/TemplateVersion (2026.1)

DiFaR：通过多样化、真实和相关的理由增强多模态错误信息检测

Herun Wan[1], Jiaying Wu[2], Minnan Luo✉[1], Xiangzheng Kong[1], Zihan Ma[1], Zhi Zeng[1]
[1]Xi'an Jiaotong University, [2]National University of Singapore
wanherun@stu.xjtu.edu.cn, minnluo@xjtu.edu.cn

Abstract

从大型视觉语言模型（LVLMs）生成文本推理以支持可训练的多模态虚假信息检测器已成为一种有前景的范式。然而，其有效性受制于三个核心挑战：（i）生成的推理缺乏多样性，（ii）由于幻觉导致的事实不准确，以及（iii）引入噪声的无关或冲突内容。我们介绍了 DiFaR，一个检测器无关的框架，能够产生多样、真实且相关的推理以增强虚假信息检测。DiFaR 使用五个链式思维提示，从 LVLMs 中引导出不同的推理路径，并结合一个轻量级的事后过滤模块，根据句子级事实性和相关性评分选择推理句。在四个流行的基准上进行的大量实验表明，DiFaR 比四个基线类别最多高出 5.9 %，并最多提升现有检测器 8.7 %。自动化指标和人类评估都确认 DiFaR 在所有三个维度上显著提高了推理质量。[1]

1 引言

大型视觉-语言模型（LVLMs）在多种多模态任务中取得了显著的性能，这是由其强大的推理和表示能力驱动的。然而，在多模态虚假信息检测（MMD）这一需要精确事实基础和细粒度、任务特定推理的任务中，其效果仍然有限（Liu et al. 2025a; Li et al. 2025b）。

为了利用大规模视觉语言模型（LVLMs）在识别多模态虚假信息方面的潜力，最近的工作提出了一种协作范式，将 LVLMs 与可训练检测器（Zheng et al. 2025）相结合，我们称之为 LVLM-as-Enhancer。在这一框架中，LVLMs 被提示生成文本推理（即可解释的理由或解释），然后与原始新闻文章配对，并传递给下游可训练检测器。此设计旨在利用 LVLMs 的泛化能力，同时保持任务特定模型的适应性。

虽然这一范式初见成效（Tahmasebi, Müller-Budack, and Ewerth 2024; Hu et al. 2024），但我们识别出三大核心局限性（如图 1 所示），这些局限性阻碍了其充分发挥潜力：

- 多样性有限。大多数现有研究集中于可训练检测器（Wang et al. 2024）的架构创新，而对生成理由的质量和变化关注较少。这些方法通常依赖于固定的提示，这限制了所捕获的视角范围。因此，它们未能利用可以丰富新闻内容解释的多样推理信号（Wan et al. 2024）。
- 有限的事实性。LVLMs 容易出现幻觉（Ji et al. 2023），并且经常生成与已验证事实（Mallen et al. 2023）不符的内容。因此，生成的依据可能会引入事实错误，从而降低下游检测器的可靠性（Pan et al. 2023）。
- 有限的相关性。生成的推理经常包含松散相关或与主题无关的信息，这可能会淡化甚至与原文的主张相冲突（Zheng et al. 2025; Xu et al. 2024）。这种不一致性降低了解释的实用性，并可能影响对真实性的评估。

[1]可在 https://github.com/whr000001/DiFaR 获取。

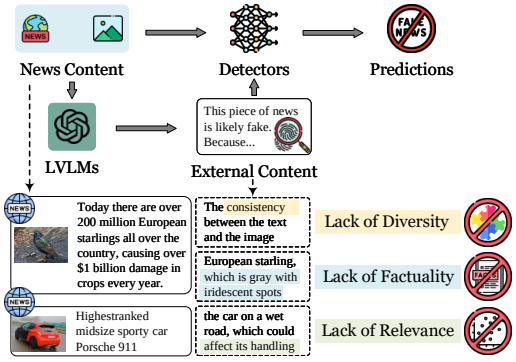

Figure 1: 在用于多模态检测（MMD）的 LVLM-as-Enhancer 范式中，说明了三个关键挑战，即多样性、事实性和相关性，其中 LVLMs 被提示生成解释性理由以支持下游检测器。

为了解决这些限制，我们提出了 DiFaR，这是一种 MMD 框架，旨在生成多样、真实且相关的理由。DiFaR 兼容任何现有的可训练检测器且无需改变架构即可运行。为了增强多样性，DiFaR 结合了从五个链式思维（CoT）提示中派生出的多种理由，每个提示针对内容的不同方面，包括文本细节、视觉特征和跨模态一致性。这种多提示策略允许更丰富和更细微的推理。

为了进一步提高推理质量，DiFaR 引入了一个事后优化模块，该模块根据事实性和相关性过滤单个推理句子。对于事实性，模块从维基百科等结构化知识库中检索证据，并将其与生成的内容进行比较（Min et al. 2023）。对于相关性，它使用基于表示的指标（Lewis et al. 2020）计算推理与源文章之间的语义相似度。得分较低的句子将被删除，从而得到一个精炼且可信的推理集。

我们在四个多模态虚假信息数据集上进行广泛实验，涵盖了人类撰写和机器生成的新闻文章。在所有数据集中，DiFaR 始终优于四个具有代表性的强基线类别，最

高可提高 5.9 % 的相对准确性。此外，将 DiFaR 整合到现有检测器中可实现高达 8.7 % 的性能提升。消融研究证实，DiFaR 的每个组件对其效能都有意义的贡献。进一步分析，包括人工评估，验证了 DiFaR 提升了生成论据的多样性、真实性和相关性。

