The Fortieth AAAI Conference on Artificial Intelligence (AAAI-26)

From Blind Transfer to Wise Selection: Prototype-Driven Neighbor-Domain Adaptation for Fake News Detection

Wayne Lu[1*], **Yiheng Li**[2†*]

[1]Independent Researcher
[2]University of International Business and Economics, Beijing, China
lu.wayne0603@gmail.com, 23349096@suibe.edu.cn

Abstract

Multimodal fake news detection across different domains is hampered by the critical challenge of negative transfer, which arises from the indiscriminate fusion of knowledge from all available source domains. Existing methods attempt to learn domain-invariant features or leverage external knowledge but often aggregate information from all domains equally. However, these approaches largely ignore the asymmetric relationships between domains, leading to performance degradation when irrelevant or conflicting knowledge is introduced. To address this, we propose a novel **PANDA** (**P**rototype-driven **A**symmetric **N**eighbor-**D**omain **A**daptation) framework that dynamically selects and integrates knowledge from only the most beneficial domains. Initially, PANDA employs a **D**omain-**a**ware **M**odal **P**rompt **G**eneration (DMPG) module to learn transferable knowledge representations for each domain. We then introduce a novel **P**rototype-based **A**symmetric **D**istance (PAD) to quantify directional domain transferability, which guides a **G**umbel-based **N**eighbor **S**elector (GNS) to identify the most relevant neighbor domains. Subsequently, a **D**omain-**C**ollaborative **A**ttention (DCA) module adaptively fuses the selected knowledge to enhance the target domain's representation. Extensive experiments on three benchmarks demonstrate PANDA's superiority, outperforming state-of-the-art baselines with an F1-score improvement of 1.5% on the Weibo-21 dataset.

Code — https://github.com/lu-wayne/panda

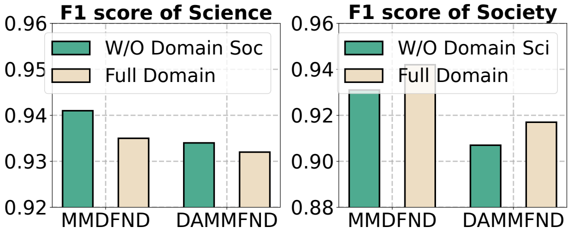

Figure 1: The performence of MMDFND(Tong et al. 2024) and DAMMFND(Lu, Tong, and Ye 2025) with selected different domains on Weibo21 Datasets.

Introduction

The rampant spread of fake news poses a profound threat to social stability and public trust, making its automatic detection a critical area of research with widespread applications (Segura-Bedmar and Alonso-Bartolome 2022; Zeng et al. 2024; Li et al. 2025h). In the contemporary media landscape, news is predominantly multimodal, composed of both textual articles and associated imagery. This characteristic renders traditional unimodal detection methods inadequate, as they fail to capture the subtle yet crucial cross-modal inconsistencies often exploited in sophisticated disinformation campaigns. Consequently, multimodal fake news detection, which jointly analyzes textual and visual content, has emerged as a more effective and necessary paradigm (Jing et al. 2023). Furthermore, news span a multitude of domains, such as politics, technology, and health. These domains exhibit distinct characteristics and patterns of falsehood. The significant domain shift between them often cripples models trained on one domain when applied to another, which has spurred the development of multi-domain multimodal fake news detection to enhance model generalization and robustness (Wang et al. 2023; Li et al. 2025e; Cui et al. 2025a).

To tackle the challenge of domain shift, existing multi-domain fake news detection research has largely progressed along two main avenues (Zhang, Zhang, and Yuan 2024; Zeng et al. 2025a). The first category consists of *domain adaptation-aware methods* (Wang et al. 2025; Li et al. 2023a; Chen et al. 2024). These approaches typically focus on learning domain-invariant features by aligning the feature distributions of different domains (Zeng et al. 2025b; Li et al. 2023b, 2024), often using techniques like adversarial training or moment matching. The goal is to create a unified representation that is robust across domains. The second category involves leveraging *domain-specific external knowledge* (Hu et al. 2021; Li et al. 2025b). These methods enrich the model's understanding by injecting external information, such as knowledge graphs tailored to specific domains (e.g., a medical knowledge base for health news), to provide contextual cues for veracity checking (Li et al. 2025d). While both lines of work have advanced the field, they implicitly assume that knowledge transfer from all available source domains is beneficial, an assumption that does not always hold true (Li et al. 2025c; Xu et al. 2025; Li et al. 2025a,g).

---

[*]These authors contributed equally.

[†]Corresponding author.

Despite significant progress, existing multi-domain methods face a core and persistent challenge: the **negative transfer problem** (Li et al. 2025f, 2023c; Lu and Yin 2025; Mo et al. 2024, 2025). This issue arises because current models do not effectively model the complex and often asymmetric relationships between domains (Cui et al. 2025b; Tong et al. 2025; Li et al. 2025c, 2023d). **When knowledge conflicts exist or when a domain is of low quality or relevance, the undifferentiated use of information from all available domains can introduce noise and bias, severely degrading detection performance on the target domain**. For instance, consider the task of detecting fake news in the "Science" domain (target domain). A model that indiscriminately incorporates knowledge from the "Society" domain (source domain) may suffer from negative transfer. The "Society" domain's fake news patterns, such as sensational language and digitally altered social photos, are largely irrelevant for identifying falsified scientific data or misinterpreted clinical trials in a "Science" article. As illustrated in Figure 1, this relationship is often asymmetric: while knowledge from the "Science" domain (e.g., identifying statistical fallacies) might help debunk a person's pseudoscientific science claims, the reverse is rarely true. Blindly transferring knowledge from "Society" would thus confuse the model and harm its performance in the "Science" domain.

To overcome the aforementioned limitation, we propose a novel framework: **PANDA** (**P**rototype-driven **A**symmetric **N**eighbor-**D**omain **A**daptation for Fake News Detection). PANDA is designed to explicitly combat negative transfer by dynamically selecting and integrating knowledge from only the most beneficial domains. First, to capture the unique essence of each domain, we introduce a **D**omain-**a**ware **M**odal **P**rompt **G**eneration (DMPG) mechanism, which generates learnable prompts to represent the core modal-level knowledge for each specific domain. These prompts inform the learning of representative *domain prototypes*, which serve as the foundation for our **P**rototype-based **A**symmetric **D**istance (PAD) metric, a novel measure designed to effectively quantify the directional transferability from a source domain to a target domain. Second, guided by the PAD metric, we develop a **G**umbel-based **N**eighbor **S**elector (GNS), which performs a differentiable top-k selection to identify the most relevant source domains for a given news instance. Finally, we design a **D**omain-**C**ollaborative **A**ttention (DCA) module. This module uses a cross-attention mechanism where the target domain's prompts act as the query, and the prompts from the selected neighbor domains serve as keys and values, enabling an adaptive and nuanced fusion of beneficial knowledge.

