Dual-Expert Collaborative Network for Fake News Detection with External Knowledge Integration

Wenxin Luo
wxluo0903@bjtu.edu.cn
Zhu Teng✉
zteng@bjtu.edu.cn
Wei Zhang
wzhang0926@bjtu.edu.cn
Baopeng Zhang
bpzhang@bjtu.edu.cn

School of Computer Science
Beijing Jiaotong Univeristy
Beijing, China

Abstract

The pervasive dissemination of misinformation has imposed a profound negative impact on societal structures, making multi-modal fake news detection an urgent research priority. Traditional approaches mainly focus on evaluating the authenticity and consistency of the original news modalities, often neglecting the contextual understanding necessary to avoid cognitive biases in complex scenarios. In this paper, we propose a Dual-Expert Collaborative Network (DECNet) for fake news detection, a novel approach that integrates original news content with external knowledge. Specifically, we leverage multi-scale images to extract visual features and design a Content-Driven Expert (CDE) that performs hierarchical fusion with textual data, enhancing semantic alignment and uncovering deeper insights of news content. Simultaneously, inspired by the reasoning capabilities of Large Vision-Language Models (LVLMs), we propose a Knowledge-Augmentation Expert (KAE) that leverages generated explanations alongside the inherent multi-modal information of the news to capture subtle cues indicative of fake news. To facilitate interaction between these two branches, a gating network is designed for seamless collaboration. Extensive experiments on two widely-employed datasets demonstrate that our method outperforms current state-of-the-art approaches.

1 Introduction

With the rapid evolution of the Internet and social media platforms, the dissemination of fake news has accelerated at an unprecedented pace, spreading quickly and widely [26]. This phenomenon is particularly concerning in domains such as politics and public security, where misinformation related to events like the US elections [15] and the COVID-19 pandemic [14] has significantly destabilized social structures and eroded public trust in the media. In parallel, the emergence of Large Vision-Language Models (LVLMs) such as ChatGPT [16] has introduced new risks, as these models may be intentionally exploited to propagate false narratives or, due to their propensity to 'hallucinate' information [29], inadvertently contribute to the spread of inaccurate content. As such, there is an urgent need to develop robust and automated systems for the effective detection of fake news.

Early research on fake news detection primarily focused on text-based unimodal approaches [13, 18, 26]. However, as news content increasingly encompasses multi-modal elements such as text, images, and videos, multi-modal fake news detection has emerged as a dominant research direction. Traditional methods typically employ Deep Neural Networks (DNNs) to integrate multi-modal features and analyze the consistency between textual and visual modalities [3, 36]. Nevertheless, cross-modal consistency is not always a decisive factor in fake news detection [30]. For example, additional cues, such as image manipulation [24] and emotional features conveyed through text [34], can also serve as effective indicators. To address these complexities, multi-view approaches have been proposed [30, 31], which integrate diverse perspectives to improve detection performance. Nevertheless, these methods still largely focus on internal content features and often struggle to handle complex real-world scenarios.

The intricate nature of fake news creation poses significant challenges, as fabricators often employ diverse strategies to manipulate news content, making it difficult to accurately identify fake news based solely on content analysis [7]. Existing methods link news to background knowledge through search engines [6] or knowledge graphs [22, 33]. While these approaches enhance the breadth and richness of news content, they still fall short of providing a deep understanding and analysis of the content itself. Effective fake news detection is a complex task, particularly when it requires a profound understanding and reasoning of common-sense knowledge. In recent years, Large Vision-Language Models (LVLMs) have demonstrated significant reasoning and analytical abilities in tasks that combine text and images, such as sarcasm detection [11], harmful meme detection [12], and metaphor identification [28]. Furthermore, Hu *et al.* [7] have shown that Large Language Models (LLMs) can provide multi-dimensional and detailed reasoning and evidence in the context of unimodal fake news detection. We argue that rich background knowledge provided by LVLMs plays a key role in multi-modal fake news identification, offering valuable evidence to support the detection process and strengthen the reliability of the system.

