Dialectical Structured Reasoning for Explainable Multimodal Fake News Detection

Ruichao Yang[1†], **Yufan Bian**[1†], **Wei Gao**[2], **Bo-Wen Zhang**[1*], **Jing Ma**[3*], **Hongzhan Lin**[3], **Ziyang Luo**[3], **Xiaobin Zhu**[1], **Xu-Cheng Yin**[1]
[1] University of Science and Technology Beijing,
[2] Singapore Management University, [3] Hong Kong Baptist University
{yangruichao,bowenzhang}@ustb.edu.cn, M202510648@xs.ustb.edu.cn
weigao@smu.edu.sg, {majing,cshzlin,cszyluo}@comp.hkbu.edu.hk

Abstract

Current multimodal fake news detectors predominantly function as opaque classifiers, offering limited deductive transparency and little insight into how conflicting evidence is reconciled. To address this limitation, we propose Dialectical Structured Reasoning (DSR), a framework modeling fake news detection as an explicit dialectical process over multimodal social context. DSR instantiates two opposing agents: a *Verifier*, which constructs evidence paths supporting semantic consistency, and a *Debunker*, which actively explores exposing logical or factual contradictions. Then a differentiable *Judge* agent adjudicates between these competing perspectives by integrating local evidence with global parametric knowledge. Experiments on three benchmarks demonstrate that DSR achieves state-of-the-art performance while producing transparent, dialectically grounded explanations that closely mirror human reasoning process.[1]

1 Introduction

Fake news increasingly combines text and images and spreads through fragmented social interactions, making reliable veracity assessment difficult for both humans and automated systems. Recent viral claims, such as reports that "Asteroid 2024 YR4 could hit Earth in seven years"[2] accompanied by dramatic visuals, exemplify how multimodal presentation can lend false narratives a veneer of plausibility, underscoring the urgent need for robust and explainable detection models.

Most existing multimodal fake news detection models follow a feature-fusion paradigm (Wu et al., 2021; Jing et al., 2023; Zhou et al., 2023; Singhal et al., 2022). They encode textual and visual information using RNNs or transformers and subsequently fuse these representations to train an end-to-end classifier (Qi et al., 2021; Chen et al., 2022; Wu et al., 2023; Wang et al., 2024b; Zhang et al., 2025; Liu et al., 2025a; Tong et al., 2024). However, by compressing rich multimodal and social context into a single vector, these models largely overlook explicit notions of *evidence paths* and *reasoning process* for truth seeking, which are capabilities central to human cognition (Mercier and Sperber, 2017) and critical for accurate and explainable multimodal fake news detection.

Nevertheless, explicitly modeling evidence paths and the associated reasoning process is challenging, as social media discourse is inherently fragmented. Driven by confirmation bias (Nickerson, 1998) and selective exposure (Zollo et al., 2017), users often self-segregate into information cocoons (Zuiderveen Borgesius et al., 2016), leading to structurally polarized discussions. As illustrated in Figure 1, a fake news claim that "empty grocery store shelves reflect the effects of America's 'Build Back Better' plan" triggers divergent responses: some users reinforce the narrative with plausibly aligned details, while others attempt to debunk it by identifying inconsistencies, such as temporal mismatches in the visual content. Although latent semantic and social correlations exist across these disjoint threads, explicit interaction between supporting and refuting evidence remains rare. Consequently, relevant evidence is dispersed across loosely connected users, threads, and modalities, making it difficult for the model to even *observe* the complete evidentiary structure.

Given the fragmented and polarized nature of multimodal social discourse, we argue that effective fake news detection requires explicitly modeling both supportive and refuting evidence and reasoning over their global structure. This perspective naturally gives rise to three core chal-

[†]Equal contribution.
[*]Corresponding authors.
[1]Code is released at https://github.com/bianyufan/DSR.
[2]https://www.cnn.com/2025/02/15/science/asteroid-2024-yr4-earth-tracking

Findings of the Association for Computational Linguistics: ACL 2026, pages 31998–32013

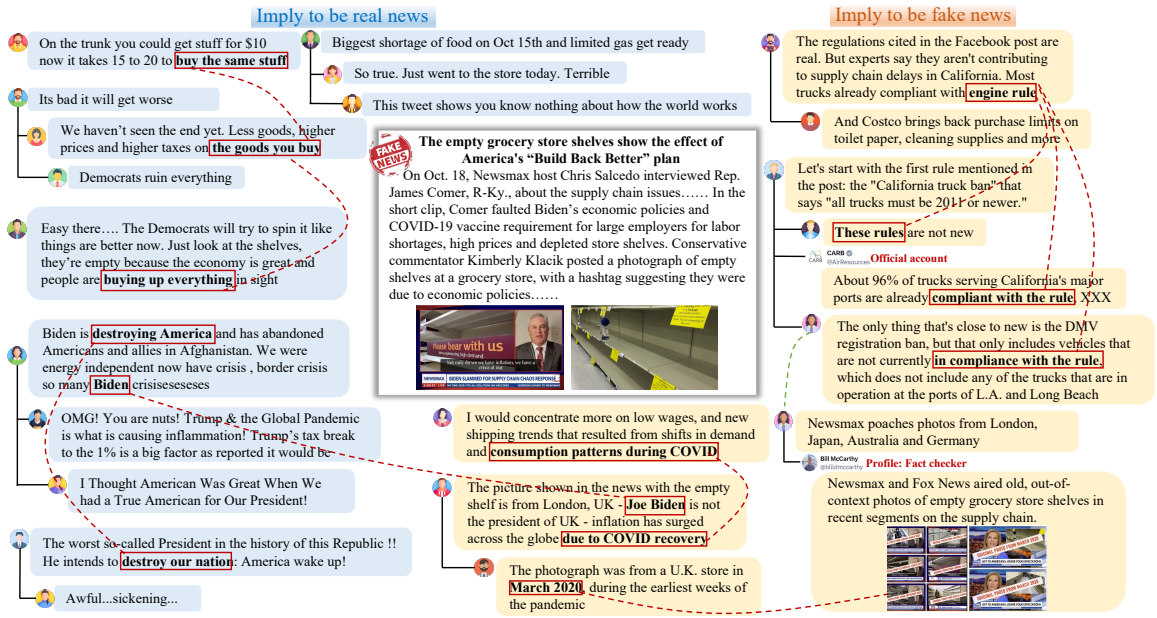

Figure 1: A motivating example of dialectical structured reasoning. Red dashed lines link posts across different threads that express semantically aligned claims or narratives, while green dashed lines denote distinct threads originating from the same user.

lenges: (1) reasoning over fragmented and heterogeneous modalities; (2) jointly exploring supportive evidence (thesis) and refuting clues (antithesis) to uncover subtle inconsistencies; and (3) synthesizing opposing perspectives for robust adjudication. While recent approaches (e.g., TED (Liu et al., 2025b)) take an initial step toward this goal by simulating pro-con debates, they primarily operate at the semantic level, and treat posts in isolation, and lack the ability to reason globally over heterogeneous evidence paths. Moreover, such debated-based approaches rely on non-differentiable, prompt-driven processes, which hinder end-to-end optimization, constrain representation learning, and incur substantial inference cost.

To address these challenges in a unified manner, we propose Dialectical Structured Reasoning (DSR), a framework in which MLLM-guided **Verifier** and **Debunker** agents perform hypothesis-driven topological traversal over multimodal propagation graphs to mine competing evidential paths, while a differentiable **Judge** agent adjudicates between these opposing chains to produce a comprehensive and explainable verdict. We summarize our contributions as follows: (1) We propose DSR, a new paradigm that formulates multimodal fake news detection as a dialectical adjudication over heterogeneous propagation graphs. (2) We design MLLM-guided Verifier and Debunker agents that perform hypothesis-driven topological traversal to mine compact and explicit supportive and refuting evidence across fragmented modalities. (3) We introduce a differentiable Judge agent that integrates opposing reasoning paths, enabling robust adjudication with intrinsic explainability. (4) We demonstrate state-of-the-art performance on three benchmark datasets while providing transparent and interpretable verdicts for complex multimodal fake news.

