As shown in Table 2, multimodal cross-domain methods (KATMF, MMDFND, and our method) outperform unimodal cross-domain methods, indicating the importance of multimodal information for cross-domain detection. Our method achieves superior performance over state-of-the-art models across eight domains on Weibo and seven domains on Weibo21. In terms of overall F1 and accuracy, our method reaches the optimal level, improving the accuracy by 0.2% over the second-best model on Weibo and by 0.4% on Weibo21.

Table 3 shows that our model also significantly outperforms single-domain multimodal methods. Even without considering domain information, our model achieves 93.6% accuracy on Weibo and 94.3% accuracy on Weibo21. This indicates that the compared methods cannot effectively capture cross-domain and domain-specific knowledge, thereby highlighting the superiority of our approach.

We attribute the effectiveness of our model to three specific design choices, each addressing a distinct challenge and making a measurable contribution to the final performance.

First, CLIP-based cross-modal alignment bridges the semantic gap between different backbone feature extractors. Traditional multimodal methods directly fuse features from BERT (for text) and Visual Geometry Group (VGG) or MAE (for images) without explicit alignment. However, these backbones are trained on different tasks (e.g., masked language modeling and masked image reconstruction), resulting in features with substantially different semantics that hinder effective fusion. We address this by introducing a CLIP-based alignment step (Eqs (1) and (2)), which maps features from both modalities into a shared semantic space learned from 400 million image–text pairs. This alignment reduces cross-modal interference, a phenomenon where misaligned features may cancel each other out during fusion, thereby degrading performance.

Second, progressive intramodality noise suppression is achieved through MLLA and MLLA-Lite. Fake news detection requires identifying subtle inconsistencies that are often hidden within irrelevant information. For text, the MLLA mechanism (Eqs (3)–(5)) focuses on informative tokens while masking out padding and stop words. For images, MLLA-Lite (Eq (9)) suppresses visual noise, such as background clutter and compression artifacts, while preserving the spatial structure critical for detecting image manipulation.

Third, domain-adaptive modality weight redistribution via SEMF prevents negative transfer. In cross-domain detection, the reliability of each modality changes systematically across domains. In text-dominated domains, textual cues are more discriminative; in domains where images are manipulated, anomalies in images are more informative; and in data-sparse domains (e.g., science and military on Weibo21), the model must rely on shared cross-domain knowledge. Standard multimodal models use static fusion weights (e.g., simple concatenation or fixed attention) that cannot adapt to these variations. SEMF addresses this issue through its squeeze-and-excitation mechanism, which learns sample-specific, domain-adaptive weights. The 24-layer deep residual architecture enables

SEMF to capture higher-order cross-modal feature interactions, which is essential for distinguishing domain-generalizable features from domain-specific noise. This dynamic weight redistribution prevents the model from over-relying on any single modality that may be less informative in a particular domain (i.e., modality collapse) and alleviates negative transfer from data-rich to data-sparse domains.

Table 2. Baseline experimental results.

