Dynamic Mixture-of-Head Attention via Feature Alignment and Conflict for Fake News Detection

Rongxin Lin[1], Meng Xing[2], Zihao Li[1], Lincheng Jiang[3], Shuohao Li[1*], Jun Zhang[1]

[1]National University of Defence Technology, Laboratory for Big Data and Decision, China
[2]Unit 32005 of the Chinese People's Liberation Army, China
[3]National University of Defence Technology, College of Advanced Interdisciplinary Studies, China

***Abstract*—Multimodal fake news detection, which fuses textual and visual information, has been proven to significantly boost the accuracy of fake news classification, thereby effectively curbing the proliferation of misinformation on social media platforms. Despite the fact that existing methods primarily focus on inter-modal and intra-modal fusion strategies, they still struggle to mitigate redundant attention patterns and lack feature learning mechanisms from the perspective of cross-modal conflicts. To address these limitations, we propose a Dynamic mixture-of-head attention via Feature Alignment and Conflict network (DFAC) for fake news detection. Specifically, we design a multi-scale feature attention to capture local fine-grained features and global correlations, enabling comprehensive extraction of discriminative information. Subsequently, we implement a dynamic expert-driven attention to adaptively select and aggregate the most discriminative sub-expert heads. Finally, we develop a feature alignment and conflict fusion module to adaptively modulate dynamic features and activate high-value semantic clues. Experiments demonstrate the superiority of our DFAC method on the Weibo, Twitter and Politifact datasets.**

***Index Terms*—multimodal fake news detection, mixture of experts, feature fusion**

I. Introduction

Under the Internet wave, social media has reshaped the logic of information dissemination, enabling the production and circulation of news to break through spatiotemporal constraints [1]. This has facilitated precise recommendations of news to specific users, but it has also caused false information to spread rapidly among target groups, disrupting public order. Authoritative institutions and social platforms have jointly tackled false information through a series of effective measures at the source.

Early approaches to fake news detection primarily concentrated on textual news, employing handcrafted features [2], while image content in news was also used to analyze the impact of rumors [3]. With the rapid advances of neural networks, fake news detection has shifted from manual feature design to data-driven feature learning. Popular methods for detecting fake news from image-text pairs can be divided into three types: consistency learning that aligns textual and visual data from pre-trained models [4], [5], feature fusion methods that address multi-scale data with varying information granularities [6], [7], and auxiliary tasks that introduce additional loss terms to learn robust news representations while considering modality coherence [8], [9].

Although significant progress has been made in relevant research, critical issues remain unsolved as follows: 1) semantic clue inadequacy. As shown in Fig. 1(a), existing methods attempt to achieve cross-modal alignment by leveraging attention mechanisms and concatenation. However, they overlook the modeling of cross-modal semantic conflicts, simply regarding the detection task as an image-text matching task. 2) feature redundancy. As illustrated in Fig. 1(b), most of these methods adopt dense feature learning schemes, yet they fail to distinguish the importance of extracted features. Traditional strategies propagate complete interactive features without refining their weight distributions.

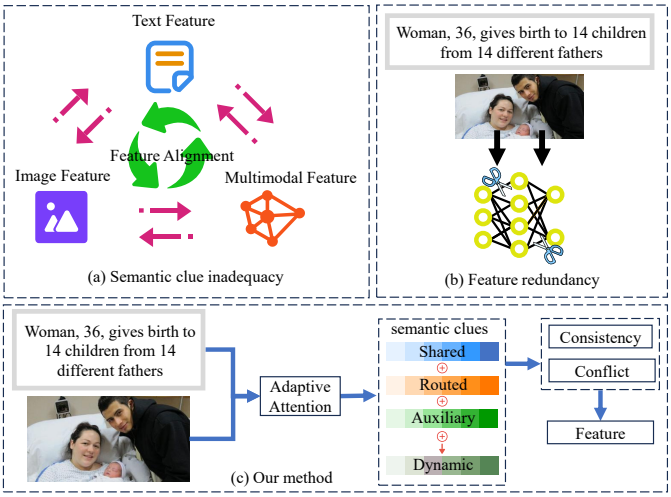


Fig. 1. Comparison of the fusion and reasoning processes. (a) Semantic clue inadequacy; (b) Feature redundancy; (c) Our method.