2 方法论

2.1 预备知识

我们将 MMD 任务视为一个二分类问题。每个新闻实例由一个文本组件和一个视觉图像组成，目标是确定新闻的真实性。形式上，让 $\mathcal{D}_{train} = \{(T_i, V_i, y_i)\}_{i=1}^{N_{train}}$ 表示一个包含 N_{train} 个已标记新闻文章的训练数据集，其中 T_i 是文本，V_i 是相关图像，而 $y_i \in \{0, 1\}$ 是真实标签。一个具有参数 θ 的可训练检测器 f 被训练以建模条件分布 $p(y \mid T, V; f, \theta)$ ，目标是最大化测试集 $\mathcal{D}_{test} = \{(T_i, V_i, y_i)\}_{i=1}^{N_{test}}$ 上的预测准确性。

给定一个特定实例 (T, V, y) （为了清晰，省略了索引），传统的可训练检测器 (Chen et al. 2022; Wang et al. 2023) 首先使用冻结的预训练编码器对模态进行编码，产生单模态表示 $\boldsymbol{t}$ （文本）和 $\boldsymbol{v}$ （图像）。这些通过模态交互模块融合为联合表示 $\boldsymbol{h}$ 。最终预测计算为 $p(y \mid T, V; f, \theta) \propto \exp(\text{MLP}(\boldsymbol{h}))$ ，其中 $\text{MLP}(\cdot)$ 是一个多层感知器。预测的标签由 $\arg\max_y p(y \mid T, V; f, \theta)$ 给出。

LVLM-as-Enhancer 范式通过利用 LVLM $\mathcal{G}$ 为输入实例生成解释性理由 $R = \mathcal{G}(T, V)$ 来扩展这一设置。然后，这些理由被编码为 $\boldsymbol{r}$ ，并通过一种专门的架构结合到检测流程中，从而计算出增强的表示 $\boldsymbol{h}$ 。虽然这种方法已经显示出早期的成功 (Hu et al. 2024)，但现有努力中使用的过度设计的集成策略可能限制了对多样化理由类型的广泛适应。例如，EFND (Wang et al. 2024) 引入了一个结构化模块，专门用于论辩推理，但可能无法推广到其他理由格式，如基于情感的新闻真实性推理 (Zhang et al. 2021) 。

2.2 DiFaR 框架

图 2 说明了 DiFaR 的整体架构，该框架旨在稳健地将多种理据集成到可训练的多模态虚假信息检测器中，同时保持通用性和可扩展性。

为了提高与各种推理的兼容性，DiFaR 保留了可训练检测器的一般流程，而无需引入特定任务的结构变化。给定一个新闻文章 (T, V, y) 和一组 M LVLM 生成的推理 $\{R_j\}_{j=1}^{M}$ ，我们首先将所有推理与原始文本输入连接以形成增强输入 $\tilde{T} = [T; R_1; \ldots; R_M]$ ，然后将其传递给检测器 f 。

虽然这种方法直观并且不依赖于特定检测器，但它提出了两个关键挑战：

- 输入长度约束。许多检测器，例如基于 CLIP (Radford et al. 2021) 的检测器，具有严格的令牌限制（例如，77 个令牌），使其无法容纳长连接的输入 (Chen et al. 2022) 。
- 顺序敏感性。串联推理的有效性可能在很大程度上取决于它们的顺序。先前的研究表明，序列的顺序显著影响上下文学习的表现 (Shi et al. 2024) 。穷举搜索所有排列在计算上是不可行的，并且不太可能产生一个普遍最佳的顺序。

为了克服这些限制，DiFaR 在表示层面增强了文本模态。具体来说，我们首先将连接的输入 $\tilde{T}$ 分割成 n 个句子 $\{\tilde{\boldsymbol{t}}_i\}_{i=1}^{n}$ 。每个句子都使用预训练的基于编码器的语言模型进行独立编码，生成的表示会被平均以形成最终表示：

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

其中 $\text{encoder}(\cdot)$ 表示句子级编码器；在我们的实现中，我们使用了 DeBERTa (He, Gao, and Chen 2023) 。

这一策略使检测器能够处理任意长度的输入，并由于对称平均算子而消除了对原因排序的敏感性。结果，DiFaR 可以稳健地结合多种理由以增强错误信息检测。

2.3 基于 CoT 的推理多样化

融合基础模型生成的新闻分析在评估新闻文章的真实性方面显示出希望 (Nan et al. 2024; Wu, Guo, and Hooi 2024) 。此外，使用多种观点可以提供互补的见解，这可能进一步有益于错误信息检测。为此，我们设计了五种思维链 (CoT) 提示，涵盖三个类别：文本内容、视觉内容和跨模态一致性，以生成一组多样化的与真实性相关的理由 $\{\bar{R}_j\}_{j=1}^{M}$ 。

每个推理 $\bar{R}_j$ 是通过结构化的多轮交互生成的。首先，LVLM 被提示从指定的角度分析新闻文章（例如，“分析这篇新闻文章的情感”），产生一个中间响应 $\bar{R}_{j(0)}$ 。然后，要求模型根据这一分析评估文章的真实性并为其判断提供理由，生成 $\bar{R}_{j(1)}$ 。完整的推理 $\bar{R}_j$ 是通过连接 $\bar{R}_{j(0)}$ 和 $\bar{R}_{j(1)}$ 形成的。

