DIFAR: Enhancing Multimodal Misinformation Detection with Diverse, Factual, and Relevant Rationales

Herun Wan¹, Jiaying Wu², Minnan Luo^{⊠ 1}, Xiangzheng Kong¹, Zihan Ma¹, Zhi Zeng¹

¹Xi'an Jiaotong University, ²National University of Singapore wanherun@stu.xjtu.edu.cn, minnluo@xjtu.edu.cn

Abstract

Generating textual rationales from large vision-language models (LVLMs) to support trainable multimodal misinformation detectors has emerged as a promising paradigm. However, its effectiveness is fundamentally limited by three core challenges: (i) insufficient diversity in generated rationales, (ii) factual inaccuracies due to hallucinations, and (iii) irrelevant or conflicting content that introduces noise. We introduce DIFAR, a detector-agnostic framework that produces diverse, factual, and relevant rationales to enhance misinformation detection. DIFAR employs five chain-of-thought prompts to elicit varied reasoning traces from LVLMs and incorporates a lightweight post-hoc filtering module to select rationale sentences based on sentence-level factuality and relevance scores. Extensive experiments on four popular benchmarks demonstrate that DIFAR outperforms four baseline categories by up to 5.9% and boosts existing detectors by as much as 8.7%. Both automatic metrics and human evaluations confirm that DIFAR significantly improves rationale quality across all three dimensions.¹

1 Introduction

Large vision-language models (LVLMs) have achieved remarkable performance across a wide array of multimodal tasks, driven by their powerful reasoning and representation capabilities. However, their effectiveness remains limited in multimodal misinformation detection (MMD), a task that demands precise factual grounding and fine-grained, task-specific reasoning (Liu et al. 2025a; Li et al. 2025b).

To harness the potential of LVLMs for identifying multimodal misinformation, recent work has proposed a collaborative paradigm that combines LVLMs with trainable detectors (Zheng et al. 2025), which we refer to as the LVLM-as-Enhancer. In this framework, LVLMs are prompted to generate textual rationales (i.e., interpretable justifications or explanations), which are then paired with the original news article and passed to a downstream trainable detector. This design aims to harness the generalization strength of LVLMs while preserving the adaptability of task-specific models.

While this paradigm has shown early promise (Tahmasebi, Müller-Budack, and Ewerth 2024; Hu et al. 2024), we identify three core limitations (illustrated in Figure 1) that hinder its full potential:

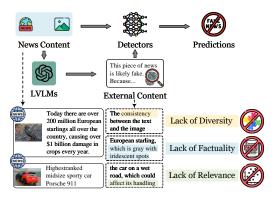


Figure 1: Illustration of three key challenges, namely, diversity, factuality, and relevance, in the LVLM-as-Enhancer paradigm for MMD, where LVLMs are prompted to generate explanatory rationales to support downstream detectors.

- Limited Diversity. Most existing works concentrate on architectural innovations within the trainable detector (Wang et al. 2024), paying limited attention to the quality and variation of generated rationales. These methods typically rely on a fixed prompt, which restricts the range of perspectives captured. As a result, they fail to exploit diverse reasoning signals that could enrich the interpretation of news content (Wan et al. 2024).
- Limited Factuality. LVLMs are prone to hallucinations (Ji et al. 2023) and often generate content that deviates from verified facts (Mallen et al. 2023). Consequently, the resulting rationales may introduce factual inaccuracies, which degrade the reliability of downstream detectors (Pan et al. 2023).
- Limited Relevance. Generated rationales frequently include loosely related or off-topic information, which may dilute or even conflict with the original article's claims (Zheng et al. 2025; Xu et al. 2024). This misalignment reduces the utility of the explanations and can compromise veracity assessment.

To address these limitations, we propose DIFAR, an MMD framework designed to produce rationales that are diverse, factual, and relevant. DIFAR is compatible with any existing trainable detector and operates without requiring architectural modifications. To enhance diversity, DI-

¹Available at https://github.com/whr000001/DiFaR.

FAR incorporates multiple rationales derived from a set of five chain-of-thought (CoT) prompts, each targeting different aspects of the content, including textual details, visual features, and cross-modal consistency. This multi-prompt strategy allows for richer and more subtle reasoning.

To further improve rationale quality, DIFAR incorporates a post-hoc refinement module that filters individual rationale sentences based on factuality and relevance. For factuality, the module retrieves evidence from structured knowledge bases such as Wikipedia and compares it with the generated content (Min et al. 2023). For relevance, it computes semantic similarity between the rationale and the source article using representation-based metrics (Lewis et al. 2020). Sentences with low scores are pruned, resulting in a distilled and trustworthy rationale set.

We conduct extensive experiments on four multimodal misinformation datasets, covering both human-written and machine-generated news articles. Across all datasets, DIFAR consistently outperforms four representative categories of strong baselines, with up to 5.9% relative accuracy improvement. Additionally, integrating DIFAR into existing detectors yields performance gains of up to 8.7%. Ablation studies confirm that each component of DIFAR contributes meaningfully to its effectiveness. Further analysis, including human evaluation, validates that DIFAR improves the diversity, factuality, and relevance of the generated rationales.

2 Methodology

2.1 Preliminaries

We consider the task of MMD as a binary classification problem. Each news instance consists of a textual component and a visual image, and the goal is to determine whether the news is real of fake. Formally, let $\mathcal{D}_{train} = \{(T_i, V_i, y_i)\}_{i=1}^{N_{train}}$ denote a training dataset of N_{train} labeled news articles, where T_i is the text, V_i is the associated image, and $y_i \in \{0,1\}$ is the ground-truth label. A trainable detector f with parameters θ is trained to model the conditional distribution $p(y \mid T, V; f, \theta)$, with the objective of maximizing predictive accuracy on a test set $\mathcal{D}_{test} = \{(T_i, V_i, y_i)\}_{i=1}^{N_{text}}$.

Given a specific instance (T,V,y) (we omit the index for clarity), conventional trainable detectors (Chen et al. 2022; Wang et al. 2023) first encode the modalities using frozen pre-trained encoders, yielding unimodal representations \mathbf{t} (text) and \mathbf{v} (image). These are fused into a joint representation \mathbf{h} via a modality interaction module. The final prediction is computed as $p(y \mid T,V;f,\theta) \propto \exp(\mathrm{MLP}(\mathbf{h}))$, where $\mathrm{MLP}(\cdot)$ is a multi-layer perceptron. The predicted label is given by $\arg\max_y p(y \mid T,V;f,\theta)$.

The LVLM-as-Enhancer paradigm extends this setup by leveraging an LVLM $\mathcal G$ to generate explanatory rationales $R=\mathcal G(T,V)$ for the input instance. These rationales are then encoded into $\mathbf r$ and incorporated into the detection pipeline through a specialized architecture that computes an enhanced representation $\mathbf h$. While this approach has shown early success (Hu et al. 2024), the over-engineered integration strategies utilized by existing efforts may limit generalizability to diverse rationale types. For example,

EFND (Wang et al. 2024) introduces a structured module tailored to reason over debates, which may not generalize to alternative rationale formats such as sentiment-based reasoning about news veracity (Zhang et al. 2021).

2.2 DIFAR Framework

Figure 2 illustrates the overall architecture of DIFAR, a framework designed to robustly integrate multiple rationales into trainable multimodal misinformation detectors while preserving generality and scalability.

To improve compatibility with diverse rationales, DIFAR preserves the general pipeline of trainable detectors without introducing task-specific architectural changes. Given a news article (T, V, y) and a set of M LVLM-generated rationales $\{R_j\}_{j=1}^M$, we first concatenate all rationales with the original textual input to form an augmented input $\tilde{T} = [T; R_1; \ldots; R_M]$, which is then passed to the detector f.

While this approach is intuitive and detector-agnostic, it presents two key challenges:

- Input length constraints. Many detectors, such as those based on CLIP (Radford et al. 2021), have strict token limits (e.g., 77 tokens), making them unable to accommodate long concatenated inputs (Chen et al. 2022).
- Ordering sensitivity. The effectiveness of concatenated rationales may depend heavily on their order. Prior work has shown that sequence ordering significantly affects incontext learning performance (Shi et al. 2024). Exhaustively searching all permutations is computationally infeasible and unlikely to yield a universally optimal order.

To overcome these limitations, DIFAR augments the textual modality at the representation level. Specifically, we first split the concatenated input \tilde{T} into n sentences $\{\tilde{\boldsymbol{t}}_i\}_{i=1}^n$. Each sentence is independently encoded using a pretrained encoder-based language model, and the resulting representations are averaged to form the final representation:

$$\tilde{\mathbf{t}} = \frac{1}{n} \sum_{i=1}^{n} \text{encoder}(\tilde{\mathbf{t}}_i), \tag{1}$$

where $\operatorname{encoder}(\cdot)$ denotes a sentence-level encoder; we use DeBERTa (He, Gao, and Chen 2023) in our implementation.