To address the above issues, we design the Dynamic mixture-of-head attention with Feature Alignment and Conflict (DFAC) for rumor detection. Specifically, for the deconstruction of features by the pre-trained model, we construct a multi-scale adaptive attention mechanism to optimize initial features from both local and global perspectives. Afterward, we develop a dynamic mixture-of-head attention module to integrate multi-dimensional shared, routed and auxiliary features, as shown in Fig. 1(c). This module captures general news semantic clues, delineates high-value discriminative features and assigns adaptive weights to trivial features, thus yielding

the overall refined semantic features. On this basis, we devise a dual-branch fusion strategy to fuse cross-modal features from the perspectives of consistency and conflict. The fused cross-modal features are further weighted via element-wise multiplication with the semantic features to generate the cross-modal enhanced semantic features, which serves to amplify the scores of critical positional information.

The main contributions are summarized as follows:

- We propose an adaptive attention mechanism to perform multi-scale feature analysis, enabling fine-grained capture of both local semantic details and global contextual patterns of news content.
- We design a dynamic mixture-of-experts module to alleviate the redundant learning of dense features, which adaptively activates the shared, routed and auxiliary experts based on the input characteristics.
- We introduce a dual-branch feature fusion mechanism from the perspectives of alignment and conflict, which effectively highlights the key discriminative information.

II. Related WORK

A. Fake News Detection

Unimodal fake news detection methods perform detection by mining features from either textual or visual data. MVNN [10] proposed a multi-domain visual neural network that fuses visual information from the frequency and pixel domainsx for dynamic fusion. AGWu-RF [11] employs a deep normalized attention to extract enriched features, and utilizes an adaptive genetic weight update-random forest classifier to achieve classification. Allein et al. [12] indirectly integrates user context by enforcing cross-modal correlations, while excluding user data during inference to prevent ethical issues of profiling. These methods have realized the shift from handcrafted design to the perception of fake semantics driven by deep learning.

Multimodal fake news detection promotes efforts for text-image integrated approaches when news evolves into the text-image format. HMCAN [13] proposes a hierarchical multi-modal contextual attention network that fuses intra-modality and inter-modality relationships and captures rich hierarchical semantics of text. CSFND [9] proposes a contextual semantic representation learning method that bridges the gap between global semantic features and decision boundaries by incorporating local context information through clustering and triplet learning. SEMFD [14] incorporates GPT-2 generated text summaries for fine-grained semantic alignment and discrepancy-weighted fusion.

B. Representation Learning

This innovative representation learning methods capture the semantic clues of news to support more accurate predictions. MRAN [15] extracts hierarchical semantic features of multi-modal features and leverages a relationship-aware attention network to generate high-order fusion features. BMR [4] optimizes multi-gate mixture-of-experts networks and then bootstraps multi-view representations for fake news detection. MIMoE-FND [5] proposes a hierarchical modality-interactive mixture-of-experts that models four modality interaction types and dynamically routes instances to specialized fusion experts. RaCMC [7] employs multi-scale feature interaction and feature-level distribution constraints to amplify discrepancies between real and fake news.

However, representation learning often fails to adequately incorporate task-specific characteristics, instead focusing on general feature alignment and semantic fusion. Here, we propose a method that explicitly accounts for distinct characteristics distinguishing real and fake news in detection tasks.

III. Methodology

In this research, we aim to describe the architecture of DFAC for fake news detection. A detailed illustration is presented in Fig. 2.

A. Unimodal Feature Extraction

To better align features with news posts, we encode the features of image-text pairs through a dual specific-unified architecture. On the one hand, we employ the pre-trained BERT [16] and ResNet-18 [17] to extract modality-specific features. Given the text content c and its corresponding image s, we have specific features $T_s = \text{BERT}(c)$ and $V_s = \text{ResNet}(s)$. On the other hand, we utilize CLIP [18] to project both textual and visual modalities into a unified semantic space, resulting in unified features $T_u = \text{CLIP}(c)$ and $V_u = \text{CLIP}(s)$. Finally, we map these four specific-unified features into the same dimension to facilitate subsequent cross-modal fusion.