2 Related Work

Early fake news detection methods mainly focused on textual information or propagation structures (Ma et al., 2016; Bian et al., 2020; Lu and Li, 2020; Lin et al., 2026). With the rise of multimodal social media content, recent approaches jointly model text and images through attention mechanisms or bilinear pooling (Wang et al., 2018; Khattar et al., 2019; Wu et al., 2023; Zhang et al., 2024a; Zhu et al., 2025b; Su et al., 2025). Mechanism-oriented methods introduce intermediate reasoning factors (Huang et al., 2024; Verga et al., 2021; Pang et al., 2025), while explanation-centric studies generate post-hoc rationales (Yang et al., 2022; Zeng et al., 2023; Ying et al., 2023; Qi et al., 2024; Lu et al., 2025). However, these methods largely provide *descriptive* explanations and lack *deductive interpretability*. We address this gap by embedding dialectical reasoning di-

rectly into the detection process.

Recent MLLMs further enable knowledge-rich (Tahmasebi et al., 2024; Le et al., 2025; Li et al., 2025a), instruction-following (Tong et al., 2025), and debate-based (Liu et al., 2025b; Han et al., 2025) methods. Nonetheless, these methods typically rely on unstructured dialogue and their debate mechanisms are often non-differentiable. In contrast, we propose a framework in which Verifier and Debunker agents perform topological traversal over propagation graphs, while a *differentiable* Judge agent integrates opposing evidence paths for end-to-end adjudication.

3 Problem Formulation

Input. Given a news item x, we define its multimodal article as $\mathcal{A}_x = (t_x, b_x, I_x)$, where t_x denotes the news title, b_x the news body, and $I_x = \{i_1, \ldots, i_n\}$ the associated images. Let $X_x = \{p_1, \ldots, p_T\}$ denote a temporally ordered set of T social media posts with message propagation paths responding to $\mathcal{A}_x$.

Output. We aim to predict a label $y_x \in \{0, 1\}$ indicating whether the news item is real ($y_x = 0$) or fake ($y_x = 1$), along with human readable path-level explanations of the predicted verdict.

4 Methodology

In this section, we present **Dialectical Structured Reasoning (DSR)**, a unified framework that integrates the reasoning capabilities of Large Language Models (LLMs) with the structural modeling strength of graph neural networks. As illustrated in Figure 2, DSR consists of four stages: (1) hypothesis-driven dialectical topology traversal, (2) multimodal path encoding, (3) dialectical structured reasoning, and (4) dialectical structured prediction. A summary of the key notations can be found in Appendix A.

4.1 Multimodal Heterogeneous Propagation Graph Construction

For each news item x, we construct a multimodal heterogeneous propagation graph $\mathcal{G}_x = (\mathcal{V}_x, \mathcal{E}_x)$, where $\mathcal{V}_x$ and $\mathcal{E}_x$ denote the node and edge sets, respectively. Each node $v \in \mathcal{V}_x$ is associated with a type $\tau(v)$ and content $c(v)$. The graph contains the following node types: (i) a **news node** v_{news} storing (t_x, b_x); (ii) **post nodes** representing posts with textual content; (iii) **user nodes** associated with profile features (e.g., verification status, follower count); (iv) **image nodes** corresponding to images in I_x or attached to posts; (v) **topic nodes** extracted from textual content of the news article and posts. The heterogeneous graph construction details are provided in Appendix B.

4.2 Hypothesis-Driven Topology Traversal

To extract dialectical evidential paths, we perform hypothesis-driven traversal over $\mathcal{G}_x$. Each node v is first assigned a topological saliency score using PageRank (Brin and Page, 1998). Traversal is conditioned on a hypothesis $H \in \{H_{\text{real}}, H_{\text{fake}}\}$, where H_{real} assumes the news is true and H_{fake} assumes it is false. An MLLM (e.g., GPT-5 (OpenAI, 2025)) acts as a semantic navigator guiding a beam search (Vijayakumar et al., 2018) process (see Algorithm 1), and prompt templates are provided in Appendix C.1.

Semantic coherence is enforced at each expansion step by conditioning the MLLM navigator jointly on hypothesis H and the accumulated path history, pruning semantically inconsistent branches to ensure each extracted path forms a logically consistent narrative. Although both agents traverse $\mathcal{G}_x$ and their subgraphs thus share v_{news} and may intersect at high-centrality nodes, the resulting paths diverge semantically due to opposing optimization objectives; this *controlled structural overlap* is architecturally intentional, furnishing the Judge with a common referential anchor for principled adjudication over competing evidence chains.

This process yields hypothesis-centric reasoning subgraphs, from which Depth-First Search (DFS) is used to extract a set of candidate evidential paths under each hypothesis: $\mathcal{P}_{\text{real}} = \{P_1^{(r)}, \ldots, P_K^{(r)}\}$, $\mathcal{P}_{\text{fake}} = \{P_1^{(f)}, \ldots, P_K^{(f)}\}$.

4.3 Multimodal Path Encoding

Given a hypothesis-specific evidential path $P = (v_0, \ldots, v_{|P|})$, where $P \in \mathcal{P}_{\text{real}}$ or $\mathcal{P}_{\text{fake}}$ and $v_0 = v_{\text{news}}$, we aim to get a representation h_P that captures its multimodal content, social cues, and structural role in the propagation graph.

Node-level Encoding. For each node v appearing in a path P, we compute a modality-aware embedding based on its type and associated content:

$$m_v = \begin{cases} \text{Enc}_{\text{text}}(c(v)), & \tau(v) \in \{\text{news}, \text{post}, \text{topic}\}, \\ \text{Enc}_{\text{img}}(c(v)), & \tau(v) = \text{image}, \\ \text{Enc}_{\text{user}}(c(v)), & \tau(v) = \text{user}, \end{cases} \quad (1)$$

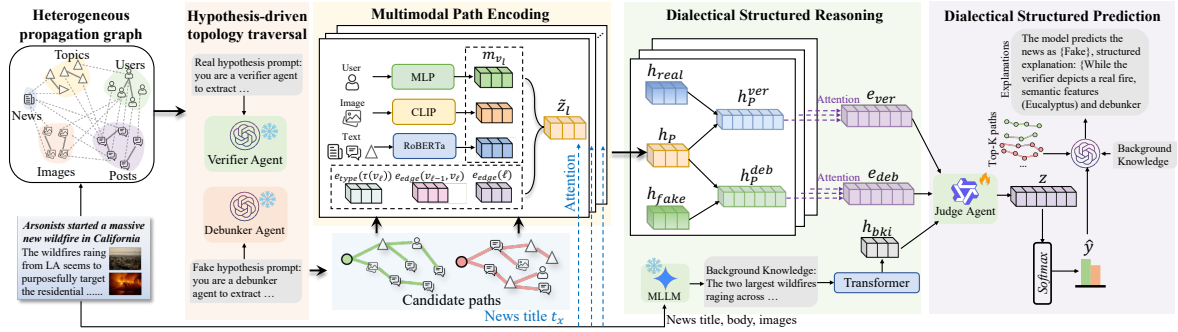

Figure 2: The overview of our DSR framework, which consists of hypothesis-driven dialectical topology traversal, multimodal path encoding, dialectical structured reasoning, and dialectical structured prediction.

Algorithm 1 MLLM-Guided Thought-on-Graph (ToG) Beam Search

Require: Graph $\mathcal{G}$, Claim $\mathcal{C}$, Max Depth L, Beam Width K
Ensure: Set of Evidence Paths $\mathcal{P}$

Function ToG_Search($\mathcal{G}, \mathcal{C}, H$) ▷ $H \in \{H_{\text{real}}, H_{\text{fake}}\}$
  Initialize $\mathcal{B} \leftarrow \{[\mathcal{C}]\}$
  for $t = 1$ **to** L **do**
    $\mathcal{B}_{\text{cand}} \leftarrow \emptyset$
    for each path $p \in \mathcal{B}$ **do**
      $v_{\text{curr}} \leftarrow p$.last_node()
      $\mathcal{N} \leftarrow$ Top_N($\mathcal{G}$.neighbors(v_{curr}), 40) ▷ PR filter
      $P \leftarrow$ Prompt($\mathcal{C}, p, \mathcal{N}, H$)
      $V_{\text{sel}} \leftarrow$ MLLM.query(P) ▷ Top-K nodes
      for each $v \in V_{\text{sel}}$ **do**
        $\mathcal{B}_{\text{cand}}$.add($p \oplus [v]$)
      end for
    end for
    $\mathcal{B} \leftarrow$ Select_Top_K($\mathcal{B}_{\text{cand}}, K$)
  end for
  return $\mathcal{B}$

Main Execution:
$\mathcal{P}_{\text{fake}} \leftarrow$ ToG_Search($\mathcal{G}, \mathcal{C}, H_{\text{fake}}$) ▷ Debunker Agent
$\mathcal{P}_{\text{real}} \leftarrow$ ToG_Search($\mathcal{G}, \mathcal{C}, H_{\text{real}}$) ▷ Verifier Agent
return $\mathcal{P}_{\text{fake}} \cup \mathcal{P}_{\text{real}}$

where $\tau(v)$ denotes the type of node v, and $c(v)$ denotes its associated content. Specifically, for news, post, and topic nodes, $c(v)$ corresponds to textual content; for image nodes, it corresponds to the raw image; and for user nodes, it corresponds to a profile features vector. Here, Enc_{text} is a pre-trained Transformer encoder (e.g., RoBERTa (Liu et al., 2019)) with mean pooling over token representations, Enc_{img} is a CLIP-style image encoder, and Enc_{user} is a Multi-Layer Perceptron (MLP). All modality-specific embeddings are linearly projected into a shared d-dimensional space.

