Do Not Guess, Verify: Logic-Guided Adaptive Reasoning for Multimodal Misinformation Detection

Kun Huang[1,2], Rui Qiu[2], Xiaoming Li[2]* and Salah Uddin[2]
[1]Macao Polytechnic University
[2]Zhejiang Yuexiu University
tiskun101@gmail.com, lxm696@tju.edu.cn

Abstract

Recent advances in Large Vision–language Models (VLMs) suggest their potential for multimodal misinformation detection. However, existing multimodal misinformation detectors often fail to effectively integrate them, relying instead on passive aggregation of multimodal features and social signals. Such correlation-driven paradigms are vulnerable to spurious associations and multimodal noise, and lack explicit verification mechanisms. In this paper, we propose **Lo**gic-**G**uided **A**daptive **R**easoning (LoGAR), a verification-oriented framework that integrates VLMs into multimodal misinformation detection through explicit rationale-guided reasoning. LoGAR leverages a VLM to generate an explicit verification rationale, which serves as a global semantic anchor to condition the entire reasoning process. Concretely, the rationale functions as an active query to guide multimodal feature fusion and as a conditioning signal to modulate message passing over heterogeneous social graphs, enabling hypothesis-aware evidence aggregation. Furthermore, LoGAR introduces an instance-aware adaptive depth mechanism that dynamically determines the required reasoning depth. Experimental results on multiple multimodal misinformation benchmarks demonstrate that LoGAR consistently outperforms state-of-the-art methods while significantly reducing computational cost.

1 Introduction

The proliferation of multimodal misinformation on social media has become a growing threat to public discourse (Yan et al., 2025; Wu et al., 2023). Unlike early-stage rumor news relying on blatant textual falsehoods (Horne and Adali, 2017), modern misinformation exploits the semantic gap between modalities—strategically pairing authentic images with misleading captions or recontextualizing historical visuals to fabricate evidence (Aneja et al., 2023; Shu et al., 2017). Detecting such content goes beyond surface-level pattern recognition and instead requires explicit logical verification across modalities.

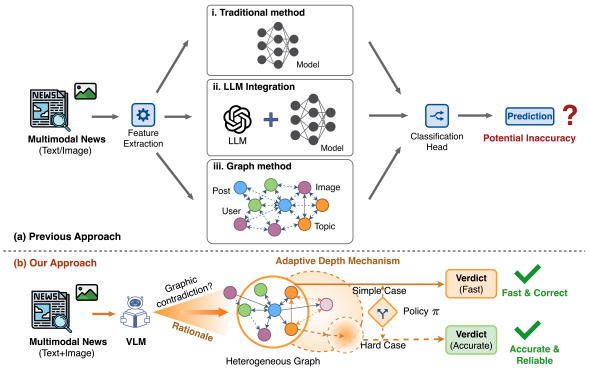


Figure 1: Comparison between our framework and previous approaches.

However, as illustrated in Figure 1, most existing detection paradigms (Yan et al., 2024; Peng et al., 2024; Li et al., 2024) still rely on passive aggregation, treating the reasoning process as an implicit black box. In multimodal settings, traditional architectures (Chen et al., 2022) and recent LLM-based detection pipelines (Tahmasebi et al., 2024) typically perform indiscriminate fusion of visual and textual features. Such blind fusion fails to distinguish evidential cues from incidental correlations, allowing irrelevant signals—such as emotional language or visual clutter—to dominate predictions and dilute critical forensic evidence (Ma et al., 2024; Chen et al., 2023). While graph-based approaches incorporate attention mechanisms or directional propagation to model social context, their weighting strategies are typically structure-driven or data-dependent, rather than verification-aware (Hu et al., 2025; Zheng et al., 2022; Lu et al., 2025). As a result, message passing is not explicitly conditioned on the specific claim or hypothesis under scrutiny, causing heterogeneous neighbors—such

*Corresponding author: Xiaoming Li

Findings of the Association for Computational Linguistics: ACL 2026, pages 4848–4861

as irrelevant discussions, adversarial users, and genuine fact-checking comments—to be aggregated without task-specific discrimination, leading to semantic drift and inefficient computation where trivial and complex cases are processed with equal depth.

To address these limitations, we propose Logic-Guided Adaptive Reasoning (LoGAR), a novel framework that shifts misinformation detection from passive learning to active, rationale-driven verification. Inspired by human fact-checking processes—where hypothesis formulation precedes evidence evaluation (De Neys and Glumicic, 2008)-LoGAR follows the principle of "Do Not Guess, Verify." Specifically, before performing multimodal or social reasoning, our model leverages a Vision-Language Model (VLM) to generate an explicit rationale, which serves as a semantic controller for the entire inference pipeline.

LoGAR incorporates this rationale into a unified heterogeneous graph framework through three key components. First, Rationale-Guided Multimodal Fusion enables content-level verification by using the rationale representation as an active query to retrieve relevant visual and textual evidence, explicitly filtering task-irrelevant noise and aligning feature representations with the verification logic. Second, Rationale-Aware Message Passing extends this guidance to social verification: the rationale globally conditions graph propagation, adaptively weighting neighbors based on their relevance to the hypothesis and effectively suppressing social noise in heterogeneous interaction graphs. Third, to improve efficiency, LoGAR introduces an Adaptive Depth Control mechanism that dynamically determines the required reasoning depth for each instance, produces fast predictions for simple cases while allocating deeper multi-hop reasoning only to ambiguous or complex posts.

In summary, our contributions are threefold:

- We introduce a logic-first verification paradigm that explicitly incorporates rationales to guide multimodal and social reasoning, mitigating semantic drift.
- We propose a unified architecture in which a single rationale vector orchestrates both fine-grained feature extraction and coarse-grained graph propagation, ensuring semantic consistency across modules.
- We present the instance-aware adaptive depth mechanism for multimodal misinformation detection, achieving state-of-the-art performance with substantially reduced computational overhead.

2 Related work

2.1 Multimodal Misinformation Detection

Early studies on misinformation detection primarily focused on textual features and propagation patterns (Shu et al., 2017; Vosoughi et al., 2018). With the increasing prevalence of visually grounded misinformation, recent work has shifted toward multimodal approaches that jointly model textual and visual signals (Khattar et al., 2019; Cui et al., 2019; Singhal et al., 2020). Subsequent models introduce attention-based fusion or pretrained vision–language encoders to enhance cross-modal interaction (Wu et al., 2021; Li et al., 2022; Qi et al., 2023; Liu et al., 2024b). Despite these advances, most multimodal detectors remain correlation-driven, relying on implicit feature aggregation to infer veracity. Even when attention mechanisms are employed, they are typically optimized to highlight statistically salient regions or tokens. Consequently, such models are vulnerable to multimodal noise and semantic drift, particularly in cases involving out-of-context images or subtle cross-modal contradictions.

2.2 Social Context and Graph-based Reasoning

To incorporate social context beyond content-level features, a line of work formulates misinformation detection as a graph reasoning problem (Zhang et al., 2024; Hu et al., 2025). Graph neural network–based models, including propagation-oriented (Sun et al., 2023) and bidirectional architectures (Bian et al., 2020), model user interactions, comment structures, and information diffusion dynamics (Jin et al., 2022; Zheng et al., 2022). Extensions to heterogeneous graphs further encode multiple node and edge types to reflect the complexity of online ecosystems (Li et al., 2025). While these models employ adaptive weighting through attention mechanisms or directional propagation, their reasoning processes are generally structure-driven or data-dependent and they tend to aggregate information from all neighbors based on feature similarity, failing to distinguish between valuable debunking comments and malicious noise.

