OVLAFND: Open-Vocabulary Logic-Aware Fake News Detection

Anonymous ACL submission

Abstract

Multimodal fake news detection is critical for maintaining the integrity of social media. Current models often rely on shallow cross-modal correlations and rigid domain labels, which limits robustness to logical inconsistency and domain shift. Existing methods attempt to handle such variance by relying on rigid, manually defined domain labels, which force continuous real-world events into discrete, artificial categories. We propose the Open-Vocabulary and Logic-Aware Fake News Detection (OVLAFND) framework to address these issues and relax traditional categorization constraints. To model fine-grained logical conflicts rather than surface relevance, our framework introduces Logical Adversarial Data Augmentation (LADA), a multimodal LLM-driven pipeline that constructs counterfactual text-image pairs as hard negatives. Then, we propose a Logic Consistency Module (LCM) that uses these adversarial signals to learn a dedicated logic-consistency subspace for assessing whether textual claims are supported by visual evidence. To improve topic flexibility, we transform the rigid domain labels into an open-vocabulary topic manifold (OVTM) to achieve highly interpretable and label-free topic reasoning. Across Weibo and Weibo-21, OVLAFND achieves competitive fixed-domain performance and better leave-one-domain-out generalization. The code can be found in the Anonymous GitHub at https://anonymous.4open.science/r/OVLAFND-A400.

1 Introduction

The proliferation of social media has fundamentally transformed information dissemination, simultaneously giving rise to an unprecedented surge in multimodal fake news. Characterized by the deceptive pairing of text and visual elements, these fabricated posts pose severe threats to public trust and social stability. Driven by the urgency of this issue, automated multimodal fake news detection (Zhu et al., 2022; Li et al., 2024a,b,c) has garnered immense academic interest. While existing detection models achieve commendable performance within single, well-defined domains, their efficacy drops precipitously in real-world scenarios. In the wild, news spans a vast and dynamic spectrum of topics, presenting a severe domain shift problem. Models trained on narrow datasets suffer from catastrophic performance degradation when confronted with unseen or overlapping events.

Recent advancements in multi-domain multimodal fake news detection (Zhu et al., 2023; Jing et al., 2023; Wu et al., 2023) have explored various adaptation and fusion strategies. MMDFND (Tong et al., 2024) incorporates a multi-modal consistency module to simultaneously capture semantic conflicts and image tampering traces, using domain-adaptive feature alignment to significantly improve cross-domain robustness. Furthermore, DAMMFND (Lu et al., 2025) mitigates the inability of predefined identifiers to adapt to complex data by extracting implicit domain information through decoupling modules. It utilizes a domain-aware multi-view discriminator to dynamically quantify the differentiated contributions of modalities within specific domains. Finally, CroMe (Choi et al., 2025) leverages a ternary Transformer with BLIP and Large Language Model (LLM) encoders, combining metric learning to optimize cross-modal similarities and widen the distribution margins between heterogeneous cross-domain samples.

Recent literature has explored domain adaptation and cross-modal fusion. However, existing methods still leave two bottlenecks unresolved. First, rigid, manually defined domain labels remain a limitation. Current multi-domain networks heavily rely on predefined categories, and this discretization forces dynamic, real-world news into mutually exclusive boxes while only partially reflecting the

continuous and overlapping nature of emerging events. Such closed-set paradigms lack extensibility and struggle to capture nuanced topical variation in misinformation. Second, many models still rely on shallow cross-modal matching. Standard cross-attention can encourage the network to memorize superficial statistical co-occurrences. As a result, when a genuine image is paired with a manipulated or counterfactual text, these models often perform poorly because they lack explicit mechanisms for fine-grained contradiction detection.

To address these gaps, we propose Open-Vocabulary and Logic-Aware Fake News Detection (OVLAFND). To reduce reliance on shallow matching and provide stronger training signals, we first design a multimodal LLM-driven Logical Adversarial Data Augmentation (LADA) pipeline that yields fact-based counterfactual hard negatives. Supervised by these negatives and a negation adversarial loss, we then propose a Logic Consistency Module (LCM) that projects inputs into a logic-consistency subspace to explicitly estimate premise-hypothesis self-consistency. Finally, to relax rigid categorization and generalize this reasoning capability across topics, we introduce an Open-Vocabulary Topic Manifold (OVTM). By explicitly anchoring learnable topic prototypes to a frozen LLM's semantic space, this memory mechanism effectively complements our paradigm, enabling zero-shot adaptation to unseen topics. The core contributions are summarized as follows:

- We design the LADA pipeline to construct counterfactual hard negatives, optimized jointly with a negation adversarial loss to balance fact alignment and dynamic logical contradiction.

- Empowered by these adversarial signals, we propose the LCM to transcend shallow keyword-matching, explicitly capturing fine-grained logical mutual exclusivity in a dedicated logic-consistency subspace.

- We pioneer a shift from rigid domain labels to a continuous topic manifold via the proposed OVTM. Anchoring dynamic clusters directly to frozen LLM embeddings achieves highly interpretable, label-free topic routing without requiring fixed domain IDs.

- Extensive experiments on Weibo and Weibo-21 datasets confirm that OVLAFND achieves state-of-the-art performance. Our reasoning-centric co-design immunizes the model against sophisticated tampering and exhibits better generalization on unseen domains.

2 Related Works

2.1 Multimodal Fake News Detection

Visual and textual information are critical in fake news detection. While early methods relied on simple feature fusion, recent advancements leverage large language models and knowledge enhancement. FKA-Owl (Liu et al., 2024) utilizes a knowledge-enhanced vision-language model pipeline for dual-branch consistency verification to prevent attention shifts. Sniffer (Qi et al., 2024) and TRUST-VL (Yan et al., 2025) recast multimodal misinformation detection as explainable out-of-context reasoning, while MAFT (Kakizaki et al., 2025) textualizes multimodal inputs for automated fact-checking. FACTGUARD (He et al., 2026) leverages LLMs for event-centric content extraction, employing a dynamic usability mechanism to mitigate writing style biases. Furthermore, the MMFakeBench study (Liu et al., 2025) introduces MMD-Agent and a mixed-source benchmark, highlighting complex distortion patterns that existing models struggle with.

2.2 Multi-domain Fake News Detection

Domain shift presents a significant challenge in adapting models to emerging news. MDFEND (Nan et al., 2021) employs a domain gating network to adapt expert weights. M^3DFEND (Zhu et al., 2023) introduces a memory bank to handle incomplete domain labels, while MMDFND (Tong et al., 2024) progressively decouples domain-specific and shared knowledge to overcome semantic biases. DAMMFND (Lu et al., 2025) further mitigates negative transfer by extracting implicit domain representations via disentanglement. These recent studies suggest that domain-aware modeling is useful, but relying on predefined, rigid domain labels fundamentally limits out-of-distribution adaptability to unseen events.