Path Token Encoding. Each node v_ℓ ($0 \le \ell \le |P|$) along the path P is treated as a heterogeneous token. We construct its token embedding as:

$$z_\ell = m_{v_\ell} + e_{\text{type}}(\tau(v_\ell)) + e_{\text{edge}}(v_{\ell-1}, v_\ell) + e_{\text{pos}}(\ell), \quad (2)$$

where $m_{v_\ell} \in \mathbb{R}^d$ is the node-level embedding obtained from Equation 1, $e_{\text{type}} \in \mathbb{R}^d$ is a learnable embedding indicating the node type $e_{\text{edge}} \in \mathbb{R}^d$ encodes the type of the edge (e.g., *Social_follow*; see Appendix B), and $e_{\text{pos}}(\ell) \in \mathbb{R}^d$ is a positional embedding indicating the node's order along the path. The positional embedding can distinguish evidence appearing at different stages of the propagation, which is crucial for reasoning over information cascades where early and late responses often play different roles.

Path-level Encoding. We feed the token embedding sequence $\{z_\ell\}_{\ell=0}^{|P|}$ for path P into a Transformer encoder:

$$\{\tilde{z}_\ell\}_{\ell=0}^{|P|} = \text{Transformer}_\phi(z_0, z_1, \ldots, z_{|P|}), \quad (3)$$

where each z_ℓ is defined in Equation 2.

Rather than treating all nodes along the path equally, we apply an attention-based aggregation mechanism to emphasize diagnostically informative nodes. Specifically, attention is conditioned on the news title representation $s_x = \text{Enc}_{\text{text}}(t_x)$:

$$\alpha_\ell = \frac{\exp(s_x^\top \tilde{z}_\ell)}{\sum_{j=0}^{|P|} \exp(s_x^\top \tilde{z}_j)}, \quad h_P = \sum_{\ell=0}^{|P|} \alpha_\ell \tilde{z}_\ell, \quad (4)$$

where α_ℓ is the weight of node v_ℓ within the path, and h_P is the resulting path-level embedding.

4.4 Dialectical Structured Reasoning

DSR performs dialectical reasoning by contrasting how evidential path support the competing hypotheses, i.e., real or fake, and aggregating these signals under a global contextual prior.

Hypothesis-aware Path Encoding. We introduce two learnable hypothesis embeddings, h_{real} and $h_{\text{fake}} \in \mathbb{R}^d$, corresponding to the *news*

is true and the *news is fake* hypotheses, respectively. Given a path-level embedding h_P (Equation 4), we construct hypothesis-aware path representations as:

$$\begin{aligned} h_P^{\text{ver}} &= \text{concat}(h_P, h_{\text{real}}), \quad P \in \mathcal{P}_{\text{real}}, \\ h_P^{\text{deb}} &= \text{concat}(h_P, h_{\text{fake}}), \quad P \in \mathcal{P}_{\text{fake}}, \end{aligned} \tag{5}$$

where concat denotes vector concatenation.

Dialectical Path Aggregation. To aggregate path-level evidence under each hypothesis, we apply an attention mechanism to estimate the importance of individual paths:

$$\beta_P = \frac{\exp\left(w^\top h_P^{\text{ver}}\right)}{\sum_{P' \in \mathcal{P}_{\text{real}}} \exp\left(w^\top h_{P'}^{\text{ver}}\right)}, \quad e_{\text{ver}} = \sum_{P \in \mathcal{P}_{\text{real}}} \beta_P \, h_P^{\text{ver}}, \tag{6}$$

where w is a learnable vector, β_P denotes the importance of path P, and e_{ver} is the aggregated evidence representation supporting the real hypothesis. The debunking evidence representation e_{deb} is computed analogously over $\mathcal{P}_{\text{fake}}$.

Background Knowledge Injection. Graph-based evidence is inherently local and may be biased by echo chambers. To provide a global contextual prior, we introduce *Background Knowledge Injection* (BKI). Given the news title t_x, body text b_x, and associated images I_x, a *frozen* MLLM (Gemini3 (Google, 2025)) is prompted to generate a concise background description t_{bki} summarizing salient entities, publicly available facts, and common manipulation patterns (see Appendix C.2). This background description is encoded as:

$$h_{\text{bki}} = \text{Transformer}_\psi(t_{\text{bki}}), \tag{7}$$

which yields a contextual representation.

We adopt a frozen MLLM for BKI over online retrieval for three reasons: (i) DSR is self-contained, pairing parametric world knowledge with real-time social propagation signals without dependence on external API availability; (ii) for breaking news, corrective signals frequently surface in social media discourse *before* retrievable fact-checking reports emerge, enabling earlier intervention; and (iii) the BKI module is modularly replaceable with a retrieval-augmented generation (RAG) component without modifying the dialectical adjudication pipeline.

4.5 Dialectical Structured Prediction

Multimodal Fake News Detection. We fuse the aggregated verifier and debunker representations with the background contextual prior via a conflict-aware fusion module (implemented with Qwen2.5 (Yang et al., 2025)):

$$\begin{aligned} g_{\text{conf}} &= \sigma(W_c(e_{\text{ver}} \odot e_{\text{deb}}) + b_c), \\ z &= \text{Qwen}(\text{concat}(e_{\text{ver}}, e_{\text{deb}}, g_{\text{conf}}, h_{\text{bki}}); \theta), \end{aligned} \tag{8}$$

where $\odot$ denotes element-wise product, $\sigma(\cdot)$ is the sigmoid function, g_{conf} captures the conflict degree between supportive and refuting evidence, and θ denotes trainable parameters adapted via LoRA (Hu et al., 2021). Specifically, background text t_{bki} is encoded via Transformer ψ into h_{bki} (Eq. 7), residing in the same d-dimensional space as e_{ver} and e_{deb}; all four components are then concatenated and processed by the LoRA-fine-tuned Judge, with the conflict gate g_{conf} (Eq. 8) providing an explicit signal of evidential tension that governs how the Judge dynamically weighs local graph evidence against the global parametric prior.

Then, we apply a fully connected softmax layer to obtain the prediction:

$$\hat{y} = \text{Softmax}(W_o z + b_o), \tag{9}$$

where $\hat{y} \in [0, 1]$ denotes the prediction of the news, and W_o and b_o are trainable parameters.

Explanation Generation. DSR provides path-level explanations derived from its internal dialectical reasoning process. Given a news item x, verifier paths $\mathcal{P}_{\text{real}}$, and debunker paths $\mathcal{P}_{\text{fake}}$, we select the top-$K$ most informative paths based on attention weights:

$$\begin{aligned} \mathcal{P}_{\text{ver}}^\star &= \text{Top-}K_{\beta_P}\{\mathcal{P}_{\text{real}}\}, \\ \mathcal{P}_{\text{deb}}^\star &= \text{Top-}K_{\beta_P}\{\mathcal{P}_{\text{fake}}\}, \end{aligned} \tag{10}$$

where β_P is the path attention weight defined in Equation 6. An MLLM then synthesizes a concise explanation conditioned on $\mathcal{P}_{\text{ver}}^\star$, $\mathcal{P}_{\text{deb}}^\star$, the predicted label $\hat{y}$, and the contextual background t_{bki}. This explanation faithfully reflects the internal dialectical reasoning of DSR.