The main contributions of this paper are as follows:

- We identify and formulate the negative transfer problem in multi-domain fake news detection as a failure to model asymmetric domain relationships, and propose a new paradigm, PANDA, to explicitly address this issue. To the best of our knowledge, this is the first work to tackle this problem through prototype-driven asymmetric knowledge transfer.
- We design a novel Prototype-based Asymmetric Distance (PAD) metric. By learning compact domain prototypes from domain-specialized features and calculating a directional distance, PAD provides a more accurate and interpretable measure of domain relevance for knowledge transfer compared to conventional symmetric metrics.
- We propose a dynamic knowledge transfer architecture consisting of a Gumbel-based Neighbor Selector (GNS) and a Domain-Collaborative Attention (DCA) module. This architecture, powered by domain-aware modal prompts, enables the model to adaptively select the most beneficial source domains and integrate their knowledge, effectively mitigating negative transfer.
- We conduct extensive experiments on several public benchmark datasets. The results demonstrate that PANDA achieves new state-of-the-art performance and significantly outperforms existing methods, confirming the effectiveness and robustness of our proposed approach.

Methodology

To address the challenge of negative transfer, we propose PANDA, a novel framework designed for wise domain selection and adaptive knowledge fusion. The overall architecture of PANDA is illustrated in Figure 2. It consists of four main components: (1) a Domain-aware Multi-view Feature Extractor that generates domain-specialized representations; (2) a Domain-aware Modal Prompt Generator that learns domain-specific knowledge cues; (3) a Prototype-based Asymmetric Distance (PAD) metric calculator that measures directional domain similarity; (4) a Dynamic Neighbor Selector and Collaborator that identifies and fuses knowledge from the most relevant domains.

Domain-aware Multi-view Feature Extraction

Given a news article $\mathcal{N}_i = (T_i, V_i)$ with domain label d_i, we first extract its multimodal features.

Multi-view Feature Extraction We employ a suite of pre-trained encoders to capture features from three distinct views: textual, visual, and fused multimodal. A text encoder (e.g., BERT (Devlin et al. 2019)) processes the text T_i to obtain a textual feature vector $\mathbf{e}_{t,i} \in \mathbb{R}^{d_t}$. A vision encoder (e.g., ViT (Dosovitskiy et al. 2020)) processes the image V_i to yield a visual feature vector $\mathbf{e}_{v,i} \in \mathbb{R}^{d_v}$. Additionally, a multimodal encoder (e.g., CLIP (Radford et al. 2021)) takes both T_i and V_i as input to produce a fused cross-modal feature $\mathbf{e}_{mm,i} \in \mathbb{R}^{d_{mm}}$. These features are projected into a common latent space and concatenated to form an initial comprehensive representation: $\mathbf{e}_{m,i} = [\text{Proj}_t(\mathbf{e}_{t,i}); \text{Proj}_v(\mathbf{e}_{v,i}); \text{Proj}_{mm}(\mathbf{e}_{mm,i})] \in \mathbb{R}^{d_m}$.

Domain-aware Mixture-of-Experts To capture domain-specific nuances, we utilize a Mixture-of-Experts (MoE) architecture, which we term the Domain-gated Expert Network (DExNet). DExNet consists of a set of L_{exp} shared expert networks $\{E_l\}_{l=1}^{L_{exp}}$ and a domain-specific gating network G. Each expert E_l is a small feed-forward network. The gating network G takes the comprehensive representation $\mathbf{e}_{m,i}$ and a learnable domain embedding $\mathbf{d}_i \in \mathbb{R}^{d_{dom}}$

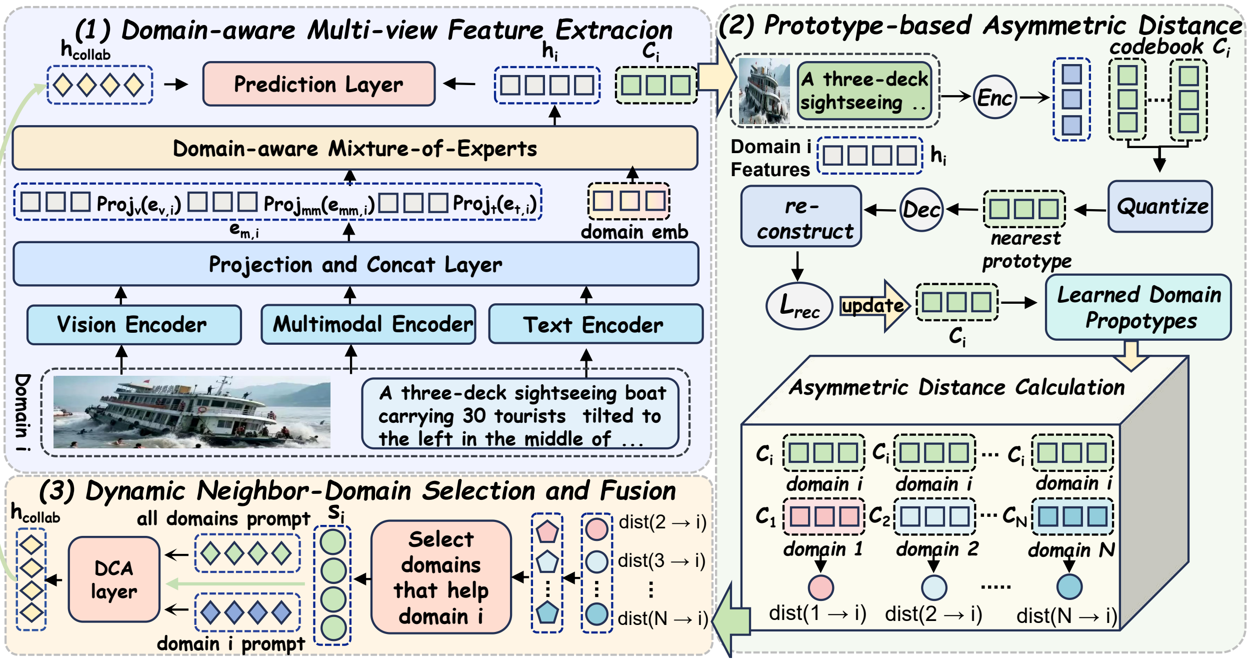

Figure 2: The architecture of PANDA framework. Domain-aware Multi-view Feature Extraction extracts and fuses features across modalities and domains. Then, the "Prototype" mechanism is employed to quantify domain differences. Finally, neighbor domains are dynamically selected by similarity, and combined with domain prompts, to achieve positive information transfer.