To this end, we propose a novel **D**ual-**E**xpert **C**ollaboration **Net**work (DECNet) for fake news detection. We design two experts that fully interact with each other's perspectives: one expert focuses on analyzing the news content, while the other integrates external knowledge for reasoning. These perspectives are then consolidated through a gating network. Specifically, the Content-Driven Expert (CDE) interacts with multi-scale images and text to capture semantic relationships, and optimizes semantic alignment using KL divergence. The Knowledge-Augmentation Expert (KAE) leverages LVLMs to generate background knowledge related to the news, enriching the textual information. This augmented knowledge is then integrated with the news image to capture implicit cues.

In summary, our contributions are three-fold:

1) We propose a novel Dual-Expert Collaborative Network (DECNet) for fake news detection, which integrates a Content-Driven Expert (CDE), a Knowledge-Augmentation Expert (KAE), and a gating network to enhance multi-modal representation and decision-making. By enriching textual and visual embeddings, CDE and KAE collaboratively promote a deeper and more holistic understanding of news content. The gating network dynamically balances the opinions of both experts, enhancing both accuracy and interpretability.

2) To better capture complex contexts, we exploit multi-resolution visual embeddings

guided by textual cues and optimize semantic alignment via KL divergence. Furthermore, we leverage LVLMs to inject external knowledge, enhancing the model's reasoning ability and interpretability.

3) We conduct extensive experiments on two widely-used benchmarks, our method significantly outperforms existing state-of-the-art approaches on these benchmarks, demonstrating our superior detection and generalization performance in fake news detection.

2 Related Work

Content-based Fake News Detection. Given that news content often includes multi-modal information such as images and text, the field of multi-modal fake news detection has gained increasing scholarly attention. Traditional methods for multi-modal fake news detection primarily rely on modal inconsistency to assess the authenticity of news. For example, CAFE [3] measures the ambiguity between different modalities using KL divergence and adaptively aggregates unimodal features and cross-modal correlations through ambiguity-aware mechanisms. FSRU [10] transforms spatial features of text and images into spectral features, capturing indicators of fake news authenticity in the frequency domain. However, focusing solely on semantic correlation and inconsistency between modalities may overlook other crucial cues within the multi-modal content. To tackle this issue, Wang *et al.* [24] detect fake news by analyzing image manipulation and its underlying intent, while BMR [30] generates multi-view representations from text, image patterns, and image semantics. Despite these advances, content-based methods still face significant limitations. They often lack access to external background knowledge, which hinders their ability to distinguish subtle, carefully crafted fabrications within the news content.

Knowledge-enhanced Fake News Detection. Knowledge-enhanced fake news detection has emerged as a promising research direction. Current approaches utilize a variety of tools, such as search engines, knowledge graphs, and LVLMs, to incorporate external knowledge relevant to news content, thereby enriching the explanatory framework. For instance, NSLM [6] uses reverse image search to retrieve contextual information from the web, helping to identify image repurposing in fake news. Other studies focus on extracting entities from both text and images, using knowledge graphs to augment the news content. Sun *et al.* [21] address both text-image semantic inconsistency and content-knowledge layer inconsistency, while Zhang *et al.* [33] propose learning adaptive knowledge subgraphs and integrating knowledge with news content via heterogeneous news graphs. Though effective, these methods often lack a dynamic understanding of the deep semantics of news content, which can hinder their ability to uncover latent information, especially when dealing with complex cross-modal information and reasoning tasks. Hu *et al.* [7] demonstrate that LLMs can provide multi-dimensional reasoning guidance for fake news detection. However, their work remains limited to unimodal scenarios. In contrast, our approach leverages LVLMs to enhance multi-modal news understanding by incorporating external background knowledge. By utilizing LVLMs for deep semantic understanding, we not only enrich the knowledge base but also improve interpretability.