为了鼓励推理的多样性，我们设计了五个提示。对于文本内容，这些提示引导 LVLM 检查与虚假信息相关的语言信号，包括 (i) 情感分析 (Toughrai, Langlois, and Smaïli 2025) 和 (ii) 宣传策略 (Piskorski et al. 2023) 。对于视觉内容，提示主要关注理解伴随的图像，特别是通过 (iii) 物体识别 (Ma et al. 2024b) 和 (iv) 图像描述 (Abdali, Shaham, and Krishnamachari 2024) 。最后，为了捕捉跨模态的一致性，我们包括了 (v) 一个评估文本信息和视觉信息之间一致性的提示，参考之前的工作 (Liu et al. 2025a) 。详细的提示在附录中提供。

我们并不打算穷尽地探索提示设计空间，而是有意选择这五个具有代表性的提示，以展示推理多样性和 DiFaR 的适应性潜力。这些提示在文本、视觉和跨模态维度上引发了互补的视角，可以很容易地扩展或定制以适应其他推理需求或领域。

2.4 事后理由完善

虽然可以直接将每个原始理由 $\bar{R}$ 视为最终输入，但这种方法存在两个关键问题：缺乏事实性 (Pan et al. 2023) 和缺乏相关性 (Zheng et al. 2025) 。为了解决这些问题，我们采用了一个句子级别的筛选程序。具体来说，我们首先将 $\bar{R}$ 分割成 $\bar{m}$ 个句子 $\{\bar{\boldsymbol{r}}_k\}_{k=1}^{\bar{m}}$ ，并应用两种筛选策略来获得一个精炼的集合 $\{\boldsymbol{r}_k\}_{k=1}^{m}$ ，该集合仅包含事实性和相关性句子。最终理由 R 是通过连接这些筛选过的句子得到的。

由于幻觉，LVLMs 可能生成带有事实错误的句子，从而降低下游检测性能。为了应对这一问题，我们计算了每个句子的事实性评分 $s_f(\boldsymbol{r})$ ，并丢弃那些得分较低的

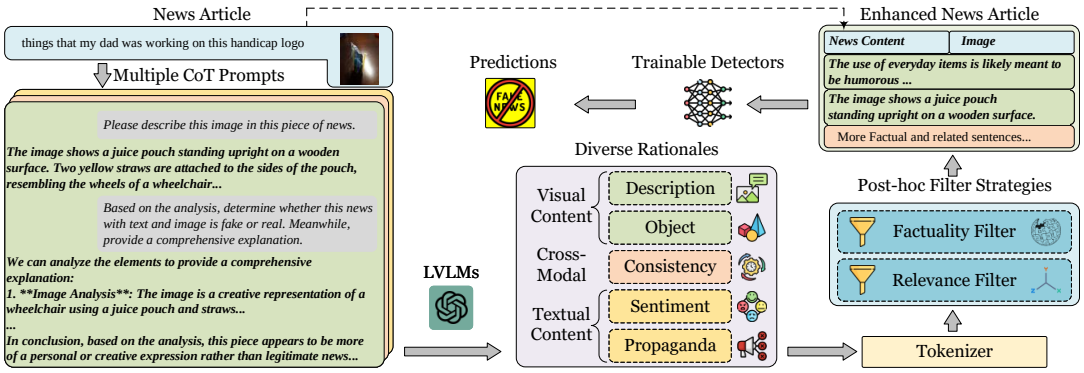

Figure 2: DiFaR 概述。该框架集成了一种简单但有效的结构，支持任意数量和长度的推理依据。它采用五种思维链提示来促进推理多样性，并使用两个后验精炼策略以确保事实性和相关性。

