A Serial Two-Stage Framework for Robust Multimodal Fake News Detection via Adaptive Reasoning

Maolin Wang
Hong Kong Institute of AI for Science, City University of Hong Kong
Hong Kong SAR, China
morin.wang@cityu.edu.hk

Ziting Mai
Hong Kong Institute of AI for Science, City University of Hong Kong
Hong Kong SAR, China
zitingmai2-c@my.cityu.edu.hk

Zichun Liu
Hong Kong Institute of AI for Science, City University of Hong Kong
Hong Kong SAR, China
zichunliu6-c@my.cityu.edu.hk

Beining Bao
Hong Kong Institute of AI for Science, City University of Hong Kong
Hong Kong SAR, China
beininbao2-c@my.cityu.edu.hk

Hongyu Chen
Hong Kong Institute of AI for Science, City University of Hong Kong
Hong Kong SAR, China
hychan2355-c@my.cityu.edu.hk

Junjie Liu
Hong Kong Institute of AI for Science, City University of Hong Kong
Hong Kong SAR, China
junjiliu6-c@my.cityu.edu.hk

Yunbo Zhang
Hong Kong Institute of AI for Science, City University of Hong Kong
Hong Kong SAR, China
yunbzhang3-c@my.cityu.edu.hk

Bingkun Zhao
Hong Kong Institute of AI for Science, City University of Hong Kong
Hong Kong SAR, China
bingkzhao2-c@my.cityu.edu.hk

Tianshuo Wei
Hong Kong Institute of AI for Science, City University of Hong Kong
Hong Kong SAR, China
tianshwei2-c@my.cityu.edu.hk

Jian Liu
Beihang University
Beijing, China
liujian2024@buaa.edu.cn

Chenbin Zhang
Independent Researcher
Beijing, China
aleczhang13@gmail.com

Haoran Yang*
Central South University
Changsha, China
yhr.cse@csu.edu.cn

Abstract

The proliferation of social media has created fertile ground for misinformation, a challenge further intensified by recent advances in generative artificial intelligence. Modern fake news increasingly takes the form of sophisticated multimodal campaigns, where synthetic images and stylistically manipulated text are jointly employed to evade existing detection systems. Despite substantial progress, real-world deployment of multimodal fake news detection models remains constrained by an "impossible triangle" of accuracy, inference efficiency, and robustness. To address these challenges, we propose DAR-Lite, a serial two-stage framework that rethinks the detection pipeline through explicit decoupling of representation denoising and contextual reasoning. In the first stage, a variational information bottleneck distills compact and noise-invariant semantic anchors from raw image–text inputs, reducing sensitivity to nuisance factors. In the second stage, an adaptive reasoning engine integrates retrieval-augmented verification, social credibility propagation, and dynamic propagation signals via gated cross-attention to perform structured reasoning over heterogeneous contexts. An auxiliary logical fallacy detection task further encourages reasoning beyond surface-level pattern matching. Extensive experiments on multiple large-scale benchmarks demonstrate that DAR-Lite consistently outperforms state-of-the-art methods, particularly under low false-positive constraints critical for real-world applications. Efficiency analyses further show that the proposed serial architecture achieves a favorable balance between detection performance and computational cost, making DAR-Lite suitable for practical, large-scale misinformation detection. Our anonymized code is available at https://github.com/MorinW/DAR-Lite.

CCS Concepts

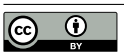
• **Information systems** → **Information systems applications**.

Keywords

Multimodal fake news detection, Robustness and efficiency

ACM Reference Format:
Maolin Wang, Ziting Mai, Zichun Liu, Beining Bao, Hongyu Chen, Junjie Liu, Yunbo Zhang, Bingkun Zhao, Tianshuo Wei, Jian Liu, Chenbin Zhang, and Haoran Yang. 2026. A Serial Two-Stage Framework for Robust Multimodal Fake News Detection via Adaptive Reasoning. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770855.3817957

*Corresponding author.



This work is licensed under a Creative Commons Attribution 4.0 International License.

KDD '26, Jeju Island, Republic of Korea

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3817957

1 Introduction

The proliferation of social media has democratized information dissemination but has simultaneously cultivated a fertile ground for the spread of misinformation [16, 38]. This challenge is currently

being exacerbated by the advent of Generative Artificial Intelligence (AIGC) [24, 37, 45]. Modern fake news is no longer limited to textual fabrication; it has evolved into sophisticated multimodal campaigns where high-fidelity synthetic images and stylistically mimicked text are combined to evade traditional detection systems [15, 28].

Consequently, the field of fake news detection is undergoing a paradigm shift from unimodal text classification to complex Multimodal Fake News Detection (MFND) [36, 39]. However, despite the proliferation of deep learning models in this domain, real-world deployment faces a persistent **"Impossible Triangle"**: the difficulty of simultaneously achieving high **Accuracy**, inference **Efficiency**, and adversarial **Robustness** [12, 33].

Existing State-of-the-Art (SOTA) solutions generally fall into two categories, each compromising one vertex of this triangle [4, 47]. The first category, *Target-Egocentric* methods, focuses heavily on deep fusion of the news content itself [15, 31, 39]. While efficient, these models suffer from "Knowledge Paucity" because they lack the external context required to debunk well-crafted lies that are internally consistent but factually false [10, 47]. The second category, *Retrieval-Enhanced* methods, attempts to bridge this gap by querying external databases or constructing massive heterogeneous graphs [13, 27, 30]. While these methods improve accuracy, they often incur prohibitive computational costs due to parallel processing of massive retrieval sets [3, 26]. Furthermore, they remain vulnerable to noise: retrieving irrelevant information can pollute the feature space, leading to model hallucinations and degraded robustness against adversarial attacks [19, 23].

To address these limitations, we propose **DAR-Lite** (Detection via Adaptive Reasoning Lite), a novel framework that fundamentally reimagines the detection workflow. Unlike traditional parallel fusion architectures, DAR-Lite introduces a **Serial Two-Stage** paradigm. This design decouples feature denoising from logical reasoning, allowing for a rigorous mathematical treatment of noise before expensive graph computations occur.

In the first stage, we introduce a **Variational Information Bottleneck (VIB)** mechanism. Instead of using deterministic embeddings that are prone to overfitting high-frequency noise (such as image background clutter), we map multimodal features into a probabilistic latent space. By minimizing the mutual information between the input and the latent representation while maximizing predictiveness, we distill a robust "semantic anchor" that is impervious to nuisance variables.

In the second stage, we equip the model with an **Adaptive Reasoning Engine**. Rather than issuing online web queries at inference time, DAR-Lite reasons over a fixed, pre-indexed evidence pool and combines two complementary paths: a VIB-guided evidence attention branch for noise-suppressed semantic matching, and a lexical-preserving branch that bypasses the bottleneck to retain high-frequency entity cues. We further introduce a training-time *Logical Fallacy Detection* auxiliary head that regularizes the final representation toward reasoning-aware decision boundaries without adding inference overhead.

Our contributions are summarized as follows:

- **Innovative Serial Architecture:** We propose a serial VIB-GNN framework that compresses multimodal inputs into probabilistic anchors before contextual reasoning, explicitly decoupling denoising from graph reasoning to prevent noisy features from polluting downstream fusion.
- **Robust Reasoning Mechanism:** We introduce dual-path retrieval verification coupled with social credibility propagation, dynamic propagation encoding, and gated cross-attention, enabling adaptive weighing of social credibility against external evidence while preserving residual content semantics.
- **Comprehensive Empirical Validation:** Experiments on Fakeddit, Weibo, and FakeNewsNet show that DAR-Lite consistently outperforms unimodal, multimodal, graph-aware, and LLM-based baselines, with notable gains under low false positive rates.

2 Methodology

This section presents the DAR-Lite framework for multimodal fake news detection, built as a two-stage serial architecture that first distills robust content semantics and then integrates heterogeneous contextual signals for reasoning.

2.1 Problem Formulation

In the traditional paradigm of multimodal fake news detection, the problem is often oversimplified as binary classification of isolated image-text pairs [1]. However, this closed-perspective approach overlooks a fundamental reality: news items do not exist as static, isolated data points but are embedded within complex, dynamically interacting heterogeneous information ecosystems. While fake news often exhibits high deceptiveness at the content level, it frequently exposes vulnerabilities in propagation patterns, social feedback, and external factual consistency. To capture these cross-dimensional subtle cues, we formalize each news event to be detected as a tuple containing multi-source heterogeneous signals: $x_i = \{T_i, I_i, \mathcal{G}_i, \mathcal{K}_i, S_i\}$.