	Method	Sci.	Mil.	Edu.	Soc.	Pol.	Hlth.	Fin.	Ent.	Dis./Int	overall		
											F1	Acc	Auc
Weibo	MMoE	0.578	0.911	0.851	0.885	0.735	0.826	0.813	0.830	0.883	0.874	0.874	0.950
	MoSE	0.793	0.738	0.834	0.912	0.764	0.859	0.791	0.844	0.883	0.890	0.890	0.954
	MDFEND	0.774	0.911	0.897	0.902	0.763	0.878	0.808	0.881	0.874	0.904	0.904	0.965
	M^{3}FEND	0.792	0.903	0.923	0.912	0.765	0.863	0.899	0.899	0.876	0.928	0.928	0.969
	KATMF	0.831	0.908	0.924	0.895	0.823	0.898	0.903	0.904	0.894	0.929	0.930	0.969
	MMDFND	0.824	0.911	**0.941**	0.939	0.735	0.913	0.917	0.917	0.888	0.934	0.934	0.972
	Ours	**0.826**	**0.938**	0.936	**0.977**	**0.897**	**0.952**	**0.985**	**0.982**	**0.897**	**0.936**	**0.936**	**0.990**
Weibo21	MMoE	0.875	0.911	0.870	0.875	0.862	0.936	0.856	0.888	0.877	0.894	0.894	0.954
	MoSE	0.850	0.885	0.881	0.872	0.880	0.917	0.867	0.891	0.867	0.893	0.894	0.954
	MDFEND	0.830	0.938	0.891	0.898	0.886	0.940	0.895	0.906	0.900	0.913	0.913	0.970
	M^{3}FEND	0.829	0.950	0.899	0.908	0.882	**0.946**	0.900	0.931	0.889	0.921	0.921	0.975
	KATMF	0.914	0.928	0.913	0.895	0.902	0.914	0.871	0.937	0.898	0.923	0.928	0.975
	MMDFND	0.937	**0.953**	0.852	0.945	0.965	0.920	0.884	0.959	0.919	0.939	0.939	0.977
	Ours	**0.938**	0.870	**0.931**	**0.977**	**0.970**	0.937	**0.936**	**0.982**	**0.919**	**0.943**	**0.943**	**0.982**

Table 3. Comparison of our method with multimodal single-domain fake news detection methods.

	Method	Accuracy	F1 score	
			Real News	Fake News
Weibo	SpotFake	0.892	0.932	0.739
	CAFE	0.840	0.842	0.837
	CMC	0.893	0.899	0.907
	BMR	0.918	0.914	0.904
	MRHFR	0.922	0.918	0.910
	Ours	**0.936**	**0.934**	**0.937**
Weibo21	SpotFake	0.851	0.828	0.866
	CAFE	0.882	0.885	0.876
	CMC	0.897	0.903	0.912
	BMR	0.929	0.927	0.925
	MRHFR	0.932	0.928	0.926
	Ours	**0.943**	**0.942**	**0.943**

Figure 3 provides a T-distributed stochastic neighbor embedding (T-SNE) visualization of the features learned by our model. The boundaries between different clusters are clear, intuitively demonstrating the powerful discriminative ability of the learned features.

4.4. Ablation study

To validate the effectiveness of each component, we conducted a comprehensive ablation study (Table 4). We evaluated the impact of removing the following components: MLLA, Text/Image Attention (T/I_att), UFD, SEMF, MLLA-Lite, and CLIP.

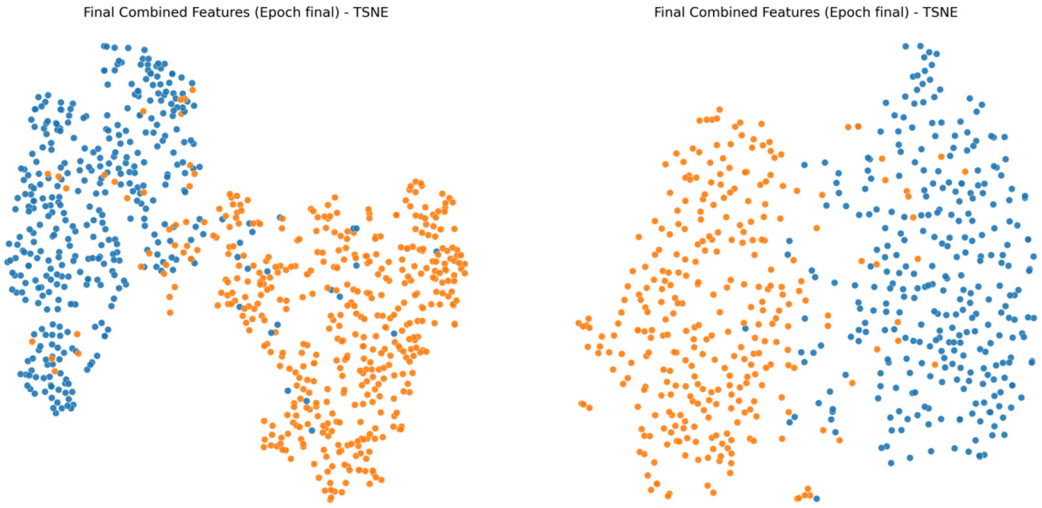

Figure 3. T-SNE visualization of learned features at convergence. Left: Weibo, Right: Weibo21.