B. Adaptive Multi-Scale Attention

News exhibits high complexity and diversity with various semantic clues that convey its authenticity. To address the limited representation capability of single-scale features, we design an adaptive multi-scale attention (AMSA) module that captures both local details and global semantics simultaneously. Denoting the input feature $\boldsymbol{X}$, we employ convolutional kernels of different scales 1×1, 3×3 and 11×11 to extract multi-scale features, which are responsible for identity mapping, local fine-grained information and long-sequence contextual dependencies. The feature process of each branch is formulated as:

$$\begin{aligned} \boldsymbol{F}_1 &= \text{GELU}\left(\text{BN}\left(\text{Conv}_{1\times1}(\boldsymbol{X})\right)\right), \\ \boldsymbol{F}_2 &= \text{GELU}\left(\text{BN}\left(\text{Conv}_{3\times3}(\boldsymbol{X})\right)\right), \\ \boldsymbol{F}_3 &= \text{GELU}\left(\text{BN}\left(\text{Conv}_{11\times11}(\boldsymbol{X})\right)\right), \end{aligned} \tag{1}$$

where BN denotes batch normalization. We aggregate multi-scale features from all branches to obtain $\boldsymbol{F}_{\text{sum}} = \sum_{i=1}^{3} \boldsymbol{F}_i$. Then compute the global context descriptor by applying adaptive average pooling:

$$\boldsymbol{G} = \text{AdaptiveAvgPool}(\boldsymbol{F}_{\text{sum}}). \tag{2}$$

We generate adaptive weighting coefficients for each feature branch on the global descriptor $\boldsymbol{W} = \text{Softmax}(\boldsymbol{G})$. Finally, the fused multi-scale feature is obtained by element-wise multiplication of each branch feature with its corresponding weight, followed by summation: $\boldsymbol{F} = \sum_{i=1}^{3} \boldsymbol{F}_i \odot \boldsymbol{W}_i$.

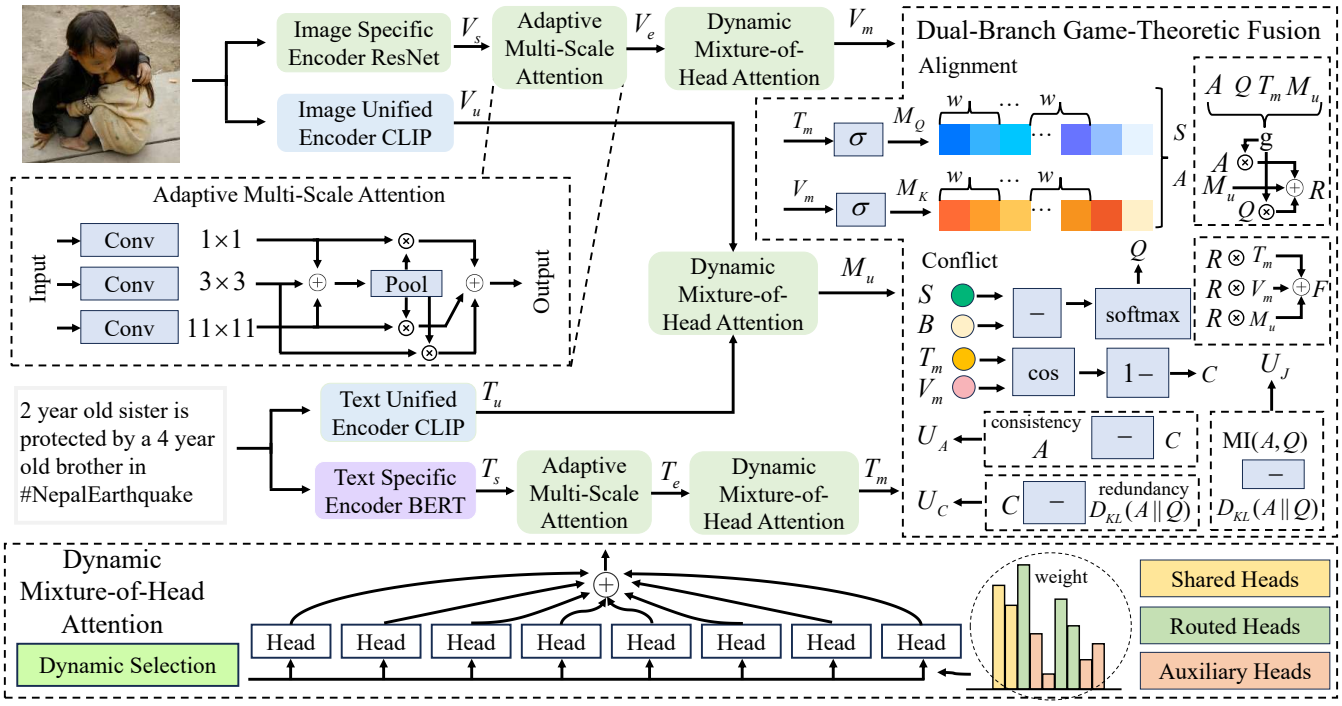


Fig. 2. The overall architecture of DFAC for cross-modal fake news detection. The framework unifies image and text features via adaptive multi-scale attention and dynamic mixture-of-head attention, then fuses them using a dual-branch game-theoretic strategy.