Here, $T_i \in \mathbb{R}^{l_t}$ and $I_i \in \mathbb{R}^{h \times w \times 3}$ represent the textual sequence and visual image of the news, respectively, constituting the "content core" of the news item. $\mathcal{G}_i = (\mathcal{V}_i, \mathcal{E}_i)$ maps the propagation topology of the news within social networks, including publishers, reposters, and their interaction relationships, aimed at revealing potential malicious propagation communities. $\mathcal{K}_i = \{d_1, d_2, \ldots, d_k\}$ represents external factual knowledge documents retrieved from a fixed evidence pool, providing objective verification benchmarks. Meanwhile, $S_i \in \mathbb{R}^{d_s}$ is a set of dynamic statistical features reflecting propagation velocity and breadth. Our core objective is to learn a robust mapping function $f : x_i \to \hat{y}_i$, where $\hat{y}_i \in [0, 1]$ represents the probability that the news is fake.

2.2 Framework Overview

Existing multimodal models predominantly adopt end-to-end parallel fusion strategies, but these face two severe challenges: first, the enormous noise introduced by heterogeneous data makes model convergence difficult; second, high computational costs limit their application in real-time detection scenarios. To address these issues, DAR-Lite adopts a **two-stage serial architecture**.

The key point is that this ordering is not a simple module concatenation. Phase I explicitly compresses raw multimodal observations into a probabilistic semantic anchor before any graph or retrieval interaction occurs, so nuisance information is suppressed before

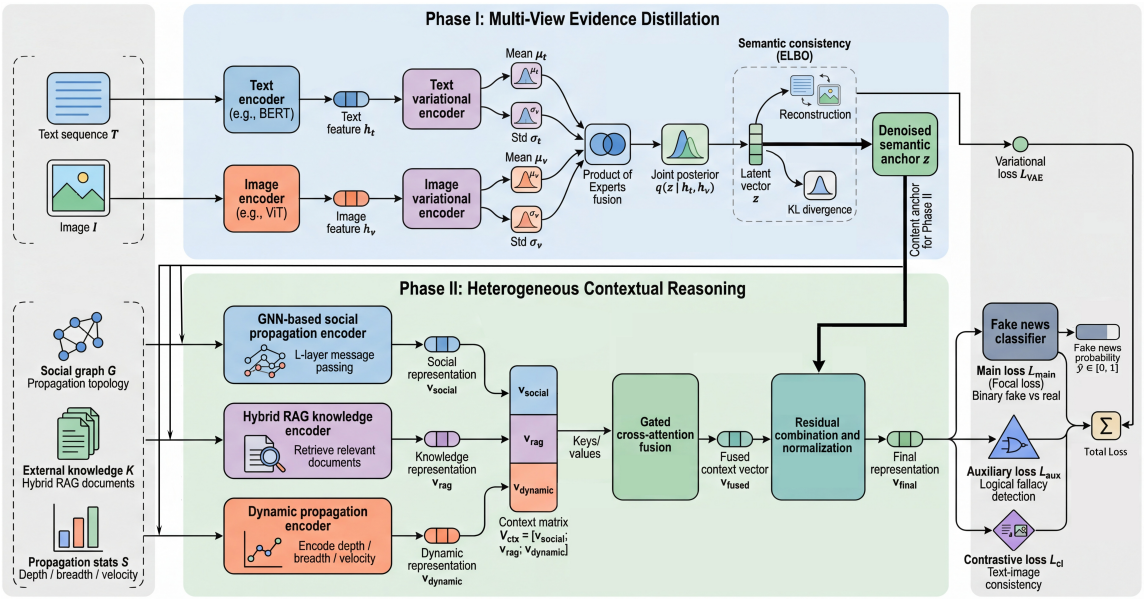


Figure 1: Overview of the DAR-Lite framework. Phase I distills robust semantic anchors from noisy multimodal content, while Phase II performs structured reasoning over heterogeneous contexts.

structured reasoning begins. Phase II then reasons over that compact anchor together with social, evidence, and dynamic signals. This layered pipeline effectively decouples "feature denoising" from "structured reasoning," thereby enhancing model robustness and reducing redundant computation. Formally, we define the overall framework as:

$$\hat{y}_i = f_{\text{Phase-II}}(z_i, \mathcal{G}_i, \mathcal{K}_i, S_i) \tag{1}$$

where $z_i = f_{\text{Phase-I}}(T_i, I_i)$ is the denoised semantic anchor output from the first stage.

2.3 Phase I: Multi-View Evidence Distillation

In the first stage, our primary task is to address the "high-dimensional noise" and "modality gap" problems inherent in raw multimodal data. Direct concatenation of pretrained model outputs often introduces substantial stylistic information irrelevant to fact-checking (such as image filters or emotional rhetoric in text), causing models to overfit easily in low-resource scenarios. Therefore, we design an information bottleneck layer based on variational inference.

2.3.1 Variational Encoding for Robust Representation. For the text and image modalities, we first extract initial features through pretrained encoders (such as BERT and ViT). However, to prevent the model from overly relying on specific, non-robust feature dimensions, we do not directly use these deterministic embeddings. Instead, we map the features to Gaussian distributions in latent space. For modality $m \in \{t, v\}$ with encoder output $h_m \in \mathbb{R}^{d_m}$, the variational encoding process is defined as:

$$\mu_m = W_\mu^{(m)} h_m + b_\mu^{(m)}, \quad \log \sigma_m^2 = W_\sigma^{(m)} h_m + b_\sigma^{(m)} \tag{2}$$

where $W_\mu^{(m)}$ and $W_\sigma^{(m)}$ are learnable projection matrices, $b_\mu^{(m)}$ and $b_\sigma^{(m)}$ are learnable bias vectors, and μ_m, σ_m denote the mean and standard deviation of modality m in latent space. This yields a diagonal Gaussian posterior $q(z \mid h_m) = \mathcal{N}(\mu_m, \text{diag}(\sigma_m^2))$. The stochastic representation introduces uncertainty modeling, forcing the model to ignore minor perturbations in the input and retain only the most robust semantic core.

2.3.2 Consensus via Product of Experts. After obtaining the distribution for each modality, how to effectively fuse them becomes a critical question. Simple concatenation or weighted summation cannot handle information conflicts between modalities. To this end, we introduce the **Product of Experts (PoE)** mechanism. PoE mathematically assumes that the joint posterior distribution is proportional to the product of individual view posteriors:

$$q(z|h_t, h_v) \propto q(z|h_t) \cdot q(z|h_v) \tag{3}$$

This essentially constructs a strict "consensus mechanism": only semantic information that is significantly present in both text and image (i.e., high-probability regions where the two distributions overlap) can be retained in the fused distribution, while modality-specific noise is mutually canceled out. Since each modality follows a Gaussian distribution, the joint distribution parameters can be analytically computed as:

$$\begin{aligned} \mu_{\text{joint}} &= \left(\sum_{m \in \{t,v\}} \sigma_m^{-2} \right)^{-1} \sum_{m \in \{t,v\}} \sigma_m^{-2} \mu_m, \\ \sigma_{\text{joint}}^2 &= \left(\sum_{m \in \{t,v\}} \sigma_m^{-2} \right)^{-1} \end{aligned} \tag{4}$$

Finally, through the reparameterization trick, we sample a low-dimensional latent vector z_i from the joint distribution $\mathcal{N}(\mu_{\text{joint}}, \sigma_{\text{joint}}^2)$.

This z_i, serving as a "minimal sufficient statistic," not only dramatically reduces feature dimensionality but also provides a clean semantic anchor for reasoning in the second stage.

2.3.3 Semantic Consistency Objective. To ensure that the compressed z_i still retains the core semantics of the original image-text pair and prevents information loss, we introduce the variational lower bound (ELBO) as the training objective:

$$\mathcal{L}_{\text{VAE}} = -\mathbb{E}_{q(z|h_t,h_v)}[\log p(h_t, h_v|z)] + \beta \cdot \text{KL}(q(z|h_t, h_v)\|p(z)) \quad (5)$$

The first reconstruction loss term ensures semantic reversibility, while the second KL divergence regularization term constrains the latent distribution to approximate a standard normal prior, thereby regularizing the geometric structure of the latent space.