4.6 Model Training

We train the model using a combination of sample-level classification loss and path-level dialectical supervision. The primary objective is binary cross-entropy over the news labels:

$$\mathcal{L}_{\text{cls}} = -\sum_{n=1}^{N} \left(y_n \log \hat{y}_n + (1 - y_n) \log(1 - \hat{y}_n)\right), \tag{11}$$

where $y_n \in \{0, 1\}$ and $\hat{y}_n \in [0, 1]$ denote the ground-truth label and predicted probability for the n-th news item, respectively.

For verifier and debunker candidate paths $\mathcal{P}_{\text{real}}$ and $\mathcal{P}_{\text{fake}}$, we also obtain a soft prediction $\hat{y}_n^{\text{dia}}$ from an external MLLM, which serves as weak supervision at the path level. The corresponding loss is defined as:

$$\mathcal{L}_{\text{dia}} = -\sum_{n=1}^{N} \left(y_n^{\text{dia}} \log \hat{y}_n^{\text{dia}} + (1 - y_{\text{dia}}) \log(1 - \hat{y}_n^{\text{dia}}) \right), \tag{12}$$

where $y_n^{\text{dia}} = 1$ if $P \in \mathcal{P}_{\text{fake}}$ and $y_n^{\text{dia}} = 0$ if $P \in \mathcal{P}_{\text{real}}$. The final training objective is a weighted combination of the two losses:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{dia}}, \tag{13}$$

where λ is the trade-off coefficient.

5 Experiments and Results

5.1 Experimental Setup

Datasets. We evaluate **DSR** on three widely used benchmarks for multimodal fake news detection: **PolitiFact** and **GossipCop** (Shu et al., 2020), and **Fakeddit** (Nakamura et al., 2020)[3]. These datasets span diverse domains, including politics and entertainment, and contain a mixture of visual evidence and social context. Dataset statistics are reported in Appendix F.

Baselines. We compare **DSR** against 13 competitive baselines, grouped into four categories: **(1) MLLMs**, which includes **Llama 4** (Meta, 2025), **Gemini 3** (Google, 2025), and **GPT-5** (OpenAI, 2025); **(2) Unimodal Methods**, which perform fake news detection using a single modality (typically textual content) often augmented with structured or debate-style reasoning, including **L-Defense** (Wang et al., 2024a), **D2D** (Han et al., 2025), and **TED** (Liu et al., 2025b); **(3) Multimodal Methods**, which rely on visual-textual content fusion without incorporating social context, including **MMDFND** (Tong et al., 2024), **SDML** (Jing et al., 2023), and **MiMoE-FND** (Liu et al., 2025a); and **(4) Social Context Enhanced Methods**, which explicitly incorporate social interactions as additional evidence, including **FORM** (Li et al., 2024), **KEN** (Zhu et al., 2025a), **HML** (Li et al., 2025b), and **RPPG-Fake** (Zhang et al., 2024b). For all baselines, we adopt the hyperparameter settings reported in the original papers and further tune key parameters (e.g., learning rate and embedding dimensionality) on validation sets to ensure fair comparison. Implementation details are provided in Appendix E and Appendix G.

Evaluation metrics. Following prior work on fake news detection (Zhang et al., 2024b), we report Accuracy (Acc), macro F1-score (F1), Precision (Pre), and Recall (Rec) on the test set for each dataset.

5.2 Multimodal Fake News Detection Results

Table 1 reports the experiment results of DSR against four groups of baselines on the PolitiFact, GossipCop, and Fakeddit datasets. As shown in Group 1, MLLMs' zero-shot performance is suboptimal (e.g., GPT-5 achieves only 53.1% accuracy on PolitiFact), which suggests that generalist MLLMs fail to capture the subtle forensic nuances required for verification. Unimodal baselines in Group 2 generally perform worse than multimodal methods in Group 3. For instance, MIMoE-FND (Group 3) surpasses the text-centric method TED (Group 2) by 3.9% in accuracy on PolitiFact, which demonstrates that visual modalities provide complementary clues essential for debunking complex misinformation. Furthermore, the social context enhanced multimodal methods (Group 4) yields the most significant performance gains across all baselines. Methods like RPPG-Fake and HML achieve accuracies above 84% across datasets, indicating that the "wisdom of crowds" and propagation paths are critical indicators for veracity assessment. RPPG-Fake stands out as the strongest baseline by simulating propagation paths via reinforcement learning.

Our proposed **DSR** framework consistently outperforms all baselines. Specifically, for accuracy metric, DSR surpasses RPPG-Fake by margins of 2% / 2.2% / 5% on PolitiFact, GossipCop, and Fakeddit datasets, respectively. We attribute this improvement to the *dialectical structure reasoning* of our approach. While baselines like RPPG-Fake focus on generating propagation paths, they often treat evidence aggregation as a linear process. In contrast, DSR employs distinct *Verifier* and *Debunker* agents to explicitly mine conflicting evidence chains. This allows DSR to identify logical contradictions between the multimodal content and social feedback, leading to more robust decisions even in cases where propagation patterns are sparse or noisy.

[3]We retain samples with more than three associated posts.

Category	Method	PolitiFact				GossipCop				Fakeddit			
		Acc	F1	Pre	Rec	Acc	F1	Pre	Rec	Acc	F1	Pre	Rec
MLLMs	Llama 4	0.413	0.452	0.463	0.441	0.402	0.438	0.448	0.429	0.401	0.427	0.436	0.419
	Gemini 3	0.468	0.479	0.490	0.468	0.451	0.464	0.479	0.450	0.446	0.462	0.473	0.451
	GPT-5	0.531	0.538	0.548	0.529	0.519	0.527	0.544	0.511	0.503	0.519	0.510	0.528
Unimodal	L-Denfense	0.713	0.705	0.726	0.685	0.696	0.685	0.698	0.672	0.681	0.677	0.688	0.667
	D2D	0.732	0.723	0.734	0.713	0.719	0.711	0.723	0.699	0.705	0.698	0.709	0.688
	TED	0.785	0.779	0.787	0.772	0.761	0.743	0.756	0.731	0.758	0.753	0.764	0.743
Multimodal	MMDFND	0.788	0.791	0.805	0.777	0.763	0.765	0.787	0.744	0.762	0.774	0.787	0.762
	SDML	0.806	0.814	0.833	0.796	0.787	0.799	0.815	0.783	0.763	0.781	0.796	0.766
	MIMoE-FND	0.824	0.835	0.849	0.821	0.810	0.826	0.836	0.817	0.797	0.806	0.823	0.789
Social Context Enhanced Multimodal	FORM	0.829	0.836	0.851	0.822	0.814	0.820	0.834	0.806	0.804	0.815	0.825	0.806
	KEN	0.837	0.842	0.852	0.833	0.831	0.838	0.848	0.828	0.819	0.823	0.829	0.817
	HML	0.841	0.853	0.868	0.839	0.826	0.842	0.850	0.835	0.820	0.828	0.841	0.816
	RPPG-Fake	0.873	0.892	0.915	0.871	0.859	0.878	0.891	0.866	0.834	0.845	0.856	**0.834**
Ours	**DSR**	**0.893**	**0.904**	**0.936**	**0.872**	**0.881**	**0.892**	**0.902**	**0.882**	**0.884**	**0.874**	**0.939**	0.817

Table 1: Multimodal fake news detection results on PolitiFact, GossipCop, and Fakeddit datasets.

Method	I	S	R	M	D
TED	3.06	3.14	3.46	3.22	3.09
D2D	3.73	3.58	4.07	2.53	2.51
L-Defense	4.01	4.12	4.29	2.10	2.02
DSR	**4.36**	**4.45**	**4.58**	**1.75**	**1.67**

Table 2: Explanation evaluation using GPT-5 scoring (1-5 scale). **I**: Informativeness, **S**: Soundness, **R**: Readability, **M**: Misleadingness, **D**: Discrepancy. For I, S, R, higher is better; for M and D, lower is better.

5.3 Evaluations on Explanations

A robust forensic framework must not only detect fake news but also provide human-interpretable justifications. Following the evaluation protocols established by Wang et al. (2024a), we conduct a comparative assessment of the explanations generated by DSR against existing baselines on **Fakeddit** dataset. It is important to note that most existing multimodal and social-enhanced methods (e.g., Group 3 and Group 4 in Table 1) are designed as purely *discriminative* frameworks. They utilize binary classification heads to predict veracity labels without generating human-readable justifications. To establish a meaningful benchmark for interpretability, we compare DSR against leading **unimodal** approaches specifically tailored for the *Explainable Fake News Detection (EFND)* task: **L-Defense** (Wang et al., 2024a), **TED** (Liu et al., 2025b), and **D2D** (Han et al., 2025).