3 Method

The task of multi-modal fake news detection aims to predict the authenticity of a news article by combining information from both the image I and text T. Specifically, given a news piece

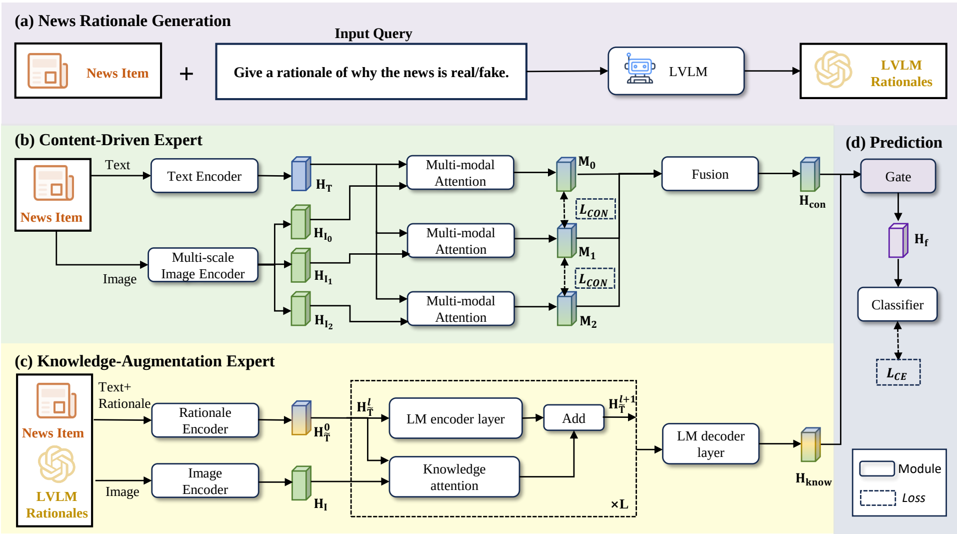

Figure 1: The architecture of the proposed DECNet. (a) News Rationale Generation: LVLM's comprehension ability is harnessed to generate knowledge, enabling broader exploration, as an input to the KAE module. (b) Content-Driven Expert: The text and multi-scale images are combined to extract insights from the news content. (c) Knowledge-Augmentation Expert: News and Rationales from LVLMs are leveraged to uncover deeper insights. (d) Prediction: Two experts to perform modality interaction gated fusion followed by a final classifier.

$X = \{I, T\}$, the goal is to determine whether the news is real or fake, producing a binary classification label $y \in \{real, fake\}$.

To address this task, we propose a Dual-Expert Collaboration Network (DECNet), as illustrated in Figure 1. It comprises two complementary branches: a Content-Driven Expert, which captures multi-modal semantics by hierarchically integrating multi-scale visual features and textual information; and a Knowledge-Augmentation Expert, which leverages Large Vision-Language Models (LVLMs) to generate explanatory knowledge that complements the original content and reveals deeper semantic cues. Finally, a gating network adaptively fuses the outputs from both experts to produce the final classification decision.

3.1 Content-Driven Expert (CDE)

In this section, we delineate the joint learning of semantic relationships within news by leveraging multi-scale image features and textual information, along with the representation of multi-modal information from news content for effective fake news detection.

Unimodal Encoder. Given a news piece consisting of both text T and image I, we begin by downsampling the image to perform multi-scale operations, where $I = \{I_i \mid i = 1, \ldots, N\}$. The text encoder and image encoder separately extract text representation $H_T \in \mathbb{R}^{n \times d}$ and image representation $H_{I_i} \in \mathbb{R}^{m \times d}$, using a fully frozen pre-trained BERT model [9] and a Transformer model [23], respectively.

Multi-modal Attention Network. To enable efficient interaction between multi-modal information, we employ a multi-modal contextual attention network [17]. This network comprises two contextual transformers, each focusing on different types of information: one focuses on the interactions from text to vision, while the other focuses on interactions from vision to text.