2.4 Phase II: Heterogeneous Contextual Reasoning

Entering the second stage, the model loads the semantic anchor z_i generated in Phase I and situates it within a broader external context for structured reasoning. The core motivation of this stage is to break the "content isolation," leveraging external knowledge, social structures, and propagation dynamics to aid in discrimination.

2.4.1 Social Credibility Propagation. Fake news often propagates within specific low-credibility communities. To capture this homophily characteristic, we construct a heterogeneous interaction graph $\mathcal{G}_i$ of publishers and news items. We employ Graph Neural Networks (GNNs) for message passing, aiming to aggregate credibility features from publishers and their neighbors. For node u, the feature update rule is:

$$h_u^{(l+1)} = \sigma\left(W^{(l)}\left[h_u^{(l)} \| \text{AGG}(\{h_v^{(l)} : v \in \mathcal{N}(u)\})\right]\right) \quad (6)$$

where $W^{(l)}$ is the learnable weight matrix of the l-th GNN layer,

$$\text{AGG}(\{h_v^{(l)} : v \in \mathcal{N}(u)\}) = \frac{1}{|\mathcal{N}(u)|}\sum_{v\in\mathcal{N}(u)} h_v^{(l)}$$

is mean aggregation over the one-hop neighborhood. We use mean rather than sum or max because it naturally normalizes heterogeneous degree distributions and keeps the scale of neighborhood features stable across sparse and dense propagation graphs. After L layers of propagation, the final embedding v_{social} of the publisher node effectively encodes the credibility of their entire social circle. This means that even if the news content itself appears flawless, if its propagation source resides within a low-credibility community structure, the model can still capture anomalous signals.

2.4.2 Epistemic Validation via Retrieval. To ground prediction in external facts while keeping inference deterministic and reproducible, Phase II does not perform live web search. Instead, each dataset is paired with a fixed, pre-indexed evidence pool from which we retrieve a bounded document set $\mathcal{K}_i$. We then use the semantic anchor z_i to re-rank those candidates through attention:

$$\alpha_j = \frac{\exp(z_i^\top e_j)}{\sum_{j'=1}^{k}\exp(z_i^\top e_{j'})}, \quad v_{\text{evi}} = \sum_{j=1}^{k}\alpha_j e_j \quad (7)$$

where e_j is the embedding of retrieved document d_j. In addition to this VIB-guided evidence path, we preserve lexical cues that the bottleneck may intentionally suppress by projecting the frozen text representation through an independent linear map:

$$v_{\text{lex}} = W_{\text{lex}} h_t, \quad v_{\text{rag}} = v_{\text{evi}} + v_{\text{lex}}$$

The resulting evidence branch is therefore "hybrid" in the sense of combining a noise-suppressed semantic path with a lexical-preserving complementary path over a fixed knowledge pool, rather than combining a static knowledge base with online search. This step thereby provides a fact-checking signal independent of social propagation

2.4.3 Dynamic Propagation Encoding. Fake news typically exhibits viral propagation characteristics (such as explosive growth within short timeframes). To capture this temporal dynamics, we map the statistical features S_i of the propagation tree (such as depth, breadth, and velocity) into a dynamic representation v_{dynamic}. This enables the model to identify "clickbait" content that appears normal but exhibits anomalous propagation behavior.

2.4.4 Gated Cross-Attention Fusion. Faced with three types of heterogeneous information—social, knowledge, and dynamic—the model needs a mechanism to dynamically determine which information is most critical for the current discrimination task. To this end, we propose a gated cross-attention mechanism. We stack all contextual features as V_{ctx} and again use the content anchor z_i as the Query:

$$a = \text{softmax}\left(\frac{W_q z_i \cdot (W_k V_{\text{ctx}})^\top}{\sqrt{d_k}}\right), \quad v_{\text{fused}} = a \cdot (W_v V_{\text{ctx}}) \quad (8)$$

Here, W_q, W_k, and W_v are learnable query, key, and value projection matrices, and d_k is the key dimension used for scaled dot-product normalization. This design endows the model with the capability of "active inquiry": conditioned on the semantic needs of the input news content, it can dynamically determine whether to rely more on social credibility signals or retrieve additional external factual evidence.

2.4.5 Global Residual Connection. Finally, to prevent the model from getting lost in excessive external context and neglecting the news content itself (i.e., "putting the cart before the horse"), we introduce a global residual connection:

$$v_{\text{final}} = \text{LayerNorm}(v_{\text{fused}} + W_z z_i) \quad (9)$$

where W_z is a learnable projection matrix that aligns the semantic anchor with the contextual fusion space. This simple additive operation is crucial: even when external retrieval fails or social networks are sparse, the model can still fall back to content-based discrimination through z_i.

2.5 Multi-Task Collaborative Learning

To further enhance the model's generalization capability, we believe that merely performing binary classification of true/false is insufficient—the model needs to understand "why it is fake." Therefore, we design a multi-task learning objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{main}} + \lambda_1\mathcal{L}_{\text{aux}} + \lambda_2\mathcal{L}_{\text{cl}} + \lambda_3\mathcal{L}_{\text{VAE}} \quad (10)$$

Here, the main task $\mathcal{L}_{\text{main}}$ uses the dataset-specific classification loss described in Appendix B; the auxiliary task $\mathcal{L}_{\text{aux}}$ is implemented as a 2-layer MLP over v_{final} that predicts five fallacy categories (Appeal

to Emotion, Strawman, False Dichotomy, Ad Hominem, and Red Herring); $\mathcal{L}_{\text{cl}}$ is the cross-modal contrastive loss, used to enhance image-text consistency detection. When a dataset does not provide fallacy annotations, we attach a zero supervision mask so that $\mathcal{L}_{\text{aux}}$ contributes zero gradient rather than conflating "unannotated" with "fallacy-free." The fallacy head therefore serves primarily as a training-time structural regularizer rather than a standalone explanation module. Through this multi-task collaboration, DAR-Lite learns not only to classify but also to reason, while the auxiliary head is discarded at inference and adds no deployment overhead.

We further analyze the time and space complexity of DAR-Lite in **Appendix** A, demonstrating that the serial two stage design significantly reduces redundant computation compared to heavy graph based baselines.

3 Experiments

In this section, we evaluate the performance and efficiency of DAR-Lite through extensive experiments.

3.1 Experimental Setup

We conduct experiments on three widely used multimodal fake news detection benchmarks: **Fakeddit** [25], **Weibo** [21], and **FakeNewsNet** [35], covering both English and Chinese scenarios. These datasets differ substantially in terms of linguistic characteristics, content length, and social context density, providing a comprehensive evaluation of model generalization across heterogeneous real-world settings. For fair comparison, we strictly follow the official train/validation/test splits provided by the dataset authors.

All experiments are implemented in PyTorch and conducted on an NVIDIA GPU server. We adopt CLIP-ViT-base-patch32 as the visual encoder for all datasets due to its strong cross-modal alignment capability. For text encoding, we use BERT-base-uncased for English datasets and BERT-base-chinese for the Weibo dataset to accommodate language-specific characteristics. To model social context, we construct heterogeneous interaction graphs and apply a 2-layer Graph Convolutional Network (GCN) for representation refinement. For the FakeNewsNet dataset, where social structures are relatively sparse, we additionally employ Node2Vec to generate structural embeddings and alleviate graph sparsity. The retrieval branch operates on a fixed, pre-indexed evidence pool for each dataset, so all latency numbers correspond to offline evidence access rather than live network queries.

We compare DAR-Lite with a diverse set of representative baselines, including **Text-only**, **Image-only**, **EANN** [39], **EXDR** [8], **CAFE** [5], and **GAMED** [34]. Text-only uses BERT-base-uncased on English data and BERT-base-chinese on Weibo with a linear classifier over the `[CLS]` token, while Image-only uses CLIP-ViT-base-patch32 with a linear classifier; neither baseline uses social graphs or external evidence. To strengthen the comparison, we further re-implement three recent competitive MFND baselines (QMFND, HCCIN, and FSRU) under the same data split and preprocessing protocol, and we benchmark a strong multimodal LLM (Qwen2.5-VL-72B-Instruct) under zero-shot, chain-of-thought, and ReAct prompting. We employ a two-stage training strategy consisting of MVEB pre-training followed by downstream fine-tuning of the reasoning layers. The model is optimized using the Adam optimizer with a CosineAnnealingLR learning rate scheduler to ensure stable convergence. For evaluation, we report Accuracy, Macro-F1, AUC, and standardized partial AUC under low false-positive-rate (FPR) constraints, which are particularly relevant for real-world deployment scenarios. Detailed experimental settings, hyperparameter configurations, release protocol, and supervision details are provided in **Appendix** B.