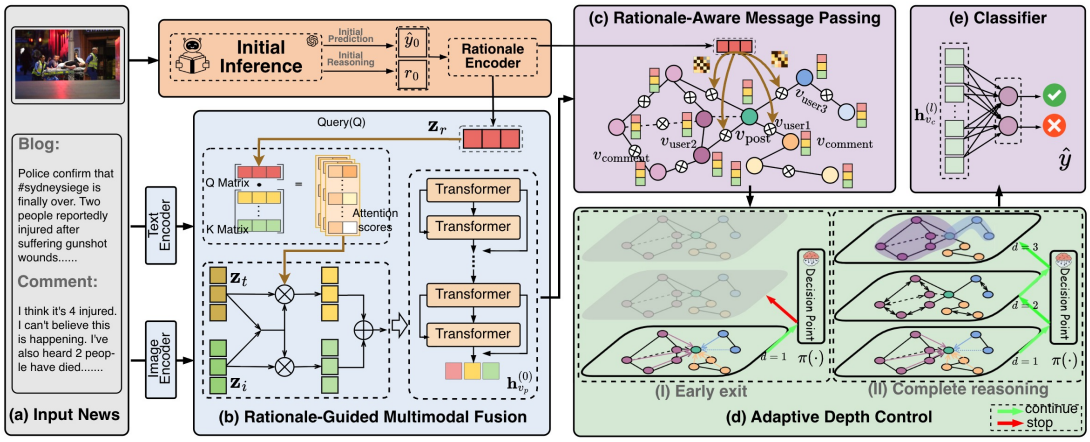


Figure 2: **Overview of the Logic-Guided Adaptive Reasoning (LoGAR) framework.** A vision–language model first generates a verification rationale $\mathbf{z}_r$ from multimodal input, which globally conditions both multimodal feature fusion and heterogeneous graph reasoning. The rationale guides cross-attention to filter task-relevant visual and textual evidence and modulates graph message passing to prioritize hypothesis-consistent neighbors. An adaptive depth controller dynamically adjusts the reasoning depth for each instance before final classification.

2.3 Large Language Models for Misinformation Detection

Recent studies have explored the use of large language models (LLMs) for misinformation detection, leveraging their strong semantic understanding and reasoning capabilities (Hu et al., 2024; Liu et al., 2024a; Hu et al., 2025; Zhang et al., 2025; Tong et al., 2025). Existing efforts primarily fall into two categories: prompt-based classification, where LLMs directly predict veracity from raw inputs (Guan et al., 2024; Qi et al., 2024), and hybrid pipelines, where LLM-generated features or explanations are integrated into downstream classifiers (Huang et al., 2025; Luo et al., 2024). Although LLMs exhibit promising zero-shot or few-shot performance, most current approaches treat them as black-box predictors or auxiliary feature extractors. The reasoning traces or explanations generated by LLMs are seldom integrated into the model's internal decision-making process, and thus do not actively guide multimodal fusion or social reasoning. In contrast, explicit utilization of LLM-generated rationales as control signals for structured reasoning remains underexplored.

3 Methodology

3.1 Problem Formulation

Rather than treating multimodal misinformation detection as a purely correlation-driven classification problem, we frame it as a logic-guided verification process. As shown in Figure 2, given a multimodal news item, the model first generates a rationale that specifies what to verify, and then aggregates multimodal and social evidence in a manner conditioned on this rationale to assess its validity.

Formally, let $\mathcal{D} = \{(C_i, y_i)\}_{i=1}^{N}$ denote a dataset of multimodal news instances, where each item $C_i = (T_i, V_i)$ consists of textual content T_i and visual content V_i, and $y_i \in \{0, 1\}$ denotes the corresponding veracity label (real or fake). The objective is to learn a mapping function $\mathcal{F} : (T, V) \rightarrow y$ that predicts the credibility of a news item.

3.2 Rationale Generation via VLM

To enable hypothesis-conditioned verification, we employ a Vision–Language Model (VLM) to generate a high level verification rationale for each multimodal news instance. Given an input $C = (T, V)$, the VLM is prompted to produce an initial verdict $\hat{y}_0$ and a textual rationale r_0:

$$(\hat{y}_0, r_0) = \arg\max_{\hat{y}, r} P_{\text{VLM}}(\hat{y}, r \mid T, V; \Theta_{pre}) \quad (1)$$

where Θ_{pre} denotes the pre-trained parameters of the VLM. The generated rationale r_0 serves as an explicit verification objective that specifies *what to verify*, such as discrepancies between described events and visual evidence.

3.3 Rationale-Guided Multimodal Fusion

For each multimodal news instance, the rationale r_0 generated by the VLM is retained and transformed

into a semantic guidance signal that conditions subsequent evidence aggregation. Rather than serving as an auxiliary feature, the rationale explicitly specifies the verification objective, directing the model to focus on evidence relevant to the hypothesized inconsistency.

Multi-grained Feature Encoding. We employ three parallel encoders to extract representations from different modalities:

$$\mathbf{z}_r = \text{Enc}(r_0) \in \mathbb{R}^{d_r} \quad \text{(Guidance Signal)} \quad (2)$$

$$\mathbf{z}_t = \text{Enc}(T) \in \mathbb{R}^{d_t} \quad \text{(Textual Content)} \quad (3)$$

$$\mathbf{z}_i = \text{Enc}(V) \in \mathbb{R}^{d_i} \quad \text{(Visual Evidence)} \quad (4)$$

where $\text{Enc}(\cdot)$ denotes modality-specific encoders. The rationale representation $\mathbf{z}_r$ encodes a high-level semantic condition that determines *what to verify*, while $\mathbf{z}_t$ and $\mathbf{z}_i$ provide factual textual and visual evidence.

Rationale-Guided Initialization. Unlike prior multimodal approaches that indiscriminately concatenate features, we perform rationale-guided multimodal fusion to obtain a noise-filtered initialization for the central post node. Specifically, the rationale vector $\mathbf{z}_r$ is used as a query to selectively attend to multimodal evidence from textual and visual features:

$$\mathbf{h}_{v_\text{p}}^{(0)} = \mathcal{F}_\text{attn}\big(Q = \mathbf{z}_r,\ K = [\mathbf{z}_t|\mathbf{z}_i],\ V = [\mathbf{z}_t|\mathbf{z}_i]\big), \quad (5)$$

where $\mathcal{F}_\text{attn}(\cdot)$ denotes a cross-attention-based modality interaction module. This design ensures that the initial node representation emphasizes evidence relevant to the verification objective, mitigating the influence of irrelevant or misleading multimodal signals.

By conditioning feature fusion on the rationale, LoGAR addresses the "garbage in, garbage out" issue commonly encountered in graph-based reasoning: without such guided initialization, subsequent message passing may propagate noisy or irrelevant information, even if graph aggregation weights are carefully learned.