Each contextual transformer consists of two modified transformers, each designed for specific tasks. The first module, CT_a, independently processes the input unimodal embedding using self-attention to capture the unimodal internal contextual information. The second module, CT_b, facilitates cross-modal feature interaction. The attention mechanism is formulated in Eq. (1), where $\{Q_I, K_I, V_I\} = \{H_I W_Q, H_I W_K, H_I W_V\}$, and $\{W_Q, W_K, W_V\} \in \mathbb{R}^{d \times d_k}$ are trainable weights.

$$\mathrm{Att}(Q,K,V) = \mathrm{softmax}\left(Q_I K_I^T / \sqrt{d_k}\right) V_I \tag{1}$$

According to Eq. (1), the multi-modal representation is computed using Eq. (2), where $x \in \{I, T\}$, $y \in \{CT_a, CT_b\}$. The inputs K_T and V_T to CT_b are derived from the outputs of CT_a. FFN consists of two fully connected layers, with nonlinear activation functions applied between them. The representation H_{CT_a} captures the internal structure of the unimodal data. Consequently, CT_b enhances the representation of the current modality by integrating features from the other modality. Through the cross-modal attention network, this process facilitates mutual enhancement between image and text features, leading to richer multi-modal representations.

$$\begin{aligned} H_y^{'} &= \mathrm{LayerNorm}\left(H_x + \mathrm{Att}(Q_x, K_x, V_x)\right) \\ H_y &= \mathrm{LayerNorm}\left(H_y^{'} + \mathrm{FFN}(H_y^{'})\right) \end{aligned} \tag{2}$$

Finally, H_{CT_a} and H_{CT_b} are concatenated to form the final multi-modal representation M_{IT}. Another contextual transformer, following the same structure, generates the multi-modal representation M_{TI}. The outputs of both contextual transformers are then summed to obtain the final output M. Multi-scale images and text undergo multi-modal feature extraction, producing $M = \{M_1, M_2, ..., M_N\}$, which generates news features at different levels with enhanced semantic expressiveness.

Multi-scale Consistent Alignment. This module enhances the exploration of multi-modal representations extracted from images at various scales. To improve the efficiency and flexibility of multi-modal fusion, we design a simple mixture of experts (MOE) module as described in Eq. (3), which facilitates more precise extraction of latent information from different resolutions.

$$H_{con} = \sum_{i=1}^{N} W_{g_i} \cdot M_i \tag{3}$$

Simultaneously, to enhance the semantic consistency of multi-modal fusion, we design consistency constraints by minimizing the KL divergence (as shown in Eq. (4)) between multi-modal embeddings at different scales. This ensures that semantic information remains consistent across modalities, thereby improving the model's stability and performance. Subsequently, we impose pairwise constraints among the multi-modal representations at different scales and formulate the consistency loss in Eq. (5).

$$\begin{aligned} \mathrm{D}(M_i, M_j) = \frac{1}{2}&(\mathrm{KL}(\mathrm{Softmax}(M_i) || \mathrm{Softmax}(M_j)) \\ &+ \mathrm{KL}(\mathrm{Softmax}(M_j) || \mathrm{Softmax}(M_i))) \end{aligned} \tag{4}$$

$$\ell_{con} = \frac{2}{N(N-1)} \sum_{i<j, i,j \in \{1,\ldots,N\}} \mathrm{D}(M_i, M_j) \tag{5}$$

3.2 Knowledge-Augmentation Expert (KAE)

Considering that LVLMs possess rich prior knowledge and a profound understanding of the world, Hu *et al*. [7] demonstrated that LLMs can offer extensive external knowledge for unimodal fake news detection. Based on this, we extend it to multimodal fake news detection. The Knowledge-Augmentation Expert is designed to enhance learning and understanding, thereby improving the accuracy of fake news detection.

News Rationale Generation. Given a news piece, consisting of both text T and image I, we design a prompt template p^* and instruct the Qwen [1] to generate explanations r. The detailed prompt template is as follows:

> *"Given a news that consists of a text and an image, please give me a rationale of why the news is real or fake, no more than 100 words, tweet text: {T}, tweet image: {I}"*

Cross-modal Knowledge Interaction. Owing to the inherent limitations of LVLMs, such as susceptibility to hallucinations [11, 12], we leverage prior knowledge gained from LVLMs and news content to fine-tune a smaller Language Model (LM), thereby enhancing its ability to detect fake news.