3.2 Main Results and Discussion

Table 1 reports the overall performance comparison between DAR-Lite and representative baselines across four benchmarks. DAR-Lite consistently achieves the strongest performance across all datasets and metrics, demonstrating robust generalization under diverse languages, text lengths, and social context densities.

First, multimodal fusion brings a clear advantage. Compared with unimodal baselines, DAR-Lite achieves substantially better results, confirming that single-modality approaches cannot reliably handle cross-modal manipulation such as misleading captions paired with unrelated images. By jointly modeling textual and visual cues, DAR-Lite forms a more robust decision boundary for detecting subtle inconsistencies.

Second, DAR-Lite demonstrates strong cross-modal reasoning capability. Compared with multimodal baselines such as EANN and CAFE, our framework consistently achieves superior performance, suggesting that explicitly learning discriminative and noise-invariant representations is more effective than relying on adversarial feature removal or ambiguity-based weighting alone.

Third, the comparison with graph-aware baselines validates DAR-Lite's lightweight social modeling strategy. While GNN-based methods exploit social structures through expensive message passing, DAR-Lite achieves better results with a more efficient design by distilling essential social context in a parameter- and computation-efficient manner, making it suitable for latency-sensitive scenarios and robust under sparse social connectivity.

Overall, these results confirm that DAR-Lite effectively combines robust multimodal fusion, discriminative representation learning, and efficient social context utilization, providing a reliable and scalable solution for real-world misinformation detection.

3.3 Additional Baseline Analysis

Table 2 shows that DAR-Lite remains clearly ahead of recent specialized MFND baselines even after they are re-run under the same preprocessing protocol. The margin is especially large on *spAUC*, indicating that the proposed serial reasoning pipeline improves low-FPR reliability rather than only average-case classification accuracy. Table 3 further positions DAR-Lite against a strong 72B multimodal LLM. Prompt engineering improves the LLM from zero-shot to ReAct, but the gap to DAR-Lite remains substantial on both Fakeddit and Weibo. The main reason is structural rather than parametric: prompting-based MLLMs operate primarily on content, while DAR-Lite explicitly reasons over social propagation graphs, offline evidence indices, and dynamic diffusion statistics.

Table 1: Performance comparison on four benchmarks. We report Macro-F1 (Mac-F1), Accuracy (Acc), AUC, and standardized partial AUC (*spAUC*). The best results are highlighted in bold.

Method	Fakeddit				FakeNewsNet-PolitiFact				FakeNewsNet-GossipCop				Weibo			
	Mac-F1	Acc	AUC	spAUC	Mac-F1	Acc	AUC	spAUC	Mac-F1	Acc	AUC	spAUC	Mac-F1	Acc	AUC	spAUC
GAMED	0.6240	0.6231	0.6592	0.5457	0.5120	0.5840	0.5315	0.5120	0.5210	0.7120	0.5140	0.5085	0.8374	0.8375	0.8906	0.7852
Text-only	0.8471	0.8471	0.9236	0.7928	0.6250	0.6480	0.6850	0.5920	0.6140	0.7650	0.6210	0.5840	0.8270	0.8270	0.8804	0.7852
Image-only	0.8620	0.8622	0.9378	0.8591	0.4580	0.5910	0.5210	0.5050	0.4850	0.7510	0.5120	0.5020	0.3402	0.5155	0.5000	0.5000
EANN	0.8988	0.8989	0.9644	0.9048	0.7245	0.7410	0.7850	0.6940	0.6850	0.7820	0.7150	0.6540	0.8378	0.8381	0.8913	0.7972
CAFE	0.8978	0.8978	0.9642	0.9035	0.7380	0.7520	0.7910	0.7020	0.7010	0.7940	0.7320	0.6680	0.8450	0.8460	0.8928	0.8120
EXDR	0.8498	0.8498	0.9253	0.7966	0.6850	0.6920	0.7150	0.6240	0.6520	0.7710	0.6840	0.6150	0.8357	0.8258	0.8921	0.7954
DAR-Lite (Ours)	**0.9298**	**0.9297**	**0.9816**	**0.9421**	**0.8916**	**0.8962**	**0.9483**	**0.8611**	**0.8040**	**0.8598**	**0.8885**	**0.7607**	**0.9065**	**0.9065**	**0.9632**	**0.9126**

Table 2: Additional comparison with recent competitive baselines on Fakeddit under our unified preprocessing protocol.

Method	Mac-F1	Acc	AUC	spAUC
QMFND	0.8610	0.8615	0.9340	0.6920
HCCIN	0.8725	0.8740	0.9425	0.6815
FSRU	0.8695	0.8710	0.9385	0.6780
DAR-Lite (Ours)	**0.9298**	**0.9297**	**0.9816**	**0.9421**

Table 3: Comparison with Qwen2.5-VL-72B-Instruct under different prompting strategies.

Method	Fakeddit Mac-F1	Weibo Mac-F1
Zero-shot	0.8012	0.7834
CoT	0.8267	0.8045
ReAct	0.8431	0.8192
DAR-Lite (Ours)	**0.9298**	**0.9065**

Table 4: Efficiency analysis on the Fakeddit dataset. We report training time per epoch, inference time per epoch, and model parameter size.

Model	Train (s/epoch)	Inference (s/epoch)	Params (MB)
Text-only	45.88	37.89	0.76
Image-only	45.12	25.14	0.76
EANN	70.80	26.08	0.88
CAFE	77.14	28.49	0.88
EXDR	67.28	27.95	0.88
GAMED	556.12	237.94	297.28
DAR-Lite (Ours)	185.60	34.36	302.44

3.4 Efficiency and Scalability Analysis

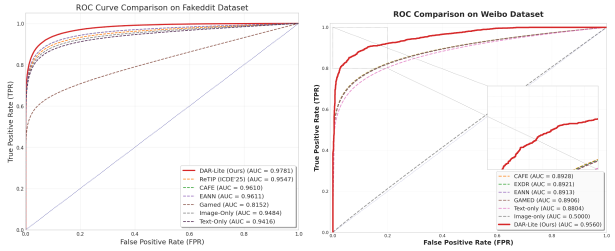
In addition to detection performance, real-world deployment of fake news detection systems requires a careful balance between accuracy and computational efficiency. Table 4 presents a comprehensive comparison of training time, inference latency, and model size on the large-scale Fakeddit benchmark. Because Phase II uses a fixed offline evidence pool, the reported latency isolates model computation and excludes external network I/O for both fairness and reproducibility.

Table 5: Retrained major-branch ablation on the Fakeddit dataset. We evaluate the impact of removing the social graph, textual modality, and visual modality.

Variant	Binary Classification		6-Way Classification	
	Macro-F1	Δ	Macro-F1	Δ
Full Model (DAR-Lite)	**0.9269**	-	**0.8287**	-
w/o GNN	0.7952	-0.1317	0.3946	-0.4341
w/o Text	0.9268	-0.0001	0.8029	-0.0258
w/o Image	0.9264	-0.0005	0.8241	-0.0046

Superior Inference Speed compared to Graph Baselines. While heavy graph-based methods like GAMED achieve competitive accuracy, they suffer from prohibitive computational costs (237.94 s/epoch for inference) due to the complex message-passing operations required on massive heterogeneous graphs. In contrast, our proposed **DAR-Lite reduces inference latency by approximately 7× (34.36 s/epoch)**. This significant speedup validates the effectiveness of our serial architecture: by decoupling feature denoising from structured reasoning, we successfully eliminate the bottleneck of heavy graph computations, making the framework highly suitable for latency-sensitive industrial applications.

Competitive with Lightweight Baselines. Compared to lightweight baselines such as EANN and CAFE, DAR-Lite incurs a moderate increase in training time and parameter size (302.44 MB). This is an expected trade-off for incorporating the retrieval verification branch and training-time reasoning regularization. Critically, however, our **inference speed remains in the same order of magnitude** as these lightweight models. This demonstrates that DAR-Lite achieves the strongest accuracy in Table 1 without sacrificing the rapid response capability required for high-throughput social media platforms. Additional cross-dataset latency trends in Appendix **??** show the same pattern on Weibo and FakeNewsNet, where DAR-Lite remains roughly 5–7× faster than GAMED. In summary, DAR-Lite successfully navigates the "Impossible Triangle," delivering the robust reasoning capability typically associated with heavy graph models while maintaining the inference efficiency of lightweight fusion architectures.