This strategy enables the detector to process inputs of arbitrary length and removes sensitivity to rationale ordering due to the symmetric averaging operator. As a result, DI-FAR can robustly incorporate multiple rationales to enhance misinformation detection.

2.3 CoT-Based Rationale Diversification

Incorporating foundation model generated news analyses has shown promise in assessing the veracity of news articles (Nan et al. 2024; Wu, Guo, and Hooi 2024). Moreover, using multiple perspectives can provide complementary insights, which may further benefit misinformation detection. To this end, we design five chain-of-thought (CoT) prompts across three categories, **textual content**, **visual content**, and **cross-modal consistency**, to generate a diverse set of veracity-related rationales $\{\bar{R}_j\}_{j=1}^M$.

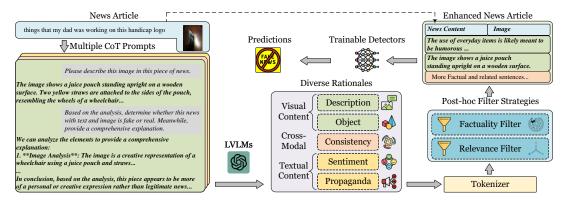


Figure 2: Overview of DIFAR. The framework integrates a simple yet effective structure that supports rationales of an arbitrary number and length. It employs five chain-of-thought prompts to promote reasoning diversity and two post-hoc refinement strategies to ensure factuality and relevance.

Each rationale \bar{R}_j is generated through a structured multiturn interaction. First, the LVLM is prompted to analyze the news article from a designated perspective (e.g., "analyze the sentiment of this news article"), yielding an intermediate response $\bar{R}_{j(0)}$. Then, the model is asked to assess the veracity of the article based on this analysis and to justify its judgment, producing $\bar{R}_{j(1)}$. The full rationale \bar{R}_j is formed by concatenating $\bar{R}_{j(0)}$ and $\bar{R}_{j(1)}$.

To encourage reasoning diversity, we design five prompts. For textual content, the prompts guide the LVLM to examine linguistic signals relevant to misinformation, including (i) *sentiment analysis* (Toughrai, Langlois, and Smaïli 2025) and (ii) *propaganda tactics* (Piskorski et al. 2023). For visual content, the prompts focus on understanding the accompanying image, specifically through (iii) *object identification* (Ma et al. 2024b) and (iv) *image description* (Abdali, Shaham, and Krishnamachari 2024). Lastly, to capture cross-modal consistency, we include (v) a prompt that evaluates the alignment between textual and visual information, following prior work (Liu et al. 2025a). Detailed prompts are provided in the Appendix.

Rather than aiming to exhaustively explore the prompt design space, we intentionally select these five representative prompts to demonstrate the potential of reasoning diversity and the adaptability of DIFAR. The prompts elicit complementary perspectives across textual, visual, and cross-modal dimensions and can be readily extended or customized to suit other reasoning needs or domains.

2.4 Post-Hoc Rationale Refinement

While it is possible to directly treat each original rationale \bar{R} as the final input, this approach suffers from two key issues: lack of factuality (Pan et al. 2023) and lack of relevance (Zheng et al. 2025). To mitigate these challenges, we apply a sentence-level filtering procedure. Specifically, we first split \bar{R} into \bar{m} sentences $\{\bar{r}_k\}_{k=1}^{\bar{m}}$ and apply two filtering strategies to obtain a refined set $\{r_k\}_{k=1}^m$ containing only factual and relevant sentences. The final rationale R is obtained by concatenating these filtered sentences.

Factuality Filter. Due to hallucinations (Ji et al. 2023), LVLMs may generate sentences with factual errors that degrade downstream detection performance. To address this, we compute a factuality score $s_f(r)$ for each sentence r and discard those with low scores. Following Min et al. (2023), we rely on an external knowledge source (specifically, Wikipedia) to support this evaluation. For each sentence r, we retrieve p candidate documents $\{d_i\}_{i=1}^p$ and define the factuality score as:

$$s_f(\mathbf{r}) = \max_{1 \le i \le p} \text{fact}(\mathbf{r} \mid \mathbf{d}_i), \tag{2}$$

where $fact(r \mid d_i)$ quantifies the factual alignment between r and document d_i . We compute this score by averaging two complementary signals: a stance classifier (Schuster, Fisch, and Barzilay 2021), which evaluates how strongly the document supports the sentence, and a summarization precision model (Feng et al. 2023), which assesses how well the sentence summarizes the document:

$$fact(\mathbf{r} \mid \mathbf{d}_i) = \frac{1}{2} (stance(\mathbf{r}, \mathbf{d}_i) + summary(\mathbf{r}, \mathbf{d}_i)).$$
 (3)

Both scores are normalized to the range [0, 1], with higher values indicating stronger factual consistency. We retain the top 50% of sentences based on the factuality scores:

$$\bar{\boldsymbol{r}}_k \in R \quad \text{if} \quad k \in \text{top-}50\%_{\ell}(s_f(\bar{\boldsymbol{r}}_{\ell})). \tag{4}$$

Relevance Filter. In some cases, the LVLM-generated rationales may diverge from the prompt and include content that is tangential or irrelevant to the source article. To ensure rationale relevance, we assess sentence-level relevance to the input text T by computing semantic similarity. Specifically, we adopt a widely used approach (Lewis et al. 2020) that utilizes an encoder-based language model to obtain sentence embeddings and measure similarity via cosine distance:

$$s_r(T, \mathbf{r}) = \cos\left(\operatorname{encoder}(T), \operatorname{encoder}(\mathbf{r})\right),$$
 (5)

where $\operatorname{encoder}(\cdot)$ is instantiated as MPNet (Song et al. 2020). Higher values indicate stronger semantic alignment between the rationale and the input article, where we also retain the top 50% of sentences.