3 Method

In this section, we present the proposed OVLAFND framework. The architecture is designed to address the limitations of discrete domain labels and shallow multimodal correlation. As illustrated in Figure 1, the OVLAFND framework

consists of six interconnected modules: LADA, Multi-Perspective Modality Encoding (MPME) , OVTM, LCM, Logic-Aware Multi-View Discriminator (LAMVD), and Semantic Anchor-Driven Decision Layer (SADDL).

3.1 Problem Formulation

3.1.1 Problem Definition

Given a multimodal fake news dataset $\mathcal{D} = \{(T_j, I_j, y_j)\}_{j=1}^{N}$, where each instance consists of a text-image pair (T, I) and a binary label $y \in \{0, 1\}$ indicating whether the news is real ($y = 0$) or fake ($y = 1$), the goal is to learn a predictor that outputs a veracity score $\hat{y} \in [0, 1]$ for each input pair. To model the overlapping and evolving topical structure of real-world news, we represent each input in a continuous open-vocabulary topic manifold parameterized by K latent semantic centers in a d-dimensional feature space. These centers are aligned with K offline semantic anchors derived from natural-language topic phrases, providing semantic guidance for topic-aware reasoning and prediction.

3.1.2 Preliminaries on Foundation Models

To extract foundational representations, we leverage several frozen, pre-trained models. For unimodal feature extraction, we utilize BERT (Devlin et al., 2019) and Masked Autoencoders (MAE) (He et al., 2022) to obtain the base textual and visual embeddings, formulated as $h_t = \mathrm{BERT}(T)$ and $h_i = \mathrm{MAE}(I)$. Furthermore, to capture perfectly aligned cross-modal semantics, we employ the pre-trained Vision-Language Model CLIP (Radford et al., 2021). CLIP projects both modalities into a shared semantic space, yielding aligned features $z_t = \mathrm{CLIP}_T(T)$ and $z_i = \mathrm{CLIP}_I(I)$. These pre-extracted foundational features serve as the core inputs for our subsequent encoding and reasoning modules. More details can be found in the Appendix (**?**)

3.2 Logical Adversarial Data Augmentation

Conventional multimodal models often suffer from unimodal bias, where they predict news veracity based on frequent keywords or visual patterns without performing logical cross-checking. To encourage the network to learn deep visual-textual reasoning, we design the LADA pipeline utilizing a multimodal LLM to generate adversarial counterfactuals. This data-and-loss co-design improves the model's robustness to sophisticated, maliciously tampered news.

Given a genuine multimodal instance comprising text T and image I, LADA constructs a hard-negative counterfactual text T_{ct} through a rigorous three-step logical synthesis:

1. **Plausible Object Inference**: The multimoda LLM analyzes the semantic context of T to infer a physical object O that is contextually likely to appear but is explicitly absent from the current text.
2. **Visual Verification**: A visual-language model checks the image I to confirm that the object O is indeed not present in the visual evidence.
3. **Contradictory Rewriting**: The LLM rewrites T into T_{ct} by asserting the presence of a specific state of object O, thereby creating an explicit logical conflict between T_{ct} and I.

The overall generation process is defined in Equation 1:

$$T_{ct} = \Phi_{adv}(T \mid O, I), \tag{1}$$

where Φ_{adv} denotes the logic-preserving rewriting function. This results in an adversarial pair (T_{ct}, I) which serves as a crucial negative signal for downstream logic consistency learning. Further details are provided in Appendix A.2.

3.3 Multi-Perspective Modality Encoding

To capture comprehensive multimodal evidence, the MPME module directly builds upon the aforementioned foundational embeddings (h_t, h_i, z_t, z_i). To explicitly model the intrinsic semantic correlation between modalities, we compute a multimodal base feature h_{mm} weighted by the cosine similarity between the aligned CLIP features. To simplify the formulation, we define it as:

$$h_{mm} = \left(\frac{z_t \cdot z_i^\top}{\|z_t\| \|z_i\|} \right) \cdot \mathrm{MLP}([z_t, z_i]), \tag{2}$$

where $[\cdot, \cdot]$ denotes concatenation.

For each individual view $v \in \{t, i, mm\}$, we deploy parallel multi-channel CNNs (Zheng et al., 2014; Yang et al., 2015) to capture diverse local semantic subspaces. The outputs from these CNN experts are then aggregated via an attention-guided gating mechanism:

$$a^v = \mathrm{Softmax}(W_a h_v + b_a), \quad f_v = \sum_{j=1}^{k} a_j^v r_j^v. \tag{3}$$