(a) ROC on Fakeddit Dataset **(b) ROC on Weibo Dataset**

Figure 2: Diagnostic Performance Analysis. (a) On Fakeddit, DAR-Lite (AUC=0.9781) shows superior low-FPR reliability compared to GAMED. (b) On Weibo, DAR-Lite (AUC=0.9560) robustly handles visual noise.

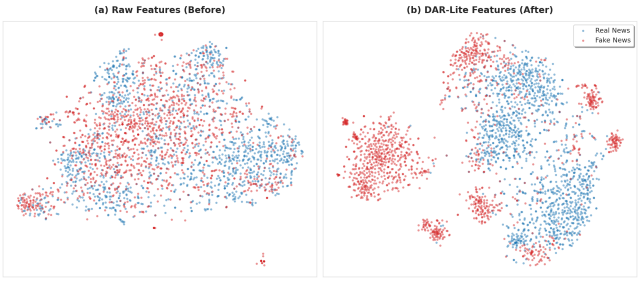


Figure 3: t-SNE visualization of feature distributions on the Fakeddit dataset. (a) Raw features show severe class overlap. (b) DAR-Lite features exhibit clear separation between real (blue) and fake (red) news, demonstrating the model's ability to learn a discriminative manifold.

3.5 Diagnostic Performance Analysis

To strictly evaluate the robustness of DAR-Lite in distinguishing deceptive content from genuine news, we visualize the Receiver Operating Characteristic (ROC) curves on the Weibo and Fakeddit datasets, as illustrated in Figure 2.

Robustness against Unimodal Failure. On the Weibo dataset, a distinct "performance gap" is observed. As shown in Figure 2(a), the Image-only baseline (grey dashed line) hovers near 0.50, indicating that visual signals in this dataset are extremely noisy or weakly correlated with veracity. Despite this unimodal failure, DAR-Lite (red solid curve) successfully leverages its Adaptive Reasoning Engine to extract discriminative features from the textual and social contexts, achieving a remarkable AUC of **0.9560**. This significantly outperforms state-of-the-art multimodal methods like CAFE and EANN, which cluster around an AUC of 0.89, validating our framework's capability to handle modality-imbalanced scenarios.

Superiority in Industrial Constraints. On the large-scale Fakeddit benchmark, DAR-Lite continues to lead with an AUC of **0.9781**. More critically, for real-world content moderation where user experience is paramount, the False Positive Rate (FPR) serves as a key bottleneck. Figure 2(b) reveals that DAR-Lite exhibits the **steepest initial ascent** in the low-FPR region ($FPR < 0.05$). This indicates that our model maintains a high True Positive Rate while minimizing false alarms. In contrast, the heavy graph-based baseline GAMED (brown curve) shows a sluggish rise (AUC=0.8152), suggesting it struggles to distinguish ambiguous samples effectively. This empirical evidence confirms that DAR-Lite's serial architecture not only significantly improves computational efficiency but also enhances detection precision in high-stakes environments.

3.6 Robustness under Visual Perturbation

We further evaluate the contribution of the VIB stage under additive visual perturbations. On clean data, the stochastic bottleneck incurs a small cost because compression removes a small amount of high-frequency detail together with nuisance features. However, the robustness curves cross at approximately $\sigma \approx 0.2$, after which the VIB branch consistently outperforms deterministic features. At severe corruption ($\sigma = 1.0$), the gain reaches +0.1652 Binary F1 ($p < 0.001$), indicating that constraining $I(X;Z)$ makes the semantic anchor substantially less sensitive to image noise. Comparable robustness evaluation under text- or graph-level attacks is left for future work, as discussed in Appendix **??**.

Table 6: Inference-time component probes on the full Fakeddit test set. Components are zeroed without retraining, so the absolute values are not directly comparable to Table 5; the goal is to isolate relative dependence on each module.

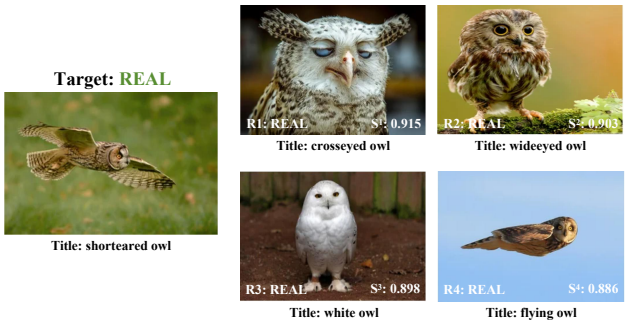
Variant	Binary F1		6-Way F1		AUC	
	Score	Δ	Score	Δ	Score	Δ
Full model	**0.8609**	–	**0.5463**	–	**0.9452**	–
w/o Graph	0.8474	-0.0135	0.4506	-0.0957	0.9331	-0.0121
w/o RAG	0.8537	-0.0072	0.4842	-0.0621	0.9382	-0.0070
w/o VIB	0.8513	-0.0096	0.5017	-0.0446	0.9367	-0.0085
w/o Image	0.8549	-0.0060	0.5099	-0.0364	0.9408	-0.0044
w/o $\mathcal{L}_{\text{aux}}$	0.8572	-0.0037	0.5198	-0.0265	0.9421	-0.0031
w/o $\mathcal{L}_{\text{cl}}$	0.8581	-0.0028	0.5291	-0.0172	0.9430	-0.0022

3.7 Ablation Study

To rigorously verify the contribution of each critical component in DAR-Lite, we conduct an ablation study on the Fakeddit dataset. We report both Binary Macro-F1 (for fake vs. real detection) and 6-Way Macro-F1 (for fine-grained categorization). The results are summarized in Table 5.

Impact of Social Context Modeling. As observed in Table 5, the removal of the Graph Neural Network module (w/o GNN) results in the largest performance drop. Specifically, the Binary F1 score decreases by **13.17%**, while the 6-Way F1 score drops by over **43%**. This empirical evidence strongly suggests that on social media platforms like Reddit (the source of Fakeddit), the *propagation structure* and *community interaction patterns* are the most dominant indicators of veracity. Without the social context captured by our Phase II reasoning engine, relying solely on content features is insufficient for robust detection.

Necessity of Multimodal Content. Removing the textual (w/o Text) or visual (w/o Image) branches leads to varying degrees of performance degradation. Notably, while the binary classification performance remains relatively stable without text, the fine-grained classification (6-Way F1) suffers a significant drop of **2.58%** (from

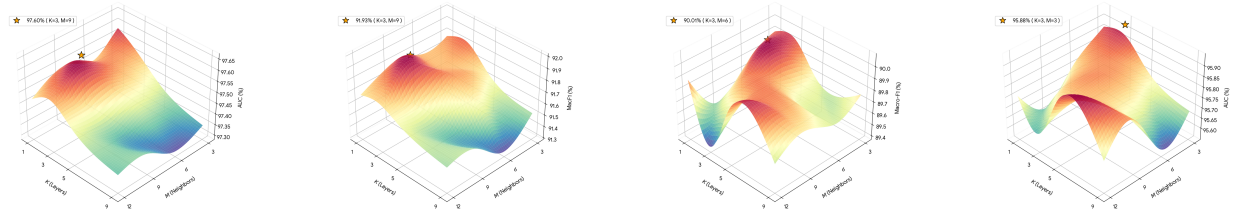


(a) Nearest-neighbor retrieval for a real news query. Retrieved samples are semantically consistent with high similarity scores, indicating fine-grained alignment in latent space.



(b) Nearest-neighbor retrieval for a fake news query. Retrieved samples exhibit strong stylistic and structural similarities (e.g., meme templates), reflecting learned manipulation patterns.

Figure 4: Case studies of nearest-neighbor retrieval for real and fake news queries. The model captures both semantic alignment and stylistic manipulation patterns in the latent space.



(a) Fakeddit (AUC) (b) Fakeddit (Macro-F1) (c) Weibo (AUC) (d) Weibo (Macro-F1)

Figure 5: Model performance with varying numbers of retrieved news instances (K) and participants (M). The orange star indicates the optimal hyperparameter setting for each dataset.