Given a piece of news $X = \left\{I, \tilde{T}\right\}$, where $\tilde{T} = \{T, r\}$ contains rationales generated by LVLMs, we employ Flan-T5 [5] as the LM. We then utilize a text embedding layer to convert the input text $\tilde{T}$ into a vector representation $H_{\tilde{T}}^{0} \in \mathbb{R}^{n \times d}$. For the image representation, we denote it as $H_{I_0} \in \mathbb{R}^{m \times d}$.

The encoder of Flan-T5 is composed of multiple stacked Transformer encoder layers. Assuming the encoder consists of L layers, each layer processes the output from the previous layer to capture higher-level semantic information. To effectively leverage both the original news content and external knowledge obtained from the large model, we incorporate a fusion mechanism at each encoder layer. This mechanism enables deep interaction and integration of the news content and external knowledge, thereby enhancing the model's comprehension and judgment in the task of fake news detection.

Before fusion, we independently apply a self-attention mechanism to each modality to fully capture the global dependencies within each modality. This operation enhances the internal consistency and expressive power of the modality-specific embeddings. As a result, the features of each modality are optimized within their respective domains, providing a more robust foundation for subsequent cross-modal interaction. Afterward, we fuse the embeddings from each modality using a cross-modal interaction mechanism, guided by the background knowledge, to generate semantically enriched news representations:

$$H_{\tilde{T}_{attn}}^{l} = \mathrm{Att}(\mathrm{Att}(H_{\tilde{T}}^{l}), \mathrm{Att}(H_{I_0}), \mathrm{Att}(H_{I_0})) \tag{6}$$

where $Att(\cdot)$ indicates that Q, K, and V are the same. $H_{\tilde{T}}^{l}$ represents the textual embedding input to the l^{th} LM encoder layer, and $H_{\tilde{T}_{attn}}^{l}$ is the attended textual embedding output. Then,

$H^l_{\tilde{T}_{attn}}$ is added to the output of the l^{th} LM encoder layer LME^l and used as the input for the next layer LME^{l+1}, as formulated by Eq. (7). The learned knowledge representation H_{know} is defined as the final embedding $H^L_{\tilde{T}}$.

$$H^{l+1}_{\tilde{T}} = \mathrm{LME}^{\mathrm{l}}(H^l_{\tilde{T}}) + H^l_{\tilde{T}_{attn}} \tag{7}$$

3.3 Prediction and Optimization

Prediction. Through the aforementioned processes, we obtain the content-based representation H_{con} and knowledge-enhanced representation H_{know} from the Content-Driven Expert and Knowledge-Augmentation Expert, respectively. These two representations are then consolidated using a gating network, as outlined in Eq. (8). Specifically, the gating network concatenates the outputs from both experts, applies a softmax function to compute the gate weights, and performs a weighted sum of the representations to generate the final fusion output. This mechanism enables the model to dynamically integrate information from both content and knowledge sources, enhancing its decision-making capability.

$$\hat{y} = \mathrm{Gate}(H_{con}, H_{know}) \tag{8}$$

Optimization. The total loss function consists of two components: a binary decision loss to classify whether a news piece is real or fake, and a consistency loss that penalizes content distances across different scales (Eq. (5)). The binary decision is made using the cross-entropy loss and the overall loss function can be written as:

$$\ell = \text{Cross-Entropy}(\hat{y}, y) + \alpha \ell_{con} \tag{9}$$

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate the proposed DECNet on two real-world datasets: Weibo [8] and Pheme [37]. The Weibo dataset contains 4,749 fake news articles and 4,779 real news articles. The Pheme dataset includes 1,972 fake news articles and 3,830 real news articles. We preprocess these datasets following previous works [20, 35] by manually removing logos from the articles and discarding those that contain only images or text.

Comparison Methods. To demonstrate the superiority of the proposed model, we compare it with a range of state-of-the-art models from two categories: 1) Content-based methods, including EANN [25], SpotFake [19], MCAN [27], HMCAN [17] and CAFE [3]; 2) Knowledge-enhanced methods, including MKEMN [32], AKA-Fake [33], and NSLM [6].