Methods		Fakeddit MiF.	Fakeddit MaF.	FakeNewsNet MiF.	FakeNewsNet MaF.	FineFake MiF.	FineFake MaF.	MMFakeBench MiF.	MMFakeBench MaF.
Vanilla LVLMs	InternVL Zero-Shot	$70.1_{\pm 2.4}$	$68.9_{\pm 2.9}$	$74.5_{\pm 3.4}$	$69.9_{\pm 3.3}$	$70.8_{\pm 2.1}$	$70.7_{\pm 2.1}$	$77.5_{\pm 2.5}$	$77.5_{\pm 2.4}$
	InternVL Few-Shot	$70.1_{\pm 3.6}$	$69.6_{\pm 3.8}$	$77.0_{\pm 2.2}$	$72.0_{\pm 3.1}$	$71.4_{\pm 2.7}$	$71.4_{\pm 2.7}$	$72.4_{\pm 2.7}$	$72.2_{\pm 2.6}$
	InternVL Retrieval	$60.6_{\pm 3.3}$	$55.4_{\pm 3.2}$	$64.1_{\pm 2.9}$	$61.1_{\pm 3.3}$	$70.5_{\pm 1.3}$	$70.4_{\pm 1.3}$	$63.6_{\pm 2.9}$	$62.1_{\pm 3.0}$
	InternVL Self-Refine	$59.7_{\pm 3.0}$	$57.9_{\pm 3.0}$	$68.9_{\pm 2.8}$	$65.5_{\pm 2.3}$	$67.3_{\pm 2.8}$	$67.3_{\pm 2.8}$	$64.6_{\pm 1.7}$	$64.6_{\pm 1.7}$
	GPT-4o Zero-Shot	$78.1_{\pm 1.5}$	$78.0_{\pm 1.4}$	$84.0_{\pm 1.8}$	$77.3_{\pm 3.0}$	$75.3_{\pm 3.4}$	$74.5_{\pm 3.6}$	$80.9_{\pm 3.4}$	$80.7_{\pm 3.4}$
	GPT-4o Few-Shot	$78.9_{\pm 2.3}$	$78.8_{\pm 2.3}$	$80.3_{\pm 1.4}$	$72.7_{\pm 3.1}$	$\underline{77.3}_{\pm 0.9}$	$\underline{76.8}_{\pm 0.9}$	$82.3_{\pm 3.1}$	$82.2_{\pm 3.0}$
	GPT-4o Retrieval	$64.1_{\pm 3.6}$	$63.4_{\pm 3.9}$	$81.8_{\pm 1.7}$	$74.8_{\pm 2.3}$	$74.5_{\pm 1.8}$	$73.9_{\pm 1.8}$	$72.9_{\pm 3.4}$	$72.9_{\pm 3.4}$
	GPT-4o Self-Refine	$77.6_{\pm 0.4}$	$77.5_{\pm 0.4}$	$81.2_{\pm 0.6}$	$73.8_{\pm 1.5}$	$73.2_{\pm 3.0}$	$72.2_{\pm 2.9}$	$78.2_{\pm 3.3}$	$78.1_{\pm 3.3}$
Enhanced LVLMs	MMD-Agent	$68.9_{\pm 1.9}$	$68.8_{\pm 1.9}$	$67.4_{\pm 4.2}$	$60.4_{\pm 4.5}$	$64.1_{\pm 2.4}$	$64.1_{\pm 2.4}$	$75.3_{\pm 4.2}$	$75.1_{\pm 4.1}$
	Knowledge Card	$52.1_{\pm 3.7}$	$42.4_{\pm 3.6}$	$73.8_{\pm 3.4}$	$67.7_{\pm 5.3}$	$64.5_{\pm 2.2}$	$64.4_{\pm 2.2}$	$57.3_{\pm 2.3}$	$56.3_{\pm 2.7}$
Trainable Detectors	CLIP	$86.0_{\pm 2.4}$	$85.9_{\pm 2.4}$	$86.6_{\pm 1.6}$	$82.3_{\pm 1.4}$	$75.7_{\pm 3.3}$	$75.5_{\pm 3.4}$	$84.3_{\pm 2.4}$	$84.2_{\pm 2.4}$
	CAFE	$\underline{87.4}_{\pm 2.1}$	$\underline{87.4}_{\pm 2.1}$	$86.8_{\pm 0.8}$	$82.8_{\pm 1.2}$	$76.2_{\pm 2.7}$	$76.0_{\pm 2.7}$	$\underline{85.4}_{\pm 2.7}$	$\underline{85.4}_{\pm 2.7}$
	COOLANT	$86.4_{\pm 2.3}$	$86.3_{\pm 2.3}$	$85.7_{\pm 1.7}$	$81.3_{\pm 1.9}$	$76.2_{\pm 2.1}$	$76.1_{\pm 2.1}$	$83.2_{\pm 2.1}$	$83.1_{\pm 2.1}$
LVLM-as-Enhancer	EARAM	$82.6_{\pm 1.9}$	$82.5_{\pm 1.9}$	$82.9_{\pm 2.9}$	$77.5_{\pm 3.0}$	$73.8_{\pm 2.4}$	$73.7_{\pm 2.3}$	$78.9_{\pm 2.0}$	$78.8_{\pm 2.1}$
	EFND	$80.3_{\pm 1.3}$	$80.2_{\pm 1.2}$	$\underline{87.6}_{\pm 1.0}$	$\underline{84.1}_{\pm 2.0}$	$75.9_{\pm 2.3}$	$75.7_{\pm 2.3}$	$76.5_{\pm 2.5}$	$76.1_{\pm 3.2}$
DiFaR		$90.8_{\pm 2.1}$	$90.8_{\pm 2.1}$	$89.3_{\pm 1.9}$	$85.5_{\pm 2.7}$	$81.2_{\pm 1.6}$	$81.1_{\pm 1.7}$	$90.4_{\pm 1.0}$	$90.4_{\pm 1.0}$

Table 1: DiFaR 和基线在四个广泛使用的多模态错误信息检测数据集上的表现。“MiF.” 和 “MaF.” 分别表示微平均和宏平均 F1 分数。粗体表示最佳性能，underline 表示第二佳性能。DiFaR 在最先进的基线之上实现了一致的改进，增幅高达 5.9 %。

句子。根据 Min et al. (2023)，我们依赖外部知识来源（特别是维基百科）来支持这一评估。对于每个句子 $\boldsymbol{r}$，我们检索 p 个候选文档 $\{\boldsymbol{d}_i\}_{i=1}^{p}$，并定义事实性评分为：

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

，其中 $\text{fact}(\boldsymbol{r} \mid \boldsymbol{d}_i)$ 量化了 $\boldsymbol{r}$ 与文档 $\boldsymbol{d}_i$ 之间的事实对齐。我们通过平均两个互补信号来计算该评分：一个立场分类器 (Schuster, Fisch, and Barzilay 2021)，用于评估文档对句子的支持程度，以及一个总结精度模型 (Feng et al. 2023)，用于评估句子总结文档的效果：

$$\text{fact}(\boldsymbol{r} \mid \boldsymbol{d}_i) = \frac{1}{2}\left(\text{stance}(\boldsymbol{r}, \boldsymbol{d}_i) + \text{summary}(\boldsymbol{r}, \boldsymbol{d}_i)\right). \tag{3}$$

。两个分数都被规范化到范围 $[0, 1]$，较高的值表明更强的事实一致性。我们根据事实性评分保留前 50 % 的句子：

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

在某些情况下，LVLM 生成的推理可能与提示不一致，并包含与源文章无关或偏离的内容。为了确保推理的相关性，我们通过计算语义相似性来评估句子与输入文本 T 的相关性。具体来说，我们采用了一种广泛使用的方法 (Lewis et al. 2020)，该方法利用基于编码器的语言模型获取句子嵌入，并通过余弦距离测量相似性：