To strengthen the specific features, we feed these features through the AMSA module. This module captures both local fine-grained details and global semantic dependencies, yielding enhanced specific features as follows:

$$\begin{aligned} \boldsymbol{T}_e &= \text{AMSA}(\boldsymbol{T}_s), \\ \boldsymbol{V}_e &= \text{AMSA}(\boldsymbol{V}_s), \end{aligned} \tag{3}$$

where $\boldsymbol{T}_e$ and $\boldsymbol{V}_e$ denote the multi-scale enhanced textual and visual specific features.

C. Dynamic Mixture-of-Head Attention

Multi-head attention utilizes a standard summation operation to compute the scaled dot-product attention [19]. Drawing inspiration from mixture-of-heads attention [20], which optimizes multi-head attention via adaptive head selection, we propose a dynamic mixture-of-head attention (DMoH) module for feature refinement. To enable complex feature interaction in the specific-unified encoder, we achieve balanced expert-guided interaction through rational head composition. For h heads $H = \{H_1, H_2, \ldots, H_h\}$, the weighted selection of heads is implemented via a dynamic selector formulated as:

$$\text{DMoH}(X, Y) = \sum_{i=1}^{h} g_i H_i W_i, \tag{4}$$

where X and Y represent the input features, and $g_i \in [0, 1]$ denotes the weight of the i-th attention head.

Moreover, DMoH incorporates h_s shared heads to capture domain-shared semantic information. From the remaining activated $(h - h_s)$ heads after excluding the shared ones, we adopt a Top-K selection strategy to adaptively select high-value expert-routed heads, while assigning the rest as auxiliary heads to capture potentially overlooked secondary information. Collectively, DMoH is composed of three types of heads: shared heads, routed heads, and auxiliary heads. In the selection stage, we calculate the respective selection scores g_i for each head to determine their activation weights.

In the first stage, we compute the weighted sum of contributions from the activated heads:

$$[\alpha_1, \alpha_2, \alpha_3] = \text{Softmax}(W_h x_t), \tag{5}$$

where α_1, α_2 and α_3 denote the contribution weights of shared heads, routed heads and auxiliary heads. In the second stage, we calculate the scores g_i for each attention head as:

$$g_i = \begin{cases} \alpha_1 \cdot \text{Softmax}(W_r x_t)_i, & \text{if } 1 \le i \le h_s, \\ \alpha_2 \cdot \text{Softmax}(W_r x_t)_{i-h_s}, & \text{if Head } i \text{ is routed}, \\ \alpha_3 \cdot \beta \cdot \text{Softmax}(W_r x_t)_{i-h_s}, & \text{if Head } i \text{ is auxiliary}. \end{cases} \tag{6}$$

where W_r is the learnable matrix and β controls the dynamic weight of auxiliary heads.

In the third stage, we output the coefficient β for auxiliary experts by quantifying the uncertainty of experts via the entropy calculation. For a probability vector $\boldsymbol{P} = [P_1, P_2, ..., P_E] = \text{Softmax}(F(X))$, the entropy is calculated as $\mathcal{H}(\boldsymbol{P}) = -\sum_{e=1}^{E} P_e \cdot \log P_e$. We normalize it to the range $[0, 1]$ using the maximum entropy as $\hat{\mathcal{H}} = \frac{\mathcal{H}(\boldsymbol{P})}{\log(E)}$. Then employ scaled dot-product attention to capture the cross-modal relationship $Z = Attention(X, Y, Y)$. To sparsify gate values,

we generate raw gate values by applying masked logits as $\boldsymbol{G} = \sigma(\text{Linear}(Z))$. Finally, we implement uncertainty-aware sparsification by the normalized entropy $\hat{\mathcal{H}}$:

$$\beta = \boldsymbol{G} \odot \sigma\left(\left(\boldsymbol{G} - \theta_0 \cdot (1 - \hat{\mathcal{H}})\right)\right) \tag{7}$$

where $\theta_0 \in \mathbb{R}^E$ is a threshold vector.

We adopt the load balance loss L_{DMoH} of the DMoH to train each sub-expert uniformly:

$$L_{\text{DMoH}} = \sum\nolimits_{i=h_s+1}^{h} P_i f_i, \tag{8}$$

where P_i denotes the average selection probability of head i and f_i represents the average frequency of that head.