Methods		Fake MiF.	eddit MaF.	FakeN MiF.	ewsNet MaF.	Fine MiF.	Fake MaF.	MMFak MiF.	eBench MaF.
Vanilla LVLMs	InternVL Zero-Shot InternVL Few-Shot InternVL Retrieval InternVL Self-Refine GPT-40 Zero-Shot GPT-40 Few-Shot GPT-40 Retrieval GPT-40 Self-Refine	$\begin{array}{c} 70.1_{\pm 2.4} \\ 70.1_{\pm 3.6} \\ 60.6_{\pm 3.3} \\ 59.7_{\pm 3.0} \\ 78.1_{\pm 1.5} \\ 78.9_{\pm 2.3} \\ 64.1_{\pm 3.6} \\ 77.6_{\pm 0.4} \end{array}$	$\begin{array}{c} 68.9 \pm 2.9 \\ 69.6 \pm 3.8 \\ 55.4 \pm 3.2 \\ 57.9 \pm 3.0 \\ 78.0 \pm 1.4 \\ 78.8 \pm 2.3 \\ 63.4 \pm 3.9 \\ 77.5 \pm 0.4 \end{array}$	$\begin{array}{c} 74.5{\pm}3.4 \\ 77.0{\pm}2.2 \\ 64.1{\pm}2.9 \\ 68.9{\pm}2.8 \\ 84.0{\pm}1.8 \\ 80.3{\pm}1.4 \\ 81.8{\pm}1.7 \\ 81.2{\pm}0.6 \end{array}$	$\begin{array}{c} 69.9{\pm}3.3 \\ 72.0{\pm}3.1 \\ 61.1{\pm}3.3 \\ 65.5{\pm}2.3 \\ 77.3{\pm}3.0 \\ 72.7{\pm}3.1 \\ 74.8{\pm}2.3 \\ 73.8{\pm}1.5 \end{array}$	$\begin{array}{ c c c c }\hline 70.8 \pm 2.1 \\ 71.4 \pm 2.7 \\ 70.5 \pm 1.3 \\ 67.3 \pm 2.8 \\ 75.3 \pm 3.4 \\ \hline 77.3 \pm 0.9 \\ \hline 74.5 \pm 1.8 \\ \hline 73.2 \pm 3.0 \\\hline \end{array}$	$70.7_{\pm 2.1}$ $71.4_{\pm 2.7}$ $70.4_{\pm 1.3}$ $67.3_{\pm 2.8}$ $74.5_{\pm 3.6}$ $\frac{76.8}{73.9_{\pm 1.8}}$ $72.2_{\pm 2.9}$	$\begin{array}{c} 77.5{\pm}2.5 \\ 72.4{\pm}2.7 \\ 63.6{\pm}2.9 \\ 64.6{\pm}1.7 \\ 80.9{\pm}3.4 \\ 82.3{\pm}3.1 \\ 72.9{\pm}3.4 \\ 78.2{\pm}3.3 \end{array}$	$\begin{array}{c} 77.5{\scriptstyle\pm2.4} \\ 72.2{\scriptstyle\pm2.6} \\ 62.1{\scriptstyle\pm3.0} \\ 64.6{\scriptstyle\pm1.7} \\ 80.7{\scriptstyle\pm3.4} \\ 82.2{\scriptstyle\pm3.0} \\ 72.9{\scriptstyle\pm3.4} \\ 78.1{\scriptstyle\pm3.3} \end{array}$
Enhanced LVLMs	MMD-Agent Knowledge Card	$\begin{array}{ c c c c c c }\hline 68.9_{\pm 1.9} \\ 52.1_{\pm 3.7} \\ \hline \end{array}$	$68.8_{\pm 1.9} \\ 42.4_{\pm 3.6}$	$\begin{array}{ c c c c c c }\hline 67.4_{\pm 4.2} \\ 73.8_{\pm 3.4} \\ \hline \end{array}$	$60.4_{\pm 4.5} \\ 67.7_{\pm 5.3}$	$\begin{array}{ c c c c c }\hline 64.1_{\pm 2.4} \\ 64.5_{\pm 2.2} \\ \hline \end{array}$	$64.1_{\pm 2.4} \\ 64.4_{\pm 2.2}$	$\begin{array}{ c c c c c c c c c c c c c c c c c c c$	$75.1_{\pm 4.1} \\ 56.3_{\pm 2.7}$
Trainable Detectors CAFE COOLANT		$ \begin{vmatrix} 86.0_{\pm 2.4} \\ 87.4_{\pm 2.1} \\ 86.4_{\pm 2.3} \end{vmatrix} $	$\begin{array}{c} 85.9_{\pm 2.4} \\ \underline{87.4}_{\pm 2.1} \\ 86.3_{\pm 2.3} \end{array}$	$ \begin{vmatrix} 86.6_{\pm 1.6} \\ 86.8_{\pm 0.8} \\ 85.7_{\pm 1.7} \end{vmatrix} $	$\begin{array}{c} 82.3_{\pm 1.4} \\ 82.8_{\pm 1.2} \\ 81.3_{\pm 1.9} \end{array}$	$ \begin{vmatrix} 75.7_{\pm 3.3} \\ 76.2_{\pm 2.7} \\ 76.2_{\pm 2.1} \end{vmatrix} $	$75.5_{\pm 3.4} \\ 76.0_{\pm 2.7} \\ 76.1_{\pm 2.1}$	$\begin{vmatrix} 84.3_{\pm 2.4} \\ 85.4_{\pm 2.7} \\ 83.2_{\pm 2.1} \end{vmatrix}$	$\begin{array}{c} 84.2_{\pm 2.4} \\ \underline{85.4}_{\pm 2.7} \\ 83.1_{\pm 2.1} \end{array}$
LVLM-as-Enhancer EARAM EFND		$\begin{array}{ c c c c c }\hline 82.6_{\pm 1.9} \\ 80.3_{\pm 1.3} \\ \hline \end{array}$	$82.5_{\pm 1.9} \\ 80.2_{\pm 1.2}$	$\begin{array}{ c c c c c c }\hline 82.9_{\pm 2.9} \\ 87.6_{\pm 1.0} \\ \hline \end{array}$	$77.5_{\pm 3.0} \\ \underline{84.1}_{\pm 2.0}$	$\begin{array}{ c c c c c c }\hline 73.8_{\pm 2.4} \\ 75.9_{\pm 2.3} \\\hline \end{array}$	$73.7_{\pm 2.3} \\ 75.7_{\pm 2.3}$		$78.8_{\pm 2.1} \\ 76.1_{\pm 3.2}$
DIFAR		90.8 _{±2.1}	90.8 $_{\pm 2.1}$	89.3 _{±1.9}	85.5 $_{\pm 2.7}$	81.2 _{±1.6}	81.1 _{±1.7}	90.4 _{±1.0}	90.4 _{±1.0}

Table 1: Performance of DIFAR and baselines on four widely used multimodal misinformation detection datasets. "MiF." and "MaF." denote micro- and macro-averaged F1 scores, respectively. **Bold** indicates the best performance, and <u>underline</u> indicates the second-best. DIFAR achieves consistent improvements over state-of-the-art baselines, with gains of up to 5.9%.

3 Experiments

3.1 Experimental Setup

Datasets. We evaluate DIFAR and existing baselines with four popular datasets: Fakeddit (Nakamura, Levy, and Wang 2020), FakeNewsNet (Shu et al. 2020), FineFake (Zhou et al. 2024), and MMFakeBench (Liu et al. 2025a). To obtain a robust evaluation, we conduct a five-fold evaluation and report the mean and variance of the performance. Detailed information about datasets is provided in the Appendix.

Baselines. We compare DIFAR with four types of state-of-the-art baselines: (i) Vanilla LVLMs: InternVL V3 (Zhu et al. 2025) and GPT-40 with zero-shot, few-shot, re-trieval (Lewis et al. 2020), and self-refine (Madaan et al. 2023) prompt; (ii) Enhanced LVLMs: MMD-Agent (Liu et al. 2025a) and Knowledge Card (Feng et al. 2024); (iii) Trainable detectors: CLIP (Radford et al. 2021), CAFE (Chen et al. 2022), and COOLANT (Wang et al. 2023); and (iv) LVLM-as-Enhancer: EARAM (Zheng et al. 2025) and EFND (Wang et al. 2024). Detailed information about the baselines is provided in the Appendix.

Settings. We use GPT-40 as the primary LVLM backbone for DIFAR. To ensure fair comparison, all detectors are evaluated under consistent hyperparameter settings across folds. For LVLM inference, we disable sampling by either setting the temperature to zero or configuring do_sample to False, ensuring deterministic outputs and reproducibility. Additional implementation details and experimental configurations are provided in the Appendix.

3.2 Effectiveness of DIFAR

We present the performance of DIFAR and the state-of-theart baseline in Table 1. Trainable detectors remain highly competitive. Among all baselines, supervised trainable detectors exhibit the strongest standalone performance, coming close to DIFAR across benchmarks. This highlights the effectiveness of learned representations under direct supervision and further supports the value of integrating LVLM-generated rationales into such architectures.

Existing LVLM-as-Enhancer approaches face generalization issues. Interestingly, existing LVLM-as-Enhancer methods often underperform, sometimes even falling below vanilla LVLMs. This suggests poor generalization across datasets, likely due to narrow reasoning scopes and low-quality rationales. For instance, EARAM focuses only on commonsense and complementarity, while EFND centers on debate-style veracity analysis. Both approaches lack broad perspective integration. We also observe that non-factual or off-topic content in generated rationales may mislead the detector. These findings directly motivate our design of DI-FAR, which addresses these gaps by enhancing the diversity, factuality, and relevance of LVLM-generated rationales.

DIFAR achieves state-of-the-art-performance. DIFAR consistently outperforms the strongest baseline on all four datasets, achieving gains of 1.9% to 5.9% in micro-averaged F1 score. Notably, both vanilla and enhanced LVLMs underperform on most benchmarks, suggesting that LVLMs alone struggle with factual reasoning and precise veracity assessment. These results reinforce the importance of the *LVLM-as-Enhancer* paradigm, where external reasoning is paired with trainable detectors.