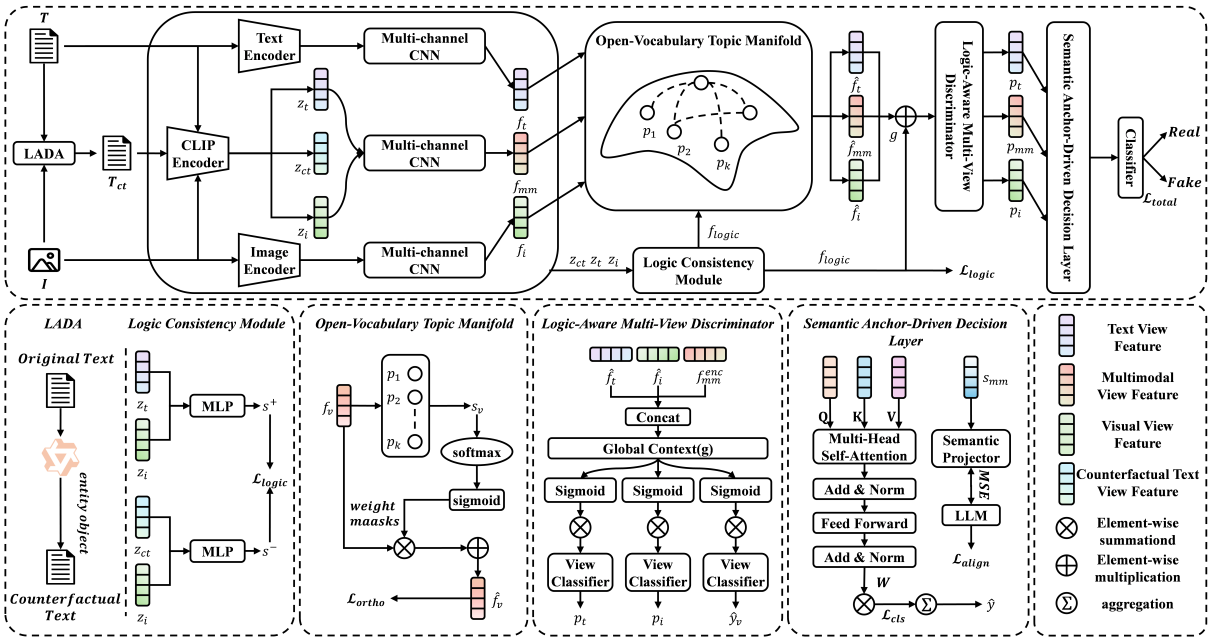

Figure 1: The network architecture of OVLAFND. Text, CLIP, and image encoders are utilized to extract different modality view representations of multimodal news. Multi-channel CNN is employed to aggregate multi-channel features from multimodal views. OVTM extracts semantic information and reduces negative transfer. The LAMVD identifies real and fake features across multimodal views. LADA and LCM evaluate reasoning chains. News authenticity is determined by the Semantic Anchor-Driven Decision Layer.

Here, r_j^v denotes the feature representation extracted by the j-th independent CNN expert for view v. The resulting aggregated features f_t, f_i, and f_{mm} represent the textual, visual, and multimodal perspectives, respectively.

3.4 Logic Consistency Module

The LCM explicitly evaluates the reasoning chain between text and image by leveraging the counterfactual data generated from LADA. It projects aligned features into a joint reasoning space to distinguish between logical coherence and contradiction.

Using the original pair (z_t, z_i) and the counterfactual pair (z_{ct}, z_i) (where $z_{ct} = \text{CLIP}_T(T_{ct})$), the LCM computes positive and negative logic scores in Equation 4:

$$\begin{aligned} s^+ &= W_s \cdot \text{MLP}_{\log}([z_i, z_t]), \\ s^- &= W_s \cdot \text{MLP}_{\log}([z_i, z_{ct}]). \end{aligned} \tag{4}$$

Simultaneously, a logic reasoning feature f_{logic} is extracted from the genuine pair and injected into the multimodal pathway to form an enhanced representation via Equation 5:

$$f_{mm}^{enc} = \text{ReLU}(W_{enc}[\hat{f}_{mm}, f_{logic}]). \tag{5}$$

By penalizing low values of s^+ and high values of s^-, the model is forced to scrutinize the factual consistency between what is said and what is shown.

3.5 Open-Vocabulary Topic Manifold

Traditional methods rely on static domain labels, which fail to capture the overlapping and evolving nature of news themes. We propose OVTM to map news features into a continuous, open-vocabulary topic manifold.

We define a learnable prototype memory matrix $P \in \mathbf{R}^{K \times d}$ containing K latent semantic centers. The effect of K is discussed in Appendix A.7. For any view feature f_v, the module calculates a soft assignment distribution $s_v \in \mathbf{R}^K$ via Equation 6:

$$s_v = \text{Softmax}\left(\frac{f_v P^\top}{\tau}\right), \tag{6}$$

where τ is the temperature scaling factor, which is set to a default value of 0.1. To enhance the features with topic-specific context, we use a residual modulation mechanism as defined in Equation 7:

$$\hat{f}_v = f_v + \sigma(W_p s_v) \odot f_v, \tag{7}$$

where σ is the Sigmoid function and $\odot$ is the Hadamard product. This mechanism allows the

model to dynamically adapt its reasoning based on the latent topic distribution rather than fixed categories.

3.6 Logic-Aware Multi-View Discriminator

The LAMVD functions as an informational sieve that modulates view-specific predictions based on the global logical context. We first construct a global state g by concatenating the refined features: $g = [\hat{f}_t, \hat{f}_i, f_{mm}^{enc}]$.

For each view v, a gating network generates a topic-conditioned attention mask g_v as shown in Equation 8:

$$g_v = \sigma(W_g^v g + b_g^v). \tag{8}$$

The mask is applied to the topic-enhanced features to retain only contextually relevant information:

$$p_v = g_v \odot \hat{f}_v. \tag{9}$$

These filtered features are then passed through view-specific classifiers to generate isolated veracity predictions $\hat{y}_v$ for each view $v \in \{t, i, mm\}$.

3.7 Semantic Anchor-Driven Decision Layer

SADDL serves as the final decision hub, utilizing a Transformer to dynamically weight the multi-view predictions. To establish a decision query Q, we combine the algebraically averaged global context with the filtered features as formulated in Equation 10 and Equation 11:

$$f_{agg} = \frac{1}{3} \sum_{v \in \{t,i,mm\}} \hat{f}_v, \tag{10}$$

$$Q = f_{agg} + \sum_{v \in \{t,i,mm\}} p_v. \tag{11}$$

A Multi-Head Self-Attention (MHSA) block processes Q against the individual view features to output dynamic decision weights $w \in \mathbf{R}^3$, where the keys and values are formed from the view features, i.e., $K = V = [\hat{f}_t, \hat{f}_i, f_{mm}^{enc}]$:

$$w = \mathrm{Softmax}(W_w \cdot \mathrm{MHSA}(Q, K, V)). \tag{12}$$

The final veracity prediction $\hat{y}$ is the weighted summation of individual views, *i.e.*, $\hat{y} = \sum_v w_v \hat{y}_v$.

To ensure the latent prototypes in P remain semantically meaningful, we project the multimodal topic distribution into a 512-dimensional semantic space:

$$f_{proj} = \mathrm{MLP}_{proj}(s_{mm} P). \tag{13}$$

This feature is aligned with high-dimensional semantic anchors derived from a frozen LLM, providing explicit interpretability to our latent manifold.