其中，encoder(·) 被实例化为 MPNet (Song et al. 2020)。较高的值表示推理与输入文章之间有更强的语义对齐度，我们还保留前 50% 的句子。

3 实验

3.1 实验设置

数据集。我们利用四个流行的数据集评估 DiFaR 和现有的基线模型：Fakeddit (Nakamura, Levy, and

Wang 2020)、FakeNewsNet (Shu et al. 2020)、FineFake (Zhou et al. 2024) 和 MMFakeBench (Liu et al. 2025a)。为了获得稳健的评价结果，我们进行五折评估并报告性能的平均值和方差。有关数据集的详细信息在附录中提供。

基线。我们将 DiFaR 与四种最先进的基线进行比较：(i) Vanilla LVLMs: InternVL V3 (Zhu et al. 2025) 和 GPT-4o，使用零样本学习、小样本学习、检索 (Lewis et al. 2020) 和自我改进 (Madaan et al. 2023) 提示；(ii) 增强的 LVLMs: MMD-Agent (Liu et al. 2025a) 和 Knowledge Card (Feng et al. 2024)；(iii) 可训练探测器: CLIP (Radford et al. 2021)、CAFE (Chen et al. 2022) 和 COOLANT (Wang et al. 2023)；以及 (iv) 作为增强器的 LVLM: EARAM (Zheng et al. 2025) 和 EFND (Wang et al. 2024)。基线的详细信息在附录中提供。

设置。我们使用 GPT-4o 作为 DiFaR 的主要 LVLM 主干。为确保公平比较，所有检测器都在各个折叠中以一致的超参数设置进行评估。对于 LVLM 推理，我们通过将温度设置为零或将 do_sample 配置为 False 来禁用采样，从而确保确定性输出和可重复性。附录中提供了额外的实现细节和实验配置。

3.2 DiFaR 的有效性

我们在表格 1 中展示了 DiFaR 和最先进的基线的性能。可训练检测器仍然具有很高的竞争力。在所有基准中，监督的可训练检测器展示出最强的独立性能，在各个基准测试中接近 DiFaR 。这突显了在直接监督下学习表示的有效性，并进一步支持将 LVLM 生成的理由整合到此类架构中的价值。

现有的 LVLM 增强方法面临泛化问题。值得注意的是，现有的 LVLM 增强方法的表现往往不如预期，有时甚至低于普通 LVLMs。这表明在数据集上的泛化能力较差，可能是由于推理范围狭窄和推理质量低所导致。例如，EARAM 仅关注常识性和补充性，而 EFND 则专注于辩论风格的真实性分析。这两种方法都缺乏广泛视角的整合。我们还观察到，生成推理中的非事实性或离题内容可能会误导检测器。这些发现直接激发了我们设计 DiFaR 的动机，旨在通过增强 LVLM 生成推理的多样性、事实性和相关性来填补这些空白。

DiFaR 达到了最先进的性能。DiFaR 在所有四个数据集上始终优于最强的基线，微平均 F1 得分取得了 1.9 % 至 5.9 % 的增益。值得注意的是，基础版本和增强版的 LVLMs 在大多数基准测试中表现不佳，这表明 LVLMs 单独在处理事实推理和精确的真实性评估方面有困难。这些结果强调了 LVLM-as-Enhancer 范式的重要性，在这个范式中，外部推理与可训练的探测器结合在一起。

3.3 适应性于多样的检测器

DiFaR 被设计为兼容任何可训练的检测器和任何 LVLM 生成的理由。为了评估其适应性，我们评估了将 DiFaR 生成的理由与各种检测器集成时的效果。

如表 2 所示，我们得出三个关键观察结果。(i) DiFaR 显著提高了所有可训练检测器的性能，增益可达 8.7 % 。这表明其在解决 LVLM-as-Enhancer 范式关键限制方面的有效性，并能在多种架构中提高检测性能。(ii) 在 COOLANT、CAFE 和 CLIP 中，性能增益逐渐减少，在 CLIP 经过增强后出现轻微的性能下降。这表明更复杂的检测器架构可能在从理由中提取信号方面更有效。

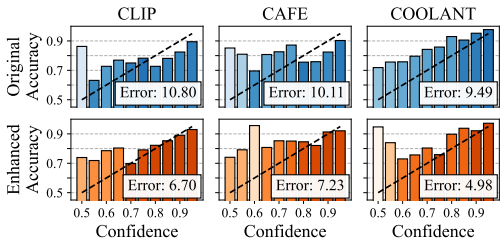

Figure 3: 现有可训练检测器在有和无 DiFaR 增强下的校准性能。“误差” 指的是期望校准误差 (ECE，×100)，较低的值表示校准较好。DiFaR 不仅提高了检测准确性，还增强了置信度估计的可靠性。

这一趋势与之前强调架构设计在推动性能方面作用的研究一致。(iii) 基准方法生成的理由往往无法提高检测器性能，甚至在某些情况下导致性能下降。这突显了理由质量的重要性，在没有足够多样性、真实性和相关性时，即使是强检测器也无法可靠地从外部解释中获益。

除了准确性之外，我们还使用 DiFaR 评估检测器在增强前后的可信度。我们通过期望校准误差 (ECE) (Guo et al. 2017) 来量化这一点。如图 3 所示，引入 DiFaR 后，检测器的校准效果得到提升，ECE 减少最多达到 47.5 %。这些结果表明，DiFaR 不仅提高了预测性能，还增强了模型置信度估计的可靠性。