Evaluation Metrics. We employ **GPT-5** as an impartial judge to evaluate the quality of generated explanations. Consistent with the methodology in L-Defense (Wang et al., 2024a), we utilize four metrics rated on a 5-point Likert scale (Joshi et al., 2015): (i) **Informativeness (I)**: whether the explanation contributes evidential details beyond the existing materials; (ii) **Soundness (S)**: logical coherence and plausibility of the explanation; (iii) **Readability (R)**: fluency and structural clarity; (iv) **Misleadingness (M)**: whether the explanation is consistent with the ground-truth label (1 = not misleading, 5 = highly misleading). In addition, following Wang et al. (2024a), we also report (v) **Discrepancy (D)** score as a label-level counterpart of misleadingness between the predicted and ground-truth labels (lower is better). The evaluator GPT-5 is prompted with the news content, the model's predicted label along with its explanation, and then returns (M, I, S, R) scores. The **Discrepancy (D)** is computed directly from predicted and ground-truth labels and does not depend on the explanation text.

Results Analysis. As shown in Table 2, L-Defense yields stronger explanations than TED and D2D, achieving higher informativeness, soundness, and readability while reducing both misleadingness and discrepancy. However, DSR further pushes the frontier. It improves the I/S score by 0.35/0.33 points over L-Defense, and attains the highest readability score (4.58). Meanwhile, DSR produces *less misleading* explanations, with M/D score reduced by 0.35/0.35 compared to L-Defense, indicating that its rationales are more tightly aligned with the ground-truth veracity labels. This is because TED and D2D often rely on generic heuristics (e.g., "the source seems unreliable"), whereas DSR tends to explicitly contrast supportive and refutational subtle clues and to ground its justifications in concrete evidence paths (e.g., user roles or cross-referenced posts). These results support our claim that dialectical, path-based reasoning on the heterogeneous prop-

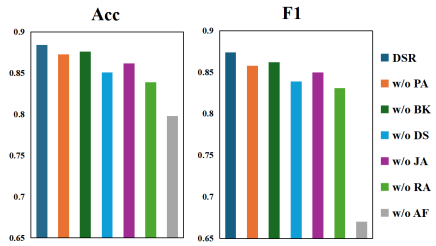

Figure 3: Ablation study of DSR on Fakeddit dataset.

agation graph not only enhances detection performance, but also yields explanations that are more informative and faithful than those existing unimodal frameworks.

5.4 Ablation Study

To get a comprehensive understanding of Dialectical Structured Reasoning, we conduct an ablation study on Fakeddit dataset to show the contribution of each component of DSR. The variants are defined as: **(1) w/o PA:** remove the path-based attention mechanism, then employ simple average pooling to combine the path embeddings. **(2) w/o BK:** remove the Background Knowledge Injection (BKI) module. **(3) w/o DS:** remove the Verifier and Debunker agents, the propagation paths are sampled randomly from the heterogeneous graph, and the Judge agent does not distinguish verify- and debunk-paths. **(4) w/o JA:** remove the Judge Agent and use a Multi-Layer Perceptron (MLP) classifier. **(5) w/o RA:** remove all reasoning agents, replace Verifier and Debunker agents with random path sampling, and replace the Judge Agent with MLP. **(6) w/o AF:** remove the entire agentic framework, including the specialized roles of Verifier, Debunker, and Judge, using a single vanilla Qwen2.5 model for direct classification on the propagation graph.

As shown in Figure 3, disabling any component consistently degrades detection performance, suggesting that each component is essential to the model's overall effectiveness. Specifically, the **w/o PA** variant highlights the contribution of the path-based attention aggregation mechanism. The performance drop observed in **w/o BK** indicates that global epistemic context is indispensable for accurate verification. Moreover, **w/o DS** suffers a severe performance drop in terms of accuracy and F1 score, which demonstrates that dialectical structure is crucial for revealing deep logical contradictions. The performance of the **w/o JA** variant also decreases, validating that a judge agent equipped with an epistemic background significantly enhances the model's predictive capability. Furthermore, the substantial performance degradation in **w/o RA** confirms that the synergy between dialectical agents and a judge agent is fundamental to the framework's success. Most notably, the **w/o AF** variant exhibits a drastic decline across all metrics. This underscores a pivotal conclusion: simply feeding heterogeneous graph data into a powerful LLM without our multi-agent dialectical workflow fails to effectively resolve complex misinformation, proving that the explicit agentic framework is the core driver of DSR's superiority.

5.5 Case Study

To intuitively exhibit the dialectical reasoning process of DSR, we select a true news sample from PolitiFact to visualize the inference process. As Figure 4 shows, the verifier (blue) confirmed the physical event mainly through visual evidence. The debunker (red) focuses on the widely circulating "arson" conspiracy theory in the posts. The Judge adopts path-based attention to assign higher weights to Verifier Agent's paths, then uses BKI to introduce official investigation results, correctly filtering out the noise from social media posts and classifying the news as true news.

Furthermore, to demonstrate the critical role of the **BKI** module in correcting potential misjudgments, we analyze a challenging false news sample from the PolitiFact dataset. Figure 5 shows that Verifier Agent falls into a "Visual Entailment" trap in high confidence since the image is not photoshopped. Debunker Agent mines the social graph but encounters high-variance noise, so the debunking signal is weak and drowned out by emotional outrage. The BKI queries world knowledge regarding the image's history and distinguishes that the image was taken in June 2014. Armed with this specific timestamp, the Judge recognizes that the image predates the current administration's policies by several years. It overrides the Verifier's clues, correctly identifying the news as false.

5.6 User Study

To evaluate the quality of the DSR output, specifically in terms of *interpretability* and *decision confidence*, we conducted a user study comparing DSR against **RPPG-Fake**, a state-of-the-art social context-enhanced baseline. We randomly select 90 samples from the Fakeddit dataset and present

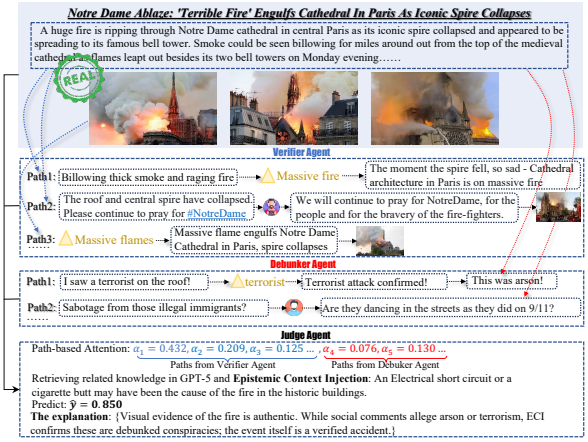

Figure 4: A real news case illustrating the dialectical structured reasoning process. The yellow triangles represent topics and the avatars represent users in the path.

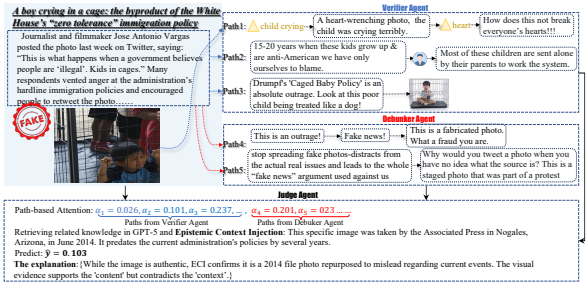

Figure 5: A fake news case illustrating the dialectical structured reasoning process and the importance of the BKI module. The yellow triangles represent topics and the avatars represent users in the path.

them in the following two settings: (1) **RPPG-Fake:** News content, images, and the propagation path generated by RPPG-Fake, which displays the hierarchical structure of user reposts. (2) **DSR:** News content, images, and DSR's dialectical reasoning output, which includes the visualized Verifying/Debunking Paths and the final Natural Language Explanation generated by the Judge Agent.

We recruited 6 participants with diverse backgrounds to determine the veracity of the news and give their confidence (Conf) in a 5-point Likert Scale (Joshi et al., 2015)[4]. Besides, we also ask people to note how helpful the model's output was in understanding the reasoning behind the verdict (IntRate, 1-5 scale, 5 denotes the most helpful).