0.8287 to 0.8029). This indicates that while social graphs effectively distinguish "fake" from "real," the *semantic content* provided by text is crucial for identifying the specific *intent* and *subtype* of misinformation (e.g., distinguishing satire from misleading content). Similarly, visual signals contribute to the overall robustness, validating the necessity of our Phase I multi-view evidence distillation.

To isolate components whose removal would otherwise require re-architecting the model, Table 6 reports inference-time zeroing probes on the full Fakeddit test split. The ranking is clear: social propagation remains the dominant veracity signal, the retrieval branch is the second most influential contextual source, and replacing VIB with deterministic embeddings degrades both Binary F1 and 6-Way F1.

The auxiliary fallacy loss and the contrastive objective each provide smaller but consistent gains, indicating that reasoning-level supervision and boundary sharpening both improve generalization.

3.8 Feature Visualization and Case Study

We visualize the feature distribution of the Fakeddit test set using t-SNE to examine the geometric structure of the learned representation space. Figure 3 compares raw multimodal features, obtained by direct concatenation of pre-trained encoders, with the refined representations produced by DAR-Lite. As shown in Figure 3(a), the raw features exhibit substantial overlap between real and fake news, suggesting that standard pre-trained embeddings are noisy and lack clear separability for veracity prediction. In contrast, Figure 3(b) shows that DAR-Lite projects samples into a more discriminative manifold, with clear inter-class separation and improved intra-class compactness. This structure indicates that the VIB-based denoising mechanism suppresses nuisance factors while preserving semantically meaningful information relevant to news authenticity.

To empirically validate the discriminative quality of the multimodal latent representations learned by DAR-Lite, we conduct a qualitative analysis via nearest-neighbor retrieval. Specifically, we query the test set and retrieve the top-K nearest neighbors from the training set based on cosine similarity in the shared latent space. Figure 4a and Figure 4b visualize the retrieval results for a real news sample and a fake news sample, respectively.

For the real news case (Figure 4a), the query image depicts a 'shorteared owl". The retrieved neighbors are not only semantically relevant but also taxonomically consistent, including species such as 'crosseyed owl" ($S = 0.915$) and "wideeyed owl" ($S = 0.903$). The consistently high similarity scores indicate that the model encodes fine-grained visual–semantic features, projecting factually consistent content into compact clusters within the manifold.

Conversely, the fake news case (Figure 4b) involves a manipulated image with casual text ('oh hey there buddy"). The retrieved neighbors exhibit strong stylistic homogeneity, sharing meme-like structures (e.g., pareidolia in objects) and phrasing patterns ('oh hey there neighbor"). Similarity scores peak at 0.958, suggesting that the model captures high-level generative artifacts and stylistic signatures intrinsic to misinformation. This contrast shows that DAR-Lite differentiates content via factual consistency for real news and stylistic patterns for fake news.

3.9 Sensitivity Analysis of Hyperparameters

To evaluate the robustness of DAR-Lite, we conduct a sensitivity analysis on two pivotal hyperparameters: the number of retrieved news instances K and the number of retrieved potential participants M. We vary K and M within the ranges of $[1, 9]$ and $[3, 12]$, respectively. The results are visualized as 3D surface plots in Figure 5.

As shown in Figure 5 (a) and (b), on the Fakeddit dataset, performance exhibits a characteristic "parabolic" trend—accuracy climbs with initial context integration but tapers off beyond an optimal threshold. The model achieves the best performance at $K = 3$ and $M = 9$, while the Macro-F1 band over the explored grid remains narrow, indicating robustness to moderate perturbations. The denser Reddit interaction graph can tolerate a larger M because there are enough genuinely informative neighbors to amortize the noise from peripheral users.

Figures 5 (c) and (d) show that the Weibo dataset demonstrates a notably sharper response to hyperparameter adjustments. Optimal performance is achieved at $K = 3$ with M maintained within a tighter range of 3 to 6. This heightened sensitivity reflects the structural volatility of the Weibo social graph, characterized by rapid, bursty engagement and more pronounced label noise, which necessitates a more parsimonious selection of context.

4 Related Works

From Content to Context: The Evolution of Paradigms. Early multimodal fake news detection largely followed a *target-egocentric* paradigm, inferring veracity from signals internal to the news item, such as textual rhetoric, visual cues, and image–text consistency [39]. Most pipelines encode each modality using pre-trained backbones and perform multimodal alignment via attention-based or bilinear fusion mechanisms [7, 32, 40]. While these end-to-end models are computationally efficient, they often suffer from *knowledge paucity*, relying on superficial intra-content inconsistencies and failing on high-level fakes that are internally coherent but externally fabricated [42, 46]. To bridge this gap, recent work incorporates external context for epistemic verification [9]. Retrieval-enhanced methods query external corpora to obtain supporting evidence and mitigate unsupported reasoning [17, 18], while social-aware approaches model dissemination through graph-based learning to capture user credibility and propagation structures [43, 44]. However, many methods rely on *parallel fusion*, treating retrieved evidence and social context as additional feature channels, which complicates joint optimization of accuracy, efficiency, and robustness (the "impossible triangle") [29], as processing large retrieval sets or heterogeneous graphs is costly and irrelevant evidence introduces feature pollution [6, 18]. Prompting-based multimodal LLMs provide a strong content-only reference point, but do not explicitly access propagation graphs or dataset-specific evidence indices, making it difficult to recover the structured social signals required by real-world misinformation detection.

Robust Representation Learning & Reasoning Mechanisms. To improve robustness under noisy context, recent studies shift from naïve fusion toward explicit denoising and robustness-oriented representation learning [34]. For high-frequency multimodal noise such as visual clutter and emotionally charged language, Information Bottleneck based approaches preserve task-relevant semantics while suppressing nuisance factors [2, 11]. In particular, the Variational Information Bottleneck projects features into a probabilistic latent space, retaining predictive semantic anchors while filtering perturbation-sensitive components [2]. Beyond representation purification, effective detection requires reasoning over heterogeneous and dynamic context [20], including aggregating static evidence and modeling propagation dynamics [41]. A central challenge lies in adaptively integrating social credibility and external factual evidence [22]; gated cross-attention and conditional routing enable dynamic context weighting and fallback to content-based cues when external signals are unreliable [22, 40]. Auxiliary objectives such as logical fallacy detection further reduce overfitting to dataset-specific artifacts [14]. Despite these advances, few methods jointly integrate VIB-style denoising and adaptive context reasoning within a unified framework for rumor detection. To address this gap, DAR-Lite adopts a serial two-stage design that decouples feature purification from contextual reasoning, reducing noise propagation and redundant computation.

5 Conclusion

This paper presents DAR-Lite, a serial two-stage framework for multimodal fake news detection that balances accuracy, inference efficiency, and robustness under real-world constraints. Unlike prior parallel fusion approaches, DAR-Lite decouples multimodal representation denoising from heterogeneous contextual reasoning, enabling a scalable detection pipeline. At the representation level, we introduce a variational information bottleneck to distill noise-invariant semantic anchors from raw image-text inputs. At the reasoning level, we design an adaptive engine that integrates social credibility propagation, fixed-pool retrieval verification, lexical-preserving evidence cues, and dynamic propagation signals through gated cross-attention. A training-time logical fallacy auxiliary head further regularizes the final representation toward reasoning-aware decision boundaries. Experiments on multiple large-scale benchmarks show that DAR-Lite consistently outperforms representative unimodal, multimodal, graph-aware, recent competitive, and prompting-based MLLM baselines, particularly under low false-positive constraints relevant to real-world deployment. Efficiency analyses and ablations confirm that the serial architecture achieves a favorable balance between performance and computational cost.

Acknowledgements

This work was supported by Central South University under Grant No. 16500-502045019. We would like to thank Yingqian Feng for her valuable support.

References

[1] Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A survey on multimodal disinformation detection. In *Proceedings of the 29th international conference on computational linguistics*. 6625–6643.

[2] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410* (2016).

[3] Pepa Atanasova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, and James Glass. 2019. Automatic fact-checking using context and discourse information. *Journal of Data and Information Quality (JDIQ)* 11, 3 (2019), 1–27.

[4] Alessandro Bondielli and Francesco Marcelloni. 2019. A survey on fake news and rumour detection techniques. *Information sciences* 497 (2019), 38–55.