3.3 Adaptability to Diverse Detectors

DIFAR is designed to be compatible with any trainable detector and any set of LVLM-generated rationales. To eval-

Models	Variants	Fakeddit	FakeNewsNet	FineFake	MMFakeBench
CLIP	Original MMD-Agent Knowledge Card EARAM EFND DIFAR	$ \begin{array}{c} 86.0_{\pm 2.4} \\ 82.5_{\pm 2.9} \; (4.1\% \downarrow) \\ 84.1_{\pm 3.6} \; (2.2\% \downarrow) \\ 82.6_{\pm 2.5} \; (4.0\% \downarrow) \\ 83.8_{\pm 2.8} \; (2.6\% \downarrow) \\ 85.3_{\pm 2.2} \; (0.8\% \downarrow) \end{array} $	$ \begin{array}{c} 86.6_{\pm 1.6} \\ 83.4_{\pm 1.7} \left(3.8\% \downarrow \right) \\ 84.5_{\pm 2.6} \left(2.5\% \downarrow \right) \\ 83.8_{\pm 2.2} \left(3.3\% \downarrow \right) \\ 84.4_{\pm 1.9} \left(2.6\% \downarrow \right) \\ 84.6_{\pm 1.9} \left(2.3\% \downarrow \right) \end{array} $	$ \begin{array}{c c} 75.7_{\pm 3.3} \\ 71.7_{\pm 2.7} (5.3\% \downarrow) \\ 74.6_{\pm 2.1} (1.5\% \downarrow) \\ 73.4_{\pm 2.3} (3.0\% \downarrow) \\ 73.4_{\pm 1.7} (3.0\% \downarrow) \\ 77.1_{\pm 2.1} (1.8\% \uparrow) \end{array} $	$ \begin{array}{c} 84.3_{\pm 2.4} \\ 82.2_{\pm 1.4} \left(2.5\% \downarrow\right) \\ 84.2_{\pm 2.5} \left(0.1\% \downarrow\right) \\ 81.8_{\pm 1.6} \left(3.0\% \downarrow\right) \\ 82.7_{\pm 1.8} \left(1.9\% \downarrow\right) \\ 85.2_{\pm 1.7} \left(1.1\% \uparrow\right) \end{array} $
CAFE	Original MMD-Agent Knowledge Card EARAM EFND DIFAR	$ \begin{array}{c} 87.4_{\pm 2.1} \\ 85.8_{\pm 2.2} \ (1.8\% \downarrow) \\ 85.4_{\pm 3.0} \ (2.3\% \downarrow) \\ 85.8_{\pm 1.3} \ (1.8\% \downarrow) \\ 86.5_{\pm 1.6} \ (1.0\% \downarrow) \\ 90.5_{\pm 2.0} \ (3.5\% \uparrow) \end{array} $	$\begin{array}{c} 86.8_{\pm 0.8} \\ 88.4_{\pm 1.1} \ (1.9\% \ \downarrow) \\ 88.5_{\pm 2.1} \ (2.0\% \ \uparrow) \\ 87.6_{\pm 0.9} \ (0.9\% \ \uparrow) \\ 88.5_{\pm 0.9} \ (2.0\% \ \downarrow) \\ 88.8_{\pm 1.6} \ (2.3\% \ \uparrow) \end{array}$	$ \begin{array}{c c} 76.2_{\pm 2.7} \\ 74.1_{\pm 2.0} & (2.8\% \downarrow) \\ 76.6_{\pm 1.8} & (0.5\% \uparrow) \\ 73.9_{\pm 2.3} & (3.0\% \downarrow) \\ 75.0_{\pm 1.8} & (1.6\% \downarrow) \\ 80.2_{\pm 1.9} & (5.2\% \uparrow) \\ \end{array} $	$ \begin{array}{c} 85.4_{\pm 2.7} \\ 84.6_{\pm 2.4} & (0.9\% \downarrow) \\ 85.3_{\pm 2.7} & (0.1\% \downarrow) \\ 84.7_{\pm 2.5} & (0.8\% \downarrow) \\ 84.9_{\pm 2.9} & (0.6\% \downarrow) \\ 88.6_{\pm 1.6} & (3.7\% \uparrow) \end{array} $
COOLANT	Original MMD-Agent Knowledge Card EARAM EFND DIFAR	$ \begin{array}{c} 86.4_{\pm 2.3} \\ 85.5_{\pm 2.5} \left(1.0\% \downarrow\right) \\ 85.6_{\pm 2.8} \left(0.9\% \downarrow\right) \\ 84.5_{\pm 2.4} \left(2.2\% \downarrow\right) \\ 86.9_{\pm 1.6} \left(0.6\% \uparrow\right) \\ 90.8_{\pm 2.1} \left(5.1\% \uparrow\right) \end{array} $	$ \begin{array}{c c} 85.7_{\pm 1.7} \\ 87.5_{\pm 1.1} & (2.1\% \uparrow) \\ 89.2_{\pm 1.7} & (4.1\% \uparrow) \\ 87.9_{\pm 1.0} & (2.6\% \uparrow) \\ 88.7_{\pm 1.3} & (3.5\% \uparrow) \\ 89.3_{\pm 1.9} & (4.3\% \uparrow) \\ \end{array} $	$ \begin{array}{c c} 76.2_{\pm 2.1} \\ 75.2_{\pm 2.3} & (1.3\% \downarrow) \\ 78.2_{\pm 2.4} & (2.6\% \uparrow) \\ 76.0_{\pm 2.7} & (0.3\% \downarrow) \\ 76.4_{\pm 2.3} & (0.3\% \uparrow) \\ 81.2_{\pm 1.6} & (6.6\% \uparrow) \end{array} $	$ \begin{array}{c} 83.2_{\pm 2.1} \\ 85.3_{\pm 2.3} \ (2.5\% \uparrow) \\ 85.0_{\pm 2.9} \ (2.2\% \uparrow) \\ 85.8_{\pm 1.5} \ (3.1\% \uparrow) \\ 85.1_{\pm 2.6} \ (2.3\% \uparrow) \\ 90.4_{\pm 1.0} \ (8.7\% \uparrow) \\ \end{array} $

Table 2: Micro-averaged F1 scores of trainable detectors enhanced with DIFAR and baseline methods. The best performance in each setting is highlighted. DIFAR improves detector performance by up to 8.7%, demonstrating its effectiveness in addressing the limitations of rationale diversity, factuality, and relevance.

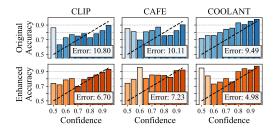


Figure 3: Calibration performance of existing trainable detectors with and without DIFAR enhancement. "Error" refers to the expected calibration error (ECE, $\times 100$), where lower values indicate better calibration. DIFAR not only improves detection accuracy but also enhances the reliability of confidence estimates.

uate its adaptability, we assess the effectiveness of DIFARgenerated rationales when integrated with various detectors.

As shown in Table 2, we draw three key observations. (i) D1FAR significantly improves the performance of all trainable detectors, achieving gains of up to 8.7%. It demonstrates its effectiveness in addressing key limitations of the LVLM-as-Enhancer paradigm and enhancing detection performance across diverse architectures. (ii) Performance gains gradually decrease across COOLANT, CAFE, and CLIP, with a slight performance drop on CLIP after enhancement. It suggests that more complex detector architectures may be more effective at extracting signal from rationales. It trend aligns with prior work emphasizing the role of architectural design in driving performance. (iii) Rationales generated by baseline methods often fail to improve detector performance and, in some cases, lead to degradation. It

highlights the importance of rationale quality, without sufficient diversity, factuality, and relevance, even strong detectors cannot reliably benefit from external explanations.

Beyond accuracy, we also evaluate the *credibility* of the detectors before and after enhancement using DIFAR. We quantify this using expected calibration error (ECE) (Guo et al. 2017). As shown in Figure 3, detectors become better calibrated after incorporating DIFAR, with reductions in ECE of up to 47.5%. These results indicate that DIFAR not only improves prediction performance but also enhances the reliability of model confidence estimates.

3.4 Ablation Study

We conduct an ablation study to evaluate the contribution of each component in DIFAR. Specifically, we consider four settings: (i) replacing the five diverse CoT prompts with a single fixed CoT prompt; (ii) removing both the factuality and relevance filtering strategies; (iii) prompting the LVLM with a generic instruction to generate rationales (i.e., "Determine whether this news with text and image is fake or real. Meanwhile, provide a comprehensive explanation."); and (iv) replacing GPT-40 with the open-sourced InternVL V3 model. We report results on the FineFake dataset in Table 6, with full results provided in the Appendix. The findings show that detectors fail to benefit from rationales generated by a vanilla prompt, underscoring the importance of rationale diversity, factuality, and relevance. Moreover, ablating any individual component of DIFAR leads to a performance drop in most cases, with decreases of up to 5.8%. These results confirm that each module in DIFAR plays an essential role in overcoming the limitations of the LVLMas-Enhancer paradigm.

Variants	CLIP	CAFE	COOLANT
DIFAR	$77.1_{\pm 2.1}$	80.2 _{±1.9}	$81.2_{\pm 1.6}$
w/o Multiple	$72.6_{\pm 3.8}$ $5.8\% \downarrow$ $72.7_{\pm 3.5}$	$78.3_{\pm 1.9}$ $2.4\% \downarrow$ $78.0_{\pm 2.3}$	$79.6_{\pm 1.4}$ $2.0\% \downarrow$
w/o Filter	$5.7\% \downarrow$	$\begin{array}{c c} 18.0\pm2.3 \\ 2.7\% \downarrow \end{array}$	$78.9_{\pm 1.4} \\ 2.8\% \downarrow$
w/ Vanilla	$73.0_{\pm 2.1} \\ 5.3\% \downarrow$	$73.8_{\pm 2.3}$ $8.0\% \downarrow$	$76.7_{\pm 2.4} \\ 5.5\% \downarrow$
w/ InternVL	$77.5_{\pm 2.2}$ $0.5\% \uparrow$	$78.8_{\pm 1.3}$ $1.7\% \downarrow$	$79.9_{\pm 2.3} \\ 1.6\% \downarrow$

Table 3: Ablation study of DIFAR. "w/o Multiple" uses a single specific CoT prompt instead of five; "w/o Filter" removes the factuality and relevance filters; "w/ Vanilla" uses rationales generated from a simple prompt; and "w/ InternVL" replaces GPT-40 with InternVL V3. Results show that each component contributes to performance gains.