Table 3 shows that 1) users determine the news veracity more accurately with DSR; 2) users show higher confidence with the results of DSR, suggesting that users tend to be more sure about their decision when verifying/debunking paths and explanations are provided; and 3) the dialectically structured reasoning paths and detailed explanations are more informative for users' decision-making.

[4]Each person is given only one setting to avoid cross influence.

	F1	Acc	Conf	IntRate
RPPG-Fake	0.796	0.805	2.816	2.305
DSR	0.953	0.967	3.781	4.443

Table 3: User study results on model outputs quality.

6 Conclusion

We propose Dialectical Structured Reasoning (DSR) to formulate multimodal fake news detection as dialectical reasoning over heterogeneous propagation graphs. DSR leverages MLLM-guided, hypothesis-driven topology traversal to first mine candidate verifying and debunking paths. It then employs a Judge agent to resolve the conflicts between supportive and refuting signals through Background Knowledge Injection (BKI) and path-based attention aggregation for producing a final decision. Extensive experiments on three benchmark datasets demonstrate that DSR consistently outperforms strong unimodal, multimodal, and social-context-enhanced baselines, indicating that structured dialectical justification leads to more robust and explainable detection of multimodal fake news. In future work, we plan to extend DSR toward finer-grained manipulation type prediction and explore settings where dialectical agents interact with human fact-checkers in the loop.

Acknowledgments

This research is supported by the National Science Fund for Distinguished Young Scholars (No. 62125601), National Natural Science Foundation of China (No. 62576031), Beijing Natural Science Foundation (No. 4264125), National Natural Science Foundation of China Young Scientists Fund (No. 62206233), and National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-RP-2024-035).

Limitations

While DSR demonstrates promising results in interpretable multimodal fake news detection, we acknowledge several limitations inherent to the current framework: (1) DSR orchestrates external MLLMs for path proposal and background

knowledge. Compared to end-to-end discriminative baselines, this architecture incurs extra computational overhead while significantly enhancing explainability. Future work may explore model distillation or token-efficient prompting strategies to mitigate this cost. (2) Our Verifier agent primarily focuses on *semantic* and *structure* (e.g., detecting if an image logically matches the text/context), and is less specialized in detecting pixel-level manipulations. Therefore, we can integrate low-level signal processing modules to handle high-fidelity neural fabrications robustly in the future. (3) The traversal and BKI modules employ proprietary MLLM APIs (GPT-5, Gemini-3) in our primary experiments; however, DSR's core pipeline remains fully functional with locally hosted open-weight models (e.g., Qwen2-VL-7B), though performance may vary across model families. To ensure reproducibility, we fix temperature at $T{=}0$ and specify exact model snapshots for all evaluations. Prompt sensitivity analysis (Appendix I) further confirms accuracy variance below 0.8% across semantically equivalent prompt variants, indicating robustness to surface-level prompt engineering. (4) DSR incurs additional preprocessing latency (9-14s per instance) due to MLLM-guided traversal and BKI generation, yet this is a *one-time offline* cost whose outputs are cacheable, and is substantially lower than RPPG-Fake ($\approx$25s). Real-time online inference via the LoRA-optimized Judge requires sub-0.1s per instance, matching standard classifiers. Future work may explore knowledge distillation to derive lightweight surrogate navigators.

References

Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 549–556.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117.

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.

Google. 2025. Build with our next generation ai systems.

Chen Han, Wenzhen Zheng, and Xijin Tang. 2025. Debate-to-detect: Reformulating misinformation detection as a real-world debate with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15125–15140, Suzhou, China. Association for Computational Linguistics.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Runsheng Huang, Liam Dugan, Yue Yang, and Chris Callison-Burch. 2024. MiRAGeNews: Multimodal realistic AI-generated news detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16436–16448, Miami, Florida, USA. Association for Computational Linguistics.

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

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396.

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*, pages 2915–2921.

Huy Hoan Le, Van Sy Thinh Nguyen, Thi Le Chi Dang, Vo Thanh Khang Nguyen, Truong Thanh Hung Nguyen, and Hung Cao. 2025. Multimedia verification through multi-agent deep research multimodal large language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 14034–14040.

Hui Li, Ante Wang, Kunquan Li, Zhihao Wang, Liang Zhang, Delai Qiu, Qingsong Liu, and Jinsong Su. 2025a. A multi-agent framework with automated decision rule optimization for cross-domain misinformation detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5720–5736.

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

Mingxin Li, Yuchen Zhang, Haowei Xu, Xianghua Li, Chao Gao, and Zhen Wang. 2025b. Learning complex heterogeneous multimodal fake news via social latent network inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 433–441.

Hongzhan Lin, Zixin Chen, Zhiqi Shen, Ziyang Luo, Zhen Ye, Jing Ma, Tat-Seng Chua, and Guandong Xu. 2026. Towards comprehensive stage-wise benchmarking of large language models in fact-checking. *arXiv preprint arXiv:2601.02669*.

Yifan Liu, Yaokun Liu, Zelin Li, Ruichen Yao, Yang Zhang, and Dong Wang. 2025a. Modality interactive mixture-of-experts for fake news detection. In *Proceedings of the ACM on Web Conference 2025*, pages 5139–5150.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yuhan Liu, Yuxuan Liu, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2025b. The truth becomes clearer through debate! multi-agent systems with large language models unmask fake news. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 504–514.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *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.

Yi-Ju Lu and Cheng-Te Li. 2020. Gcan: Graph-aware co-attention networks for explainable fake news detection on social media. In *ACL*.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *IJCAI*.

Hugo Mercier and Dan Sperber. 2017. *The enigma of reason*. Harvard University Press.

Meta. 2025. Llama 4: Leading intelligence. unrivaled speed and efficiency.

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*, pages 6149–6157.

Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220.

OpenAI. 2025. Gpt-5 system card.

Hui Pang, Chaozhuo Li, Litian Zhang, Senzhang Wang, and Xi Zhang. 2025. Beyond text: Fine-grained multi-modal fact verification with hypergraph transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6389–6397.

Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. 2021. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. In *Proceedings of the 29th ACM international conference on multimedia*, pages 1212–1220.

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.

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):171–188.

Shivangi Singhal, Tanisha Pandey, Saksham Mrig, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. 2022. Leveraging intra and inter modality relationship for multimodal fake news detection. In *Companion proceedings of the Web conference 2022*, pages 726–734.

Xinqi Su, Zitong Yu, Yawen Cui, Ajian Liu, Xun Lin, Yuhao Wang, Haochen Liang, Wenhui Li, Li Shen, and Xiaochun Cao. 2025. Dynamic analysis and adaptive discriminator for fake news detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 8164–8173.

Sahar Tahmasebi, Eric Müller-Budack, and Ralph Ewerth. 2024. Multimodal misinformation detection using large vision-language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2189–2199.

Yu Tong, Weihai Lu, Xiaoxi Cui, Yifan Mao, and Zhejun Zhao. 2025. Dapt: Domain-aware prompt-tuning for multimodal fake news detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 7902–7911.

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.

Pat Verga, Haitian Sun, Livio Baldini Soares, and William Cohen. 2021. Adaptable and interpretable

neural MemoryOver symbolic knowledge. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3678–3691, Online. Association for Computational Linguistics.

Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. Diverse beam search for improved description of complex scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Bo Wang, Jing Ma, Hongzhan Lin, Zhiwei Yang, Ruichao Yang, Yuan Tian, and Yi Chang. 2024a. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM Web Conference 2024*, pages 2452–2463.

Jiandong Wang, Hongguang Zhang, Chun Liu, and Xiongjun Yang. 2024b. Fake news detection via multi-scale semantic alignment and cross-modal attention. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 2406–2410.

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*, pages 849–857.

Lianwei Wu, Pusheng Liu, Yongqiang Zhao, Peng Wang, and Yangning Zhang. 2023. Human cognition-based consistency inference networks for multi-modal fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, 36(1):211–225.

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*, pages 2560–2569.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Zhiwei Yang, Jing Ma, Hechang Chen, Hongzhan Lin, Ziyang Luo, and Yi Chang. 2022. A coarse-to-fine cascaded evidence-distillation neural network for explainable fake news detection. In *Proceedings of the 29th international conference on computational linguistics*, pages 2608–2621.

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*, volume 37, pages 5384–5392.