3.4 Adaptive Evidence Aggregation

Heterogeneous Graph Construction. Based on the noise-filtered initialization $\mathbf{h}_{v_\text{post}}^{(0)}$, we construct a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to model both content and social context. The node set $\mathcal{V}$ includes the central post node v_post, as well as auxiliary nodes representing users (v_u) and comments (v_c). These nodes form a structurally heterogeneous graph that captures diverse evidence sources. Rather than introducing type-specific reasoning operators, we adopt a unified rationale-aware attention mechanism, allowing the model to verify the same hypothesis across heterogeneous neighbors in a consistent manner. Edges in $\mathcal{E}$ encode heterogeneous relations such as authorship, interaction, and semantic association. Message passing is performed over this graph to aggregate evidence from multi-source neighbors.

Rationale-Aware Message Passing We conceptualize the rationale vector $\mathbf{c}$ as a "query" that filters neighbor information. For a node v and its neighbor $u \in \mathcal{N}(v)$ at layer l, the attention coefficient $e_{vu}^{(l)}$ is computed not just by node features, but by their compatibility with the rationale:

$$e_{vu}^{(l)} = \text{LeakyReLU}\left(\mathbf{a}_{\tau(u)}^T \left[\tilde{\mathbf{h}}_v^{(l)} \parallel \tilde{\mathbf{h}}_u^{(l)} \parallel \lambda\mathbf{c}\right]\right) \quad (6)$$

where $\tilde{\mathbf{h}}_v^{(l)} = \mathbf{W}^{(l)}\mathbf{h}_v^{(l)}$ represents the linearly transformed node features and $\mathbf{a}_{\tau(u)}$ is a type-specific attention vector for node type $\tau(u)$. The $\parallel$ denotes concatenation and λ is a gating factor. Meanwhile, the $\mathbf{c} \in \mathbb{R}^d$ is a global conditioning vector projected from the rationale embedding, defined as $\mathbf{c} = \mathbf{W}_c\mathbf{z}_r$, with $\mathbf{W}_c$ being a learnable parameter.

The normalized attention weight $\alpha_{vu}^{(l)}$ is obtained via Softmax:

$$\alpha_{vu}^{(l)} = \frac{\exp\left(e_{vu}^{(l)}\right)}{\sum_{k\in\mathcal{N}(v)} \exp\left(e_{vk}^{(l)}\right)} \quad (7)$$

A high $\alpha_{vu}^{(l)}$ implies that neighbor u constitutes a critical piece of the chain of evidence, ensuring the aggregation focus is logic-aligned. Finally, the rationale-guided node representation is updated by aggregating the weighted neighbor features:

$$\mathbf{h}_v^{(l+1)} = \sigma\left(\sum_{u\in\mathcal{N}(v)} \alpha_{vu}^{(l)}\tilde{\mathbf{h}}_u^{(l)}\right) + \mathbf{h}_v^{(l)} \quad (8)$$

Intuitively, by injecting the global condition $\mathbf{c}$ into the attention calculation, Eq. (8) explicitly penalizes neighbors that are semantically dissonant with the VLM's reasoning. This effectively prunes irrelevant structural noise at the feature level, ensuring the graph reasoning remains strictly aligned with the verification intent.

3.5 Adaptive Depth Control

Unlike message passing, which determines how evidence is aggregated, adaptive depth control focuses on deciding when sufficient evidence has been accumulated to verify the hypothesis. We model the reasoning depth selection in an MDP-inspired decision process, where a policy network π_θ dynamically determines whether further graph expansion is necessary.

At each reasoning layer l, the policy observes a state $\mathbf{s}_l$ constructed by jointly considering the current central node representation $\mathbf{h}_{v_p}^{(l)}$, the global rationale vector $\mathbf{c}$, and the initial uncertainty p_0 produced by the VLM:

$$\mathbf{s}_l = \mathit{MLP}_{\text{state}}(\mathbf{h}_{v_p}^{(l)} \| \mathbf{c} \| \psi(p_0)), \tag{9}$$

where $\psi(\cdot)$ denotes a sinusoidal embedding that maps the scalar uncertainty into a high-dimensional space and p_0 denotes the maximum probability of the VLM's initial prediction. The policy outputs an action distribution:

$$P(a_l \mid \mathbf{s}_l) = \text{Softmax}(\mathbf{W}_p \mathbf{s}_l), \tag{10}$$

with actions $a_l \in \{\texttt{STOP}, \texttt{CONTINUE}\}$.

If $a_l = \texttt{STOP}$, the reasoning process terminates early, and the current representation $\mathbf{h}_{v_p}^{(l)}$ is used for final prediction:

$$\hat{y} = \text{Softmax}(\text{MLP}_{\text{cls}}(\mathbf{h}_{v_p}^{(l)})). \tag{11}$$

Otherwise, the model proceeds to the next reasoning layer to aggregate additional evidence. By explicitly conditioning the policy on the rationale and uncertainty, LoGAR avoids redundant reasoning on trivial cases while preserving sufficient depth for ambiguous instances.

3.6 Optimization and Training

To ensure the framework achieves accurate verification while maintaining efficiency and logical consistency, we employ a multi-task learning objective. The total loss $\mathcal{L}_{total}$ comprises three components: classification accuracy, policy optimization, and semantic alignment.

Classification Loss: For the final verification task, we minimize the standard Cross-Entropy loss between the predicted probability $\hat{y}$ and the ground truth y:

$$\mathcal{L}_{cls} = -\mathbb{E}_{(C,y)\sim\mathcal{D}} \left[y \log \hat{y} + (1-y) \log(1-\hat{y}) \right] \tag{12}$$

Policy Gradient Loss: To optimize the depth control policy π_θ, we subtract a moving-average baseline from the reward, following standard REINFORCE practice. We define a hybrid reward function R that balances accuracy and computational cost:

$$R = \mathbb{I}(\hat{y} = y) - \lambda_{cost} \left(\frac{l_{used}}{l_{max}} \right) \tag{13}$$

where $\mathbb{I}(\cdot)$ is the indicator function for correct prediction, and l_{used} is the reasoning depth. The policy loss is defined to maximize the expected reward:

$$\mathcal{L}_{rl} = -\mathbb{E}_{a_l \sim \pi_\theta} [R \cdot \log \pi_\theta(a_l \mid \mathbf{s}_l)] \tag{14}$$

Semantic Alignment Constraint: A core risk in GNN reasoning is semantic drift, where the node representation diverges from the initial logical premise after multiple aggregation hops. To mitigate this, we treat the projected rationale $\mathbf{c} = \mathbf{W}_c \mathbf{z}_r$ as a semantic anchor. We impose a cosine embedding loss to force the final graph representation $\mathbf{h}_{v_p}^{(final)}$ to remain directionally consistent with the VLM's priors:

$$\mathcal{L}_{align} = \left| \frac{\mathbf{h}_{v_p}^{(final)}}{|\mathbf{h}_{v_p}^{(final)}|_2} - \frac{\mathbf{c}}{|\mathbf{c}|_2} \right|_2^2 \tag{15}$$

The final objective function effectively synchronizes the "intuition" of the VLM with the "deliberation" of the GNN:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \gamma_1 \mathcal{L}_{rl} + \gamma_2 \mathcal{L}_{align} \tag{16}$$

This constraint ensures that the structural reasoning refines, rather than contradicts, the large model's linguistic reasoning.