4 Rationale Quality Analysis

We further evaluate whether DIFAR effectively addresses the challenges of limited diversity, factuality, and relevance in generated rationales.

4.1 Overall Helpfulness Evaluation

We begin by assessing the overall quality of rationales generated by DIFAR. To this end, we conduct a human evaluation with three experts in misinformation-related topics. Four types of rationales are selected for comparison: (i) **DIFAR**: rationales generated by DIFAR before filtering; (ii) **Baseline**: rationales generated by EFND; (iii) **Single**: rationales generated from a single randomly selected perspective; and (iv) **Filtered**: rationales produced by DIFAR after applying factuality and relevance filters.

We conduct four pairwise comparisons: DIFAR vs. Baseline, DIFAR vs. Single, Filtered vs. Baseline, and Filtered vs. Single. For each pair, experts are asked to judge which rationale is more helpful for verifying the veracity of the news article, or to indicate if the two are indistinguishable. Final decisions are determined via majority vote. Details of the evaluation protocol are provided in the Appendix. Fleiss' Kappa across all judgments is 0.34, indicating a fair level of inter-rater agreement.

Results in Figure 4 show that DIFAR significantly outperforms both the Baseline and Single settings, demonstrating that it produces more useful rationales for human misinformation assessment. However, after post-hoc filtering, the advantage of DIFAR is reduced and, in some cases, performs worse than the Single baseline. We speculate that while filtering improves factuality and relevance, it may also degrade the fluency and coherence of the rationales, limiting their interpretability for human readers.

4.2 Fine-Grained Quality Evaluation

Diversity. DIFAR employs five chain-of-thought (CoT) prompts to capture signals from diverse reasoning perspectives. As a baseline, we compare against EFND, which generates rationales from only two perspectives. We begin by

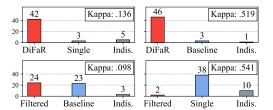


Figure 4: Human voting results of generated rationales via pairwise comparison. "Kappa" denotes the Fleiss' Kappa score among three experts, and "Indis." indicates the proportion of instances where the two rationales were judged indistinguishable. DIFAR produces the most helpful rationales for veracity assessment.

Dataset	Orig.	Consis.	Textual	Visual
Fakeddit	$90.8_{\pm 2.1}$	$\begin{array}{ c c c } 89.5_{\pm 1.7} \\ 1.4\% \downarrow \end{array}$	$\begin{array}{ c c c c } 88.8_{\pm 3.1} \\ 2.2\% \downarrow \end{array}$	$\begin{array}{ c c } 90.2_{\pm 1.5} \\ 0.7\% \downarrow \end{array}$
FakeNewsNet	89.3 _{±1.9}	$\begin{array}{ c c c c } 88.8_{\pm 1.6} \\ 0.6\% \downarrow \end{array}$	$\begin{vmatrix} 88.9_{\pm 2.2} \\ 0.5\% \downarrow \end{vmatrix}$	$\begin{array}{ c c } 89.3_{\pm 2.0} \\ 0.0\% \downarrow \end{array}$
FineFake	81.2 _{±1.6}	$\begin{array}{ c c c } 80.0_{\pm 1.7} \\ 1.5\% \downarrow \end{array}$	$\begin{array}{ c c } 79.4_{\pm 1.7} \\ 2.2\% \downarrow \end{array}$	$\begin{array}{ c c } 80.8_{\pm 1.5} \\ 0.5\% \downarrow \end{array}$
MMFakeBench	90.4 _{±1.0}	$\begin{array}{ c c }\hline 91.0_{\pm 1.4} \\ 0.7\% \uparrow \\ \end{array}$	$88.1_{\pm 2.8}$ $2.5\% \downarrow$	$89.6_{\pm 2.0}$ $0.9\% \downarrow$

Table 4: Ablation study on CoT prompt categories, where only rationales from a single prompt type are retained. "Orig." denotes the original performance of DIFAR using all five prompts, and "Consis." refers to the variant using only the cross-modal consistency prompt. We report micro-averaged F1 scores and the corresponding performance changes. Results show that diverse rationales generally outperform those from any single perspective.

computing the proportion of distinct tokens in rationales generated on the FineFake dataset. DIFAR achieves a distinct token ratio of 0.904, substantially higher than EFND's 0.406, suggesting broader lexical coverage.

To further quantify diversity, we analyze token frequency and inter-prompt similarity. Specifically, we use infinigram (Liu et al. 2024a) to measure token frequency and BERTScore (Zhang et al. 2020) to compute pairwise similarity between rationales generated from different prompts. Lower values on both metrics indicate greater diversity. As shown in the Appendix, DIFAR achieves a lower average similarity (0.57 vs. 0.66) and lower token frequency $(2.41\times10^6~{\rm vs.}~2.85\times10^6)$, confirming that it produces more lexically diverse outputs.

We also conduct an ablation study with COOLANT to evaluate the impact of different CoT prompt types. Table 4 shows that removing any single prompt category leads to performance drop of up to 2.5%, validating the importance of incorporating multiple reasoning perspectives for misinformation detection. These findings support our design of rationale diversification as a core component of DIFAR.

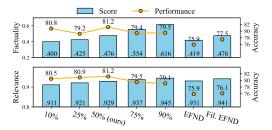


Figure 5: Performance of DIFAR under different filtering thresholds, along with corresponding factuality and relevance scores. "Fil. EFND" denotes the EFND baseline after filtering out the bottom 25% of sentences based on scores. It shows that moderate filtering improves performance, while overly aggressive filtering may reduce effectiveness.

Factuality and Relevance. DIFAR employs two post-hoc filtering strategies to improve the factuality and relevance of generated rationales. To investigate their effect, we vary the filtering threshold and evaluate both model performance and average factuality/relevance scores, using EFND as a baseline for comparison. Figure 5 shows that increasing the filtering threshold results in higher factuality and relevance scores. However, the overall detection performance of DIFAR does not continue to increase proportionally. We speculate that overly aggressive filtering, while improving quality scores, may remove semantically rich content, thereby weakening the enhancement signal provided to detectors.

Interestingly, we observe that EFND and DIFAR with a low filtering threshold (e.g., 10%) achieve similar factuality and relevance scores, yet their detection performance differs significantly. We attribute this to the presence of low-quality sentences in EFND's rationales, which likely mislead the detector. To test this, we apply our filtering strategy to EFND and remove the bottom 25% of its sentences by factuality and relevance score. This leads to a measurable performance gain, supporting the effectiveness of our filtering design. Moreover, unlike EFND, DIFAR generates rationales from diverse prompts, offering broader perspectives and richer semantic coverage. This diversity allows DIFAR to retain informative content even after filtering, contributing to its superior performance.

4.3 Case Study

We analyze a representative example from the dataset to illustrate how DIFAR contributes to improved misinformation detection. The generated rationales and corresponding model predictions are shown in Figure 6. In this case, the original COOLANT model produces an incorrect prediction. Although a single rationale correctly identifies the misinformation, incorporating it into COOLANT does not lead to a correct prediction. In contrast, when COOLANT is enhanced with the full set of diverse rationales and non-factual or irrelevant sentences are filtered out, the model successfully detects the misinformation. This example demonstrates that DIFAR provides richer and more focused semantic signals, which meaningfully support downstream detection.

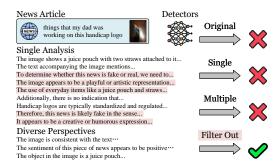


Figure 6: Case study of predictions using different variants of DIFAR with COOLANT as the base detector. The original detector fails to identify the misinformation, while enhancement with DIFAR enables a correct prediction.