We discuss how DMoH achieves specific and unified information interaction. For a unimodal perspective, we apply the equation (4) with T_e as both X and Y to get the fine-grained specific textual feature T_m. Similarly, we obtain the self-interaction feature V_m from the original features V_e. For a multimodal perspective, we employ the same equation with T_u as X and V_u as Y to get the dynamic representation M_u.

We record the load balance losses $L_{\text{T-DMoH}}$, $L_{\text{V-DMoH}}$ and $L_{\text{M-DMoH}}$ from the above structures. Specifically, during the backward pass, the gradient of the load balance loss is explicitly injected into the gradient flow of the model parameters, while the forward pass returns the original feature unmodified to preserve the integrity of the primary computation.

D. Dual-Branch Game-Theoretic Fusion

Real news can be considered semantically aligned, while fake news exhibits distinct semantic conflicts. To leverage this inherent trait, we present a dual-branch game-theoretic fusion (DFG) that formalizes multimodal integration as a cooperative game with the alignment branch and the conflict branch.

The alignment branch aims to maximize semantic correspondence. Given a context window w, we derive contextual features by computing weights of neighboring elements:

$$\begin{aligned} &M_Q = \sigma(W_Q T_m), \quad M_K = \sigma(W_K V_m) \\ &Q_{\text{ctx}} = \sum_{i=-w}^{w} \alpha_i \cdot M^{(i)}_{Q-\text{neighbors}}, \quad K_{\text{ctx}} = \sum_{i=-w}^{w} \gamma_i \cdot M^{(i)}_{K-\text{neighbors}} \end{aligned} \tag{9}$$

where α_i, γ_i are context weights. The context-aware similarity incorporates both direct and contextual information: $S = \text{sigmoid}(T_m \cdot V_m^\top + Q_{\text{ctx}} \cdot K_{\text{ctx}}^\top)$. With the attention function, the aligned feature is obtained as:

$$A = \text{softmax}\left(\frac{S}{\sqrt{D}}\right) \cdot V \tag{10}$$

The conflict branch identifies and leverages discriminative discrepancies. We compute the text-image semantic conflict score $C = \mu \cdot (1 - \cos(T_m, V_m))$, where μ scales the importance of semantic conflicts. Next compute an inverse attention based on the alignment similarity to get conflict features:

$$Q = \text{softmax}\left(\frac{B - S}{\sqrt{D}}\right) \tag{11}$$

where B is a learnable scaling matrix and D represents the hidden dimension of the input feature embedding.

The core innovation of our approach is formulating multimodal fusion as a cooperative game with three utility functions. Alignment utility measures intra-modal coherence as:

$$\begin{aligned} U_A &= \alpha \cdot \text{consistency} - \beta \cdot C, \\ \text{consistency} &= \frac{1}{L} \sum_{l=1}^{L} \cos\left(A, \overline{A}\right). \end{aligned} \tag{12}$$

Conflict utility quantifies information overlap between alignment and conflict features based on the KL divergence as:

$$\begin{aligned} U_C &= \gamma \cdot C - \delta \cdot \text{redundancy} \\ \text{redundancy} &= D_{\text{KL}}\left(A \parallel Q\right). \end{aligned} \tag{13}$$

The joint utility is designed to balance the information synergy and divergence between aligned and conflict features as:

$$U_J = \eta \cdot \text{MI}(A, Q) - \theta \cdot D_{\text{KL}}\left(A \parallel Q\right) \tag{14}$$

where MI means the mutual information to quantify their information correlation.

Given the multimodal feature M_u, the fusion gate g is designed to optimize the combined utility, and is used to fuse the features F as follows:

$$\begin{aligned} g &= \sigma(W_g[A; Q; T_m; M_u]) \\ R &= \text{LayerNorm}(W_f[g \cdot A + (1 - g) \cdot Q; M_u]) \end{aligned} \tag{15}$$

Finally, we perform weighted element-wise multiplication on the fused features and aggregate the results to obtain the final feature, which is formulated as:

$$F = R \times T_m + R \times V_m + R \times M_u. \tag{16}$$

E. Task Predictor

The model is optimized using a multi-task objective that combines the prediction loss with a utility-driven regularization term:

$$\begin{aligned} \mathcal{L}_{\text{pre}} &= -\left[y \cdot \log(p) + (1 - y) \cdot \log(1 - p)\right] \\ \mathcal{L} &= \mathcal{L}_{\text{pre}} - \lambda \cdot U_{\text{total}} \end{aligned} \tag{17}$$

where $U_{\text{total}} = (U_A + U_C) + U_J$ denotes the total utility defined from a game-theoretic perspective, encouraging the model to learn representations that maximize the game-theoretic utility.