for the domain d_i as input, and produces a set of weights for the experts:

$$\alpha_{i,1},\ldots,\alpha_{i,L_{exp}} = \text{Softmax}(G([\mathbf{e}_{m,i};\mathbf{d}_i])). \tag{1}$$

The final domain-specialized feature representation $\mathbf{h}_{d_i}$ is the weighted sum of the outputs from all experts:

$$\mathbf{h}_{d_i} = \sum_{l=1}^{L_{exp}} \alpha_{i,l} \cdot E_l(\mathbf{e}_{m,i}). \tag{2}$$

This allows the model to learn a customized feature representation for each news article that is sensitive to its domain context.

Domain-aware Modal Prompt Generation

To explicitly model the transferable knowledge of each domain, inspired by (Wen et al. 2025), we introduce a set of learnable domain-aware modal prompts. For each domain $d \in \mathbb{D}$ and each modality view $m \in \{t, v, mm\}$, we initialize a set of L_p prompt vectors:

$$\mathbf{\Pi}_{d,m} = \{\boldsymbol{\pi}_{d,m,l} \in \mathbb{R}^{D_p}\}_{l=1}^{L_p} \tag{3}$$

where $\mathbf{\Pi}_{d,m}$ is the collection of prompts for domain d and modality view m, and D_p is the dimension of each prompt vector. These prompts are learnable parameters, designed to distill the core characteristics of each modality view within a specific domain. They serve as condensed representations of domain-specific knowledge and will later be used as queries to seek collaborative information. For each domain d, we form a unified domain prompt set by concatenating the prompts from all three views, $\mathbf{\Pi}_d = [\mathbf{\Pi}_{d,t}; \mathbf{\Pi}_{d,v}; \mathbf{\Pi}_{d,mm}] \in \mathbb{R}^{3L_p \times D_p}$.

Prototype-based Asymmetric Distance

To quantify the transferability between domains, we first learn a set of representative prototypes for each domain and then compute an asymmetric distance metric based on them.

Domain Prototype Learning For each domain $d \in \mathbb{D}$, we aim to learn a set of M prototypes $\mathcal{C}_d = \{\mathbf{c}_{d,j}\}_{j=1}^{M}$, where each $\mathbf{c}_{d,j} \in \mathbb{R}^{d_h}$ encapsulates a salient characteristic of that domain. These prototypes are learned from the set of all domain-specialized features $\{\mathbf{h}_{d_i} | d_i = d\}$ belonging to domain d. We employ a vector quantization approach within an autoencoder framework (Cao et al. 2023). An encoder maps $\mathbf{h}_{d_i}$ to a latent representation, which is then quantized by finding the nearest prototype in the codebook $\mathcal{C}_d$. A decoder then reconstructs the original feature from the quantized representation. The prototypes (the codebook vectors) are updated via the reconstruction loss $\mathcal{L}_{rec}$:

$$\mathcal{L}_{rec} = \mathbb{E}_{i:d_i=d}\left[||\text{Dec}(\text{Quantize}(\text{Enc}(\mathbf{h}_{d_i}))) - \mathbf{h}_{d_i}||_2^2\right]. \tag{4}$$

This process ensures that the prototypes $\mathcal{C}_d$ become compact and representative summaries of domain d.

Asymmetric Distance Calculation With the learned prototypes, we define the **P**rototype-based **A**symmetric **D**istance (PAD) from a source domain s to a target domain t, denoted as $dist(s \rightarrow t)$. It measures how well the knowledge from domain s can be applied to domain t. This is calculated as the average minimum distance from each prototype of the source domain to the set of prototypes in the target domain:

$$dist(s \to t) = \frac{1}{M} \sum_{j=1}^{M} \min_{\mathbf{c}_{t,l} \in \mathcal{C}_t} ||\mathbf{c}_{s,j} - \mathbf{c}_{t,l}||_2. \quad (5)$$

This distance is inherently asymmetric, i.e., $dist(s \to t) \neq dist(t \to s)$, reflecting the directional nature of knowledge transfer. A lower $dist(s \to t)$ indicates higher relevance of domain s to domain t.

Dynamic Neighbor-Domain Selection and Fusion

Guided by the PAD metric, we dynamically select the most beneficial neighbor domains and fuse their knowledge using an attention mechanism.

Gumbel-based Neighbor Selector (GNS) For a news item from a target domain t, we first compute its transferability score with respect to all other potential source domains $s \in \mathbb{D}$. The similarity is the inverse of the distance, $sim(s \to t) = 1/(dist(s \to t) + \epsilon)$, where ϵ is a small constant for stability. To select the top-S most beneficial domains in a differentiable manner, we employ the Gumbel-Softmax trick (Jang, Gu, and Poole 2016). We generate a selection vector $\mathbf{s}_t \in \mathbb{R}^K$:

$$\mathbf{s}_t = \text{Gumbel-Softmax}(\log(\text{sim}(\cdot \to t))), \quad (6)$$

where $\text{sim}(\cdot \to t)$ is the vector of similarity scores from all domains to domain t. The Gumbel-Softmax outputs a sparse probability distribution, effectively performing a soft selection of the top domains, which becomes a hard one-hot selection at test time.