5 Related Work

Multimodal misinformation detectors typically encode textual content and visual content using pretrained encoders, followed by architectures designed to model cross-modal interactions (Tonglet, Moens, and Gurevych 2024; Tong et al. 2024; Zhang et al. 2024; Lu, Tong, and Ye 2025; Cao et al. 2025; Li et al. 2025c; Yu et al. 2025; Feng et al. 2025). With the rise of LVLMs, early work directly employs LVLMs as backbones to identify misinformation (Lucas et al. 2023; Gabriel et al. 2024; Huang et al. 2024; Liu et al. 2025b; Chen and Zhang 2025; Li et al. 2025a; Wu et al. 2025). However, these models often suffer from hallucinations and lack factual grounding (Hu et al. 2024), limiting their effectiveness. Thus, LVLM-as-Enhancer paradigm is proposed. This paradigm first design prompts to generate external textual content, namely, explanations or rationales, to provide rich semantic information (Saha and Srihari 2024; Liu et al. 2024b), such as stance (Choi et al. 2025), propagation (Liu et al. 2024c), and entities (Ma et al. 2024a). They then design a trainable module to capture the semantic information to enhance performance (Zhang et al. 2025; Zhou et al. 2025). In this work, we identify key limitations of existing LVLM-generated rationales, specifically, their lack of diversity, factuality, and relevance, and propose DIFAR, a general framework that addresses these challenges through multiperspective prompting and post-hoc filtering.

6 Conclusion

We propose DIFAR, a simple yet effective framework under the *LVLM-as-Enhancer* paradigm that seamlessly adapts to multiple rationales without requiring structural changes to the detector. It employs five chain-of-thought prompts to encourage diverse reasoning and two post-hoc filtering strategies to ensure factuality and relevance. Extensive experiments show that DIFAR achieves state-of-the-art performance and could significantly enhance existing trainable detectors. Further analyses, including human evaluations, confirm that DIFAR successfully enhances rationale diversity, factuality, and relevance.

References

- Abdali, S.; Shaham, S.; and Krishnamachari, B. 2024. Multi-modal misinformation detection: Approaches, challenges and opportunities. *ACM Computing Surveys*, 57(3): 1–29.
- Cao, B.; Wu, Q.; Cao, J.; Liu, B.; and Gui, J. 2025. External Reliable Information-enhanced Multimodal Contrastive Learning for Fake News Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 31–39.
- Chen, C.; and Zhang, S. 2025. RetrieverGuard: Empowering Information Retrieval to Combat LLM-Generated Misinformation. In *Findings of the Association for Computational Linguistics: NAACL* 2025, 4399–4411.
- Chen, Y.; Li, D.; Zhang, P.; Sui, J.; Lv, Q.; Tun, L.; and Shang, L. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM web conference* 2022, 2897–2905.
- Choi, E. C.; Balasubramanian, A.; Qi, J.; and Ferrara, E. 2025. Limited effectiveness of llm-based data augmentation for covid-19 misinformation stance detection. In *Companion Proceedings of the ACM on Web Conference* 2025, 934–937.
- Feng, S.; Balachandran, V.; Bai, Y.; and Tsvetkov, Y. 2023. FactKB: Generalizable Factuality Evaluation using Language Models Enhanced with Factual Knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, 933–952.
- Feng, S.; Shi, W.; Bai, Y.; Balachandran, V.; He, T.; and Tsvetkov, Y. 2024. Knowledge Card: Filling LLMs' Knowledge Gaps with Plug-in Specialized Language Models. In *ICLR*.
- Feng, Y.; Li, W.; Wang, Y.; Wang, J.; Liu, F.; and Han, Z. 2025. Contradicted in Reliable, Replicated in Unreliable: Dual-Source Reference for Fake News Early Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 23896–23904.
- Gabriel, S.; Lyu, L.; Siderius, J.; Ghassemi, M.; Andreas, J.; and Ozdaglar, A. 2024. MisinfoEval: Generative AI in the Era of "Alternative Facts". In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 8566–8578.
- Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. In *International conference on machine learning*, 1321–1330. PMLR.
- He, P.; Gao, J.; and Chen, W. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023.*
- Hu, B.; Sheng, Q.; Cao, J.; Shi, Y.; Li, Y.; Wang, D.; and Qi, P. 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, 22105–22113.

- Huang, Y.; Shu, K.; Yu, P. S.; and Sun, L. 2024. From creation to clarification: ChatGPT's journey through the fake news quagmire. In *Companion Proceedings of the ACM Web Conference* 2024, 513–516.
- Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12): 1–38.
- Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474.
- Li, D.; Li, F.; Song, B.; Tang, L.; and Zhou, W. 2025a. IM-RRF: Integrating Multi-Source Retrieval and Redundancy Filtering for LLM-based Fake News Detection. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 9127–9142.
- Li, F.; Wu, J.; He, C.; and Zhou, W. 2025b. CMIE: Combining MLLM Insights with External Evidence for Explainable Out-of-Context Misinformation Detection. In *Findings of the Association for Computational Linguistics: ACL 2025*, 9342–9354.
- Li, M.; Zhang, Y.; Xu, H.; Li, X.; Gao, C.; and Wang, Z. 2025c. Learning complex heterogeneous multimodal fake news via social latent network inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 433–441.
- Li, Y.; Guerin, F.; and Lin, C. 2024. Latesteval: Addressing data contamination in language model evaluation through dynamic and time-sensitive test construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18600–18607.
- Liu, J.; Min, S.; Zettlemoyer, L.; Choi, Y.; and Hajishirzi, H. 2024a. Infini-gram: Scaling Unbounded n-gram Language Models to a Trillion Tokens. In *First Conference on Language Modeling*.
- Liu, X.; Li, P.; Huang, H.; Li, Z.; Cui, X.; Liang, J.; Qin, L.; Deng, W.; and He, Z. 2024b. Fka-owl: Advancing multimodal fake news detection through knowledge-augmented lylms. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 10154–10163.
- Liu, X.; Li, Z.; Li, P.; Huang, H.; Xia, S.; Cui, X.; Huang, L.; Deng, W.; and He, Z. 2025a. MMFakeBench: A Mixed-Source Multimodal Misinformation Detection Benchmark for LVLMs. In *The Thirteenth International Conference on Learning Representations, ICLR* 2025, Singapore, April 24-28, 2025.
- Liu, Y.; Chen, X.; Zhang, X.; Gao, X.; Zhang, J.; and Yan, R. 2024c. From Skepticism to Acceptance: Simulating the Attitude Dynamics Toward Fake News. In *IJCAI*.
- Liu, Z.; Zhang, X.; Yang, K.; Xie, Q.; Huang, J.; and Ananiadou, S. 2025b. Fmdllama: Financial misinformation detection based on large language models. In *Companion Proceedings of the ACM on Web Conference 2025*, 1153–1157.

- Lu, W.; Tong, Y.; and Ye, Z. 2025. DAMMFND: Domain-Aware Multimodal Multi-view Fake News Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 559–567.
- Lucas, J.; Uchendu, A.; Yamashita, M.; Lee, J.; Rohatgi, S.; and Lee, D. 2023. Fighting Fire with Fire: The Dual Role of LLMs in Crafting and Detecting Elusive Disinformation. In 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, 14279–14305. Association for Computational Linguistics (ACL).
- Ma, X.; Zhang, Y.; Ding, K.; Yang, J.; Wu, J.; and Fan, H. 2024a. On fake news detection with LLM enhanced semantics mining. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 508–521.
- Ma, Z.; Luo, M.; Guo, H.; Zeng, Z.; Hao, Y.; and Zhao, X. 2024b. 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)*, 5809–5821.
- Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36: 46534–46594.
- Mallen, A.; Asai, A.; Zhong, V.; Das, R.; Khashabi, D.; and Hajishirzi, H. 2023. When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9802–9822.
- Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; Yih, W.-t.; Koh, P.; Iyyer, M.; Zettlemoyer, L.; and Hajishirzi, H. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 12076–12100.
- Nakamura, K.; Levy, S.; and Wang, W. Y. 2020. Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020,* 6149–6157. European Language Resources Association.
- Nan, Q.; Sheng, Q.; Cao, J.; Hu, B.; Wang, D.; and Li, J. 2024. Let silence speak: Enhancing fake news detection with generated comments from large language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 1732–1742.
- Pan, Y.; Pan, L.; Chen, W.; Nakov, P.; Kan, M.-Y.; and Wang, W. 2023. On the Risk of Misinformation Pollution with Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1389–1403.
- Piskorski, J.; Stefanovitch, N.; Da San Martino, G.; and Nakov, P. 2023. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, 2343–2361.

- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.
- Saha, S.; and Srihari, R. K. 2024. Integrating argumentation and hate-speech-based techniques for countering misinformation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 11109–11124.
- Schuster, T.; Fisch, A.; and Barzilay, R. 2021. Get Your Vitamin C! Robust Fact Verification with Contrastive Evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 624–643.
- Shi, W.; Min, S.; Lomeli, M.; Zhou, C.; Li, M.; Lin, X. V.; Smith, N. A.; Zettlemoyer, L.; Yih, W.-t.; and Lewis, M. 2024. In-Context Pretraining: Language Modeling Beyond Document Boundaries. In *ICLR*.
- Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 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.
- Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33: 16857–16867.
- Tahmasebi, S.; Müller-Budack, E.; and Ewerth, R. 2024. Multimodal misinformation detection using large vision-language models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 2189–2199.
- Tong, Y.; Lu, W.; Zhao, Z.; Lai, S.; and Shi, T. 2024. MMDFND: Multi-modal multi-domain fake news detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 1178–1186.
- Tonglet, J.; Moens, M. F.; and Gurevych, I. 2024. "Image, Tell me your story!" Predicting the original meta-context of visual misinformation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 7845–7864.
- Toughrai, Y.; Langlois, D.; and Smaïli, K. 2025. Fake News Detection via Intermediate-Layer Emotional Representations. In *Companion Proceedings of the ACM on Web Conference* 2025, 2680–2684.
- Wan, H.; Feng, S.; Tan, Z.; Wang, H.; Tsvetkov, Y.; and Luo, M. 2024. DELL: Generating Reactions and Explanations for LLM-Based Misinformation Detection. In *Findings of the Association for Computational Linguistics ACL* 2024, 2637–2667.
- Wang, B.; Ma, J.; Lin, H.; Yang, Z.; Yang, R.; Tian, Y.; and Chang, Y. 2024. Explainable fake news detection with large language model via defense among competing wisdom. In *Proceedings of the ACM Web Conference* 2024, 2452–2463.
- Wang, L.; Zhang, C.; Xu, H.; Xu, Y.; Xu, X.; and Wang, S. 2023. Cross-modal contrastive learning for multimodal fake

- news detection. In *Proceedings of the 31st ACM international conference on multimedia*, 5696–5704.
- Wu, J.; Guo, J.; and Hooi, B. 2024. Fake news in sheep's clothing: Robust fake news detection against LLM-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, 3367–3378.
- Wu, J.; Li, F.; Kan, M.-Y.; and Hooi, B. 2025. Seeing Through Deception: Uncovering Misleading Creator Intent in Multimodal News with Vision-Language Models. *arXiv* preprint arXiv:2505.15489.
- Xu, R.; Qi, Z.; Guo, Z.; Wang, C.; Wang, H.; Zhang, Y.; and Xu, W. 2024. Knowledge Conflicts for LLMs: A Survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 8541–8565.
- Yu, X.; Sheng, Z.; Lu, W.; Luo, X.; and Zhou, J. 2025. Racmc: Residual-aware compensation network with multigranularity constraints for fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 986–994.
- Zhang, C.; Feng, Z.; Zhang, Z.; Qiang, J.; Xu, G.; and Li, Y. 2025. Is LLMs Hallucination Usable? LLM-based Negative Reasoning for Fake News Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1031–1039.
- Zhang, Q.; Liu, J.; Zhang, F.; Xie, J.; and Zha, Z.-J. 2024. Natural language-centered inference network for multi-modal fake news detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, 2542–2550.
- Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.
- Zhang, X.; Cao, J.; Li, X.; Sheng, Q.; Zhong, L.; and Shu, K. 2021. Mining dual emotion for fake news detection. In *Proceedings of the web conference 2021*, 3465–3476.
- Zheng, X.; Zeng, Z.; Wang, H.; Bai, Y.; Liu, Y.; and Luo, M. 2025. From predictions to analyses: Rationale-augmented fake news detection with large vision-language models. In *Proceedings of the ACM on Web Conference* 2025, 5364–5375.
- Zhou, Z.; Zhang, X.; Tan, S.; Zhang, L.; and Li, C. 2025. Collaborative evolution: Multi-round learning between large and small language models for emergent fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 1210–1218.
- Zhou, Z.; Zhang, X.; Zhang, L.; Liu, J.; Wang, S.; Liu, Z.; Zhang, X.; Li, C.; and Yu, P. S. 2024. Finefake: A knowledge-enriched dataset for fine-grained multi-domain fake news detection. *arXiv* preprint *arXiv*:2404.01336.
- Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Prompts of DIFAR

For the diverse perspectives of news articles, we employ the following prompts (assuming that we first provide the textual and visual content to LVLMs):

- Sentiment. Please analyze the sentiment of this piece of news.
- **Propaganda.** Please analyze the propaganda tactics utilized in this piece of news.
- Consistency. Please analyze the consistency between the text and the image of this piece of news.
- **Object.** Please analyze the object that appears in the image of this piece of news.
- Description. Please describe this image in this piece of news.

Finally, we prompt LVLMs to judge the veracity by prompt Based on the analysis, determine whether this news with text and image is fake or real. Meanwhile, provide a comprehensive explanation.

B Datasets

We evaluate DIFAR and existing baselines with four widely used multimodal misinformation detection datasets, where each news article contains text content and an image, including human-written and machine-generated multimodal news articles:

- Fakeddit (Nakamura, Levy, and Wang 2020) is a multimodal dataset consisting of over 1 million samples from multiple categories of fake news, where the source is Reddit. Each instance is labeled according to 2-way, 3-way, and 6-way classification categories through distant supervision. We employ 2-way labels, namely, consider it a binary classification task.
- FakeNewsNet (Shu et al. 2020) contains two comprehensive data sets: Politifact, which includes political news, and Gossipcop, which includes entertainment news. Each instance contains diverse features such as news content, social context, and spatiotemporal information. We only employ the textual content and visual content in news articles.
- FineFake (Zhou et al. 2024) includes 16,909 news articles covering six semantic topics and eight platforms.
 Each instance contains multi-modal content, potential social context, semi-manually verified common knowledge, and fine-grained annotations that surpass conventional binary labels. We only employ the textual content and visual content in news articles, and leverage the binary labels.
- MMFakeBench (Liu et al. 2025a) includes three critical sources: textual veracity distortion, visual veracity distortion, and cross-modal consistency distortion, along with 12 sub-categories of misinformation forgery types. It covers machine-generated news articles, including the generated textual content and visual content. We leverage the binary labels.

Datasets	# Instances	# Fake	# Real	
Fakeddit	1,000	500	500	
FakeNewsNet	985	275	710	
FineFake	1,000	500	500	
MMFakeBench	1,000	500	500	

Table 5: The statistics of datasets.

To ensure a fair comparison, we sample from the original datasets to create balanced subsets with a similar number of real and fake instances. The statistics of the datasets are shown in Table 5. Additionally, to enhance the robustness of the results, we randomly partition each dataset into five equal folds for cross-validation.

C Baselines

We compare DIFAR with four types of state-of-the-art baselines.

Vanilla LVLMs employs simple prompts to prompt GPT-40 and InternVL V3 to conduct misinformation detection. We employ the following prompt style:

 Zero-shot directly prompts LVLMs to obtain the results, where the prompt is as follows:

Text: Text Image: Image

Based on the above text and image, please judge whether this piece of news is real or fake. Just output real or fake without any explanations.

 Few-shot additional provides a random sample of instances with the label to LVLMs, where the prompt is as follows:

Text: Text Image: Image Label: Label

Based on the above examples, please judge whether the following piece of news is real or fake. Just output real or fake without any explanations.

Text: Text Image: Image Label:

 Retrieval (Lewis et al. 2020) first retrieves the three most related news from BBC news resource (Li, Guerin, and Lin 2024) using bm25. It then provided the related news to LVLMs as external content, where the prompt is as follows:

Related News:

Retrieval news articles

Text: Text Image: Image

Based on the above text and image, please judge whether this piece of news is real or fake. Just output real or fake without any explanations.

• Self-Refine (Madaan et al. 2023) prompts LVLMs to check whether the answer is correct and refine the predictions, where the prompt is as follows:

Text: Text Image: Image

Based on the above text and image, please judge whether this piece of news is real or fake. Just output real or fake without any explanations.

The first prediction

Is the answer correct? If not, give your answer.

Enhanced LVLMs design a framework containing multiple prompts to enhance the ability of LVLMs, including:

- MMD-Agent (Liu et al. 2025a) could integrate the reasoning, action, and tool-use capabilities of LVLM agents to enhance the generalization and improve the detection performance.
- Knowledge Card (Feng et al. 2024) proposes a modular framework to plug in new factual and relevant knowledge into large language models. We employ the bottom-up approach and the advised cards to enhance LVLMs.

Trainable detectors represent the traditional multimodal misinformation detectors that require to learn the parameters, including:

- CLIP (Radford et al. 2021) is a widely used backbone to encode textual and visual information. We employ MLP layers to classify misinformation after obtaining the representations.
- CAFE (Chen et al. 2022) is an ambiguity-aware multimodal fake news detection method, containing a crossmodal alignment module, a cross-modal ambiguity learning module, and a cross-modal fusion module.
- COOLANT (Wang et al. 2023) is a cross-modal contrastive learning framework for multimodal fake news detection, including an auxiliary task, a cross-modal fusion module, and an attention mechanism with an attention guidance module.