Zhi Zeng, Mingmin Wu, Guodong Li, Xiang Li, Zhongqiang Huang, and Ying Sha. 2023. An Explainable Multi-view Semantic Fusion Model for Multimodal Fake News Detection . In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1235–1240, Los Alamitos, CA, USA. IEEE Computer Society.

Fanrui Zhang, Jiawei Liu, Jingyi Xie, Qiang Zhang, Yongchao Xu, and Zheng-Jun Zha. 2024a. Escnet: Entity-enhanced and stance checking network for multi-modal fact-checking. In *Proceedings of the ACM Web Conference 2024*, pages 2429–2440.

Litian Zhang, Xiaoming Zhang, Ziyi Zhou, Xi Zhang, Senzhang Wang, Philip S Yu, and Chaozhuo Li. 2024b. Early detection of multimodal fake news via reinforced propagation path generation. *IEEE Transactions on Knowledge and Data Engineering*.

Tianlin Zhang, En Yu, Yi Shao, and Jiande Sun. 2025. Multimodal inverse attention network with intrinsic discriminant feature exploitation for fake news detection. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 7940–7948. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. 2023. Multimodal fake news detection via clip-guided learning. In *2023 IEEE international conference on multimedia and expo (ICME)*, pages 2825–2830. IEEE.

Peican Zhu, Yubo Jing, Le Cheng, Keke Tang, and Yangming Guo. 2025a. Ken: Knowledge augmentation and emotion guidance network for multimodal fake news detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 1793–1801.

Ye Zhu, Yunan Wang, and Zitong Yu. 2025b. Multimodal fake news detection: Mfnd dataset and shallow-deep multitask learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 8012–8020. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Fabiana Zollo, Alessandro Bessi, Michela Del Vicario, Antonio Scala, Guido Caldarelli, Louis Shekhtman, Shlomo Havlin, and Walter Quattrociocchi. 2017. Debunking in a world of tribes. *PloS one*, 12(7):e0181821.

Frederik J Zuiderveen Borgesius, Damian Trilling, Judith Möller, Balázs Bodó, Claes H De Vreese, and Natali Helberger. 2016. Should we worry about filter bubbles? *Internet policy review*, 5(1).

A Notation Summary

Table 4 provides a summary of the key notations used in this paper.

Symbol	Description
x	Instance
$\mathcal{A}_x = (t_x, b_x, I_x)$	News article (title, body, images)
t_x	News title
b_x	News body
$I_x = \{i_1, \ldots, i_n\}$	Associated images
$X_x = \{p_1, \ldots, p_T\}$	Social posts responding to x
$\mathcal{G}_x = (\mathcal{V}_x, \mathcal{E}_x)$	Propagation graph
$v \in \mathcal{V}_x$	Graph node
$\tau(v)$	Node type
$c(v)$	Node content
v_{news}	Root news node
$P = (v_0, \ldots, v_{\lvert P \rvert})$	Evidential path
v_ℓ	Node at path position ℓ
$\mathcal{P}_{\text{real}}$	Supportive (verifier) paths
$\mathcal{P}_{\text{fake}}$	Refuting (debunker) paths
m_v	Node embedding
z_ℓ	Path token embedding
h_P	Path representation
s_x	Title representation
$H_{\text{real}}, H_{\text{fake}}$	Real / fake hypotheses
$h_{\text{real}}, h_{\text{fake}}$	Hypothesis embeddings
$h_P^{\text{ver}}, h_P^{\text{deb}}$	Hypothesis-aware path repr.
β_P	Path attention weight
$e_{\text{ver}}, e_{\text{deb}}$	Aggregated evidence repr.
t_{bki}	Background text
h_{bki}	Contextual prior embedding
g_{conf}	Evidence conflict signal
$\hat{y}$	Predicted fake probability
y	Ground-truth label
$\mathcal{L}_{\text{cls}}$	Classification loss
$\mathcal{L}_{\text{dia}}$	Path-level loss
λ	Loss weight

Table 4: Summary of notation used in the paper.

B Heterogeneous Propagation Graph Construction

Topic nodes are constructed using Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to identify latent topics summarizing dominant viewpoints in the discourse. These nodes act as semantic bridges between the news article and user-generated content.

We represent each news item x as a multimodal heterogeneous graph $\mathcal{G}_x = (\mathcal{V}_x, \mathcal{E}_x)$. Their internal connections are conducted by different relationships. *Social_follow* (user→user), *Reply/Repost* (posts→posts), *Co-occurrence* (keywords↔topic in the same topic, image↔image in a news/posts). As for the way the five nodes are connected to each other. News is connected with the posts, users, topics, and images through edges of mention. *Mention_by* (news→post, news→users, news→topic, news→images). The posts are connected with users, topics, and images by authorship and mention relationship. *Authroship*(post→user), *mention* (post↔topic, post↔image). And users are also connected with topics and images through mention type edges, *mention* (user→topic, user→image). This design allows DSR to jointly model textual, visual, and social signals within a unified heterogeneous graph.

C Prompt Template

C.1 Hypothesis-driven Topology Traversal Prompt

When conditioned on H_{real}, the model acts as a *Verifier*, prioritizing supportive evidence; when conditioned on H_{fake}, it acts as a *Debunker*, prioritizing refuting evidence. The following is a Verifier agent prompt template:

You are a Verifier agent operating under the real hypothesis: the target news is true.

You will be given the following inputs: - News headline and body: <News_text> - Current reasoning note: <Note_so_far> - Current node and its neighbors along with salient score: <Current_node>, <Neighbor1, score1>, <Neighbor2, score2>, ...

Your goal is to identify the most promising next-hop nodes in the propagation graph that could provide supporting evidence for the real hypothesis with beam search. Prioritize nodes and neighbor types that: - increase factual credibility (e.g., verified users, authoritative sources, fact-consistent reports), - reduce uncertainty about time, place, or key entities, - maintain semantic consistency with the news content.

You need to output: 1. Your selected next-hop nodes. 2. The reason why you chose these nodes. 3. Decide whether to CONTINUE exploring or STOP extending this path, given the current evidence.

Do NOT decide whether the news is ultimately true or fake; only guide where to explore under the real hypothesis.

The prompt template for the Debunker agent is structurally isomorphic to that of the Verifier, necessitating only a logical inversion from "real" to "fake".

C.2 Background Knowledge Retrieve Prompt

The following is a prompt template for background knowledge retrieve:

You are a fact-checking assistant. Based on the news title, text, and image provided, please retrieve the background knowledge of this event. Do not judge True/Fake yet. Just describe the relevant facts, notable entities involved, and the typical context of such claims. Answer concisely within 100 words.

D MLLM I/O and Graph Traversal Interface

To clarify how the MLLM interacts with the symbolic graph traversal (e.g., Beam Search), we define a strict **"Menu-Selection" interface protocol**, which bridges the semantic reasoning of the MLLM and the topological constraints of the graph:

Input Format (Graph $\rightarrow$ MLLM): Instead of feeding the entire graph, the interface serializes the local subgraph into a structured prompt. At traversal step t, the input comprises: (i) the multimodal root claim (text + images); (ii) the current path history; and (iii) a **Candidate Menu** constructed by the traversal algorithm. The menu lists top-N valid next-hop neighbors filtered by PageRank (e.g., `- [ID: post_123] (PR: 0.05, Trust: 0.3) 'PETA are a bunch of... '`).

Output Format (MLLM Reasoning): The MLLM acts as a semantic navigator. It is constrained via system prompts to output strictly formatted JSON objects without conversational padding. The output specifies the selected nodes to expand the beam, e.g., `{"paths": ["post_123", "post_456"], ["image_0"]}`.

Interface Validation & Grounding (MLLM $\rightarrow$ Graph): The core interface mechanism lies in **ID-based routing**. Once the JSON is parsed, the graph traversal algorithm intercepts the selected IDs and performs a rigorous grounding check:

- **Verification:** It verifies if the generated IDs perfectly match the candidates provided in step t's menu.
- **State Update:** If valid, the graph algorithm physically appends these nodes to the current beam paths $\mathcal{B}$ and retrieves the next-hop neighbors for step $t+1$.
- **Anti-Hallucination:** If the MLLM hallucinates an invalid ID not present in the local topology, the interface rejects it, enforcing structural fidelity.

Through this tightly coupled interface, DSR seamlessly marries the MLLM's open-ended reasoning capabilities with the rigid, non-differentiable topological rules of social propagation graphs.

E Baseline Models

We provide brief descriptions of all baseline models evaluated in this work, organized according to the four categories used in Section 5.1.