Table 4. Ablation studies of our model on two datasets.

Module						Weibo			Weibo21		
Clip	MLLA	MLLA-Lite	T/I_att	UFD	SEMF	Acc	F1 score		Acc	F1 score	
							Real	Fake		Real	Fake
√	√	√	√	√	√	**0.936**	**0.934**	**0.938**	**0.943**	**0.942**	**0.943**
√		√	√	√	√	0.922	0.916	0.937	0.918	0.919	0.918
√	√		√	√	√	0.923	0.917	0.928	0.917	0.917	0.917
√	√	√		√	√	0.916	0.909	0.922	0.913	0.912	0.914
√	√	√	√		√	0.922	0.915	0.927	0.926	0.927	0.926
√	√	√	√	√		0.909	0.904	0.913	0.913	0.913	0.914
	√	√	√	√	√	0.818	0.817	0.813	0.826	0.828	0.826
√	√	√	√			0.898	0.895	0.901	0.907	0.902	0.912
√			√	√	√	0.925	0.921	0.928	0.922	0.914	0.929
√				√	√	0.902	0.897	0.905	0.915	0.904	0.923

As shown in Table 4, our full model consistently outperforms all ablated variants, demonstrating that each proposed component makes a unique contribution. Specifically, removing either the local optimization modules (MLLA or MLLA-Lite) or the global feature extraction modules (T/I_attention) severely impairs performance, indicating that both local feature refinement and robust global representation are necessary prior to aggregation. Notably, removing SEMF (i.e., the domain-adaptive modality weighting module) leads to the most significant performance drop: 2.7% on Weibo and 3.0%

on Weibo21. This result strongly verifies the critical role of SEMF's dynamic modality weighting capability. Furthermore, the removal of CLIP-based cross-modal alignment yields the most dramatic collapse in performance, with F1 scores plunging to approximately 0.81–0.83 on both datasets, underscoring that aligning BERT and MAE features into a shared semantic space is a fundamental prerequisite without which all subsequent fusion stages become ineffective. Removing the entire feature refinement stage results in a substantial overall performance decline, further confirming the necessity of purifying features before fusion. In contrast, removing the UFD yields a smaller drop, 1.4% on Weibo and 1.7% on Weibo21, suggesting that its primary role may be to accelerate model convergence and prevent modality features from being diluted during fusion. Finally, the variant that removes both UFD and SEMF confirms that simple feature concatenation performs poorly. Collectively, these components exhibit synergistic effects, validating the rationality of our architectural design.

4.5. Modality weight allocation analysis of SEMF

To verify that SEMF indeed learns domain-adaptive modality weights, we record the averaged (a_t, a_i, a_f) output by SEMF for samples from each domain on the test set. As illustrated in Figure 4, the weight patterns differ markedly across domains. For example, in text-dominated domains such as Finance ($a_t = 0.45$, $a_i = 0.20$) and Education ($a_t = 0.44$, $a_i = 0.23$), text features receive significantly higher weights than image features; Conversely, in the domain Disasters ($a_t = 0.19$, $a_i = 0.49$), where image manipulation is the primary falsification method, image features are assigned the highest weights. This confirms that SEMF dynamically adjusts modality importance based on domain characteristics.