4 Experiments

4.1 Experiment Setting

Datasets Following prior work (Hu et al., 2025; Ma et al., 2024), we evaluate LoGAR on two widely used multimodal rumor detection benchmarks: **PHEME** (Zubiaga et al., 2017) (English) and **Weibo** (Song et al., 2019) (Chinese). PHEME consists of Twitter discussions surrounding breaking news events, while Weibo is collected from the Sina Weibo platform. Both datasets provide aligned textual content, images, and user comments, enabling multimodal and social-context reasoning. Additional dataset statistics and preprocessing details are provided in Appendix B.

Baselines We compare LoGAR with several state-of-the-art baselines covering traditional multimodal models, LLM-based methods, and large vision–language models (LVLMs). Traditional methods (e.g., ENDEF, EANN, MVAE, FND-CLIP) primarily focus on text–image fusion, while MFAN further models social graphs but relies on fixed-depth propagation. LLM-based methods (ARG, GLPN) introduce generative reasoning but lack adaptive inference control. Recent LVLMs (Qwen2/3-VL, LLaMA-3.2-V, InternVL2.5) perform end-to-end multimodal prediction without explicit structural verification.

In contrast, LoGAR uniquely unifies multimodal perception, heterogeneous graph reasoning, and instance-adaptive depth control. Detailed comparisons are deferred to Appendix C.

4.2 Overall Performances

Table 1 reports the performance comparison on the Weibo and PHEME datasets. Overall, LoGAR consistently achieves the best results across all evaluation metrics, indicating the effectiveness of incorporating rationale-guided and adaptive graph reasoning for multimodal misinformation detection. Unless otherwise stated, we use Qwen3-VL as the default rationale generator, as it achieves the best overall performance in our ablation study (Appendix D).

- We first observe that general-purpose large vision-language models (e.g., Qwen2-VL, Llama-3.2-V, InternVL2.5) exhibit limited performance when directly applied to veracity classification, particularly on PHEME, where Qwen2-VL attains only 40.89% accuracy. Although scaling to stronger models (e.g., Qwen3-VL) yields noticeable improvements, these models remain substantially behind graph-based detectors. This suggests that relying solely on holistic multimodal reasoning, without explicitly modeling social context or evidence propagation, is insufficient for robust misinformation verification. In contrast, LoGAR integrates VLM-generated rationales as guidance signals rather than final decision-makers, enabling downstream modules to ground the reasoning process in structured evidence.

- Comparing traditional multimodal classifiers (e.g., FND-CLIP, EANN) with graph-based methods (e.g., MFAN, GLPN), we observe that incorporating user interaction and propagation structure generally leads to stronger performance. For example, GLPN improves accuracy over FND-CLIP by 2.54% on Weibo, highlighting the importance of contextual evidence beyond content-level cues. However, these gains are achieved with fixed aggregation strategies, which may still introduce irrelevant or redundant information during reasoning.

- Finally, when compared with the strongest baseline GLPN, LoGAR shows consistent improvements across all metrics, with particularly notable gains in AUC (+5.40% on Weibo and +5.18% on PHEME). We attribute this improvement to LoGAR's adaptive reasoning mechanism, which dynamically adjusts the depth of evidence aggregation based on instance difficulty and conditions message passing on the global rationale. This design allows the model to selectively accumulate relevant evidence while avoiding unnecessary aggregation, resulting in more stable and discriminative representations.

4.3 Ablation Studies

To assess the contribution of each component in LoGAR, we perform ablation studies by systematically removing individual modules: (1) **w/o Rationale**, which removes the VLM-generated rationale; (2) **w/o Guided Fusion**, which replaces rationale-guided fusion with standard multimodal fusion; (3) **w/o Hetero**, which removes the heterogeneous graph structure; (4) **w/o Adaptive**, which disables adaptive depth control; and (5) **w/o Visual**, which excludes visual features. The comparison results are visualized in Figure 3.

Across both datasets, removing the rationale guidance (**w/o Rationale**) leads to the most substantial degradation across all metrics, particularly in Acc and F1. This indicates that, without an verification hypothesis, the graph reasoner struggles to distinguish evidence relevant to veracity from topically related but uninformative neighbors. In contrast, the full model benefits from using the rationale as a global semantic condition, which consistently stabilizes performance across different metrics.

The **w/o Guided Fusion** variant also shows a noticeable decline, though less severe than removing the rationale entirely. This suggests that not only

Table 1: Performance comparison on Weibo and PHEME datasets. Mean ± standard deviation over 5 runs with different random seeds. The best results are highlighted in bold.

Methods	Weibo					PHEME				
	Acc	Prec	Rec	macF1	AUC	Acc	Prec	Rec	macF1	AUC
Qwen2-VL	49.48 ± 3.12	57.97 ± 3.25	55.69 ± 3.05	49.86 ± 3.18	58.86 ± 4.95	40.89 ± 4.30	56.14 ± 3.22	54.51 ± 3.15	40.20 ± 3.28	55.70 ± 3.02
Llama-3.2-V	59.11 ± 3.92	61.90 ± 3.98	59.74 ± 3.88	59.65 ± 3.95	59.93 ± 3.82	30.75 ± 4.10	49.27 ± 4.05	49.88 ± 4.12	44.93 ± 4.08	49.89 ± 4.95
InternVL2.5	61.56 ± 3.85	63.12 ± 3.88	61.56 ± 3.82	61.97 ± 3.86	61.29 ± 3.78	59.38 ± 3.92	57.28 ± 3.95	59.38 ± 3.89	56.42 ± 3.93	62.57 ± 3.85
Qwen3-VL	68.17 ± 3.75	73.35 ± 3.79	70.81 ± 3.72	68.31 ± 3.76	74.85 ± 3.68	63.65 ± 3.82	61.54 ± 3.85	53.74 ± 3.78	55.70 ± 3.81	55.27 ± 3.75
ENDEF	71.53 ± 1.62	74.35 ± 1.65	69.28 ± 1.58	70.83 ± 1.60	76.35 ± 1.55	65.82 ± 1.68	64.69 ± 1.71	65.39 ± 1.65	64.15 ± 1.67	69.40 ± 1.61
EANN	75.28 ± 0.95	76.13 ± 0.88	71.25 ± 0.82	73.26 ± 0.86	77.85 ± 0.89	70.17 ± 1.22	71.28 ± 1.05	67.36 ± 0.99	69.10 ± 1.01	74.20 ± 0.96
MVAE	77.16 ± 1.51	75.32 ± 1.54	87.83 ± 2.48	80.14 ± 1.52	82.45 ± 2.45	81.37 ± 1.58	79.53 ± 1.61	81.22 ± 1.55	79.43 ± 1.59	83.60 ± 2.52
FND-CLIP	87.93 ± 1.38	87.62 ± 1.41	86.59 ± 1.35	86.83 ± 1.39	89.75 ± 1.32	86.85 ± 1.42	86.37 ± 1.45	86.14 ± 1.38	87.49 ± 1.43	87.49 ± 1.36
MMDFND	76.66 ± 0.95	76.32 ± 0.98	75.41 ± 0.92	76.37 ± 0.96	79.92 ± 0.90	81.25 ± 1.02	79.22 ± 1.15	79.11 ± 1.29	79.69 ± 1.21	82.55 ± 1.08
MFAN	88.35 ± 0.85	88.61 ± 0.78	87.53 ± 0.82	88.33 ± 0.86	90.62 ± 0.79	87.93 ± 0.89	87.47 ± 0.92	84.65 ± 0.85	86.14 ± 0.88	89.93 ± 0.84
ARG	89.53 ± 0.82	89.89 ± 0.85	90.22 ± 0.80	90.57 ± 0.83	91.67 ± 0.78	88.28 ± 0.76	88.12 ± 0.89	85.29 ± 0.83	87.53 ± 0.85	89.15 ± 0.81
GLPN	90.47 ± 1.28	89.35 ± 0.51	92.02 ± 0.56	91.68 ± 0.59	92.10 ± 0.54	88.68 ± 0.53	88.27 ± 0.56	86.11 ± 1.30	86.32 ± 1.32	90.05 ± 1.28
LoGAR	**93.36** ± 0.57	**92.82** ± 0.57	**93.52** ± 0.73	**93.10** ± 0.60	**97.50** ± 0.25	**90.18** ± 0.18	**89.98** ± 0.19	**84.87** ± 0.21	**86.71** ± 0.26	**95.23** ± 0.23