IV. EXPERIMENTS

A. Dataset

To evaluate the performance of DFAC, we conduct intensive experiments on three benchmark datasets, namely Weibo [21], Twitter [22] and Politifact [23]. Weibo is a large-scale Chinese microblog corpus constructed for rumor detection, where rumor samples are collected from official rumor debunking system and non-rumor samples are sourced from tweets verified by Xinhua News Agency. Twitter focuses on serving as a popular news-sharing platform for governments, politicians, and the public, hosting numerous posts linked to various events. Politifact focuses on political news, featuring high-quality expert annotations and rich political context information. The detailed statistics are presented in Table I. The detailed statistics are presented in Table I.

TABLE I
DATASET STATISTICS OF WEIBO, TWITTER AND POLITIFACT.

Dataset	Train			Test		
	Real	Fake	Total	Real	Fake	Total
Weibo	2807	3347	6154	835	864	1699
Twitter	5993	7336	13329	474	643	1117
Politifact	166	130	296	42	32	74

TABLE II
MAIN EXPERIMENTAL RESULTS BETWEEN OURS AND OTHER METHODS ON WEIBO, TWITTER AND POLITIFACT DATASETS.

Dataset	Models	Acc	Fake News			Real News		
			Pre	Recall	F1	Pre	Recall	F1
Weibo	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	SAFE	0.763	0.833	0.659	0.736	0.717	0.868	0.785
	CAFE	0.840	0.855	0.830	0.842	0.825	0.851	0.837
	MRAN	0.903	0.904	**0.908**	0.906	0.897	0.892	0.894
	CSFND	0.895	0.899	0.895	0.897	0.892	0.896	0.894
	DFAC	**0.912**	**0.920**	0.905	**0.913**	**0.903**	**0.919**	**0.911**
Twitter	MVAE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	SAFE	0.766	0.777	0.795	0.786	0.752	0.731	0.742
	CAFE	0.806	0.807	0.799	0.803	0.805	0.813	0.809
	LIIMR	0.831	0.825	0.830	0.827	**0.836**	0.832	**0.830**
	CSFND	**0.833**	**0.899**	0.799	0.846	0.763	**0.878**	0.817
	DFAC	**0.833**	0.844	**0.869**	**0.857**	0.815	0.783	0.799
Politifact	SAFE	0.874	0.851	0.830	0.840	0.889	0.903	0.896
	CAFE	0.864	0.724	0.778	0.750	0.895	0.919	0.907
	RaCMC	0.922	0.833	0.926	0.877	0.967	0.921	0.943
	DFAC	**0.946**	**0.912**	**0.967**	**0.940**	**0.975**	**0.929**	**0.951**

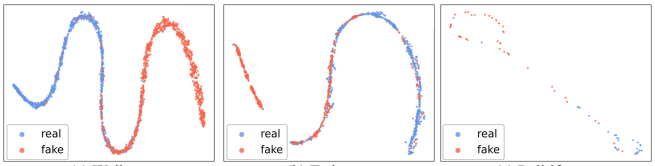
B. Implementation Details

For fairness, we adopt the identical data processing pipeline for each dataset, which includes removing special characters, converting all text to lowercase, and eliminating news samples without either textual content or images.

We set the Chinese and English versions used in the CLIP configuration [18] during the model initialization. For the DMoH, we set the total number of expert heads to 8, with 2 shared heads configured and top-2 for remaining activated heads to decide routed heads and auxiliary heads. The dropout rate is 0.5. The learning rate is 1e-4. We utilize the AdamW optimizer to optimize our model, with batch sizes of 256, 64, and 8 for each datasets. The weight decay is 1e-3. We set the constant coefficient $\lambda = 1$ to weight the total utility loss.

C. Baselines

We conduct a comparative analysis against multimodal methods in fake news detection. MVAE [24] proposes a multimodal variational autoencoder to learn shared representations between textual and visual modalities. SAFE [25] presents a similarity-aware method to investigate cross-modal similarity relationships. CAFE [26] adaptively aggregates optimizes feature fusion weights by quantifying the distribution characteristics of textual and visual features. LIIMR [27] captures intra-modality relationships and inter-modality relationships to suppress weak modality information. MRAN [15] leverages a relationship-aware attention network to model intra-modal and cross-modal relationships for effective multimodal feature fusion. CSFND [9] addresses the inconsistency between the semantic and decision spaces via unsupervised context learning, and fuses these to capture discriminative contextual semantic features. RaCMC [7] employs a multiscale residual-aware compensation to interact cross-modal features with consistency constraints with multi-granularity constraints. Compared with other baseline models, DFAC achieves competitive performance on all three datasets.