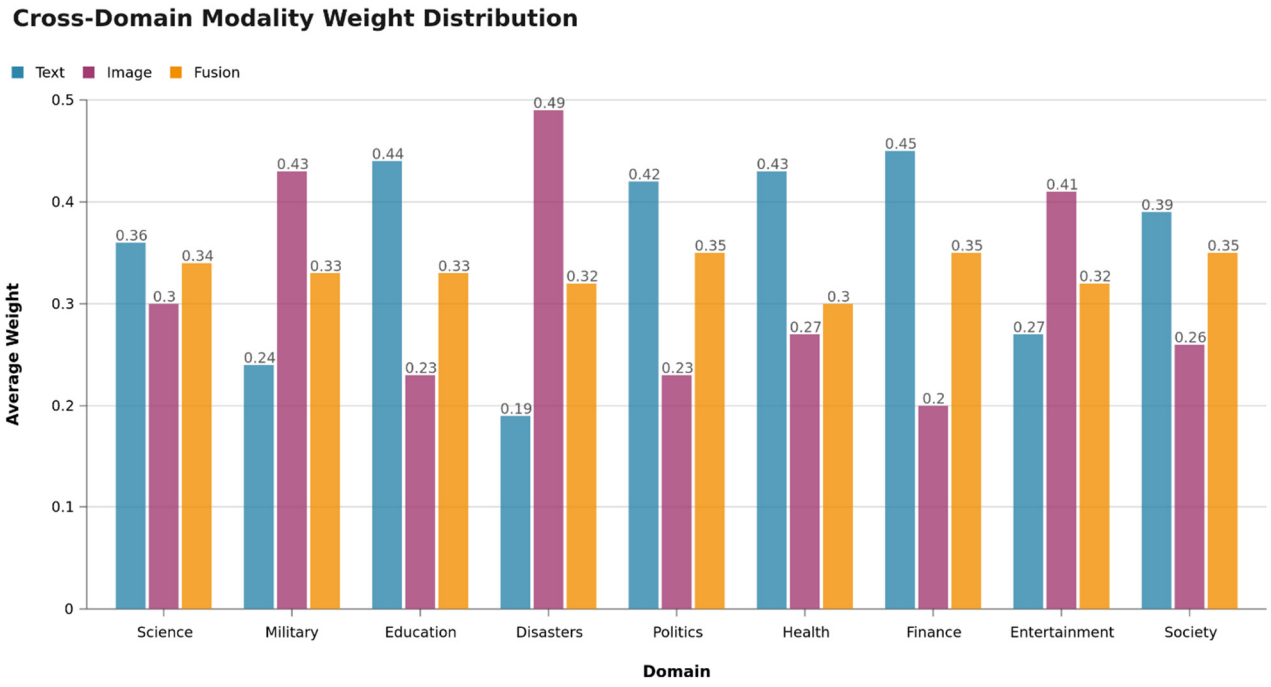

Figure 4. Cross-domain modality weight distribution.

4.6. Parameter sensitivity analysis

To evaluate the stability and robustness of the proposed model, we performed a sensitivity analysis on the hyperparameters in the multiobjective loss function (Eq (21)). The coefficient values are in the same order as in the formula, including the options for average weighting and setting the weight coefficient of a certain loss to 0 (i.e., removing that loss). The results are shown in Table 5.

Table 5. λ Hyperparameter experiments.

	λ_1 / λ_2 / λ_3 / λ_4	Acc	Auc	F1
Weibo	0.4/0.2/0.2/0.2	**0.936**	**0.990**	**0.936**
	0.25/0.25/0.25/0.25	0.927	0.966	0.927
	0.7/0.1/0.1/0.1	0.922	0.959	0.921
	0.8/0.2/0/0	0.911	0.941	0.911
	0.8/0/0.2/0	0.919	0.937	0.920
	0.8/0/0/0.2	0.933	0.981	0.933
Weibo21	0.4/0.2/0.2/0.2	**0.943**	**0.982**	**0.943**
	0.25/0.25/0.25/0.25	0.931	0.989	0.931
	0.7/0.1/0.1/0.1	0.929	0.976	0.930
	0.8/0.2/0/0	0.915	0.951	0.915
	0.8/0/0.2/0	0.927	0.973	0.927
	0.8/0/0/0.2	0.934	0.981	0.934

Experiments show that when the coefficients are set to 0.4/0.2/0.2/0.2, the model achieves optimal performance on both datasets. When the weight assignment becomes extreme, model performance degrades significantly. Equal weight assignment yields performance between the optimal combination and the extreme ones. We attribute the optimal performance under the 0.4/0.2/0.2/0.2 weighting to the following reasons: The fusion auxiliary loss, which directly acts on the multimodal fused features, is most relevant to the final classification task and thus contributes the most; the textual and visual auxiliary losses have relatively smaller impacts, yet removing any one of them leads to performance degradation, indicating that together they constitute an effective multitask learning framework.