3.8 Training and Inference

Training. The framework is optimized via a joint loss function. The primary classification objective is the cross-entropy ℓ calculated over the final and auxiliary predictions in Equation 14:

$$\mathcal{L}_{cls} = \ell(\hat{y}, y) + \frac{1}{3} \sum_v \ell(\hat{y}_v, y). \tag{14}$$

To train the LCM, we utilize a margin-based contrastive logic loss to separate original and counterfactual scores by a dynamic margin γ, as shown in Equation 15. Here, s^+ and s^- denote unnormalized logic scores, BCE denotes binary cross-entropy, and y_{smooth} denotes the label-smoothed target for genuine pairs:

$$\mathcal{L}_{logic} = \mathrm{BCE}(\sigma(s^+), y_{smooth}) + \max(0, \gamma - s^+ + s^-). \tag{15}$$

To preserve prototype diversity and prevent topic collapse, we enforce an orthogonality constraint on the prototype matrix P in Equation 16:

$$\mathcal{L}_{ortho} = \|PP^\top - I_K\|_2^2. \tag{16}$$

Finally, we align the latent space with real-world semantics by minimizing the distance between the projected features and the pre-extracted offline LLM semantic anchors h_{LLM} in Equation 17, the details of which are provided in Appendix A.4.

$$\mathcal{L}_{align} = \|f_{proj} - h_{LLM}\|_2^2. \tag{17}$$

The total optimization loss function is given by Equation 18:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{logic} + \lambda_2 \mathcal{L}_{ortho} + \lambda_3 \mathcal{L}_{align}, \tag{18}$$

where $\lambda_1, \lambda_2, \lambda_3$ are balancing hyperparameters, which are evaluated in Appendix A.7.

Inference. During inference, we bypass the LADA pipeline and the semantic anchoring process, using them only for training supervision. For unseen samples, the optimized OVTM directly computes soft routing weights in the continuous topic manifold without real-time LLM intervention. Concurrently, the LCM evaluates the original input to extract positive logical features, inherently

discarding the counterfactual comparison branch. Finally, the decision layer integrates these topic-modulated features and logical representations to infer news authenticity.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluated our model using two real-world benchmark datasets: Weibo (Wang et al., 2018) and Weibo-21 (Nan et al., 2021). The detailed statistics of both datasets are summarized in Appendix A.1.

Baselines. We categorize our baselines into three groups: (1) Single-modal multi-domain methods, including: MMoE (Ma et al., 2018), MoSE (Qin et al., 2020), MDFEND (Nan et al., 2021), and M^{3}DFEND (Zhu et al., 2023). (2) Multimodal single-domain methods, including: SpotFake (Singhal et al., 2019), CAFE (Chen et al., 2022), CMC (Wei et al., 2022), BMR (Ying et al., 2023), and MRHFR (Wu et al., 2023). (3) Multimodal multi-domain methods, including: KATMF (Song et al., 2021), MMDFND (Tong et al., 2024), and DAMMFND (Lu et al., 2025).

4.2 Overall Performance

To evaluate the effectiveness of our proposed OVLAFND framework, we conducted extensive experiments using the Weibo and Weibo-21 datasets. We evaluated the model under both conventional fixed-domain settings and an open-set zero-shot setting (Leave-One-Domain-Out test). Overall, OVLAFND outperforms the baseline models across the reported scenarios.

4.2.1 Fixed-Domain Performance

As demonstrated in Table 1 and Table 2, OVLAFND significantly outperforms both multi-domain and multimodal baselines, achieving peak accuracies of 95.2% on Weibo and 96.9% on Weibo-21. Among multi-domain methods, KATMF, MMDFND, and DAMMFND utilize explicit domain labels for adaptive gating. However, their rigid categorization enforces hard boundaries, causing them to struggle with cross-domain semantic overlaps and emerging unlabeled events. Furthermore, while multimodal methods like SpotFake and MRHFR excel at feature extraction, they primarily rely on shallow semantic alignment. This makes them inherently vulnerable to sophisticated visual-textual logical traps and less adaptable to domain variations. In contrast, Our OVLAFND framework continuous topic manifold circumvents these bottlenecks by mapping instances into a shared semantic space, exhibiting exceptional robustness across all predefined domains.

4.2.2 Leave-One-Domain-Out Performance

To further validate the model's robustness in real-world scenarios where a domain is held out at test time, we conducted a leave-one-domain-out (LODO) evaluation, as presented in Table 3. In this setting, the model is trained on 7 domains, validated on 1 domain, and tested on the remaining unknown domain.

As shown in Table 3, On Weibo, OVLAFND outperforms the strongest baseline, DAMMFND, by 2.6% F1 on average (87.0% vs. 84.4%), with especially clear gains in Social and Political news. A more revealing pattern appears on Weibo-21: the original baseline already surpasses the existing domain-aware SOTA models, suggesting that rigid domain modeling can become unstable under a strict held-out-domain protocol. OVLAFND further improves on top of this stronger reference point, reaching 74.4% average F1. These results highlight OVLAFND's stronger ability to handle unseen and dynamically shifting topics under cross-domain evaluation.

4.3 Ablation Study

To demonstrate the effectiveness of each component in the OVLAFND framework, we conducted an ablation study by systematically removing specific modules and performing comparative analysis. Table 4 presents our ablation study results. The original OVLAFND outperforms all variants, demonstrating each component's critical effectiveness. Key findings include:

- **Removing LADA** significantly decreases performance, leading to F1-score decreases of 0.7% and 1.0% on the two datasets, respectively. This highlights the critical role of LADA in shifting the model's paradigm from mere cross-modal "relevance matching" to rigorous "logical verification."
- **Removing the LCM** substantially reduces performance, causing an accuracy drop of 3.0% and 1.9%. This confirms the necessity of the LCM in explicitly evaluating the reasoning chain between different modalities.
- **Removing the OVTM** and reverting to mutually exclusive domain labels decreases accuracy by

	Method	Sci.	Mil.	Edu.	Soc.	Pol.	Hlth.	Fin.	Ent.	Dis./Int.	Overall		
											F1	Acc	AUC
Weibo	MMoE	0.578	0.911	0.851	0.885	0.735	0.826	0.813	0.830	0.883	0.874	0.874	0.950
	MoSE	0.793	0.738	0.834	0.912	0.764	0.859	0.791	0.844	0.883	0.890	0.890	0.954
	MDFEND	0.774	0.911	0.897	0.902	0.763	0.878	0.808	0.881	0.874	0.904	0.904	0.965
	M^3DFEND	0.792	0.903	0.923	0.912	0.765	0.863	0.899	0.899	0.876	0.928	0.928	0.969
	KATMF	0.831	0.908	0.924	0.895	0.823	0.898	0.903	0.904	0.894	0.929	0.930	0.969
	MMDFND	0.824	0.911	0.941	0.939	0.735	0.913	0.917	0.917	0.888	0.934	0.934	0.972
	DAMMFND	0.853	0.911	0.956	0.943	0.822	0.939	**0.937**	**0.956**	**0.928**	0.943	0.944	0.983
	OVLAFND	**0.925**	**1.000**	**0.972**	**0.948**	**0.852**	**0.957**	0.936	0.930	0.889	**0.952**	**0.952**	**0.986**
Weibo-21	MMoE	0.875	0.911	0.870	0.875	0.862	0.936	0.856	0.888	0.877	0.894	0.894	0.954
	MoSE	0.850	0.885	0.881	0.872	0.880	0.917	0.867	0.891	0.867	0.893	0.894	0.954
	MDFEND	0.830	0.938	0.891	0.898	0.886	0.940	0.895	0.906	0.900	0.913	0.913	0.970
	M^3DFEND	0.829	0.950	0.899	0.908	0.882	0.946	0.900	0.931	0.889	0.921	0.921	0.975
	KATMF	0.914	0.928	0.913	0.895	0.902	0.914	0.871	0.937	0.898	0.923	0.928	0.975
	MMDFND	0.937	0.953	0.852	0.945	0.965	0.920	0.884	0.959	0.919	0.939	0.939	0.977
	DAMMFND	0.932	0.953	0.932	0.917	0.982	0.919	0.941	**0.990**	0.945	0.947	0.947	0.985
	OVLAFND	**1.000**	**0.953**	**0.965**	**0.950**	**1.000**	**0.952**	**0.954**	0.980	**1.000**	**0.969**	**0.969**	**0.992**