the presence of the rationale, but its explicit use as a query during multimodal fusion, is important for filtering task-irrelevant visual and textual signals before graph reasoning.

Disabling the adaptive reasoning mechanism (**w/o Adaptive**) results in consistent performance drops, most clearly reflected in Rec on both Weibo and PHEME. This observation aligns with the intuition that fixed-depth reasoning cannot accommodate the varying complexity of misinformation cases: shallow aggregation may miss long-range evidence, while deeper aggregation can introduce unnecessary noise for simpler instances.

Finally, removing the heterogeneous graph (**w/o Hetero**) or visual modality (**w/o Visual**) also degrades performance, confirming the importance of structural context and visual evidence. However, the relatively larger drop observed in **w/o Rationale** compared to **w/o Visual** suggests that high-level semantic guidance plays a more central role than any single modality alone.

Overall, these results indicate that LoGAR's performance gains stem from the complementary interaction between logic-guided representation learning and adaptive evidence aggregation, rather than from any isolated component.

4.4 Adaptive Reasoning Analysis

Adaptive vs. Fixed Depth To assess whether adaptive reasoning is necessary beyond fixed-depth heuristics, we compare LoGAR with static variants using $k \in 1, 2, 3$. As shown in Fig. 4, no single fixed depth consistently dominates across datasets or metrics. Shallow reasoning ($k = 1$) fails to aggregate sufficient evidence for interaction-rich scenarios, resulting in noticeable performance degradation on Weibo. Conversely, deeper reasoning

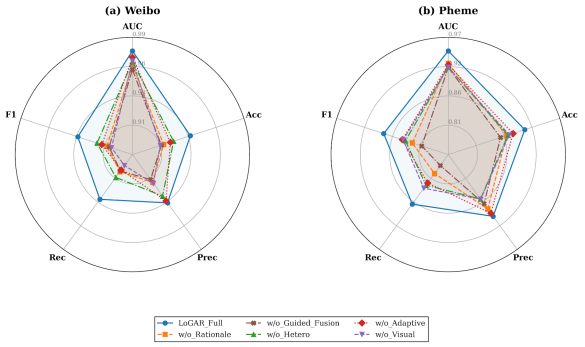


Figure 3: Performance comparison of ablation variants on Weibo and PHEME via radar charts.

($k = 3$) introduces redundant or noisy evidence, leading to precision drops on the noise-sensitive PHEME dataset.

LoGAR resolves this structural limitation by dynamically adjusting reasoning depth at the instance level. With an average depth of $d = 1.83$ on Weibo and $d = 1.85$ on PHEME, LoGAR consistently outperforms the strongest fixed-depth baseline across all metrics. This demonstrates that adaptive depth selection is not merely a computational optimization, but a necessary mechanism for handling heterogeneous reasoning demands.

Distribution of Inference Depths To further understand why fixed-depth reasoning is suboptimal, we analyze the distribution of inference depths selected by LoGAR across test samples. As illustrated in Figure 5, the required reasoning depth varies substantially among instances, reflecting heterogeneous verification difficulty.

On Weibo, while a majority of samples can be resolved with shallow reasoning ($k = 1$), a considerable portion requires deeper evidence propagation ($k = 3$) to reach a reliable decision. In

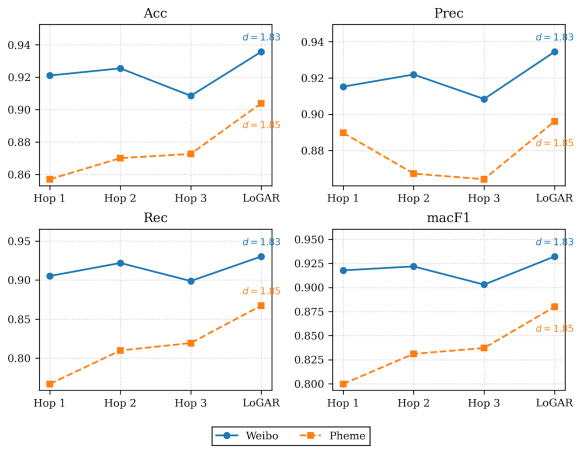


Figure 4: Performance comparison between fixed reasoning depths (Hop 1-3) and LoGAR's adaptive policy. The average reasoning depths (d) of LoGAR are annotated.

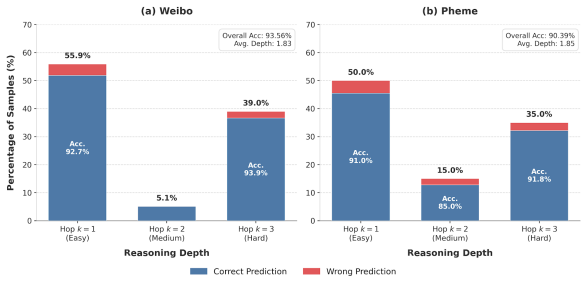


Figure 5: Distribution of inference depths selected by LoGAR across Weibo and PHEME datasets.

contrast, PHEME exhibits a higher concentration of medium-depth cases, indicating greater sensitivity to noise accumulation. This heterogeneous depth demand explains why no single fixed-depth strategy performs consistently well across datasets.

Efficiency Analysis. We further report wall-clock latency and throughput for the verification stage on a single NVIDIA RTX 4090 GPU, with VLM rationale generation treated as offline preprocessing. To reflect per-instance verification latency, all measurements are conducted with an inference batch size of 1. As shown in Table 2, compared with fixed full-depth reasoning, LoGAR reduces relative FLOPs from $1.00\times$ to $0.57\times$, decreases latency from 18.5 to 10.8 ms/sample, and improves throughput from 54.1 to 92.0 samples/s. Overall, this corresponds to an approximate 43% reduction in FLOPs and a 41% speedup in inference latency, showing that adaptive depth control not only improves performance but also yields clear computational benefits in practice.

Table 2: Wall-clock efficiency comparison between fixed full-depth reasoning and LoGAR's adaptive-depth policy on the verification stage. VLM rationale generation is treated as offline preprocessing.