[5] 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*. 2897–2905.

[6] Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 719–729.

[7] 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, volume 1 (long and short papers)*. 4171–4186.

[8] Guoxuan Ding, Yuqing Li, Ziyan Zhou, Zheng Lin, Daren Zha, and Jiangnan Li. 2026. ExDR: Explanation-driven Dynamic Retrieval Enhancement for Multimodal Fake News Detection. *arXiv preprint arXiv:2601.15820* (2026).

[9] Tayasan MH Gedara, Rosario Gaeta, Stefania Tomasiello, and Radwa El Shawi. 2025. Toward Explainable Multimodal Fake News Detection: A Fusion of Prompt Learning and Automated Machine Learning. *IEEE Access* 13 (2025), 188150–188169.

[10] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the association for computational linguistics* 10 (2022), 178–206.

[11] Linfeng Han, Xiaoming Zhang, Tianbo Wang, Yun Liu, and Zhiqiang Dong. 2026. Enhancing large language model for fake news video detection via cross-modal retrieval. *Information Processing & Management* 63, 2 (2026), 104471.

[12] Linmei Hu, Siqi Wei, Ziwang Zhao, and Bin Wu. 2022. Deep learning for fake news detection: A comprehensive survey. *AI open* 3 (2022), 133–155.

[13] Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 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.

[14] Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022. Logical Fallacy Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 7180–7198. doi:10.18653/v1/2022.findings-emnlp.532

[15] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*. 2915–2921.

[16] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.

[17] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.

[18] Hao Liao, Jiahao Peng, Zhanyi Huang, Wei Zhang, Guanghua Li, Kai Shu, and Xing Xie. 2023. Muser: A multi-step evidence retrieval enhancement framework for fake news detection. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*. 4461–4472.

[19] Fanzhen Liu, Sharif Abuadbba, Kristen Moore, Surya Nepal, Cecile Paris, Jia Wu, Jian Yang, and Quan Z Sheng. 2025. Adversarial attacks against automated fact-checking: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 22979–23001.

[20] Hui Liu, Wenya Wang, and Haoliang Li. 2023. Interpretable Multimodal Misinformation Detection with Logic Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 9781–9796. doi:10.18653/v1/2023.findings-acl.620

[21] DONG Luo, NHM Nor, KG Bai, AZIAN MUHAMAD Adzmi, and SITI SYUHADA ABDUL Rais. 2023. Fake news sharing among Weibo users in China. *Jurnal Komunikasi: Malaysian Journal of Communication* 39, 4 (2023).

[22] Hongzhen Lv, Wenzhong Yang, Fuyuan Wei, Jiaren Peng, and Haokun Geng. 2024. Mdf: A dynamic fusion model for multi-modal fake news detection. *arXiv preprint arXiv:2406.19776* (2024).

[23] Jing Ma, Wei Gao, Shafiq Joty, and Kam-Fai Wong. 2019. Sentence-level evidence embedding for claim verification with hierarchical attention networks. Association for Computational Linguistics.

[24] Yisroel Mirsky and Wenke Lee. 2021. The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR)* 54, 1 (2021), 1–41.

[25] Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the twelfth language resources and evaluation conference*. 6149–6157.

[26] Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. *arXiv preprint arXiv:2103.07769* (2021).

[27] Van-Hoang Nguyen, Kazunari Sugiyama, Preslav Nakov, and Min-Yen Kan. 2020. Fang: Leveraging social context for fake news detection using graph representation. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 1165–1174.

[28] Sophie J Nightingale and Hany Farid. 2022. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences* 119, 8 (2022), e2120481119.

[29] Idza Aisara Norabid, Noor Hafhizah Abd Rahim, and Masita Abdul Jalil. 2025. Comparative Study of Fusion Techniques for Multimodal Fake News Detection. In *2025 14th International Conference on Information Technology in Asia (CITA)*. IEEE, 54–59.

[30] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. Declare: Debunking fake news and false claims using evidence-aware deep learning. *arXiv preprint arXiv:1809.06416* (2018).

[31] Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 153–162.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[33] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. 2020. Green ai. *Commun. ACM* 63, 12 (2020), 54–63.

[34] Lingzhi Shen, Yunfei Long, Xiaohao Cai, Imran Razzak, Guanming Chen, Kang Liu, and Shoaib Jameel. 2025. Gamed: Knowledge adaptive multi-experts decoupling for multimodal fake news detection. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*. 586–595.

[35] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data* 8, 3 (2020), 171–188.

[36] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter* 19, 1 (2017), 22–36.

[37] Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. 2023. AI model GPT-3 (dis) informs us better than humans. *Science Advances* 9, 26 (2023), eadh1850.

[38] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science* 359, 6380 (2018), 1146–1151.

[39] 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 & data mining*. 849–857.

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

[41] Zhiyuan Wu, Dechang Pi, Junfu Chen, Meng Xie, and Jianjun Cao. 2020. Rumor detection based on propagation graph neural network with attention mechanism. *Expert systems with applications* 158 (2020), 113595.

[42] Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. 2021. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management* 58, 5 (2021), 102610.

[43] Facheng Yan, Mingshu Zhang, Bin Wei, Wen Jiang, and Kelan Ren. 2023. Fake News Detection Utilizing Social Context Information with Graph Convolutional Networks and Attention Mechanisms. In *Proceedings of the 2023 7th International Conference on Electronic Information Technology and Computer Engineering*. 406–413.

[44] Litian Zhang, Xiaoming Zhang, Ziyi Zhou, Xi Zhang, Senzhang Wang, Philip S Yu, and Chaozhuo Li. 2024. Early detection of multimodal fake news via reinforced propagation path generation. *IEEE Transactions on Knowledge and Data Engineering* (2024).
[45] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–20.
[46] X Zhou, J Wu, and R Zafarani. 2020. SAFE: similarity-aware multi-modal fake news detection. arXiv. *arXiv preprint arXiv:2003.04981* (2020).
[47] Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)* 53, 5 (2020), 1–40.

A Complexity Analysis

We analyze the computational complexity of DAR-Lite by decomposing it into its two serial stages.

In Phase I, the framework performs multimodal evidence distillation by encoding textual and visual inputs using pretrained backbone models, followed by variational representation learning. Let L denote the text sequence length and d the hidden feature dimension. The dominant computational cost comes from the backbone encoders, with a complexity of $O(L \cdot d^2)$ for text encoding and $O(d^2)$ for visual encoding. The variational information bottleneck introduces lightweight operations, including linear projections, Gaussian sampling, and KL divergence computation, whose complexity scales linearly with the latent dimension. Since Phase I is executed once per news item and is independent of contextual scale, its computational overhead remains stable during inference.

In Phase II, the denoised semantic anchor guides heterogeneous contextual reasoning. Social credibility modeling is implemented using a shallow graph convolutional network with depth l, resulting in a complexity of $O(l \cdot |E|)$, where $|E|$ denotes the number of edges in the local propagation graph. Retrieval verification attends over a bounded set of retrieved documents K, yielding a complexity of $O(K \cdot d)$, while the lexical-preserving bypass contributes only a single linear projection. Dynamic propagation encoding operates on compact statistical features and incurs negligible computational cost. The gated cross attention mechanism scales linearly with the number of contextual sources and remains independent of global graph size. Overall, DAR-Lite avoids parallel processing of large heterogeneous graphs or massive retrieval sets. By serially distilling noise invariant semantic anchors before contextual reasoning, the framework reduces redundant computation and prevents noise amplification. As a result, the inference complexity scales with local context size rather than global graph density, explaining the favorable efficiency observed in practice. We also emphasize that all timing measurements in the paper are collected with an offline indexed evidence pool, so they measure model computation rather than live network I/O.

B Experimental Details

B.1 Datasets

To comprehensively evaluate the effectiveness, robustness, and linguistic generalizability of our proposed framework, we conduct extensive experiments on three widely used multimodal fake news detection benchmarks: **Fakeddit** [25], **Weibo** [21], and **FakeNewsNet** [35]. These datasets vary significantly in terms of language (English vs. Chinese), content length (short posts vs. long articles), and social context density. Detailed statistics of these datasets, including the distribution of real and fake news across different splits, are summarized in Table 7.

Table 7: Dataset statistics of Fakeddit, Weibo, and FakeNewsNet.