Implementation Details. For the CDE module, we set input text length limits of 200 characters for Weibo, and 50 characters for Pheme. We use the "bert-base-chinese" model for processing Chinese datasets and the "bert-base-uncased" model for English datasets. For news images, we resize them to three resolution levels: 224x224, 112x112, and 56x56, with $N = 3$, and use the "clip-vit-base-patch32" model for image processing. For the KAE module, we standardize the input text length limit to 512 characters across all datasets and use the "flan-t5-base" model for processing. The hidden sizes for VIT, BERT, and Flan-T5-BERT are all 768, with BERT and VIT frozen, while the last two layers of the encoder and decoder in Flan-T5-BERT are unfrozen. We use the Adam optimizer with default parameters, a batch size of 8, and a learning rate of 5e-5.

4.2 Experimental Results

Category	Method	Weibo				Pheme			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
Content based	EANN* [25]	0.782	0.790	0.780	0.780	0.681	0.693	0.707	0.721
	SpotFake* [19]	0.851	0.850	0.851	0.851	0.809	0.812	0.704	0.728
	MCAN* [27]	0.884	**0.884**	0.884	**0.884**	0.832	0.795	0.804	0.799
	HMCAN* [17]	0.830	0.830	0.831	0.830	**0.861**	0.838	0.759	0.763
	CAFE* [3]	0.840	0.840	0.841	0.840	**0.861**	0.844	0.794	0.814
Knowledge Enhanced	MKEMN [32]	0.814	0.773	0.809	0.805	0.816	0.813	0.810	0.812
	AKA-Fake [33]	-	-	-	-	0.858	**0.918**	**0.877**	**0.897**
	NSLM [6]	**0.885**	**0.884**	**0.885**	**0.884**	-	-	-	-
Proposed	DECNet w/o CDE	0.892	0.896	0.890	0.891	0.879	0.858	0.845	0.848
	DECNet w/o KAE	0.874	0.875	0.873	0.874	0.871	0.844	0.844	0.844
	DECNet	**0.912**	**0.914**	**0.911**	**0.912**	**0.886**	**0.860**	**0.867**	**0.864**

Table 1: Result of different fake news detection methods on two datasets. The best performance is highlighted in bold red, and the second-best performance is highlighted in bold blue. –: missing or unreproducible results. *: open-sourced methods that are reproduced.

We conduct a performance comparison across two datasets, with the results presented in Table 1. Our model consistently ranks first, outperforming the second-best method by 2.7% on Weibo and 2.5% on Pheme in terms of accuracy, demonstrating its superior performance.

The performance of EANN and Spotfake is comparatively lower, likely due to their reliance on relatively simple feature extractors and the fact that their multimodal features are simply concatenated from text and visual features, without fully exploring the deep correlations between the two modalities. In contrast, the superior performance of HMCAN,MCAN and CAFE suggests that by incorporating similarity measures between modalities, these methods better capture the relationships between textual and visual information, thereby improving multimodal fusion and enhancing the model's ability to detect fake news. In contrast, our method achieves the best performance among all content-based models, surpassing the second-best model MCAN by a significant margin of 2.8% on the Weibo dataset, which highlights the benefits of incorporating external knowledge for enhancing fake news detection.

For the knowledge-enhanced models, our model outperforms all others by a significant margin across all two datasets. For instance, on the Weibo dataset, our model surpasses NSLM by 2.7%. This superior performance is attributed to the dual-expert design, which explores the deeper semantic information of the news, as well as leverages the intrinsic understanding of news content provided by LVLMs.

As shown in Table 1, the proposed method achieves the highest overall accuracy on the Pheme dataset. However, this advantage does not consistently extend to precision, recall, or F1 score. This may be due to the class imbalance in the Pheme dataset, which makes the model more likely to predict the majority class. While this can increase overall accuracy,

Fusion Strategy	Acc	Pre	Rec	F1
SUM-linear	0.909	0.911	0.907	0.908
SUM-sigmoid	0.906	0.906	0.905	0.905
MUL-sigmoid	0.910	0.910	0.911	0.910
SUM-tanh	0.906	0.906	0.905	0.905
MUL-tanh	0.887	0.890	0.885	0.887
DECNet	**0.912**	**0.914**	**0.911**	**0.912**

Table 2: Impact of the gating network on the Weibo dataset.