(1) MLLMs.

- **Llama 4**: A multimodal large language model released by Meta in 2025 as part of the Llama series, extending the original Llama models with native multimodal capabilities.
- **Gemini 3**: A multimodal foundation model developed and released by Google DeepMind in 2025, following earlier versions of the Gemini family.
- **GPT-5**: A multimodal large language model released by OpenAI in 2025, representing the latest generation of the GPT series.

(2) Unimodal Methods.

- **L-Defense** (Wang et al., 2024a): A text-based fake news detection method that integrates explainability mechanisms into the verification process to identify misleading content through structured reasoning over unimodal textual evidence.
- **D2D** (Han et al., 2025): A debate-driven framework that casts fake news detection as a structured multi-agent discussion, where opposing agents articulate arguments and counterarguments to improve decision transparency.
- **TED** (Liu et al., 2025b): A truth-evolving debate model using iterative rounds of argumentation over unimodal evidence to refine credibility judgments and better resolve factual disputes.

(3) Multimodal Methods.

- **MMDFND** (Tong et al., 2024): A multimodal fake news detector that fuses text, image, and auxiliary semantic cues through unified representation learning, aiming to robustly leverage complementary signals from each modality.
- **SDML** (Jing et al., 2023): A semantic-driven multimodal learning model that enforces cross-modal semantic consistency during feature fusion to enhance discrimination between real and fake news.
- **MiMoE-FND** (Liu et al., 2025a): A mixture-of-experts architecture that dynamically directs modality-specific features to specialized expert modules, adapting fusion pathways based on interaction patterns across modalities.

(4) Social Context Enhanced Methods.

- **FORM** (Li et al., 2024): A social-aware fake news detection model (Focusing On Relevant social evidence and Multimedia) that integrates propagation dynamics and user interaction signals to enhance verification accuracy.
- **KEN** (Zhu et al., 2025a): A Knowledge-Enhanced Network that incorporates social engagement features and content semantics to improve multimodal fact-checking robustness.
- **HML** (Li et al., 2025b): A heterogeneous multimodal learning framework with social latent network inference, which builds latent social graphs to capture relationships between diverse news attributes and dissemination patterns.
- **RPPG-Fake** (Zhang et al., 2024b): A framework for early fake news detection that models temporal patterns of repost propagation graphs to capture deceptive diffusion behaviors and enable timely interventions.

F Datasets Statistics

Table 5 shows the statistics of PolitiFact, GossipCop, and Fakeddit datasets.

	Stat.	PolitiFact	GossipCop	Fakeddit
Train	# True	168	1,584	7,391
	#Fake	236	1,502	7,391
Test	# True	77	675	3,169
	#Fake	115	630	3,169
Total	–	596	4,391	21,120
#Avg. Paths	# True	2.54	2.81	3.09
	#Fake	2.63	2.82	3.12
#Avg. Path Len	# True	2.28	2.30	2.78
	#Fake	2.38	2.40	2.78

Table 5: Statistics of the datasets used.

G Implementation Details

Encoding and Dimensions: For our DSR (Dialectical Structured Reasoning) framework, all textual nodes (including news titles, contents, and social media posts) are encoded using a pretrained RoBERTa-base model ($d = 768$). Visual nodes are initialized with image embeddings from CLIP-ViT-B/32, which are subsequently projected into a 768-dimensional space via a linear layer to align with textual features. The internal knowledge embeddings are pre-cached as 3584-dimensional vectors, corresponding to the hidden size of the LLM.

Model Configuration: The Judge Agent is built upon the Qwen2.5-7B-Instruct large language model. To optimize computational efficiency, we load the model with 4-bit NormalFloat (NF4) quantization and apply LoRA (Low-Rank Adaptation). The LoRA rank is set to $r = 16$ with a scaling factor $\alpha = 32$. We target all linear projections in the transformer layers for adaptation, including query, key, value, output, and MLP gate/up/down projections. The path-based attention mechanism utilizes 8 attention heads to aggregate dialectical evidence.

Hyperparameters and Optimization: We set the maximum path length $T_{max} = 5$ and the maximum number of evidence paths $K = 5$ per agent. The models are trained for a maximum of 100 epochs, with an Early Stopping mechanism (patience of 5 epochs) monitored by the validation loss. We use a fixed random seed of 3407 for all experiments.

The models are optimized using the AdamW (Loshchilov and Hutter) optimizer with a weight decay of 0.01. We implement a two-tier learning rate strategy: the learning rate for LoRA adapters is $lr_{lora} = 2 \times 10^{-5}$, and the classifier heads and projection layers use $lr_{head} = 5 \times 10^{-5}$. A linear learning rate scheduler with a warmup ratio of 0.1 is employed. The loss trade-off coefficient is set to $\lambda = 0.3$.

Method	Avg. Total Latency (per instance)	Computational complexity	Explainability
RPPG-Fake	25s	O(SN)	no
DSR (Ours)	14s	O(N)	important path + text

Table 6: Efficiency comparison. S denotes the number of RL exploration steps per instance in RPPG-Fake; N denotes graph size. DSR's preprocessing overhead is strictly bounded by $T_{\max}$=5 and beam width B=5.

Dataset	Training Samples	Training Time	Memory
PolitiFact	404	0.7 h	13 GB
GossipCop	3,086	3.0 h	13 GB
Fakeddit	14,782	15.0 h	14 GB

Table 7: Training cost statistics per dataset.

Hardware: All experiments are implemented using PyTorch and the Accelerate DDP framework. Training is conducted on 5 NVIDIA TITAN RTX (24GB) GPUs using FP16 mixed precision. We use a per-GPU batch size of 1 and set gradient accumulation steps to 16. This configuration results in an effective total batch size of 80 (1 batch $\times$ 16 steps $\times$ 5 GPUs) per optimization step.

H Computational Efficiency Analysis

Inference Latency. The per-instance latency breakdown for DSR is as follows: (i) hypothesis-driven traversal requires 10–11s for PolitiFact and GossipCop due to their complex textual and visual contexts, while the traversal process for Fakeddit requires approximately 5s; (ii) Background Knowledge Injection (BKI) takes approximately 3s, involving a single MLLM invocation; and (iii) Judge inference achieves sub-0.1s latency through a single forward pass with the LoRA-optimized Qwen2.5-7B. Overall, the total average latency is approximately 14s for PolitiFact and GossipCop, and 9s for Fakeddit.

As noted in the Limitations section, the traversal and BKI stages constitute a one-time offline preprocessing cost whose outputs are cacheable; real-time inference via the LoRA-optimized Judge requires < 0.1s. Detailed per-dataset breakdowns are summarized in Tables 6 and 7.

Training Cost. All experiments were conducted on 5 NVIDIA TITAN RTX GPUs (24GB VRAM each) using FP16 mixed precision with 4-bit NF4 quantization (QLoRA). Peak memory consumption remained below 15 GB per GPU across all datasets, making DSR deployable on standard consumer-grade hardware.

Prompt Variant	PolitiFact Acc	PolitiFact F1
Simplified	0.890	0.900
Paraphrased	0.892	0.906
Structural	0.895	0.903
DSR (default)	0.893	0.904

Table 8: Prompt sensitivity analysis on PolitiFact. Accuracy variance < 0.8% across all variants.

I Prompt Sensitivity Analysis

To assess DSR's robustness to prompt formulation, we evaluated three semantically equivalent prompt variants for each MLLM-interfacing module (Verifier, Debunker, BKI, and Judge): **Simplified** (retaining only the core task specification), **Paraphrased** (lexical substitutions of key terms, e.g., replacing "Verifier" with "Validation Assistant"), and **Structural** (converting free-text instructions to a JSON-schema format). All other experimental conditions were held constant.

As shown in Table 8, accuracy variance across prompt variants remains below **0.8%** on PolitiFact, indicating that DSR's decision-making is robust to surface-level prompt engineering. This stability arises because the MLLM primarily performs high-level semantic navigation over a pre-filtered, ID-anchored candidate set, rather than reasoning from unconstrained free-form text. The slight variation in explanation phrasing across variants does not affect the final classification outcome, which is determined by the differentiable Judge module rather than by the MLLM's generated text directly.

To ensure full reproducibility, we specify the exact model snapshots employed: `gpt-5-2025-05-01`, `gemini-3-pro-v1`, and `Qwen2.5-7B-Instruct`. All inferences were conducted at temperature $T = 0$.