3.4 消融研究

我们进行了一项消融研究，以评估 DiFaR 中每个组件的贡献。具体来说，我们考虑四种设置：(i) 用单个固定的 CoT 提示替换五个不同的 CoT 提示；(ii) 去除事实性和相关性过滤策略；(iii) 使用通用指令提示 LVLM 生成推理 (即，“确定这则包含文本和图像的新闻是真是假。同时，提供全面的解释。”)；以及 (iv) 用开源的 InternVL V3 模型替换 GPT-4o。我们在表 6 中报告了 FineFake 数据集上的结果，完整结果在附录中提供。研究结果显示，由于基础提示生成的推理，检测器未能受益，这强调了推理多样性、事实性和相关性的重要性。此外，去除 DiFaR 的任何单个组件在大多数情况下会导致性能下降，最大下降幅度可达 5.8 %。这些结果证实了 DiFaR 中的每个模块在克服 LVLM 作为增强器范式的局限性方面起到了关键作用。

我们进一步评估 DiFaR 是否有效地解决了生成的理由中有限的多样性、事实性和相关性的问题。

3.5 整体有用性评估

我们首先评估由 DiFaR 生成的推理的整体质量。为此，我们对三位在错误信息相关领域的专家进行人工评估。选择了四种类型的推理进行比较：(i) DiFaR：由 DiFaR 生成的未经过滤的推理；(ii) 基线：由 EFND 生成的推理；(iii) 单一：从单一随机选取的角度生成的推理；以及 (iv) 过滤：在应用了真实性和相关性过滤器后由 DiFaR 生成的推理。

我们进行四组成对比较：DiFaR vs. Baseline，DiFaR vs. Single，Filtered vs. Baseline，以及 Filtered vs. Single。对于每一对，专家被要求判断哪个理由对于验证新闻文章的真实性更有帮助，或者指出两者是否无法区分。最终决定通过多数投票决定。评估协议的详细信

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

Table 2: 通过使用 DiFaR 增强的可训练检测器和基线方法的微平均 F1 分数。在每个设置中性能最好的结果已被突出显示。DiFaR 将检测器性能提高了最多 8.7 %，证明了其在解决推理多样性、事实性和相关性方面局限性的有效性。

Variants	CLIP	CAFE	COOLANT
DiFaR	$77.1_{\pm 2.1}$	$80.2_{\pm 1.9}$	$81.2_{\pm 1.6}$
w/o Multiple	$72.6_{\pm 3.8}$ 5.8% ↓	$78.3_{\pm 1.9}$ 2.4% ↓	$79.6_{\pm 1.4}$ 2.0% ↓
w/o Filter	$72.7_{\pm 3.5}$ 5.7% ↓	$78.0_{\pm 2.3}$ 2.7% ↓	$78.9_{\pm 1.4}$ 2.8% ↓
w/ Vanilla	$73.0_{\pm 2.1}$ 5.3% ↓	$73.8_{\pm 2.3}$ 8.0% ↓	$76.7_{\pm 2.4}$ 5.5% ↓
w/ InternVL	$77.5_{\pm 2.2}$ 0.5% ↑	$78.8_{\pm 1.3}$ 1.7% ↓	$79.9_{\pm 2.3}$ 1.6% ↓

Table 3: 对 DiFaR 的消融研究。“w/o Multiple” 使用一个特定的 CoT 提示代替五个；“w/o Filter” 去除了事实性和关联性筛选器；“w/ Vanilla” 使用从简单提示生成的推理；“w/ InternVL” 用 InternVL V3 替换 GPT-4o。结果表明，每个组件都对性能提升有贡献。

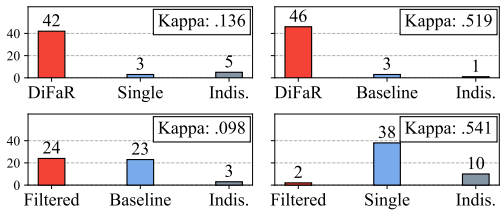

Figure 4: 通过对生成的理由进行成对比较的人类投票结果。“Kappa” 表示三位专家之间的一致性 Kappa 系数，“Indis.” 表示两种理由被判断为无法区分的实例比例。DiFaR 产生了对真实性评估最有帮助的理由。

息在附录中提供。所有判断的弗里斯 Kappa 值为 0.34，表明评分者之间具有中等水平的一致性。

图 4 中的结果显示，DiFaR 的表现远远优于 Baseline 和 Single 设置，这证明它能够为人工错误信息评估提供更有用的依据。然而，在经过事后过滤后，DiFaR 的优势降低，在某些情况下，表现甚至不如 Single 基准设置。我们推测，虽然过滤提高了事实性和相关性，但可能也会降低依据的流畅性和连贯性，从而限制了对人类读者的可解释性。

3.6 细粒度质量评估

多样性。 DiFaR 使用了五个串联思维（CoT）提示，以从多样化的推理视角中捕获信号。作为一个基准，我们与只从两个角度生成推理的 EFND 进行比较。我们首先计算在 FineFake 数据集上生成的推理中不同标识符的比例。DiFaR 实现的不同标识符比例为 0.904，明显高于 EFND 的 0.406，这表明其词汇覆盖范围更广。

为了进一步量化多样性，我们分析了标记频率和提示间的相似性。具体来说，我们使用 infini-gram（Liu et al. 2024a）来测量标记频率，并使用 BERTScore（Zhang et al. 2020）来计算不同提示生成的理由之间的成对相似性。两个指标的较低值表明更大的多样性。如附录所示，DiFaR 实现了更低的平均相似性（0.57 比 0.66）和更低的标记频率（2.41×10^6 比 2.85×10^6），确认了它产生了更多在词汇上多样化的输出。