4.7. Complexity analysis

As shown in Table 6, our MLLA-Lite module has fewer parameters than the original Mamba, leading to a corresponding reduction in computation time. Although SEMF has more parameters than the original SE-Net, the computation time does not increase significantly, and the F1 score is substantially improved compared to using the original SE-Net.

4.8. Statistical significance and stability analysis

To verify the reliability and robustness of the proposed framework, we conducted multiple independent runs on both datasets using six different random seeds. Figure 5 reports the accuracy of our model across six independent runs with different random seeds. As shown, the variance on both datasets is 0.0001, indicating that our model maintains stable performance and is not particularly sensitive to random initialization.

Table 6. Comparison of module complexity.

Module	Computational Complexity	Parameters	Latency (ms)	F1
Standard Mamba	$O(L \cdot d^2)$	1.2M	14.2	0.883
MLLA-Lite(Ours)	$O(L \cdot d \cdot \frac{d}{h})$	0.9M	11.8	0.936
Standard SE-Net	$O(C^2/r)$	0.05M	1.5	0.891
SEMF(Ours)	$O(N_{res}(C^2/r))$	0.45M	3.2	0.936

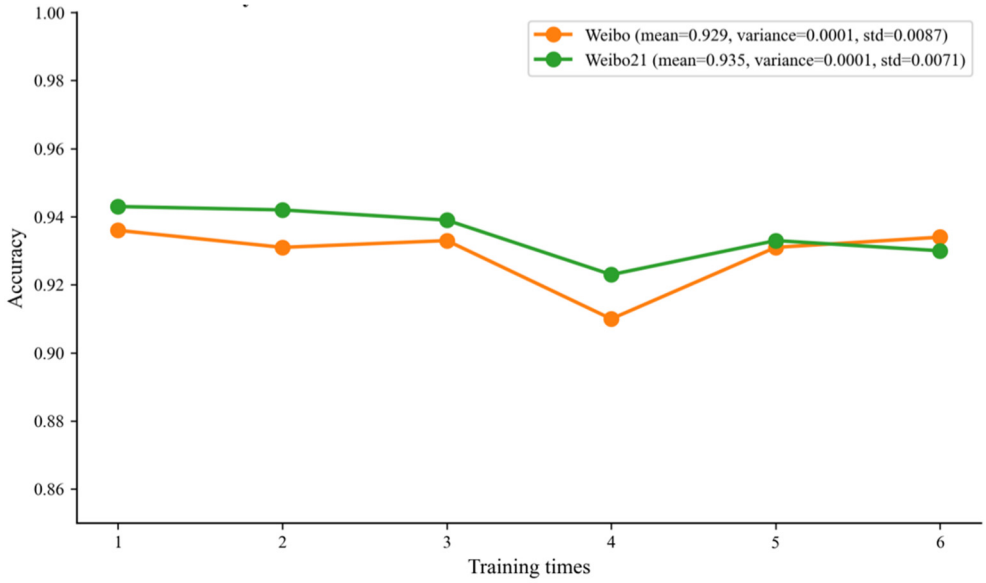

Figure 5. Accuracy stability under different random seeds.

5. Conclusions

In this paper, we have proposed an efficient multimodal fusion framework specifically designed for cross-domain fake news detection. By integrating a lightweight MLLA-Lite module for visual refinement and a masked linear attention mechanism for textual analysis, our model effectively suppresses modality-specific noise. The core contribution, the SEMF module, enables dynamic, context-aware weight assignment, successfully addressing the domain shift problem that has plagued traditional detection methods. Experimental results on multiple real-world datasets demonstrate that our approach significantly outperforms existing state-of-the-art baseline models.

Owing to its lightweight architecture and high cross-domain adaptability, our framework can be deployed on news aggregation platforms to provide automatic credibility scores for emerging events with sparse historical data. Furthermore, the interpretability of the SEMF module allows content moderators to understand which modality contributes more to a suspicious rating, thereby facilitating a more transparent fact-checking process.