Domain-Collaborative Attention (DCA) To fuse knowledge from the selected domains, we use the domain prompt sets in a cross-attention mechanism. For a target domain t, its prompt set $\mathbf{\Pi}_t$ serves as the Query. The prompt sets of all domains, concatenated into a global prompt bank $\mathbf{\Pi}_{all} = [\mathbf{\Pi}_1; \mathbf{\Pi}_2; \ldots; \mathbf{\Pi}_K] \in \mathbb{R}^{(K \cdot 3L_p) \times D_p}$, serve as the source for Keys and Values. The attention is modulated by the GNS selection vector $\mathbf{s}_t$.

$$\mathbf{Q} = \mathbf{\Pi}_t W_Q, \quad \mathbf{K} = \mathbf{\Pi}_{all} W_K, \quad \mathbf{V} = \mathbf{\Pi}_{all} W_V, \quad (7)$$

$$\text{scores} = \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} + \mathbf{M}_{mask}, \quad (8)$$

$$\mathbf{H}_{collab} = \text{Softmax}(\text{scores})\mathbf{V}, \quad (9)$$

where W_Q, W_K, W_V are learnable projection matrices. $\mathbf{M}_{mask} \in \mathbb{R}^{3L_p \times (K \cdot 3L_p)}$ is a mask matrix derived from the GNS selection vector $\mathbf{s}_t$. The mask adds a large negative value to the attention scores corresponding to non-selected domains, effectively removing their contribution. The resulting matrix $\mathbf{H}_{collab} \in \mathbb{R}^{3L_p \times D_p}$ contains the contextually enriched prompts for the target domain. We aggregate this information via mean pooling to get the final collaborated knowledge vector: $\mathbf{h}_{collab} = \text{Mean}(\mathbf{H}_{collab})$.

Prediction and Loss Function

Finally, the collaborated knowledge is combined with the instance-specific feature to make the final veracity prediction.

Final Classification The domain-collaborated knowledge $\mathbf{h}_{collab}$ is fused with the domain-specialized feature $\mathbf{h}_{d_i}$ for the given news item. This is achieved through a simple concatenation followed by a multi-layer perceptron (MLP):

$$\mathbf{h}_{final} = \text{MLP}([\mathbf{h}_{d_i}; \mathbf{h}_{collab}]). \quad (10)$$

The final representation $\mathbf{h}_{final}$ is then passed through a final classification layer with a sigmoid activation function to produce the probability of the news being fake, $\hat{y}_i = \sigma(\text{Linear}(\mathbf{h}_{final}))$.

Optimization Objective The overall model is trained by minimizing a composite loss function $\mathcal{L}_{total}$, which includes the primary classification loss and the auxiliary prototype reconstruction loss:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{rec}, \quad (11)$$

where λ is a hyperparameter to balance the two terms. The classification loss $\mathcal{L}_{cls}$ is the standard binary cross-entropy (BCE) loss:

$$\mathcal{L}_{cls} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right]. \quad (12)$$

The reconstruction loss $\mathcal{L}_{rec}$ is the sum of the prototype learning losses across all domains, ensuring the quality of the learned prototypes. By optimizing this joint objective, PANDA learns to not only classify news but also to understand the intricate relationships between domains for effective knowledge transfer. Furthermore, we provide detailed theoretical proofs to demonstrate that our scheme can effectively alleviate the problem of negative transfer between domains, which can be found in the supplementary materials.

Experiments

In this section, we conduct extensive experiments on three public datasets to answer the following research questions:

- **(Effectiveness) RQ1:** Does our proposed PANDA outperform state-of-the-art multi-domain and multimodal fake news detection baselines?
- **(Ablation) RQ2:** How do the key components within PANDA contribute to its overall performance?
- **(Sensitivity) RQ3:** How sensitive is PANDA to its key hyperparameters?
- **(Analysis) RQ4:** Can PANDA effectively mitigate negative transfer from irrelevant domains?
- **(Efficiency) RQ5:** What is the computational efficiency of PANDA in terms of parameters and inference time?
- **(Visualization) RQ6:** Can PANDA learn discriminative and domain-adaptive representations for detection?

Experimental Settings

Datasets Following previous work (Tong et al. 2024), our model is evaluated on three real-world datasets: two Chinese multi-domain datasets, Weibo (Wang et al. 2018) and Weibo-21 (Nan et al. 2021), and one English multi-domain dataset, FineFake (Zhou et al. 2024). The Weibo and Weibo-21 datasets are divided into nine domains, while FineFake contains six domains and one uncategorized domain. We use the same data partitioning method as MMDFND (Tong et al. 2024) and KEAN (Zhou et al. 2024), and more detailed dataset descriptions can be found in the appendix of the supplementary materials.

Baselines To comprehensively assess the performance of our proposed model, we compare it against baseline methods organized into three categories: (1) single-modal multi-domain methods, including: **MMoE** (Ma et al. 2018), **MoSE** (Qin et al. 2020), **MDFEND** (Nan et al. 2021), **M^{3}DFEND** (Zhu et al. 2022) and **PLDFEND** (Peng et al. 2023); (2) multi-modal multi-domain methods, including: **KATMF** (Song et al. 2021), **MMDFND** (Tong et al. 2024) and **DAMMFND** (Lu, Tong, and Ye 2025); Descriptions of baseline methods and more relate works can be found in the appendix of the supplementary materials.

Implementation Details We implement our model using PyTorch (Paszke et al. 2019) and conduct all experiments on NVIDIA RTX 4090 GPUs. For text processing, we use a pre-trained BERT (Devlin et al. 2019) model, setting the maximum sequence length to 197. For image processing, we use a pre-trained ViT (Dosovitskiy et al. 2020) model with input images resized to 224×224 pixels. For the initial multimodal representations, we utilize CN-CLIP (Yang et al. 2022) for the Chinese datasets and CLIP (Radford et al. 2021) for the English dataset. In our Domain-gated Expert Network (DExNet), the number of experts L_{exp} is set to 8. For the Domain-aware Modal Prompt Generator (DMPG), we use $L_p = 4$ prompts per modality view. The number of domain prototypes M for the PAD metric is set to 16. The Gumbel-based Neighbor Selector (GNS) is configured to select the top-$S = 2$ neighbor domains. We train the model using the AdamW optimizer with a learning rate of 1e-4 and a batch size of 32. The balancing hyperparameter λ for the reconstruction loss is set to 0.1. Performance is evaluated using Accuracy, F1-score, and AUC.

Overall Performance (RQ1)

Table 1 presents the overall performance comparison between PANDA and various baseline methods. The results clearly demonstrate the superiority of our proposed approach across all three benchmark datasets.

(1) PANDA achieves new state-of-the-art performance. As shown, PANDA consistently outperforms all baselines in terms of overall F1, Accuracy, and AUC. For instance, on Weibo-21, PANDA achieves an F1-score of 0.958, surpassing the previous best method, DAMMFND, by 1.5%. This significant improvement is attributed to PANDA's ability to mitigate negative transfer by wisely selecting and integrating knowledge from the most beneficial neighbor domains, leading to more robust and generalized representations.