Metric	**Fixed full-depth**	**LoGAR**
Max depth	3	3
Relative FLOPs	$1.00\times$	$\mathbf{0.57}\times$
Latency (ms/sample)	18.5	**10.8**
Throughput (samples/s)	54.1	**92.0**

4.5 Case Study

Figure 6 provides a qualitative analysis of LoGAR's ability to correct LLM hallucinations using the Sydney Siege case from PHEME. Despite the post's claim of gunshot victims being supported by an image of paramedics, Panel (a) shows that the VLM's reliance on parametric intuition leads to a false-positive 'Real' verdict. This confirms that semantic consistency between modalities does not always equate to truthfulness.In Panel (b), LoGAR's adaptive reasoning policy triggers a deeper investigation ($d = 2.0$) upon detecting high uncertainty. Our Logic-Guided Attention allows for highly selective evidence aggregation. While other static models might be distracted by emotional responses, LoGAR assigns significantly higher weights to verified factual corrections (Comment 3). By suppressing irrelevant social noise and focusing on dissonant structural cues, LoGAR successfully rectifies the prediction to FAKE. This demonstrates LoGAR's capability to ground generative semantics in external structural verification, effectively bridging the gap between intuition and logic. Additional case studies on both datasets are provided in the appendix E.

5 Conclusion

We proposed LoGAR, a logic-guided adaptive reasoning framework for multimodal misinformation detection. By reframing detection as an abductive verification process, LoGAR leverages VLM-generated rationales as global semantic guidance to filter multimodal noise and regulate graph-based evidence aggregation. Moreover, an adaptive depth policy dynamically determines the sufficient reasoning depth under uncertainty, improving efficiency without sacrificing accuracy. Experiments across multiple benchmarks demonstrate that LoGAR consistently outperforms strong multimodal and graph-based baselines, while achieving favorable accuracy–efficiency trade-offs compared

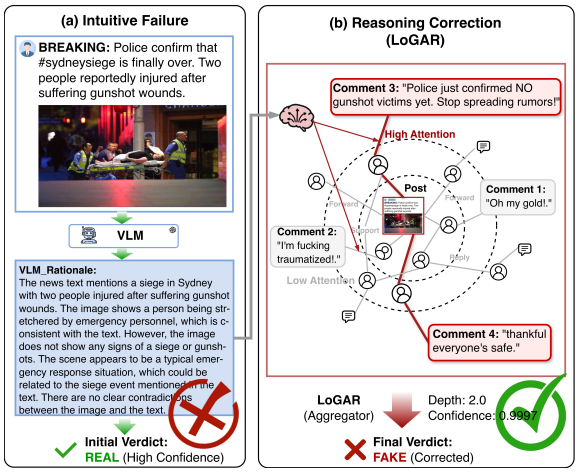


Figure 6: Case Study — An example where LoGAR corrects an intuition-driven prediction.

to fixed-depth reasoning. This work highlights the potential of integrating large vision-language models with structured graph reasoning for effective and interpretable multimodal verification.

Acknowledgments

This work was supported by the National Science Foundation of China (62272311) and the Shaoxing Municipal Higher Education Research Project (SGJ2026026).

Limitations

Despite its effectiveness, LoGAR has several limitations that warrant further investigation.

- The quality of rationale generation remains a critical dependency. LoGAR assumes that the VLM-generated rationale provides a meaningful high-level hypothesis to guide subsequent graph reasoning. Although our experiments demonstrate that rationale-guided fusion and reasoning consistently improve performance, erroneous or overly vague rationales may mislead the evidence aggregation process. While the proposed semantic alignment and adaptive depth control partially mitigate this risk, improving the robustness of rationale generation itself remains an open challenge.

- The construction of heterogeneous graphs relies on predefined entity types and relations (e.g., users, comments, and posts). While this design improves interpretability and verification structure, it may limit scalability to platforms where metadata is sparse, noisy, or unavailable. Extending LoGAR to dynamically discover entity types or operate under weaker structural assumptions is a promising direction for future work.

- The current framework focuses on post-level verification and does not explicitly model long-term temporal evolution of misinformation. Although adaptive depth control enables selective evidence aggregation across interaction hops, integrating temporal reasoning or cross-event dependencies may further strengthen the model's ability to capture evolving rumor dynamics.

We leave these extensions to future work and believe that addressing them will further enhance the generality and robustness of logic-guided adaptive reasoning frameworks.

Ethics Statement

This work follows the ACM Code of Ethics and Professional Conduct. We conduct experiments on publicly available benchmark datasets (Weibo and PHEME) that have been widely used in prior research, without collecting new data or identifying individual users. All user-related information is anonymized, and the proposed method focuses on content, visual, and structural patterns rather than user profiling. Proper attribution is given to all datasets, models, and toolkits used. This work aims to support safer online information ecosystems through improved misinformation detection. We emphasize that any downstream deployment of the proposed method should follow applicable data protection regulations and platform policies.

References

Shivangi Aneja, Chris Bregler, and Matthias Nießner. 2023. Cosmos: catching out-of-context image misuse using self-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 14084–14092.

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.

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.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 21 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *CoRR*, abs/2412.05271.

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 627–638.

Limeng Cui, Suhang Wang, and Dongwon Lee. 2019. Same: sentiment-aware multi-modal embedding for detecting fake news. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*, pages 41–48.

Wim De Neys and Tamara Glumicic. 2008. Conflict monitoring in dual process theories of thinking. *Cognition*, 106(3):1248–1299.

Matthias Fey and Jan Eric Lenssen. 2019. Fast graph representation learning with pytorch geometric. *arXiv preprint arXiv:1903.02428*.

Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2024. Language models hallucinate, but may excel at fact verification. In *Proceedings of the 2024 conference of the North American chapter of the association for computational linguistics: human language technologies (volume 1: long papers)*, pages 1090–1111.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Benjamin Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 759–766.

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

Shuguo Hu, Jun Hu, and Huaiwen Zhang. 2025. Synergizing llms with global label propagation for multimodal fake news detection. *arXiv preprint arXiv:2506.00488*.

Kun Huang, Xiaoming Li, and Salah Uddin. 2025. Enhancing fake news detection through fact-augmented llm generation with co-attention. *Journal of Intelligent Information Systems*, pages 1–19.

Yiqiao Jin, Xiting Wang, Ruichao Yang, Yizhou Sun, Wei Wang, Hao Liao, and Xing Xie. 2022. Towards fine-grained reasoning for fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 5746–5754.

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.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Mingxin Li, Yuchen Zhang, Haowei Xu, Xianghua Li, Chao Gao, and Zhen Wang. 2025. 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.

Qilei Li, Mingliang Gao, Guisheng Zhang, Wenzhe Zhai, Jinyong Chen, and Gwanggil Jeon. 2024. Towards multimodal disinformation detection by vision-language knowledge interaction. *Information Fusion*, 102:102037.

Hui Liu, Wenya Wang, Haoru Li, and Haoliang Li. 2024a. Teller: A trustworthy framework for explainable, generalizable and controllable fake news detection. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15556–15583.

Xuannan Liu, Peipei Li, Huaibo Huang, Zekun Li, Xing Cui, Jiahao Liang, Lixiong Qin, Weihong Deng, and Zhaofeng He. 2024b. Fka-owl: Advancing multimodal fake news detection through knowledge-augmented lvlms. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10154–10163.

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.