Despite its effectiveness on the two datasets evaluated, our model has certain limitations.

Currently, it primarily focuses on image–text pairs and does not account for video-based misinformation or complex social graph metadata. Future research will explore the integration of temporal propagation features and the application of large language models to enhance the framework's semantic reasoning capabilities. In addition, we plan to extend the model to handle finer-grained categories of fake news, such as satire or unintentional misinformation.

Authorship contribution statement

Qingfan Zhang: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation. Tenghui Wang: Visualization, Investigation, Data curation.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgements

We are deeply grateful to the researchers who shared their data and the students, teachers, and staff who helped us.

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. Jing J, Wu H, Sun J, Fang X, Zhang H, (2023) Multimodal fake news detection via progressive fusion networks. *Inf Process Manag* 60: 103120. https://doi.org/10.1016/j.ipm.2022.103120
2. Liu X, Nourbakhsh A, Li Q, Fang R, Shah S, (2015) Real-time rumor debunking on twitter. In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 1867–1870. https://doi.org/10.1145/2806416.2806651
3. Yu F, Liu Q, Wu S, Wang L, Tan T, (2017) A convolutional approach for misinformation identification. In: *IJCAI*, 3901–3907. https://doi.org/10.24963/ijcai.2017/545
4. Nan Q, Cao J, Zhu Y, Wang Y, Li J, (2021) MDFEND: Multi-domain fake news detection. In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 3343–3347. https://doi.org/10.1145/3459637.3482139
5. Zhang W, Deng L, Zhang L, Wu D, (2022) A survey on negative transfer. *IEEE/CAA J Autom Sin* 10: 305–329. https://doi.org/10.1109/JAS.2022.106004
6. Han D, Wang Z, Xia Z, Han Y, Pu Y, Ge C, et al. (2024) Demystify mamba in vision: A linear attention perspective. *Adv Neural Inf Process Syst* 37: 127181–127203. https://doi.org/10.52202/079017-4039
7. Zhu Y, Sheng Q, Cao J, Nan Q, Shu K, Wu M, et al. (2022) Memory-guided multi-view multi-domain fake news detection. *IEEE Trans Know Data Eng* 35: 7178–7191. https://doi.org/10.1109/TKDE.2022.3185151

8. Wu D, Tan Z, Zhao H, Jiang T, Geng N, (2024) Domain-and category-style clustering for general fake news detection via contrastive learning. *Inf Process Manag* 61: 103725. https://doi.org/10.1016/j.ipm.2024.103725
9. Lu W, Tong Y, Ye Z, (2025) Dammfnd: Domain-aware multimodal multi-view fake news detection. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 39: 559–567. https://doi.org/10.1609/aaai.v39i1.32036
10. Yang X, Wang Y, Zhang X, Wang S, Wang H, Lam K, (2025) A macro-and micro-hierarchical transfer learning framework for cross-domain fake news detection. In: *Proceedings of the ACM on Web Conference 2025*, 5297–5307. https://doi.org/10.1145/3696410.3714517
11. Wang Y, Ma F, Jin Z, Yuan Y, Xun G, Jha K, et al. (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*, 849–857. https://doi.org/10.1145/3219819.3219903
12. Wei Z, Pan H, Qiao L, Niu X, Dong P, Li D, (2022) Cross-modal knowledge distillation in multi-modal fake news detection. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, 4733–4737. https://doi.org/10.1109/ICASSP43922.2022.9747280
13. Wu L, Liu P, Zhang Y, (2023) See how you read? multi-reading habits fusion reasoning for multi-modal fake news detection. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, 37: 1373613744. https://doi.org/10.1609/aaai.v37i11.26609
14. Ying Q, Hu X, Zhou Y, Qian Z, Zeng D, Ge S, (2023) Bootstrapping multi-view representations for fake news detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, 37: 5384–5392. https://doi.org/10.1609/aaai.v37i4.25670
15. Nasir S, Wasim M, Rehman A, Ghani MU, (2025) FACT-CLIP: Fake news detection via CLIP-based cross-modal attention and transformer fusion. In: *2025 International Conference on Emerging Technologies in Electronics, Computing, and Communication (ICETECC)*, 1–6. https://doi.org/10.1109/ICETECC65365.2025.11070224
16. Qi P, Yan Z, Hsu W, Lee ML, (2024) Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13052–13062. https://doi.org/10.1109/CVPR52733.2024.01240
17. Song C, Ning N, Zhang Y, Wu B, (2021) Knowledge augmented transformer for adversarial multidomain multiclassification multimodal fake news detection. *Neurocomputing* 462: 88–100. https://doi.org/10.1016/j.neucom.2021.07.077
18. Tong Y, Lu W, Zhao Z, Lai S, Shi T, (2024) Mmdfnd: Multi-modal multi-domain fake news detection. In: *Proceedings of the 32nd ACM International Conference on Multimedia*, 1178–1186. https://doi.org/10.1145/3664647.3681317
19. Koroteev MV, (2021) BERT: A review of applications in natural language processing and understanding, preprint, arXiv:2103.11943. https://doi.org/10.48550/arXiv.2103.11943
20. He K, Chen X, Xie S, Li Y, Dollár P, Girshick R, (2022) Masked autoencoders are scalable vision learners. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16000–16009. https://doi.org/10.1109/CVPR52688.2022.01553
21. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. (2021) Learning transferable visual models from natural language supervision. In: *International Conference on Machine Learning*, 8748–8763.
22. Gu A, Dao T, (2024) Mamba: Linear-time sequence modeling with selective state spaces, In: *First Conference on Language Modeling*.