我们还进行了一项使用 COOLANT 的消融研究，以评估不同 CoT 提示类型的影响。表格 4 表明，删除任何单一的提示类别都会导致性能下降最高达 2.5 %，这验证了在虚假信息检测中结合多种推理视角的重要性。这些发现支持我们将理由多样化设计为 DiFaR 核心组件。

真实性和相关性。 DiFaR 采用了两种事后过滤策略来提高生成的理由的事实性和相关性。为了研究它们的效果，我们改变过滤阈值，并使用 EFND 作为比较基线

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

Table 4: 对 CoT 提示类别的消融研究，仅保留单一提示类型的理由。“Orig.” 表示使用所有五种提示的 DiFaR 的原始性能，“Consis.” 指仅使用跨模态一致性提示的变体。我们报告了微平均 F1 分数及相应的性能变化。结果表明，多样化的理由通常比单一视角的理由表现更好。

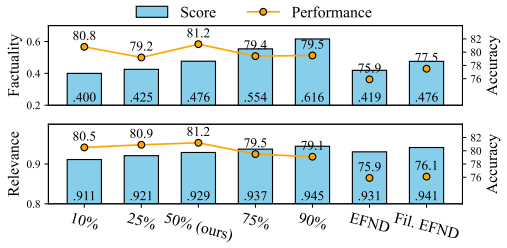

Figure 5: 不同过滤阈值下 DiFaR 的性能，以及对应的事实性和相关性得分。“Fil. EFND” 表示基于得分过滤掉底部 25 % 句子后的 EFND 基线。结果显示适度的过滤可以提高性能，而过于激进的过滤可能会降低效果。

来评估模型性能和平均事实性/相关性得分。图 5 显示，增加过滤阈值会导致更高的事实性和相关性评分。然而，DiFaR 的整体检测性能并没有以同比例继续增加。我们推测，过于激进的过滤虽然提高了质量评分，但可能会移除语义丰富的内容，从而削弱提供给检测器的增强信号。

有趣的是，我们观察到低过滤阈值（如 10 %）的 EFND 和 DiFaR 在事实性和相关性得分上表现相似，但它们的检测性能却显著不同。我们将此归因于 EFND 的理由中存在低质量句子，这可能误导检测器。为验证这一点，我们对 EFND 应用了我们的过滤策略，并通过事实性和相关性得分去除了其句子中最底部 25 % 的部分。这带来了可测量的性能提升，支持了我们过滤设计的有效性。此外，与 EFND 不同的是，DiFaR 从多样的提示中生成理由，提供更广泛的视角和更丰富的语义覆盖。这种多样性使得 DiFaR 即使经过过滤仍能保留信息丰富的内容，从而增强了其卓越的性能。

3.7 案例分析

我们分析了数据集中的一个代表性例子，以说明 DiFaR 如何有助于提高错误信息检测能力。生成的依据和相应的模型预测显示在图 6 中。在这种情况下，原始的 COOLANT 模型给出了错误的预测。尽管单个依据正确识别了错误信息，但将其融入 COOLANT 也未能得

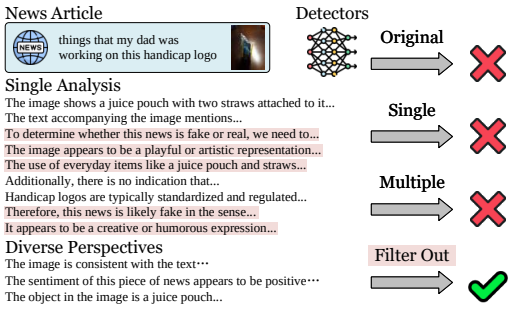

Figure 6: 使用不同变体的 DiFaR 与 COOLANT 作为基础检测器的预测案例研究。原始检测器未能识别错误信息，而通过 DiFaR 的增强可以进行正确预测。

出正确预测。相反，当 COOLANT 增强了全套多样化依据并过滤掉非事实或不相关的句子后，模型成功检测到了错误信息。这个例子表明 DiFaR 提供了更丰富和更集中的语义信号，从而有效支持下游检测。

4 相关工作

多模态虚假信息检测器通常使用预训练的编码器来编码文本内容和视觉内容，随后通过设计的架构来建模跨模态交互 (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) 。随着大语言视觉模型 (LVLMs) 的兴起，早期的工作直接采用 LVLMs 作为主干来识别虚假信息 (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) 。然而，这些模型往往由于幻觉和缺乏事实基础而受到限制，限制了它们的有效性。因此，提出了 LVLM-as-Enhancer 范式。该范式首先设计提示生成外部文本内容，即解释或理由，以提供丰富的语义信息 (Saha and Srihari 2024; Liu et al. 2024b) ，如立场 (Choi et al. 2025) 、传播 (Liu et al. 2024c) 和实体 (Ma et al. 2024a) 。然后，他们设计了一个可训练的模块来捕捉该语义信息以增强性能 (Zhang et al. 2025; Zhou et al. 2025) 。在这项工作中，我们确定了现有 LVLM 生成的理由的关键限制，具体来说，它们缺乏多样性、事实性和相关性，并提出了 DiFaR ，一个通过多视角提示和事后筛选来解决这些挑战的一般框架。

5 结论

我们提出了 DiFaR ，这是一个在 LVLM-as-Enhancer 范式下简单而有效的框架，可以无缝适应多种推理，而不需要对检测器进行结构上的改变。它使用五种思维链提示来鼓励多样化的推理，以及两种事后过滤策略以确保真实性和相关性。大量实验表明，DiFaR 达到了最新的性能水平，并可以显著增强现有的可训练检测器。进一步的分析，包括人工评估，证实了 DiFaR 成功地增强了推理的多样性、真实性和相关性。

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. IMRRF: 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 lvlms. 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 multi-granularity 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.