LVLM-as-enhancer detectors contain two representative baselines with this paradigm, including:

- EARAM (Zheng et al. 2025) could use multimodal small language models to extract useful rationales from the multi-perspective analyses of LVLMs. The LVLMs analyze the common sense and the complementarity of news articles.
- EFND (Wang et al. 2024) designs a prompt-based module that utilizes a large language model to generate justifications by inferring reasons towards two possible veracities. It also proposes a specific trainable module to capture the signals from two perspectives.

We believe these four categories of multimodal misinformation detectors encompass most existing approaches and represent the advances in this field.

D Settings

Inferences of LVLMs. We set the temperature as 0 or set the do sample as false to ensure reproducibility. For GPT-40, we employ the official API. For InternVL V3, we employ it using two RTX4090 GPUs with 24GB of memory.

Models	Variants	Fakeddit	FakeNewsNet	FineFake	MMFakeBench
CLIP	DIFAR w/o Multiple w/o Filter w/ Vanilla w/ InternVL	$ \begin{vmatrix} 85.3_{\pm 2.2} \\ 84.8_{\pm 2.2} & (0.6\% \downarrow) \\ 82.8_{\pm 3.0} & (2.9\% \downarrow) \\ 83.1_{\pm 3.2} & (2.6\% \downarrow) \\ 83.7_{\pm 3.5} & (1.9\% \downarrow) \end{vmatrix} $	$ \begin{vmatrix} 84.6_{\pm 1.9} \\ 83.6_{\pm 2.4} & (1.2\% \downarrow) \\ 83.7_{\pm 2.6} & (1.1\% \downarrow) \\ 83.7_{\pm 2.9} & (1.1\% \downarrow) \\ 84.8_{\pm 1.9} & (0.2\% \uparrow) \end{vmatrix} $	$ \begin{array}{c c} 77.1_{\pm 2.1} \\ 72.6_{\pm 3.8} \ (5.8\% \downarrow) \\ 72.7_{\pm 3.5} \ (5.7\% \downarrow) \\ 73.0_{\pm 2.1} \ (5.3\% \downarrow) \\ 77.5_{\pm 2.2} \ (0.5\% \uparrow) \end{array} $	$ \begin{vmatrix} 85.2_{\pm 1.7} \\ 81.8_{\pm 1.6} & (4.0\% \downarrow) \\ 82.0_{\pm 1.9} & (3.8\% \downarrow) \\ 81.7_{\pm 2.0} & (4.1\% \downarrow) \\ 85.2_{\pm 2.0} & (0.0\% \downarrow) \end{vmatrix} $
CAFE	DIFAR w/o Multiple w/o Filter w/ Vanilla w/ InternVL	$ \begin{array}{c} 90.5{\scriptstyle \pm 2.0} \\ 88.9{\scriptstyle \pm 2.3} \; (1.8\% \downarrow) \\ 90.8{\scriptstyle \pm 2.7} \; (0.3\% \uparrow) \\ 83.6{\scriptstyle \pm 2.6} \; (7.6\% \downarrow) \\ 85.7{\scriptstyle \pm 2.2} \; (5.3\% \downarrow) \end{array} $	$ \begin{array}{c} 88.8{\scriptstyle \pm 1.6} \\ 88.5{\scriptstyle \pm 1.7} \; (0.3\% \downarrow) \\ 88.9{\scriptstyle \pm 2.2} \; (0.1\% \uparrow) \\ 88.0{\scriptstyle \pm 1.5} \; (0.9\% \downarrow) \\ 88.7{\scriptstyle \pm 1.0} \; (0.1\% \downarrow) \end{array} $	$ \begin{array}{c} 80.2_{\pm 1.9} \\ 78.3_{\pm 1.9} \ (2.4\% \downarrow) \\ 78.0_{\pm 2.3} \ (2.7\% \downarrow) \\ 73.8_{\pm 2.3} \ (8.0\% \downarrow) \\ 78.8_{\pm 1.3} \ (1.7\% \downarrow) \end{array} $	$ \begin{array}{c} 88.6_{\pm 1.6} \\ 87.5_{\pm 2.9} \ (1.2\% \ \downarrow) \\ 87.6_{\pm 1.6} \ (1.1\% \ \downarrow) \\ 83.5_{\pm 1.9} \ (5.8\% \ \downarrow) \\ 86.4_{\pm 3.1} \ (2.5\% \ \downarrow) \end{array} $
COOLANT	DIFAR w/o Multiple w/o Filter w/ Vanilla w/ InternVL	$ \begin{vmatrix} 90.8_{\pm 2.1} \\ 87.8_{\pm 1.2} (3.3\% \downarrow) \\ 89.9_{\pm 2.5} (1.0\% \downarrow) \\ 83.4_{\pm 3.0} (8.1\% \downarrow) \\ 85.7_{\pm 1.9} (5.6\% \downarrow) \end{vmatrix} $	$ \begin{vmatrix} 89.3_{\pm 1.9} \\ 88.7_{\pm 1.3} & (0.7\% \downarrow) \\ 89.2_{\pm 2.0} & (0.1\% \downarrow) \\ 87.1_{\pm 1.6} & (2.5\% \downarrow) \\ 87.6_{\pm 1.5} & (1.9\% \downarrow) \end{vmatrix} $	$ \begin{vmatrix} 81.2_{\pm 1.6} \\ 79.6_{\pm 1.4} & (2.0\% \downarrow) \\ 78.9_{\pm 1.4} & (2.8\% \downarrow) \\ 76.7_{\pm 2.4} & (5.5\% \downarrow) \\ 79.9_{\pm 2.3} & (1.6\% \downarrow) \end{vmatrix} $	$ \begin{vmatrix} 90.4_{\pm 1.0} \\ 87.9_{\pm 3.3} & (2.8\% \downarrow) \\ 88.5_{\pm 2.2} & (2.1\% \downarrow) \\ 84.2_{\pm 3.1} & (6.9\% \downarrow) \\ 85.3_{\pm 3.0} & (5.6\% \downarrow) \end{vmatrix} $

Table 6: The ablation study of DIFAR. It illustrates that each module of DIFAR could improve the detection performance.

Hyperparameter	CLIP	CAFE	COOLANT
Optimizer	Adam	Adam	AdamW
Weight Decay	1e-5	1e-5	5e-4
Dropout	0.5	0.5	0.5
Learning Rate	1e-3	1e-3	1e-4
Batch Size	256	32	64

Table 7: The hyperparameters of the baselines.

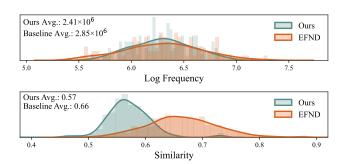


Figure 7: Token frequency and similarity distributions of DIFAR and a EFND. DIFAR presents a lower frequency and similarity, proving the diversity of the generated rationales.

Trainable Detectors To obtain a fair comparison, we set the hyperparameters the same for each detector in each fold. Every baseline can be held in one RTX4090 GPU with 24GB of memory. Meanwhile, we run each baseline five times and report the run with the best micro f1-score for each fold. Table 7 presents the hyperparameters of each baseline. We also provide the related codes in the supplementary material.

E Ablation Study

We present the complete ablation study in Table 6.

F Human Evaluation

Evaluation Guideline Document. We first provide a brief guideline document for each expert. The guideline content is as follows:

Large vision language models are proven helpful in enhancing multimodal misinformation detection.

A widely used paradigm, LVLM-as-enhancer, proposes to generate external explanations/rationales to enhance the performance of trainable detectors.

However, the explanations/rationales generated by LVLMs suffer from three challenges:

- Lack of diversity: the rationales do not provide multiple perspectives for analyzing the news articles
- Lack of factuality: the rationales might contain factual errors
- Lack of relevance: the rationales might contain noisy information that are not helpful for judging.

Thus, this evaluation aims to evaluate which rationale is better to help judge the veracity of a specific news article.

Files in 'news_articles' contain the textual content of a specific news article and two corresponding rationales.

Files in 'images' contain the visual image of a news article.

You need to enter your preferred rationale (based on diversity, factuality, and relevance) in 'answer.csv' for each news article.

- 1 for preference of explanation 1
- 2 for preference of explanation 2
- 3 for no clear preference

Please follow your subjective feelings.

Major Voting Each human evaluator needs to evaluate 200 rationale pairs, where the selected pairs are the same for all evaluators. For each pair, we employ major voting to obtain the final results. Notably, if the three experts have distinct answers, we consider the final result to be "indistinguishable".

G Diversity Analysis

Figure 7 shows the distributions of token frequency and similarity.