23. Hu J, Shen L, Sun G, (2018) Squeeze-and-excitation networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7132–7141. https://doi.org/10.1109/CVPR.2018.00745
24. Ma J, Zhao Z, Yi X, Chen J, Hong L, Chi EH, (2018) 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*, 1930–1939. https://doi.org/10.1145/3219819.3220007
25. Qin Z, Cheng Y, Zhao Z, Chen Z, Metzler D, Qin J, (2020) Multitask mixture of sequential experts for user activity streams. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3083–3091. https://doi.org/10.1145/3394486.3403359
26. Singhal S, Shah RR, Chakraborty T, Kumaraguru P, Satoh SI, (2019) Spotfake: A multi-modal framework for fake news detection. In: *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, 39–47. https://doi.org/10.1109/BigMM.2019.00-44
27. Chen Y, Li D, Zhang P, Sui J, Lv Q, Tun L, Shang L, (2022) Cross-modal ambiguity learning for multimodal fake news detection. In: *Proceedings of the ACM Web Conference 2022*, 2897–2905. https://doi.org/10.1145/3485447.3511968

Appendix

Table A.1. Symbol summary.

Symbol	Mode	Dimension	Description
e_t	Text	[B,197,768]	Text features extracted by BERT.
e_i	Image	[B,197,768]	Image features extracted by MAE.
f_{clip_t}	Text	[B,512]	Text features extracted by CLIP.
f_{clip_i}	Image	[B,512]	Image features extracted by CLIP.
e_f	Fusion	[B,768]	CLIP-based joint multimodal features.
f'_{text}	Text	[B,197,768]	Local features refined by MaskedLinearAttention.
f_{text}	Text	[B,768]	Global semantic features obtained through Text_Attention.
f'_{image}	Image	[B,197,768]	Local features refined by MLLA-Lite.
f_{image}	Image	[B,768]	Global features obtained through Image_Attention.
f_{fusion}	Fusion	[B,64]	To adapt to the subsequent fusion features of SEMF that reduce the dimensionality of e_f .
z_{text}	Text	[B,16]	Features after distillation of f_{text} .
z_{image}	Image	[B,16]	Features after distillation of f_{image} .
F	Splicing	[B,96]	The feature obtained by concatenating f_{fusion} , z_{text} , and z_{image} according to feature dimensions.
H_n	Splicing	[B,128,3]	The initial H_0 is F projected onto a unified space through a one-dimensional convolution, and the final H_n is the output of the last ResBlock.