(a) Weibo (b) Twitter (c) Politifact

Fig. 3. t-SNE Visualizations on the Weibo, Twitter, and Politifact Datasets

D. Results and Analysis

We report evaluation metrics, namely accuracy, precision, recall, and F1 score for both real and fake news in Table II. It can be observed that DFAC achieves stable accuracy by dynamic experts with feature alignment and conflict.

Under the context of rich semantics in Chinese texts, DFAC outperforms comparative models across key evaluation metrics, while filtering out reliable and valuable semantic clues. This indicates that the sub-experts decision tends to summarize the patterns of fake news and suppress misclassification into the fake category.

For distinguishing highly similar contexts on Twitter, DFAC exhibits remarkable effectiveness. On Politifact, DFAC mitigates the overfitting risk induced by insufficient training data and identifies potentially crucial relationships amid feature alignment and feature conflicts. The results demonstrate that our model achieves stable performance for English content and can extract relevant expert features that align textual content with corresponding images. Compared with baseline models, our method exhibits superior capability in processing short and highly repetitive English posts. Specifically, DFAC achieves 1.7% and 2.3% higher accuracy on the Weibo and Politifact datasets, and matches the accuracy of CSFND on Twitter. As illustrated in Fig. 3, we visualize classification results across three datasets via t-SNE visualizations. It shows that the model effectively distinguish between real news and fake news with clear distribution between the two classes in the feature space across all datasets.

E. Ablation Studies

To verify the effectiveness of each component, we conduct control experiments by removing each specific components as follows: (1) w/o A: remove the AMSA component; (2) w/o D: remove the DMoH component; (3) w/o G: remove the DGF component.

In Table III, it can be observed that all ablated variants exhibit worse performance compared to the full model. Specifically, the AMSA module preserves discriminative information

TABLE III
ABLATION STUDY RESULTS OF DFAC ON WEIBO, TWITTER AND POLITIFACT DATASETS.

Dataset	Models	Acc	Fake News			Real News		
			Pre	Recall	F1	Pre	Recall	F1
Weibo	w/o A	0.898	0.902	0.897	0.900	0.894	0.900	0.897
	w/o D	0.863	0.899	0.823	0.859	0.832	0.904	0.866
	w/o G	0.893	0.914	0.871	0.892	0.873	0.915	0.894
	DFAC	**0.912**	**0.920**	**0.905**	**0.913**	**0.903**	**0.919**	**0.911**
Twitter	w/o A	0.810	0.814	**0.869**	0.841	0.805	0.730	0.766
	w/o D	0.723	0.743	0.795	0.768	0.692	0.627	0.658
	w/o G	0.760	0.764	0.844	0.802	0.754	0.646	0.695
	DFAC	**0.833**	**0.844**	**0.869**	**0.857**	**0.815**	**0.783**	**0.799**
Politifact	w/o A	0.892	0.900	0.844	0.871	0.886	0.929	0.907
	w/o D	0.716	0.690	0.625	0.656	0.733	0.786	0.759
	w/o G	0.763	0.751	0.879	0.810	0.786	0.606	0.684
	DFAC	**0.946**	**0.912**	**0.967**	**0.940**	**0.975**	**0.929**	**0.951**

to support specific representation. The removal of DMoH impairs the ability to learn valuable semantic features, directly resulting in redundant semantics and feature sparsity. Furthermore, the DGF module, which maintains the semantic balance between aligned and conflicting information, is particularly effective for rumor detection.

V. CONCLUSIONS

In this paper, we propose DFAC that adaptively selects experts for mutual interaction, incorporates the interaction of alignment and conflict information, and activates high-value features for fake news detection. Specifically, we adopt a dual-branch feature encoding strategy to enhance multimodal semantic features, and leverage adaptive multi-scale attention to focus on specific features. Then, we utilize dynamic mixture-of-heads attention, which not only enables cross-modal fusion for specific features but also achieves self-interaction to highlight unified features of the primary modality. Finally, via the alignment branch and conflict branch, we analyze the semantic relevance of news content. Experiments confirm the effectiveness of our approach, showing significant performance compared with baseline models.

REFERENCES

[1] B. Hu, Z. Mao, and Y. Zhang, "An overview of fake news detection: From a new perspective," *Fundamental Research*, vol. 5, no. 1, pp. 332–346, 2025.