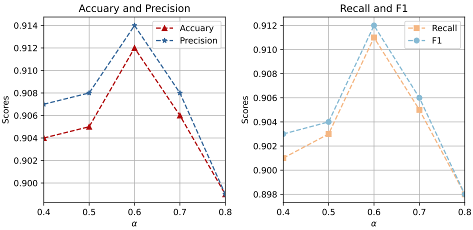

Figure 2: Sensitivity analysis of the parameter α in Eq. (9) on the Weibo dataset.

it can lower the recall for fake news and thus limit the F1 score. Additionally, the limited generalization ability of the Pheme dataset may also contribute to this phenomenon.

4.3 Ablation Study

Effectiveness of the Components. To investigate the individual contributions of each expert in our framework, we perform ablation studies by comparing different variants, as summarized in Table 1. Specifically, "w/o CDE" and "w/o KAE" denote versions of the model with the Content-Driven Expert and Knowledge-Augmentation Expert removed, respectively. Removing either expert leads to a significant decrease in performance on Weibo and Pheme, highlighting the essential role of both CDE and KAE in the model's overall effectiveness. These results not only demonstrate the independent contributions of each expert but also emphasize the synergistic effect they create when integrated into DECNet for fake news detection.

Impact of the Gating Network. To validate the effectiveness of our gating strategy, we examine the performance of different gating network configurations. Building on previous works [2, 4], we implement several configurations to integrate the content-based representation H_{con} and knowledge-enhanced representation H_{know}, as described in Eq. (10).

$$\begin{cases} \text{SUM-linear:} & \hat{y} = H_{know} + H_{con} \\ \text{SUM-sigmoid:} & \hat{y} = H_{know} + \sigma(H_{con}) \\ \text{MUL-sigmoid:} & \hat{y} = H_{know} * \sigma(H_{con}) \\ \text{SUM-tanh:} & \hat{y} = H_{know} + \tanh(H_{con}) \\ \text{MUL-tanh:} & \hat{y} = H_{know} * \tanh(H_{con}) \end{cases} \tag{10}$$

The results are shown in Table 2, where our proposed gating network outperforms all the other configurations, highlighting its effectiveness and suitability for the task.

Parameter Sensitivity Analysis. To assess the sensitivity of hyperparameters in Eq. (9), we conduct an ablation study by varying the value of α on the Weibo dataset, as shown in Figure 2. A smaller value of α (e.g., 0.5) prioritizes binary classification, potentially weakening multi-modal consistency. On the other hand, a larger value of α (e.g., 0.7) places greater emphasis on consistency, potentially undermining the direct classification task. The best performance is achieved when $\alpha = 0.6$, which strikes an optimal balance between binary loss and consistency loss, allowing the model to effectively reconcile both objectives.

5 Conclusion

In this paper, we propose the Dual-Expert Collaborative Network (DECNet) for multi-modal fake news detection, a novel approach that seamlessly integrates original news modalities with external knowledge. By combining visual embeddings from multi-scale image data and textual information through the Content-Driven Expert (CDE), and enhancing reasoning capabilities with the Knowledge-Augmentation Expert (KAE), our method uncovers deeper, more nuanced cues of fake news. Extensive experiments on two publicly available datasets (Weibo and Pheme) show that DECNet significantly outperforms existing state-of-the-art models, validating its effectiveness in tackling the complex challenges of fake news detection.

Acknowledgment. This work is sponsored by National Natural Science Foundation of China (62472025).

References

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[2] Yuyan Bu, Qiang Sheng, Juan Cao, Peng Qi, Danding Wang, and Jintao Li. Fakingrecipe: Detecting fake news on short video platforms from the perspective of creative process. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1351–1360, 2024.

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

[4] Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 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, 2023.

[5] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25 (70):1–53, 2024.

[6] Yiqi Dong, Dongxiao He, Xiaobao Wang, Youzhu Jin, Meng Ge, Carl Yang, and Di Jin. Unveiling implicit deceptive patterns in multi-modal fake news via neuro-symbolic reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8354–8362, 2024.