Yifeng Luo, Yupeng Li, Dacheng Wen, and Liang Lan. 2024. Message injection attack on rumor detection under the black-box evasion setting using large language model. In *Proceedings of the ACM Web Conference 2024*, pages 4512–4522.

Zihan Ma, Minnan Luo, Hao Guo, Zhi Zeng, Yiran Hao, and Xiang Zhao. 2024. Event-radar: Event-driven multi-view learning for multimodal fake news detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5821.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z.

Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Liwen Peng, Songlei Jian, Zhigang Kan, Linbo Qiao, and Dongsheng Li. 2024. Not all fake news is semantically similar: Contextual semantic representation learning for multimodal fake news detection. *Information Processing & Management*, 61(1):103564.

Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. 2023. Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14444–14452.

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, 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):22–36.

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

Changhe Song, Cheng Yang, Huimin Chen, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2019. Ced: Credible early detection of social media rumors. *IEEE Transactions on Knowledge and Data Engineering*, 33(8):3035–3047.

Ling Sun, Yuan Rao, Yuqian Lan, Bingcan Xia, and Yangyang Li. 2023. Hg-sl: Jointly learning of global and local user spreading behavior for fake news early detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 5248–5256.

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.

Llama Team. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Qwen Team. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

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.

Zhao Tong, Yimeng Gu, Huidong Liu, Qiang Liu, Shu Wu, Haichao Shi, and Xiao-Yu Zhang. 2025. Generate first, then sample: Enhancing fake news detection with llm-augmented reinforced sampling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24276–24290.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191.

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, and Yanning Zhang. 2023. See how you read? multi-reading habits fusion reasoning for multi-modal fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 13736–13744.

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.

Yeqing Yan, Peng Zheng, and Yongjun Wang. 2024. Enhancing large language model capabilities for rumor detection with knowledge-powered prompting. *Engineering Applications of Artificial Intelligence*, 133:108259.

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

Chaowei Zhang, Zongling Feng, Zewei Zhang, Jipeng Qiang, Guandong Xu, and Yun Li. 2025. Is llms hallucination usable? llm-based negative reasoning for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1031–1039.

Litian Zhang, Xiaoming Zhang, Ziyi Zhou, Feiran Huang, and Chaozhuo Li. 2024. Reinforced adaptive knowledge learning for multimodal fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 16777–16785.

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

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.

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

Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2017. Exploiting context for rumour detection in social media. In *International conference on social informatics*, pages 109–123. Springer.

A Implementation Details

We implemented LoGAR using the PyTorch framework (Paszke et al., 2019) and PyTorch Geometric (Fey and Lenssen, 2019). For the visual backbone, we employed ResNet-50 (He et al., 2016) pre-trained on ImageNet to extract grid features, while the textual representations were encoded using a 1D-CNN with a kernel size of 3. The input images were resized to 224×224, and the maximum text sequence length was truncated to 100 tokens. All experiments were conducted on a workstation equipped with an NVIDIA GeForce RTX 4090 GPU and Intel Xeon Platinum 8352V CPUs. To ensure transparency and reproducibility, we also provide the full prompt template used for rationale generation in Figure 7.

Table 3 outlines the specific hyperparameter settings for both Weibo and Pheme datasets. The optimal values were determined through a refined parallel random search for each dataset respectively.

Table 3: Hyperparameter settings of LoGAR on Weibo and Pheme.

Hyperparameter	**Weibo**	**Pheme**
Optimizer	Adam	Adam
Learning Rate (LR)	2.98e-4	2.60e-4
Batch Size	32	32
Max Hops (K)	3	3
Hidden Dim (d)	64	64
Fusion Heads	4	4
Reinforcement Learning Rewards		
Early Stop (r_{stop})	0.1190	0.0919
Continue (r_{cont})	0.0691	0.0859
Exploration (r_{exp})	0.0977	0.0825

B Dataset

We conduct experiments on two real-world multimodal rumor detection benchmarks: PHEME (Zubiaga et al., 2017) and Weibo (Song et al., 2019). PHEME is an English-language dataset consisting of Twitter threads related to five breaking news events, while Weibo is a Chinese dataset collected from the Sina Weibo platform.

Both datasets contain multimodal posts with associated textual content, images, and user-generated comments, making them suitable for evaluating multimodal and graph-based reasoning methods. Since LoGAR explicitly relies on joint textual, visual, and social graph features, we remove samples that lack either textual or visual information to ensure modality completeness.

After preprocessing, the PHEME dataset contains 2,018 events, including 1,428 non-rumors and 590 rumors, along with 7,388 associated comments. The Weibo dataset includes 1,467 events (877 non-rumors and 590 rumors) generated by 985 users, with a total of 4,534 comments.

C Baseline method

Table 4 summarizes the modeling capabilities of the baseline methods considered in this work. The comparison highlights differences in supported input modalities, graph modeling, and whether adaptive reasoning is explicitly incorporated during inference.

D Ablation on Vision-Language Models for Rationale Generation

We investigate the impact of different Vision-Language Models (VLMs) used for rationale generation while keeping all other components of LoGAR fixed. As shown in Table 5, we evaluate four representative VLMs—Qwen2-VL, LLaMA-3.2-V, InternVL2.5, and Qwen3-VL—on both

Rationale Generation Prompt

System Prompt

You are a multimodal fake news detection expert. Please perform rapid abductive reasoning on the given news image and text.

Output Requirements

- `rationale` (string): logical reasoning chain describing whether the visual evidence is consistent with the textual claim, including contradictions such as seasonal mismatch, location-feature mismatch, editing traces, or unsupported details.
- `verdict` (string): preliminary judgment, either `Real` or `Fake`.
- `confidence` (float): confidence score ranging from `0.0` to `1.0`.

Constraint

The model should return only valid JSON without markdown formatting or additional explanatory text.

Example Output

Real Example JSON

```
{
  "552787144373460992": {
    "rationale": "The image shows a French Police Nationale car with multiple bullet
holes in the windshield, parked in front of a building with a classic Parisian
architectural style. The text claims the car is in front of the Charlie Hebdo
headquarters. However, there is no visible signage or architectural feature
confirming that location. The building in the background appears more consistent
with a residential or commercial Parisian street facade than the known Charlie Hebdo
site. Although the seasonal cues are broadly consistent with the event timing, the
location mismatch remains a significant logical inconsistency.",
    "verdict": "Fake",
    "confidence": 0.85,
    "imgnum": "3",
    "valid": true
  }
}
```

Input: *[news image, accompanying text]*
Output: *JSON rationale prior*

Figure 7: Rationale generation prompt with example output

Table 4: Comparisons with baseline methods in terms of modeling capabilities.

Method	Text	Image	Graph	Adaptive
Traditional Learning				
ENDEF (Zhu et al., 2022)	✓	×	×	×
EANN (Wang et al., 2018)	✓	✓	×	×
MVAE (Khattar et al., 2019)	✓	✓	×	×
FND-CLIP (Zhou et al., 2023)	✓	✓	×	×
MMDFND (Tong et al., 2024)	✓	✓	×	×
MFAN (Zheng et al., 2022)	✓	✓	✓	×
LLM-based Methods				
ARG (Hu et al., 2024)	✓	×	×	×
GLPN (Hu et al., 2025)	✓	✓	✓	×
Large Vision-Language Models (LVLMs)				
Qwen2-VL (Wang et al., 2024)	✓	✓	×	×
Llama-3.2-V (Team, 2024)	✓	✓	×	×
InternVL2.5 (Chen et al., 2024)	✓	✓	×	×
Qwen3-VL (Team, 2025)	✓	✓	×	×
LoGAR (Ours)	✓	✓	✓	✓

Graph indicates explicit graph-based modeling over social or relational structures. **Adaptive** denotes instance-aware dynamic control of reasoning depth during inference.