[2] J. Ma, W. Gao, Z. Wei, Y. Lu, and K.-F. Wong, "Detect rumors using time series of social context information on microblogging websites," in *Proceedings of the 24th ACM international on conference on information and knowledge management*, 2015, pp. 1751–1754.

[3] K. Wu, S. Yang, and K. Q. Zhu, "False rumors detection on sina weibo by propagation structures," in *2015 IEEE 31st international conference on data engineering*. IEEE, 2015, pp. 651–662.

[4] Q. Ying, X. Hu, Y. Zhou, Z. Qian, D. Zeng, and S. Ge, "Bootstrapping multi-view representations for fake news detection," in *Proceedings of the AAAI conference on Artificial Intelligence*, vol. 37, no. 4, 2023, pp. 5384–5392.

[5] Y. Liu, Y. Liu, Z. Li, R. Yao, Y. Zhang, and D. Wang, "Modality interactive mixture-of-experts for fake news detection," in *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 5139–5150.

[6] Y. Wu, P. Zhan, Y. Zhang, L. Wang, and Z. Xu, "Multimodal fusion with co-attention networks for fake news detection," in *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, 2021, pp. 2560–2569.

[7] X. Yu, Z. Sheng, W. Lu, X. Luo, and J. Zhou, "Racmc: Residual-aware compensation network with multi-granularity constraints for fake news detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 1, 2025, pp. 986–994.

[8] L. Peng, S. Jian, D. Li, and S. Shen, "Mrml: Multimodal rumor detection by deep metric learning," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[9] L. Peng, S. Jian, Z. Kan, L. Qiao, and D. Li, "Not all fake news is semantically similar: Contextual semantic representation learning for multimodal fake news detection," *Information Processing & Management*, vol. 61, no. 1, p. 103564, 2024.

[10] P. Qi, J. Cao, T. Yang, J. Guo, and J. Li, "Exploiting multi-domain visual information for fake news detection," in *2019 IEEE international conference on data mining (ICDM)*. IEEE, 2019, pp. 518–527.

[11] A. M. Luvembe, W. Li, S. Li, F. Liu, and G. Xu, "Dual emotion based fake news detection: A deep attention-weight update approach," *Information Processing & Management*, vol. 60, no. 4, p. 103354, 2023.

[12] L. Allein, M.-F. Moens, and D. Perrotta, "Preventing profiling for ethical fake news detection," *Information Processing & Management*, vol. 60, no. 2, p. 103206, 2023.

[13] S. Qian, J. Wang, J. Hu, Q. Fang, and C. Xu, "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*, 2021, pp. 153–162.

[14] X. Wang, W. Shang, J. Cao, X. Bo, C. Wang, and K. Song, "Semfd: Summary-enhanced multimodal fake news detection," in *2025 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2025, pp. 1–8.

[15] H. Yang, J. Zhang, L. Zhang, X. Cheng, and Z. Hu, "Mran: Multimodal relationship-aware attention network for fake news detection," *Computer Standards & Interfaces*, vol. 89, p. 103822, 2024.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[18] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[20] P. Jin, B. Zhu, L. Yuan, and S. Yan, "Moh: Multi-head attention as mixture-of-head attention," *arXiv preprint arXiv:2410.11842*, 2024.

[21] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 795–816.

[22] C. Boididou, S. Papadopoulos, D.-T. Dang-Nguyen, G. Boato, M. Riegler, S. Middleton, A. Petlund, and Y. Kompatsiaris, "Verifying multimedia use at mediaeval 2016," in *MediaEval Benchmarking Initiative for Multimedia Evaluation*, 2015.

[23] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," *Big data*, vol. 8, no. 3, pp. 171–188, 2020.

[24] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "Mvae: Multimodal variational autoencoder for fake news detection," in *The world wide web conference*, 2019, pp. 2915–2921.

[25] X. Zhou, J. Wu, and R. Zafarani, "Safe: Similarity-aware multi-modal fake news detection," in *Pacific-Asia Conference on knowledge discovery and data mining*. Springer, 2020, pp. 354–367.

[26] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, and L. Shang, "Cross-modal ambiguity learning for multimodal fake news detection," in *Proceedings of the ACM web conference 2022*, 2022, pp. 2897–2905.

[27] S. Singhal, T. Pandey, S. Mrig, R. R. Shah, and P. Kumaraguru, "Leveraging intra and inter modality relationship for multimodal fake news detection," in *Companion proceedings of the Web conference 2022*, 2022, pp. 726–734.