(2) Multimodal multi-domain methods demonstrate superior performance. The results also indicate that multimodal methods (e.g., MMDFND, DAMMFND, PANDA) significantly outperform single-modal methods (e.g., MDFEND, M^{3}DFEND). This highlights the critical importance of leveraging cross-modal information for detecting fake news. PANDA further advances this by not only exploiting multimodal cues but also explicitly modeling the asymmetric relationships between domains, which allows for more effective knowledge transfer and ultimately leads to its superior performance.

Ablation Study (RQ2)

We conduct a series of ablation experiments to validate the effectiveness of each key component in PANDA. The results, summarized in Table 2.

Effect of Neighbor-Domain Selection. We design two variants to analyze core selection mechanism. (1) **-GNS/PAD (Blind Transfer)** removes both the GNS and the PAD metric, forcing model to aggregate information from all domains indiscriminately. (2) **-PAD (Symm. Dist.)** replaces our Prototype-based Asymmetric Distance with a symmetric one (e.g., Euclidean distance between domain prototype centroids). The significant performance drop in the "Blind Transfer" variant confirms our central hypothesis that indiscriminate knowledge fusion leads to negative transfer. The lesser, yet still notable, decline in the "-PAD (Symm. Dist.)" variant validates the importance of modeling asymmetric nature of domain relationships for valid knowledge transfer.

Effect of Knowledge Representation and Fusion. We analyze the prompt generator and the fusion module with two variants. (1) **-DMPG** removes the Domain-aware Modal Prompt Generator and uses simple learnable domain embeddings as queries for attention. (2) **-DCA (Avg. Fusion)** replaces the Domain-Collaborative Attention with a simple averaging of the selected neighbor domains' prompt features. The performance degradation in both cases underscores the benefits of our designs. The results for -DMPG show that learning rich, modal-specific prompts is superior to a single domain vector. Similarly, the results for -DCA (Avg. Fusion) demonstrate that the attention mechanism provides a more nuanced and effective fusion of knowledge compared to simple averaging.

Parameter Sensitivity Analysis (RQ3)

We analyzed the sensitivity of PANDA to four key hyperparameters on the Weibo-21 dataset: the number of domain prototypes M, the number of selected neighbor domains S, the dimension of domain prompts D_p, and the loss balancing coefficient λ. As shown in Figure 3, PANDA demonstrates robust performance across a reasonable range for all tested hyperparameters. For instance, performance peaks when the number of prototypes M is 16 and the number of selected neighbors S is 3, indicating that a moderate number of prototypes can effectively capture domain characteristics and that fusing knowledge from a few highly relevant domains is more effective than including too many. The model is relatively stable with prompt dimensions between 64 and 256. This overall robustness reduces the burden of extensive hyperparameter tuning and demonstrates the stability of our proposed framework.

Negative Transfer Mitigation Effect (RQ4)

To explicitly verify that *PANDA* can mitigate negative transfer, we designed a targeted experiment on the *Weibo-21*

	Method	Sci.	Mil. Con.	Edu. Unc.	Soc.	Pol.	Hlth.	Fin.	Ent.	Dis. Int	Overall		
											F1	Acc	AUC
Weibo	MMoE*	0.576	0.913	0.849	0.887	0.733	0.829	0.811	0.832	0.881	0.875	0.875	0.949
	MoSE*	0.795	0.736	0.837	0.910	0.766	0.857	0.793	0.842	0.885	0.889	0.889	0.955
	MDFEND*	0.772	0.914	0.895	0.905	0.761	0.876	0.810	0.879	0.876	0.903	0.903	0.966
	M^{3}DFEND*	0.794	0.901	0.925	0.910	0.768	0.861	0.901	0.897	0.878	0.927	0.927	0.970
	PLDFEND	0.840	0.915	0.882	0.927	0.921	0.920	0.926	0.898	0.447	0.926	0.929	0.975
	KATMF	0.833	0.906	0.921	0.897	0.825	0.895	0.901	0.906	0.892	0.931	0.931	0.968
	MMDFND*	0.826	0.909	0.939	0.941	0.732	0.915	0.914	0.919	0.886	0.933	0.933	0.973
	DAMMFND*	0.853	0.911	**0.956**	**0.943**	0.822	0.939	**0.937**	**0.956**	**0.928**	0.941	0.944	0.979
	PANDA	**0.935**	**0.947**	0.941	0.919	**0.942**	**0.976**	0.932	0.954	0.923	**0.951**	**0.953**	**0.987**
Weibo-21	MMoE*	0.877	0.909	0.872	0.873	0.865	0.934	0.858	0.886	0.879	0.893	0.893	0.955
	MoSE*	0.848	0.887	0.879	0.875	0.878	0.919	0.865	0.893	0.869	0.895	0.895	0.953
	MDFEND*	0.832	0.936	0.893	0.895	0.888	0.938	0.897	0.904	0.902	0.914	0.914	0.969
	M^{3}DFEND*	0.831	0.948	0.901	0.906	0.885	0.943	0.903	0.929	0.891	0.920	0.920	0.976
	PLDFEND	0.910	**0.959**	0.901	0.911	0.888	0.940	0.885	0.936	0.887	0.925	-	-
	KATMF	0.912	0.930	0.911	0.898	0.899	0.917	0.873	0.935	0.900	0.925	0.929	0.974
	MMDFND*	0.935	0.955	0.854	0.942	0.963	0.923	0.881	0.961	0.917	0.940	0.940	0.976
	DAMMFND*	0.932	0.953	**0.932**	0.917	**0.982**	0.919	0.941	**0.990**	**0.945**	0.943	0.946	0.979
	PANDA	**0.962**	0.946	0.927	**0.961**	0.913	**0.979**	**0.984**	0.942	0.923	**0.958**	**0.959**	**0.988**
FineFake	MMoE*	-	0.713	0.623	0.781	0.754	0.769	0.808	0.842	-	0.778	0.789	0.871
	MoSE*	-	0.692	0.598	0.785	0.767	**0.810**	0.820	0.799	-	0.781	0.786	0.874
	MDFEND*	-	-	-	-	-	-	-	-	-	0.781	0.788	-
	M^{3}DFEND*	-	-	-	-	-	-	-	-	-	0.772	0.781	-
	PLDFEND	-	0.724	0.638	0.785	0.766	0.777	0.829	0.835	-	0.789	0.790	0.873
	KATMF	-	0.727	0.701	0.758	0.756	0.805	0.839	0.846	-	0.782	0.786	0.869
	MMDFND*	-	**0.747**	0.740	0.760	0.768	0.767	0.831	0.839	-	0.786	0.789	0.868
	DAMMFND*	-	0.741	0.746	0.787	0.776	0.803	0.838	0.853	-	0.792	0.794	0.878
	PANDA	-	0.734	**0.757**	**0.799**	**0.792**	0.793	**0.843**	**0.861**	-	**0.803**	**0.805**	**0.884**

Table 1: Comparison between PANDA and the latest multi-domain fake news detection methods on Weibo, Weibo-21 and FineFake. *: open-source. The best results are in **bold**, and the second-best results are underlined. The t-tests validate the significance of performance improvements with p-value ≤ 0.05.