Table 1: Overall performance comparison on Weibo and Weibo-21 datasets.

	Method	Accuracy	F1 score	
			Real	Fake
Weibo	SpotFake	0.892	0.932	0.739
	CAFE	0.840	0.842	0.837
	CMC	0.893	0.899	0.907
	BMR	0.918	0.914	0.904
	MRHFR	0.922	0.918	0.910
	DAMMFND	0.944	0.945	0.943
	OVLAFND	**0.952**	**0.950**	**0.954**
Weibo-21	SpotFake	0.851	0.828	0.866
	CAFE	0.882	0.885	0.876
	CMC	0.897	0.903	0.912
	BMR	0.929	0.927	0.925
	MRHFR	0.932	0.928	0.926
	DAMMFND	0.947	0.948	0.947
	OVLAFND	**0.969**	**0.969**	**0.969**

Table 2: Performance comparison with multimodal methods.

1.3% and 1.0% on the two datasets. These results validate that modeling news as a continuous topic manifold captures semantic nuances more effectively than rigid categorization, thereby mitigating negative transfer and significantly enhancing fake news detection adaptability.

4.3.1 Visualizations

Figure 2 presents t-SNE visualizations of features learned by OVLAFND, DAMMFND, and MMDFND on the Weibo and Weibo-21 datasets. Compared to the severe feature entanglement and scattered boundary outliers seen in the baselines, OVLAFND condenses the features into denser intra-class clusters and produces a clearer separation in the visualization, which suggests a more distinguishable feature representation.

(a) t-SNE visualization on the Weibo dataset.

(b) t-SNE visualization on the Weibo-21 dataset.

Figure 2: The t-SNE visualizations of learned features for real vs. fake news.

4.3.2 Case Study

We selected three cases with healthy, constructive education and daily technology from the Weibo-21 datasets to demonstrate the interpretability of our OVLAFND framework. The quantitative results and semantic routing details are illustrated in Figure 3. OVLAFND successfully identifies the subtle logical contradictions and explicitly routes

	Method	Sci.	Mil.	Edu.	Soc.	Pol.	Hlth.	Fin.	Ent.	Dis./Int.	F1(Avg.)
Weibo	baseline	0.729	0.807	0.767	0.467	0.739	0.672	0.836	0.776	0.860	0.739
	MMDFND	0.711	0.394	0.749	0.205	0.856	0.342	0.418	0.418	0.645	0.526
	DAMMFND	0.868	0.868	0.810	0.766	0.880	**0.881**	0.864	0.847	0.815	0.844
	OVLAFND	**0.870**	**0.876**	**0.837**	**0.824**	**0.921**	0.879	**0.865**	**0.875**	**0.879**	**0.870**
Weibo-21	baseline	0.760	0.793	**0.814**	0.660	0.718	0.807	0.463	0.716	0.784	0.724
	MMDFND	0.170	0.158	0.120	0.419	0.289	0.213	0.478	0.376	0.526	0.305
	DAMMFND	**0.783**	**0.879**	0.783	0.642	0.593	0.779	0.483	0.523	0.796	0.696
	OVLAFND	0.766	0.856	0.811	**0.686**	**0.722**	**0.817**	**0.490**	**0.726**	**0.819**	**0.744**

Table 3: Open-set domain generalization performance under the LODO protocol.

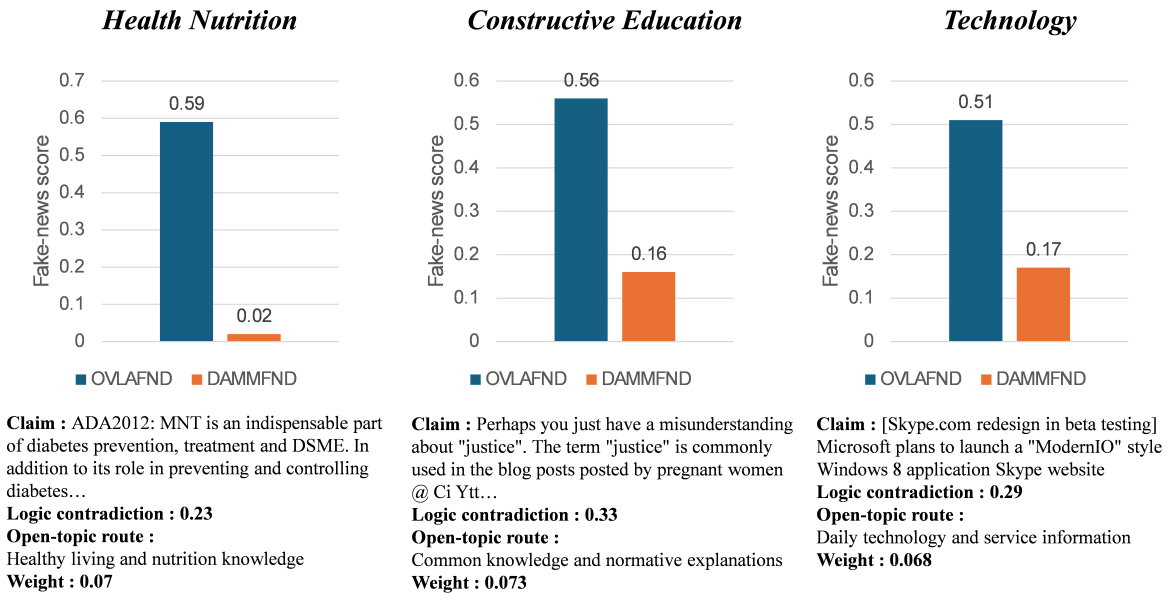

Figure 3: Interpretability case study on three benign fake news samples.