Rationale Generator	**Weibo (Acc.)**	**PHEME (Acc.)**
Qwen2-VL	90.79	87.43
LLaMA-3.2-V	91.53	88.72
InternVL2.5	92.88	89.57
Qwen3-VL	**93.56**	**90.39**

Table 5: Ablation study on different Vision-Language Models for rationale generation. All other components of LoGAR are fixed, and only the VLM used for rationale generation is varied.

Weibo and PHEME datasets.

Across both datasets, we observe a consistent performance ranking, where Qwen3-VL achieves the best accuracy, reaching **93.56%** on Weibo and **90.39%** on PHEME. Compared to Qwen2-VL, Qwen3-VL yields absolute gains of +2.77% on Weibo and +2.96% on PHEME, indicating that higher-quality rationales substantially improve downstream reasoning and verification. InternVL2.5 and LLaMA-3.2-V show moderate improvements over Qwen2-VL, but remain consistently inferior to Qwen3-VL across both benchmarks.

Notably, the relative performance trends are highly stable across the Chinese (Weibo) and English (PHEME) datasets, suggesting that the effectiveness of VLM-generated rationales generalizes across languages and domains. This ablation study demonstrates that while LoGAR is robust to different rationale generators, stronger VLMs produce more informative and reliable rationales, which in turn facilitate more effective adaptive reasoning. Based on these observations, we adopt Qwen3-VL as the default rationale generator in all main experiments.

E Additional Case Studies

To better illustrate how Logic-Guided Adaptive Reasoning (LoGAR) corrects intuitive yet flawed multimodal predictions, we present two representative case studies shown in Figure 8& 9. Both examples highlight a common failure mode of vision-language models (VLMs): over-reliance on surface-level visual–textual consistency while ignoring latent logical contradictions revealed through social context.

Crisis-Related Breaking News In the first example, the post claims that police have taken down a suspect at Martin Place, with potential relevance to an ongoing siege. The accompanying image depicts police officers restraining an individual in a public area.

The VLM predicts the post as REAL with high confidence (0.95), based on apparent alignment between the visual content (police action) and the textual description. The generated rationale focuses on descriptive consistency and the absence of obvious visual manipulation, without explicitly verifying the factual claim regarding the event's

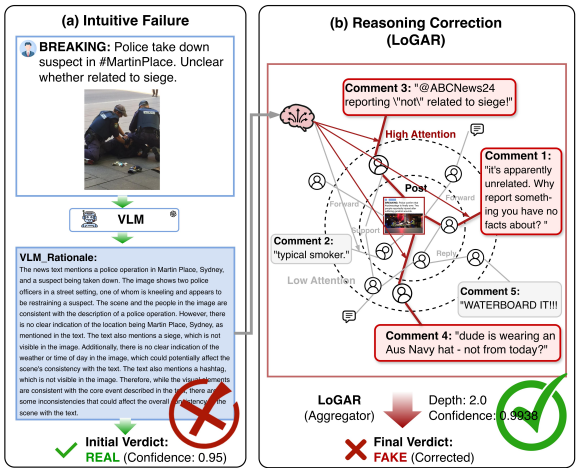


Figure 8: Case Study on the PHEME dataset. A corresponding example demonstrating LoGAR's ability to rectify intuition-driven errors.

relation to the siege.

LoGAR reformulates the task as a hypothesis verification problem and assigns higher attention to comments containing factual negation or temporal inconsistency (e.g., reports stating that the incident is not related to the siege). Through multi-hop aggregation of logically relevant comments, LoGAR identifies a consistent contradiction to the original claim and corrects the prediction to FAKE (confidence 0.9938).

This case demonstrates that visual plausibility alone is insufficient for misinformation detection in fast-evolving news scenarios.

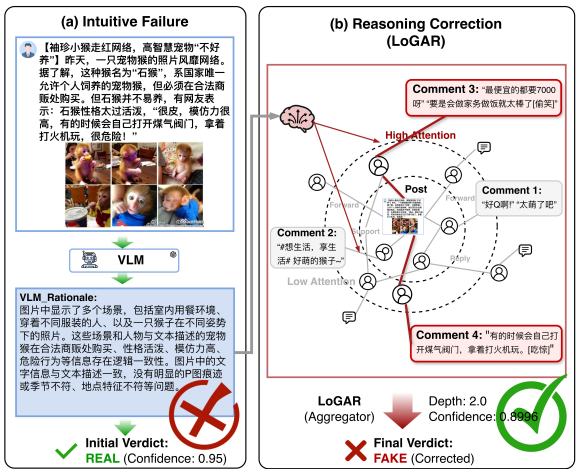


Figure 9: Case Study on the Weibo dataset. An example where LoGAR corrects an intuition-driven prediction via dynamic reasoning.

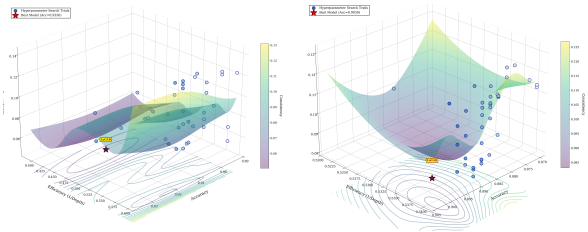
Viral Pet Ownership Claim on Social Media The second example involves a viral post asserting that a specific monkey breed is legally purchasable and suitable as a household pet, supported by multiple images of a monkey in domestic environments.

The VLM predicts the post as REAL (confidence 0.95), relying on the coherence between the images and the textual description. Positive sentiment and aesthetic appeal dominate the reasoning process, while implicit claims regarding legality and safety remain unexamined.

Across both cases, LoGAR consistently mitigates a key failure mode of VLM-based misinformation detection: over-reliance on local visual–textual alignment. By explicitly incorporating logic-guided attention and multi-step reasoning over social context, LoGAR enables systematic correction of intuitive but incorrect predictions.

F Efficiency–Performance Trade-off Analysis

To further analyze the behavior of LoGAR's adaptive reasoning policy, we investigate the trade-off between predictive performance, computational efficiency, and decision stability. Figure 10 visualizes the hyperparameter search landscape, where Efficiency is defined as the inverse of the average reasoning depth ($1/\text{Depth}$), Accuracy reflects classification performance, and Consistency corresponds to the early-stopping reward weight that regularizes policy stochasticity.



(a) Weibo (b) PHEME

Figure 10: Trade-off between accuracy, efficiency ($1/\text{Depth}$), and consistency in LoGAR. The red star indicates the selected Pareto-optimal configuration..

The resulting surface reveals a clear Pareto-optimal region: shallow reasoning leads to insufficient evidence aggregation, while excessive depth yields diminishing returns with higher computational cost. By modulating the consistency reward, LoGAR learns to terminate reasoning once sufficient evidence is accumulated, achieving a judicious balance between accuracy and efficiency.