Variant	**Weibo**	**Weibo-21**	**FineFake**
PANDA (Full Model)	**0.951**	**0.958**	**0.803**
On Neighbor-Domain Selection:			
-GNS/PAD (Blind Transfer)	0.928	0.931	0.778
-PAD (Symm. Dist.)	0.941	0.946	0.791
On Knowledge Representation and Fusion:			
-DMPG	0.944	0.949	0.794
-DCA (Avg. Fusion)	0.947	0.952	0.798

Table 2: Ablation study of PANDA on three datasets. We report the F1-score. "-GNS/PAD" denotes blind transfer, "-PAD (Symm. Dist)" uses a symmetric distance, "-DMPG" removes domain prompts, and "-DCA (Avg. Fusion)" replaces attention with averaging.

dataset. We hypothesize that the broad and noisy "*Society*" domain could negatively impact the more technical "*Science*" domain for a standard model.We compare the performance of a strong baseline, *MMDFND*, with an enhanced version, *MMDFND+Ours*, which integrates our three proposed modules: *Domain-aware Modal Prompt Generation*, *Prototype-based Asymmetric Distance*, and *Dynamic Neighbor-Domain Selection and Fusion*. We evaluate their F1-score on the "*Science*" domain under two conditions: (1) training on the full dataset, and (2) training on a dataset where the "*Society*" domain has been removed.

The results are illustrated in Figure 4. For the standard *MMDFND* model, removing the "*Society*" domain *improves* its performance on the "*Science*" domain (from 0.935 to 0.942), which is a clear indicator of negative transfer. In contrast, *MMDFND+Ours* not only achieves a higher absolute performance but also *benefits* from the inclusion of the "*Society*" domain (its performance slightly decreases when the domain is removed), demonstrating that our selective mechanism successfully filters out the noisy signals while extracting useful information. This experiment provides direct and compelling evidence for *PANDA*'s ability to turn a potential source of negative transfer into a beneficial contributor.

Efficiency Analysis (RQ5)

We assess PANDA's practical applicability by analyzing its training and inference efficiency against several baselines on the Weibo-21 dataset. As detailed in Table 3, PANDA demonstrates a highly competitive performance profile. Its training time of 65.3 seconds per epoch is on par with other advanced multimodal models, indicating that its sophisticated domain selection mechanisms do not introduce prohibitive computational overhead.

More notably, PANDA excels in inference efficiency, processing a single sample in just 12.6 ms. This is signifi-

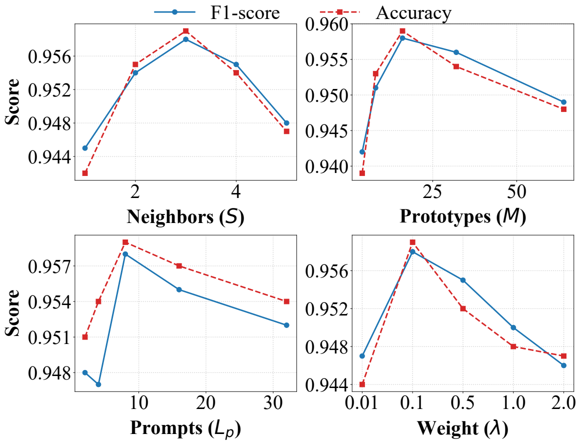

Figure 3: Parameter sensitivity analysis of four key hyperparameters on the Weibo-21 dataset, evaluated by F1-score and Accuracy.

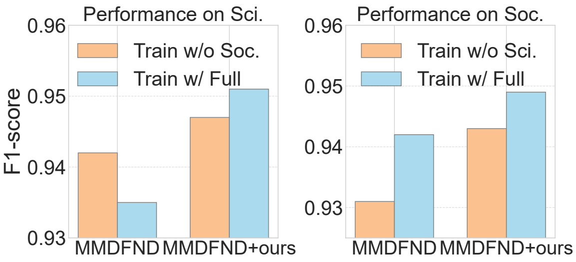

Figure 4: Analysis of negative transfer mitigation on the Weibo-21 dataset. Removing the ‘Society’ domain helps the baseline MMDFND on ‘Science’ news (negative transfer), while PANDA’s performance remains stable and high.

cantly faster than other top-performing multimodal baselines such as DAMMFND (15.2 ms/sample) and KATMF (15.8 ms/sample). This speed advantage is a direct result of our Gumbel-based Neighbor Selector, which prunes the computational graph by focusing only on the most relevant domains during knowledge fusion. This high throughput confirms that PANDA’s accuracy gains do not sacrifice deployability, making it well-suited for the rapid detection required by high-volume social media platforms like Twitter and Weibo.

Model	Modality	Train (s/epoch)	Inference (ms/sample)
MDFEND	Text	∼55.8	∼9.5
M^{3}FEND	Text	∼58.1	∼10.2
KATMF	Multimodal	∼75.5	∼15.8
MMDFND	Multimodal	∼70.4	∼14.1
DAMMFND	Multimodal	∼73.2	∼15.2
PANDA (Ours)	**Multimodal**	**65.3**	**12.6**

Table 3: Efficiency comparison on the Weibo-21 dataset.