	Method	Accuracy	F1 score	
			Real	Fake
Weibo	**OVLAFND**	**0.952**	**0.950**	**0.954**
	-w/o LADA	0.945	0.943	0.948
	-w/o LCM	0.922	0.922	0.923
	-w/o OVTM	0.939	0.937	0.941
Weibo-21	**OVLAFND**	**0.969**	**0.969**	**0.969**
	-w/o LADA	0.959	0.959	0.960
	-w/o LCM	0.950	0.948	0.951
	-w/o OVTM	0.959	0.959	0.959

Table 4: Ablation study of the proposed framework.

the samples to human-readable topics, whereas DAMMFND is misled by the positive narratives. More details are provided in Appendix A.5.

Conclusion

In this paper, we propose OVLAFND, an open-vocabulary, logic-aware framework for multimodal multi-domain fake news detection. The method is built around three components. First, LADA constructs counterfactual hard negatives that expose fine-grained text-image contradictions during training. Second, LCM uses these adversarial signals to learn a dedicated logic-consistency subspace, reducing reliance on shallow cross-modal matching. Third, OVTM transforms rigid domain labels into a continuous topic manifold, improving robustness to overlapping and unseen events. Experiments on two benchmark datasets show that OVLAFND achieves competitive performance and exhibits stronger adaptability to unseen domains under the open-set LODO setting.

Limitations

Although OVLAFND shows strong performance and generalization, there are still practical constraints. First, the quality of LADA depends on external LLM/VLM generation, and ambiguous visual scenes may occasionally produce noisy counterfactual samples. Second, our evaluation mainly focuses on Weibo-style benchmarks, so cross-lingual and cross-platform transfer should be further validated. Third, the full training pipeline introduces additional preprocessing cost due to adversarial sample construction and semantic alignment. These factors do not affect the deploy-time model structure, but they motivate future work on more efficient data generation and broader multilingual evaluation.

References

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

Eunji Choi, Jaemin Ahn, Xiaoyu Piao, and Jong-Kook Kim. 2025. CroMe: Multimodal fake news detection using cross-modal tri-transformer and metric learning. *IEEE Access*, 13:197124–197132.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*, pages 4171–4186.

Jing He, Han Zhang, Yuanhui Xiao, Wei Guo, Shaowen Yao, and Renyang Liu. 2026. Factguard: Event-centric and commonsense-guided fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 363–371.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15979–15988.

Jing Jing, Hongchen Wu, Jie Sun, Xiaochang Fang, and Huaxiang Zhang. 2023. Multimodal fake news detection via progressive fusion networks. *Information Processing and Management*, 60(1):103120.

Kazuya Kakizaki, Yuto Matsunaga, and Ryo Furukawa. 2025. MAFT: Multimodal automated fact-checking via textualization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29646–29648.

Haoxuan Li, Zhengmao Yang, Yunshan Ma, Yi Bin, Yang Yang, and Tat-Seng Chua. 2024a. Mm-forecast: A multimodal approach to temporal event forecasting with large language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2776–2785.

Jing Li, Yi Bin, Yuxi Ma, Yang Yang, Zhiqiang Huang, and Tat-Seng Chua. 2024b. Filter-based stance network for rumor verification. *ACM Transactions on Information Systems*, 42(4):108:1–108:28.

Jun Li, Yi Bin, Liang Peng, Yang Yang, Yangyang Li, Hao Jin, and Zi Huang. 2024c. Focusing on relevant responses for multi-modal rumor detection. *IEEE Transactions on Knowledge and Data Engineering*, 36(11):6225–6236.

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

Xuannan Liu, Zekun Li, Pei-Pei Li, Huaibo Huang, Shuhan Xia, Xing Cui, Linzhi Huang, Weihong Deng, and Zhaofeng He. 2025. Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms. In *Proceedings of the Thirteenth International Conference on Learning Representations*.

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

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. 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 and Data Mining*, pages 1930–1939.

Qiong Nan, Juan Cao, Yongchun Zhu, Yanyan Wang, and Jintao Li. 2021. Mdfend: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management*, pages 3343–3347.

Peng Qi, Zehong Yan, Wynne Hsu, and Mong-Li Lee. 2024. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13052–13062.

Zhen Qin, Yicheng Cheng, Zhe Zhao, Zhe Chen, Donald Metzler, and Jingzheng Qin. 2020. Multitask mixture of sequential experts for user activity streams. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 3083–3091.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763.

Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pages 39–47.

Chenguang Song, Nianwen Ning, Yunlei Zhang, and Bin Wu. 2021. Knowledge augmented transformer for adversarial multidomain multiclassification multimodal fake news detection. *Neurocomputing*, 462:88–100.

Yu Tong, Weihai Lu, Zhe Zhao, Song Lai, and Tong Shi. 2024. Mmdfnd: Multi-modal multi-domain fake news detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1178–1186.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 849–857.

Zimian Wei, Hengyue Pan, Linbo Qiao, Xin Niu, Peijie Dong, and Dongsheng Li. 2022. Cross-modal knowledge distillation in multi-modal fake news detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4733–4737.

Lianwei Wu, Pusheng Liu, and Yanning Zhang. 2023. See how you read? multi-reading habits fusion reasoning for multi-modal fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13736–13744.

Zehong Yan, Peng Qi, Wynne Hsu, and Mong-Li Lee. 2025. TRUST-VL: An explainable news assistant for general multimodal misinformation detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5588–5604, Suzhou, China. Association for Computational Linguistics.

Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiaoli Li, and Shonali Krishnaswamy. 2015. Deep convolutional neural networks on multichannel time series for human activity recognition. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 3995–4001.

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

Yi Zheng, Qi Liu, Enhong Chen, Yong Ge, and J. Leon Zhao. 2014. Time series classification using multichannels deep convolutional neural networks. In *International Conference on Web-Age Information Management*, pages 298–310.

Yongchun Zhu, Qiang Sheng, Juan Cao, Shuokai Li, Danding Wang, and Fuzhen Zhuang. 2022. Generalizing to the future: Mitigating entity bias in fake news detection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2120–2125.

Yongchun Zhu, Qiang Sheng, Juan Cao, Qiong Nan, Kai Shu, Minghui Wu, Jindong Wang, and Fuzhen Zhuang. 2023. Memory-guided multi-view multi-domain fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):7178–7191.