[7] Beizhe Hu, Qiang Sheng, Juan Cao, Yuhui Shi, Yang Li, Danding Wang, and Peng Qi. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22105–22113, 2024.

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

[9] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.

[10] An Lao, Qi Zhang, Chongyang Shi, Longbing Cao, Kun Yi, Liang Hu, and Duoqian Miao. Frequency spectrum is more effective for multimodal representation and fusion: A multimodal spectrum rumor detector. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18426–18434, 2024.

[11] Hongzhan Lin, Zixin Chen, Ziyang Luo, Mingfei Cheng, Jing Ma, and Guang Chen. Cofipara: A coarse-to-fine paradigm for multimodal sarcasm target identification with large multimodal models. *arXiv preprint arXiv:2405.00390*, 2024.

[12] Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM on Web Conference 2024*, pages 2359–2370, 2024.

[13] Yang Liu and Yi-Fang Wu. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[14] Salman Bin Naeem and Rubina Bhatti. The covid-19 'infodemic': a new front for information professionals. *Health Information & Libraries Journal*, 37(3):233–239, 2020.

[15] Kartik Narayan, Harsh Agarwal, Surbhi Mittal, Kartik Thakral, Suman Kundu, Mayank Vatsa, and Richa Singh. Desi: Deepfake source identifier for social media. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2858–2867, 2022.

[16] OpenAI. Chatgpt. https://openai.com/blog/chatgpt, 2023.

[17] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng 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*, pages 153–162, 2021.

[18] Kai Shu, Guoqing Zheng, Yichuan Li, Subhabrata Mukherjee, Ahmed Hassan Awadallah, Scott Ruston, and Huan Liu. Leveraging multi-source weak social supervision for early detection of fake news. *arXiv preprint arXiv:2004.01732*, 2020.

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

[20] Shivangi Singhal, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, Tanmoy Chakraborty, and Ponnurangam Kumaraguru. Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13915–13916, 2020.

[21] Mengzhu Sun, Xi Zhang, Jianqiang Ma, Sihong Xie, Yazheng Liu, and S Yu Philip. Inconsistent matters: A knowledge-guided dual-consistency network for multi-modal rumor detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(12): 12736–12749, 2023.

[22] Yu-Wun Tseng, Hui-Kuo Yang, Wei-Yao Wang, and Wen-Chih Peng. Kahan: knowledge-aware hierarchical attention network for fake news detection on social media. In *Companion Proceedings of the Web Conference 2022*, pages 868–875, 2022.

[23] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[24] Bing Wang, Shengsheng Wang, Changchun Li, Renchu Guan, and Ximing Li. Harmfully manipulated images matter in multimodal misinformation detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2262–2271, 2024.

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

[26] Lianwei Wu, Linyong Wang, and Yongqiang Zhao. Unified evidence enhancement inference framework for fake news detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6541–6549, 2024.

[27] Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 2560–2569, 2021.

[28] Yanzhi Xu, Yueying Hua, Shichen Li, and Zhongqing Wang. Exploring chain-of-thought for multi-modal metaphor detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 91–101, 2024.

[29] Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2733–2743, 2023.

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

[31] Zhi Zeng, Mingmin Wu, Guodong Li, Xiang Li, Zhongqiang Huang, and Ying Sha. An explainable multi-view semantic fusion model for multimodal fake news detection. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1235–1240. IEEE, 2023.

[32] Huaiwen Zhang, Quan Fang, Shengsheng Qian, and Changsheng Xu. Multi-modal knowledge-aware event memory network for social media rumor detection. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1942–1951, 2019.

[33] Litian Zhang, Xiaoming Zhang, Ziyi Zhou, Feiran Huang, and Chaozhuo Li. Reinforced adaptive knowledge learning for multimodal fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16777–16785, 2024.

[34] Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021*, pages 3465–3476, 2021.

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

[36] Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. Multimodal fake news detection via clip-guided learning. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2825–2830. IEEE, 2023.

[37] Arkaitz Zubiaga, Maria Liakata, and Rob Procter. Exploiting context for rumour detection in social media. In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part I 9*, pages 109–123. Springer, 2017.