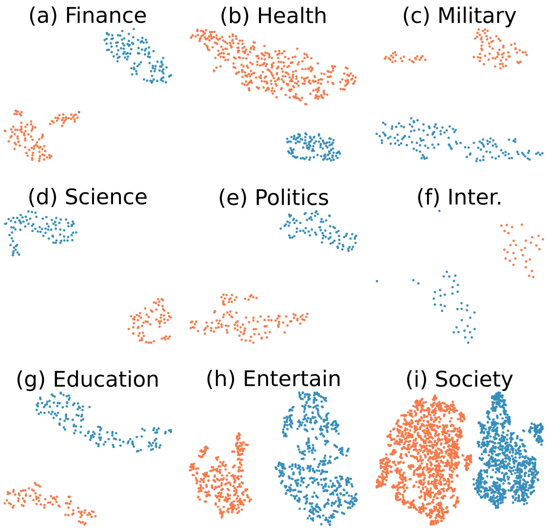

Figure 5: T-SNE showcase PANDA’s classification across various domains in the Weibo’s test dataset, where identically colored dots denote the same category labels.

Visualization Analysis (RQ6)

To verify whether PANDA can effectively alleviate the Imbalanced Domain Distribution and extract domain-adaptive representations from raw news, we visualized the representations before the output layer of the model using t-SNE (Maaten and Hinton 2008). As shown in Figure 5, despite the data sparsity issues in some domains (such as International and Science), PANDA still manages to obtain effective representations for these domains, which can be utilized by subsequent classifiers. This is because PANDA not only effectively captures the coarse-grained and fine-grained feature extraction processes across different domains but also transfers sufficient and effective information from data-rich domains (such as Health and Society) to data-sparse domains, thereby enhancing the performance of the latter.

Conclusion

In this paper, we tackled the negative transfer problem in multi-domain multimodal fake news detection, which arises from the unprincipled fusion of knowledge across domains. We proposed **PANDA**, a novel framework designed for prudent neighbor-domain selection and adaptive knowledge fusion. PANDA introduces a **P**rototype-based **A**symmetric **D**istance (PAD) to measure directional domain transferability, which in turn guides a **G**umbel-based **N**eighbor **S**elector (GNS) to identify the most relevant domains. Subsequently, knowledge from these selected domains, encapsulated as features learned by the **D**omain-aware **M**odal **P**rompt **G**eneration (DMPG) module, is adaptively integrated via a **D**omain-**C**ollaborative **A**ttention (DCA) module. Extensive experiments confirm that PANDA significantly outperforms state-of-the-art methods and effectively mitigates negative transfer.

References

Cao, B.; Bi, Z.; Hu, Q.; Zhang, H.; Wang, N.; Gao, X.; and Shen, D. 2023. Autoencoder-driven multimodal collaborative learning for medical image synthesis. *International Journal of Computer Vision*, 131(8): 1995–2014.

Chen, C.; Zhang, Y.; Li, Y.; Wang, J.; Qi, L.; Xu, X.; Zheng, X.; and Yin, J. 2024. Post-training attribute unlearning in recommender systems. *ACM Transactions on Information Systems*, 43(1): 1–28.

Cui, X.; Lu, W.; Tong, Y.; Li, Y.; and Zhao, Z. 2025a. Diffusion-based Multi-modal Synergy Interest Network for Click-through Rate Prediction. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 581–591.

Cui, X.; Lu, W.; Tong, Y.; Li, Y.; and Zhao, Z. 2025b. Multi-Modal Multi-Behavior Sequential Recommendation with Conditional Diffusion-Based Feature Denoising. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1593–1602.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. 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)*, 4171–4186.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Hu, L.; Yang, T.; Zhang, L.; Zhong, W.; Tang, D.; Shi, C.; Duan, N.; and Zhou, M. 2021. Compare to the knowledge: Graph neural fake news detection with external knowledge. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, 754–763.

Jang, E.; Gu, S.; and Poole, B. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Jing, J.; Wu, H.; Sun, J.; Fang, X.; and Zhang, H. 2023. Multimodal fake news detection via progressive fusion networks. *Information processing & management*, 60(1): 103120.

Li, Y.; Cai, H.; Kong, R.; Chen, X.; Chen, J.; Yang, J.; Zhang, H.; Li, J.; Wu, J.; Chen, Y.; et al. 2025a. Towards AI Search Paradigm. *arXiv preprint arXiv:2506.17188*.

Li, Y.; Cao, Y.; He, H.; Cheng, Q.; Fu, X.; Xiao, X.; Wang, T.; and Tang, R. 2025b. M²IV: Towards Efficient and Fine-grained Multimodal In-Context Learning via Representation Engineering. In *Second Conference on Language Modeling*.

Li, Y.; Chen, C.; Zhang, Y.; Liu, W.; Lyu, L.; Zheng, X.; Meng, D.; and Wang, J. 2023a. Ultrare: Enhancing receraser for recommendation unlearning via error decomposition. *Advances in Neural Information Processing Systems*, 36: 12611–12625.

Li, Y.; Chen, C.; Zheng, X.; Zhang, Y.; Han, Z.; Meng, D.; and Wang, J. 2023b. Making users indistinguishable: Attribute-wise unlearning in recommender systems. In *Proceedings of the 31st ACM International Conference on Multimedia*, 984–994.

Li, Y.; Feng, X.; Chen, C.; and Yang, Q. 2024. A survey on recommendation unlearning: Fundamentals, taxonomy, evaluation, and open questions. *arXiv preprint arXiv:2412.12836*.

Li, Y.; Xiong, H.; Kong, L.; Zhang, R.; Xu, F.; Chen, G.; and Li, M. 2023c. Mhrr: Moocs recommender service with meta hierarchical reinforced ranking. *IEEE Transactions on Services Computing*, 16(6): 4467–4480.

Li, Y.; Xiong, H.; Wang, Q.; Kong, L.; Liu, H.; Li, H.; Bian, J.; Wang, S.; Chen, G.; Dou, D.; et al. 2023d. COLTR: Semi-supervised learning to rank with co-training and over-parameterization for web search. *IEEE Transactions on Knowledge and Data Engineering*, 35(12): 12542–12555.

Li, Y.; Yang, J.; Shen, Z.; Han, L.; Xu, H.; and Tang, R. 2025c. CATP: Contextually Adaptive Token Pruning for Efficient and Enhanced Multimodal In-Context Learning. *arXiv preprint arXiv:2508.07871*.

Li, Y.; Yang, J.; Yang, Z.; Li, B.; He, H.; Yao, Z.; Han, L.; Chen, Y. V.; Fei, S.; Liu, D.; et al. 2025d. Cama: Enhancing multimodal in-context learning with context-aware modulated attention. *arXiv preprint arXiv:2505.17097*.