A Appendix

A.1 Datasets and Feature Processing

Datasets. The Weibo dataset contains 7,532 news articles (3,749 true, 3,783 fake) for training and 1,996 articles (996 true, 1,000 fake) for testing. While traditionally categorized into nine rigid domains, we process these texts to extract natural language topic descriptions for our open-vocabulary manifold. Weibo-21, a complex multi-domain dataset, comprises 9,127 articles (4,640 true, 4,487 fake), which we partitioned into training and test sets following established benchmark procedures. The detailed statistical breakdown of both datasets is summarized in Table 5.

Dataset	Real	Fake	Total
Weibo	4,745	4,783	9,528
Weibo-21	4,640	4,487	9,127

Table 5: Statistics of the Weibo and Weibo-21 datasets.

Semantic Feature Processing. While these datasets are traditionally categorized into rigid, predefined domains, our open-vocabulary paradigm requires dynamic natural language topic descriptions. To establish the Open-Vocabulary Topic Manifold, we design an offline semantic feature extraction pipeline to generate explicit semantic anchors. For each multimodal news instance, we utilize the Qwen3-VL-8B model to jointly analyze the original text and image, prompting it to output 3 to 5 concise topic phrases that precisely summarize

the core scene (e.g., "Traffic Accident," "Political Election"). Subsequently, we employ a pre-trained frozen CN-CLIP (ViT-B-16) text encoder to project these generated phrases into a 512-dimensional semantic space. The resulting L2-normalized vectors are cached locally as `.pt` files. This offline processing not only prevents heavy computational overhead from real-time LLM API calls during training but also ensures that our latent topic prototypes are securely anchored to stable, well-defined natural language semantics.

A.2 Counterfactual Text Generation for LADA

To construct counterfactual hard negatives, our LADA pipeline employs an open-source multimodal LLM, i.e., Qwen3-VL-8B, along with four logical attack strategies. For each multimodal instance, one strategy is randomly selected to generate the counterfactual text T_{ct}. These strategies range from basic entity denial to multi-step contradiction generation and are summarized in Table 6.

If the multimodal LLM fails to generate a sufficiently strong negation, as detected by a keyword-based filtering rule, we apply a fallback strategy. This fallback directly negates the original text predicates to preserve a clear logical contradiction and provide usable adversarial supervision.

A.3 Generation of LLM Semantic Anchors

To construct the OVTM, we perform an offline semantic feature extraction pipeline to generate the stable, high-dimensional semantic anchors. This offline processing avoids the massive computational overhead of querying large models during real-time training. The pipeline consists of two stages:

Stage 1: Topic Phrase Extraction via multimodal LLM. For each multimodal news instance, we leverage the Qwen3-VL-8B model to jointly analyze the image and the text. We instruct the model to distill the input into concise, high-level semantic concepts. The exact prompt template used is as follows:

> **System Prompt:** Carefully analyze the provided image (if any) and the following news text. Please extract 3 to 5 concise words or phrases that accurately summarize the core theme and scene of this news. Output only the words, separated by commas, without any extra explanation.
> **Input News Text:** {text}

Stage 2: Semantic Encoding via Frozen CLIP. Once the concise topic phrases (e.g., "Traffic Accident, Urban Flood") are generated, we tokenize and feed them into a pre-trained, frozen CN-CLIP (ViT-B/16) text encoder. The text encoder projects these natural language concepts into a 512-dimensional embedding space. We then apply L2-normalization to these embeddings to obtain the final semantic anchors, which are cached locally as `.pt` files and utilized to explicitly align the latent prototypes during training.

A.4 Implementation Details

For text encoding, we set the maximum input length to 197 and utilized pre-trained BERT (Devlin et al., 2019) and CLIP (Radford et al., 2021) models. For visual encoding, images were resized to 224×224 pixels and processed using pre-trained MAE (ViT-B/16) (He et al., 2022) and CLIP (ViT-B/16) models. To establish the continuous topic manifold, we employed a frozen LLM to extract 512-dimensional semantic anchors. Counterfactual augmentation in the LADA pipeline was powered by the Qwen3-VL:8B model via Ollama. All experiments were implemented in PyTorch and executed on an NVIDIA RTX 3090 GPU.

A.5 Detailed Case Analysis

We provide detailed case studies illustrated in Figure 3 to further illustrate the interpretability of our model as follows.

Case 1: Health Nutrition. In the first case (Figure 3, left), the text presents a benign medical claim. OVLAFND outputs a fake news score of 0.587 (surpassing the 0.5 threshold), correctly identifying it as a forgery. In contrast, DAMMFND outputs a remarkably low score of 0.020, failing to detect the anomaly. OVLAFND decomposes the evidence into text, image, and fusion views (scoring 0.066, 0.810, and 0.659, respectively) and yields a low logic contradiction score of 0.227, capturing the fine-grained visual-textual mismatch. Furthermore, our OVTM routes the sample to Topic T15 (weight: 0.070), which semantically anchors to *"healthy living and nutrition knowledge."*

Case 2: Constructive Education. In the second case (Figure 3, middle), involving a positive educational narrative, OVLAFND accurately predicts it as fake with a score of 0.556, while DAMMFND scores it at a mere 0.163. The explicit evidence scores from the text, image, and fusion views are 0.161, 0.731, and 0.572, respectively, accompanied

Strategy	Objective	Prompt Instruction Details (Translated)
Existence Negation	To create a visual logical trap by denying the existence of prominent objects.	**Task:** Construct an "Existence Negation" fake news. 1. Find the most prominent core object/person in the image. 2. Modify the original sentence to explicitly assert that this object is completely absent from the scene. 3. You must use explicit negation words like "none", "not found", or "empty".
Quantity/Degree Reversal	To invert the scale, density, or severity depicted in the multimodal input.	**Task:** Construct a "Quantity/Degree Reversal" fake news. 1. Focus on the quantity, density, or degree in the original sentence or image. 2. Modify it to a completely opposite logical state (e.g., from crowded to empty, from massive fire to slight smoke). 3. Keep the tone formal like a real news report, but ensure the facts are completely inverted.
State Mutually Exclusive	To introduce fine-grained contradictions in actions or physical states.	**Task:** Construct a "State Mutually Exclusive" logical trap. Modify the adjectives or verbs in the original sentence to make them logically mutually exclusive with the image content (e.g., change "fighting" to "hugging amicably", or "broken" to "intact").
Chain-of-Thought Negation	To perform step-by-step rigorous logical contradiction generation.	**Task:** You are a fabrication expert. Execute step-by-step: *Step 1 (Visual Analysis):* Carefully observe the image and identify a core object that explicitly exists. *Step 2 (Logical Reversal):* Based on this object, construct a sentence asserting it does not exist or is entirely something else. Ensure the sentence sounds natural but contradicts the existing visual evidence.