Dataset	Split	Real	Fake	Total	Text Length
Fakeddit	Train	289282	188235	477517	41.9
	Val	41326	26891	68217	41.7
	Test	82652	53782	136434	42.3
FakeNewsNet-PolitiFact	Train	869	303	1172	9.6
	Val	62	43	105	9.0
	Test	125	86	211	9.5
FakeNewsNet-GossipCop	Train	11772	3726	15498	11.2
	Val	1682	532	2214	11.0
	Test	3363	1065	4428	12.1
Weibo	Train	3803	4451	8254	67.4
	Val	543	636	1179	64.0
	Test	1087	1272	2359	66.3

Table 8: Graph statistics of different datasets.

Statistic	Fakeddit	FakeNewsNet-PolitiFact	FakeNewsNet-GossipCop	Weibo
Node nums	1641717	43162	358542	70957
Edge nums	100792812	41794	330938	23662

B.1.1 Fakeddit. Fakeddit is a large-scale, fine-grained multimodal dataset sourced from Reddit. We utilize its multimodal subset, which comprises text, images, and associated user comments. Fakeddit is characterized by high-quality visual content and complex social interactions, providing a challenging testbed for multi-element retrieval and multimodal fusion.

B.1.2 Weibo. Weibo is a widely adopted Chinese benchmark dataset collected from Sina Weibo. It contains real-world news posts verified by official sources (e.g., Xinhua News Agency) and fake news debunked by the Weibo official rumor-busting system. This dataset is highly heterogeneous and noisy, posing unique challenges regarding linguistic diversity and class imbalance.

B.1.3 FakeNewsNet. FakeNewsNet is a comprehensive dataset consisting of two sub-datasets, **PolitiFact** and **GossipCop**. Unlike Fakeddit and Weibo, FakeNewsNet focuses on full-length news articles embedded with rich social context (e.g., user retweets and profile metadata). The graph structure in FakeNewsNet is often sparser compared to social media posts, requiring more robust structural feature extraction.

B.2 Baselines

To comprehensively evaluate the performance of our proposed framework, **DAR-Lite**, we conduct comparative experiments against a diverse set of baseline methods.

B.2.1 Unimodal Baselines. We first include unimodal models that rely on a single modality as lower-bound baselines. **Text-only** adopts BERT-base-uncased for English datasets and BERT-base-chinese for Weibo, and applies a linear classifier on the `[CLS]` embedding. This baseline detects misinformation solely from linguistic patterns while ignoring visual evidence, propagation graphs, and

external evidence. **Image-only** employs CLIP-ViT-base-patch32 followed by a linear classifier, aiming to capture visual manipulation cues without considering textual or social signals.

B.2.2 Multimodal Baselines. We further compare DAR-Lite with representative multimodal fake news detection frameworks to demonstrate the superiority of our fusion strategy. **EANN** [39] is an early method designed to address domain shift by incorporating an event discriminator and applying adversarial learning to remove event-specific features, thereby encouraging event-invariant multimodal representations. **EXDR** [8] serves as a strong representative of deep multimodal representation learning approaches, typically fusing textual and visual features through concatenation or attention pooling, and provides a standard benchmark for evaluating basic cross-modal correlation modeling. **CAFE** [5] is a recent state-of-the-art model that tackles cross-modal ambiguity by explicitly measuring the inconsistency between text and image modalities, and dynamically adjusts unimodal and multimodal contributions based on an ambiguity score to effectively mitigate the negative impact of irrelevant images.

B.2.3 Graph-aware Baselines. To evaluate the contribution of social structural modeling and contrast our lightweight graph strategy, we include heavy graph-based baselines. **GAMED** [34] constructs a heterogeneous graph consisting of users, posts, and entities, and employs graph neural networks such as GCN or GAT to perform message passing. By modeling high-order propagation patterns and rich social context together with multimodal content features, GAMED represents a typical heavy-weight graph-aware alternative to our proposed DAR-Lite framework.

B.2.4 Recent Competitive Baselines. To address concerns about comparison breadth, we additionally re-implement three recent competitive methods under our standardized preprocessing and split protocol: **QMFND**, **HCCIN**, and **FSRU**. These methods strengthen the reference set beyond classic MFND baselines by covering more recent feature interaction and consistency modeling designs.

B.2.5 Prompting-based Multimodal LLMs. We also evaluate **Qwen2.5-VL-72B-Instruct** as a strong content-only reference. We test three prompting modes: direct zero-shot classification, chain-of-thought prompting, and ReAct prompting. The LLM receives only the raw post text and image, and does not access propagation graphs, dynamic statistics, or the offline evidence index used by DAR-Lite.

B.3 Implementation Details

All experiments are implemented using PyTorch and conducted on a Linux server equipped with four NVIDIA RTX 5090 GPUs.

B.3.1 Backbone Models. To ensure high-quality feature representations, we leverage pre-trained state-of-the-art models as backbones. For visual encoding, we employ the pre-trained CLIP-ViT-base-patch32 across all datasets due to its strong capability in aligning visual concepts with semantic textual representations. For text encoding, we adopt language-specific BERT models: BERT-base-uncased for English datasets (Fakeddit and FakeNewsNet) and BERT-base-chinese for the Chinese dataset (Weibo). All models are initialized using weights released on HuggingFace. In all experiments, BERT and CLIP act as frozen feature extractors and their embeddings are pre-computed offline; backbone weights are never updated during Phase II training.

B.3.2 Heterogeneous Graph Construction. We construct heterogeneous graphs to model interactions among news content, publishers, and social comments. For Fakeddit and Weibo, node features are initialized using the projected embeddings extracted from the text and visual encoders. For FakeNewsNet, to mitigate the issue of structural sparsity, we additionally employ Node2Vec to generate initial structural embeddings. These features are further refined by a 2-layer Graph Convolutional Network (GCN) with a dropout rate of 0.4 to capture high-order social propagation patterns. Graph statistics are summarized in Table 8.

B.3.3 Hyperparameter Settings. We carefully tune hyperparameters according to dataset characteristics. The latent feature dimension d is set to 128 for Fakeddit and Weibo, while for FakeNewsNet we increase it to 256 (`GLOBAL_EMBED_DIM` = 256) to better capture semantic nuances from long articles and sparse graph structures. For text input, we truncate the maximum sequence length to 512 tokens for Fakeddit to preserve detailed Reddit discussions, while setting it to 160 tokens for Weibo and FakeNewsNet. We adopt a batch size of 64 for Fakeddit and 48 for Weibo and FakeNewsNet, balancing GPU memory consumption and training stability. The evidence-depth hyperparameter K is selected from $\{1, 3, 5, 7, 9\}$ and the participant budget M from $\{3, 6, 9, 12\}$. For the auxiliary fallacy loss, we tune λ_1 over $\{0.1, 0.2, 0.3, 0.5\}$ and select $\lambda_1 = 0.3$ based on validation Macro-F1.

B.3.4 Training Strategy. We propose a two-stage training strategy. In Phase 1, we train the MVEB module on frozen encoder features to align multimodal representations into a shared latent space while suppressing redundant noise. In Phase 2, the reasoning layers, classification heads, and auxiliary objectives are jointly optimized with the multi-task loss, while the pre-trained CLIP and BERT backbones remain frozen. We use the Adam optimizer combined with a CosineAnnealingLR scheduler to ensure stable convergence. Thus, downstream optimization updates only the DAR-Lite modules on top of frozen backbone features, not the backbone encoders themselves. To address dataset-specific learning challenges, we adopt Cross-Entropy Loss for Fakeddit and FakeNewsNet. For Weibo, due to label imbalance and noisy short-post distributions, we employ Focal Loss with focusing parameter $\gamma = 2.0$ to emphasize hard-to-classify samples.

B.3.5 Evaluation Metrics. We report four standard evaluation metrics: Accuracy (Acc), Macro-F1 (Mac-F1), Area Under the ROC Curve (AUC), and Standardized Partial AUC ($spAUC_\sigma$).

Accuracy measures overall classification correctness, while Macro-F1 provides an unbiased evaluation under class imbalance by equally weighting each class. AUC reflects the ranking quality across all possible classification thresholds. Moreover, in real-world misinformation detection systems, false positives are often more detrimental than false negatives since mistakenly flagging legitimate news harms user trust and suppresses information dissemination. Therefore, we additionally report standardized partial AUC restricted to the False Positive Rate (FPR) range of $[0, 0.1]$, highlighting model reliability under strict false alarm constraints.