Li, Y.; Yang, J.; Yun, T.; Feng, P.; Huang, J.; and Tang, R. 2025e. Taco: Enhancing multimodal in-context learning via task mapping-guided sequence configuration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 736–763.

Li, Y.; Zhang, H.; Zhang, Y.; Cai, H.; Cai, M.; Wang, S.; Xiong, H.; Kong, L.; Yin, D.; and Chen, L. 2025f. Rankexpert: A mixture of textual-and-behavioral experts for multi-objective learning-to-rank in web search. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 4578–4589.

Li, Y.; Zhang, H.; Zhang, Y.; Ma, X.; Ye, W.; Song, N.; Wang, S.; Xiong, H.; Yin, D.; and Chen, L. 2025g. M 2 oERank: Multi-Objective Mixture-of-Experts Enhanced Ranking for Satisfaction-Oriented Web Search. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, 4441–4454. IEEE.

Li, Y.; Zhang, Y.; Liu, W.; Feng, X.; Han, Z.; Chen, C.; and Yan, C. 2025h. Multi-Objective Unlearning in Recommender Systems via Preference Guided Pareto Exploration. *IEEE Transactions on Services Computing*.

Lu, W.; Tong, Y.; and Ye, Z. 2025. DAMMFND: Domain-Aware Multimodal Multi-view Fake News Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 559–567.

Lu, W.; and Yin, L. 2025. DMMD4SR: Diffusion Model-based Multi-level Multimodal Denoising for Sequential Recommendation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 6363–6372.

Ma, J.; Zhao, Z.; Yi, X.; Chen, J.; Hong, L.; and Chi, E. H. 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.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov): 2579–2605.

Mo, M.; Lu, W.; Xie, Q.; Lv, X.; Xiao, Z.; Yang, H.; and Zhang, Y. 2024. MIN: Multi-stage Interactive Network for Multimodal Recommendation. In *International Conference on Web Information Systems Engineering*, 191–205. Springer.

Mo, M.; Lu, W.; Xie, Q.; Xiao, Z.; Lv, X.; Yang, H.; and Zhang, Y. 2025. One multimodal plugin enhancing all: CLIP-based pre-training framework enhancing multimodal item representations in recommendation systems. *Neurocomputing*, 637: 130059.

Nan, Q.; Cao, J.; Zhu, Y.; Wang, Y.; and Li, J. 2021. MDFEND: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 3343–3347.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Peng, Z.; Lin, N.; Zhou, Y.; Zhou, D.; and Yang, A. 2023. Prompt learning for low-resource multi-domain fake news detection. In *2023 international conference on asian Language Processing (IALP)*, 314–319. IEEE.

Qin, Z.; Cheng, Y.; Zhao, Z.; Chen, Z.; Metzler, D.; and 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.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Segura-Bedmar, I.; and Alonso-Bartolome, S. 2022. Multimodal fake news detection. *Information*, 13(6): 284.

Song, C.; Ning, N.; Zhang, Y.; and Wu, B. 2021. Knowledge augmented transformer for adversarial multidomain multiclassification multimodal fake news detection. *Neurocomputing*, 462: 88–100.

Tong, Y.; Lu, W.; Cui, X.; Mao, Y.; and Zhao, Z. 2025. DAPT: Domain-Aware Prompt-Tuning for Multimodal Fake News Detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 7902–7911.

Tong, Y.; Lu, W.; Zhao, Z.; Lai, S.; and Shi, T. 2024. MMDFND: Multi-modal multi-domain fake news detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1178–1186.

Wang, D.; Zhang, W.; Wu, W.; and Guo, X. 2023. Soft-label for multi-domain fake news detection. *IEEe Access*, 11: 98596–98606.

Wang, X.; Meng, J.; Zhao, D.; Meng, X.; and Sun, H. 2025. Fake news detection based on multi-modal domain adaptation. *Neural Computing and Applications*, 37(7): 5781–5793.

Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; and Gao, J. 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.

Wen, Y.; Liu, Y.; Xu, D.; Luo, H.; Jia, P.; Wu, Y.; Wang, S.; Liang, K.; Wang, M.; Wang, Y.; Zhuang, F.; and Zhao, X. 2025. Measure Domain's Gap: A Similar Domain Selection Principle for Multi-Domain Recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25)*. Preprint arXiv:2505.20227.

Xu, W.; Xiang, D.; Ding, T.; and Lu, W. 2025. MMM-Fact: A Multimodal, Multi-Domain Fact-Checking Dataset with Multi-Level Retrieval Difficulty. *arXiv preprint arXiv:2510.25120*.

Yang, A.; Pan, J.; Lin, J.; Men, R.; Zhang, Y.; Zhou, J.; and Zhou, C. 2022. Chinese CLIP: Contrastive Vision-Language Pretraining in Chinese. *arXiv preprint arXiv:2211.01335*.

Zeng, Z.; Luo, M.; Kong, X.; Liu, H.; Guo, H.; Yang, H.; Ma, Z.; and Zhao, X. 2024. Mitigating World Biases: A Multimodal Multi-View Debiasing Framework for Fake News Video Detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 6492–6500.

Zeng, Z.; Wu, J.; Luo, M.; Kong, X.; Ma, Z.; Dai, G.; and Zheng, Q. 2025a. Understand, Refine and Summarize: Multi-View Knowledge Progressive Enhancement Learning for Fake News Video Detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 9216–9225.

Zeng, Z.; Wu, J.; Luo, M.; Wan, H.; Kong, X.; Ma, Z.; Dai, G.; and Zheng, Q. 2025b. IMOL: Incomplete-Modality-Tolerant Learning for Multi-Domain Fake News Video Detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 30921–30933.

Zhang, G.; Zhang, S.; and Yuan, G. 2024. Bayesian graph local extrema convolution with long-tail strategy for misinformation detection. *ACM Transactions on Knowledge Discovery from Data*, 18(4): 1–21.

Zhou, Z.; Zhang, X.; Zhang, L.; Liu, J.; Wang, S.; Liu, Z.; Zhang, X.; Li, C.; and Yu, P. S. 2024. FineFake: A Knowledge-Enriched Dataset for Fine-Grained Multi-Domain Fake News Detection. *arXiv preprint arXiv:2404.01336*.

Zhu, Y.; Sheng, Q.; Cao, J.; Nan, Q.; Shu, K.; Wu, M.; Wang, J.; and Zhuang, F. 2022. Memory-guided multi-view multidomain fake news detection. *IEEE Transactions on Knowledge and Data Engineering*.