Table 6: Detailed prompt instructions were utilized in the multimodal LLM-driven LADA pipeline to generate fine-grained counterfactual text.

by a logic contradiction score of 0.335. The sample is dynamically routed to Topic T12 (weight: 0.073), corresponding to the *"Common knowledge and normative explanations"* semantic space.

Case 3: Technology. For the third case (Figure 3, right) concerning daily technology, OVLAFND successfully flags it (score: 0.511) compared to DAMMFND's failure (score: 0.174). The isolated view scores (0.048, 0.522, 0.561) and the logic contradiction score (0.291) expose the underlying discrepancy. The sample is logically routed to Topic T01 (weight: 0.068), anchored to *"Daily technology and service information."*

Overall, these cases suggest that DAMMFND is more easily misled by mild positive multimodal narratives, whereas OVLAFND captures the relevant contradictions more reliably. By combining LADA, LCM, and OVTM, OVLAFND provides traceable multi-view evidence and interpretable topic routing without relying on fixed domain labels.

A.6 Hyperparameter Settings

Due to the different data distributions of Weibo and Weibo-21, we conduct separate hyperparameter searches for the two datasets. The models are trained with AdamW on Weibo and Adam on Weibo-21, using a base learning rate of 1×10^{-4} and a batch size of 64. The number of latent topic centers is set to $K = 16$. .

For the loss balancing factors in Equation 18, the optimal weights are configured as follows to balance the primary classification and the auxiliary logical/topic constraints:

- **Logic Loss (λ_1):** Set to 0.2 for Weibo-21. For Weibo, a warm-up schedule is applied ($\lambda_1 = 0.2 \times \min(1.0, epoch/10)$) to prevent early divergence.

Parameter	Variation	Weibo Dataset		Weibo-21 Dataset	
		AUC	Macro-F1	AUC	Macro-F1
Latent Topic Centers (K)	$K = 8$	0.985	0.946	0.989	0.967
	$K = 16$ (Optimal Baseline)	**0.986**	**0.952**	0.989	**0.969**
	$K = 32$	0.980	0.935	0.989	0.963
Logic Consistency (λ_1)*	$\lambda_1 = 0.1$	0.980	0.930	0.988	0.964
	$\lambda_1 = 0.2$ (Optimal Baseline)	**0.986**	**0.952**	0.989	**0.969**
	$\lambda_1 = 0.4$	0.986	0.938	**0.991**	0.963
Orthogonal Loss (λ_2)	$\lambda_2 = 0.02$ (Weibo) / 0.05 (Weibo-21)	0.982	0.940	0.989	0.959
	Optimal Baseline ($\lambda_2 = 0.05$ / 0.1)	**0.986**	**0.952**	0.989	**0.969**
	$\lambda_2 = 0.20$	0.986	0.943	**0.990**	0.954
Semantic Alignment (λ_3)	$\lambda_3 = 0.05$ (Weibo) / 0.1 (Weibo-21)	0.985	0.941	**0.992**	0.964
	Optimal Baseline ($\lambda_3 = 0.1$ / 10.0)	**0.986**	**0.952**	0.989	**0.969**
	$\lambda_3 = 10.0$ (Weibo) / 20.0 (Weibo-21)	0.984	0.937	0.991	0.967

* Note: For the Weibo dataset, λ_1 is dynamically scaled by the warm-up schedule $\min(1.0, epoch/10)$.

Table 7: Parameter sensitivity performance across the Weibo and Weibo-21 datasets under different hyperparameter configurations (Metric denotes Macro-F1 score).

- **Orthogonal Loss (λ_2):** Set to 0.05 for Weibo and 0.1 for Weibo-21 to preserve prototype diversity.
- **Alignment Loss (λ_3):** Set to 0.1 for Weibo and 10.0 for Weibo-21, effectively adjusting for the probabilistic distribution scale variance across the two datasets.

A.7 Parameter Sensitivity Analysis

To investigate the robustness of OVLAFND, we analyze the sensitivity of four critical hyperparameters: the number of latent topic centers (K), the logic consistency loss weight (λ_1), the orthogonal loss weight (λ_2), and the semantic alignment loss weight (λ_3). The experimental results, comprehensively evaluated on both the Weibo and Weibo-21 datasets, are summarized in Table 7.

Impact of Topic Center Number (K). As shown in Table 7, reducing K to 8 slightly lowers performance, suggesting that 8 centers may be too coarse to capture the diversity of news topics. Meanwhile, increasing K to 32 also degrades performance on both datasets (Macro-F1 drops to 0.935 and 0.963, respectively), suggesting that an overly large K fragments the latent topic manifold into redundant sub-clusters. Finally, the model achieves the best overall balance at $K = 16$ (Macro-F1 of 0.952 on Weibo and 0.969 on Weibo-21).

Impact of Logic Consistency (λ_1). The logic loss weight λ_1 controls how strongly the model responds to counterfactual anomalies generated by LADA. Reducing it to 0.1 leads to a clear F1 drop (e.g., 0.930 on Weibo), suggesting that weaker logical supervision is insufficient. Increasing it to 0.4 also lowers F1 (0.938 on Weibo and 0.963 on Weibo-21), indicating that an overly strong logic penalty can interfere with the main classification objective. With $\lambda_1 = 0.2$, the model achieves the best balance in our search.

Impact of Orthogonal Disentanglement (λ_2). The orthogonal constraint ensures prototype diversity and prevents topic collapse. When λ_2 is set lower than optimal (e.g., 0.02 on Weibo), the F1 drops to 0.940 due to potential feature entanglement among the latent prototypes. On the other hand, scaling it too high ($\lambda_2 = 0.20$) forces unnatural geometric separation between closely related topics, harming the contextual semantic fusion and lowering the F1 score.

Impact of Semantic Alignment (λ_3). The parameter λ_3 controls the strength of semantic alignment between the latent clusters and the LLM-derived anchors. Weibo-21 benefits from a stronger alignment penalty ($\lambda_3 = 10.0$), whereas Weibo performs better with a milder setting ($\lambda_3 = 0.1$). Applying $\lambda_3 = 10.0$ to Weibo reduces F1 to 0.937, while lowering Weibo-21 to $\lambda_3 = 0.1$ reduces its F1 to 0.964. This suggests that semantic anchoring is useful, but its weight should be calibrated to the representation scale of each dataset.