对于新闻文章的多样化观点，我们使用以下提示（假设我们首先向 LVLMs 提供文本和视觉内容）：

- 情感。请分析这篇新闻的情感。
- 宣传。请分析这篇新闻中使用的宣传策略。
- 一致性。请分析这则新闻的文字与图像之间的一致性。
- 对象。请分析这条新闻图片中出现的对象。
- 描述。请描述这条新闻中的这张图片。

最后，我们通过提示“根据分析，判断这篇带有文本和图像的新闻是真实的还是虚假的。同时提供一个全面的解释。”来提示 LVLMs 判断其真实性。

我们使用四个广泛使用的多模态虚假信息检测数据集来评估 DiFaR 和现有基线，每篇新闻文章包括文本内容和图像，涵盖了由人工和机器生成的多模态新闻文章：

- Fakeddit (Nakamura, Levy, and Wang 2020) 是一个多模态数据集，由超过一百万个来自多个类别的虚假新闻样本组成，来源是 Reddit。每个实例通过远程监督按照两类、三类和六类分类类别进行标记。我们采用两类标签，即将其视为一个二分类任务。
- FakeNewsNet (Shu et al. 2020) 包含两个综合数据集：Politifact，包括政治新闻，和 Gossipcop，包括娱乐新闻。每个实例包含多样化的特征，例如新闻内容、社会背景和时空信息。我们仅使用新闻文章中的文本内容和视觉内容。
- FineFake (Zhou et al. 2024) 包括 16,909 篇新闻文章，涵盖六个语义主题和八个平台。每个实例包含多模态内容、潜在社会背景、半人工验证的常识知识以及超越传统二元标签的细粒度注释。我们仅使用新闻中的文本内容和视觉内容，并利用二元标签。
- MMFakeBench (Liu et al. 2025a) 包含三个关键来源：文本真实性扭曲、视觉真实性扭曲和跨模态一致性扭曲，以及 12 种错误信息伪造类型的子类别。它涵盖了机器生成的新闻文章，包括生成的文本内容和视觉内容。我们利用二元标签。

为了确保公平比较，我们从原始数据集中抽样以创建具有相似数量真实和假的实例的平衡子集。各数据集的统计数据如表 5 所示。此外，为了提高结果的稳健性，我们将每个数据集随机分成五个相等的部分进行交叉验证。

A 基线

我们将 DiFaR 与四种类型的最新基线进行比较。

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: 数据集的统计信息。

传统 LVLM 模型 使用简单提示来提示 GPT-4o 和 InternVL V3 进行错误信息检测。我们采用以下提示风格：

- 零次提示直接使用 LVLMs 获得结果，其中提示如下：

  文本：文本
  图像：图像
  基于上述文本和图像，请判断这条新闻是真实的还是假的。只输出“真实”或“假”，不需要任何解释。
- 少量样本提供了带有标签的实例的随机样本给 LVLMs，提示如下：

  文本：文本
  图片：图像
  标签：标签
  根据上述示例，请判断下面的新闻是真实的还是虚假的。仅输出真实或虚假而不进行任何解释。
  文本：文本
  图片：图像
  标签：
- 检索 (Lewis et al. 2020) 首先使用 bm25 从 BBC 新闻资源 (Li, Guerin, and Lin 2024) 中检索出三条最相关的新闻。然后将相关新闻提供给 LVLMs 作为外部内容，提示如下：

  相关新闻：
  检索新闻文章
  文本：文本
  图像：图像
  基于上述文本和图像，请判断这则新闻是真实的还是虚假的。只输出真实或虚假，无需任何解释。
- Self-Refine (Madaan et al. 2023) 提示 LVLMs 检查答案是否正确并修正预测，其中提示如下：

  文本：文本
  图像：图像
  基于上述文本和图像，请判断这则新闻是真实的还是虚假的。仅输出真实或虚假，无需任何解释。
  第一个预测
  答案正确吗？如果不正确，请给出你的答案。

增强型 LVLMs 设计了一个包含多个提示的框架来增强 LVLMs 的能力，包括：

- MMD-Agent (Liu et al. 2025a) 可以整合 LVLM 代理的推理、行动和工具使用能力，以增强泛化能力和提高检测性能。
- 知识卡 (Feng et al. 2024) 提出了一种模块化框架，可以将新的事实性和相关知识插入大型语言模型中。我们采用自下而上的方法，并使用建议的卡片来增强 LVLMs。

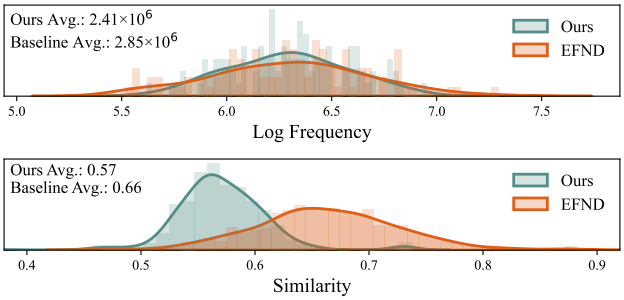

Figure 7: DiFaR 和 EFND 的标记频率和相似度分布。DiFaR 表现出较低的频率和相似度，证明生成的理由具有多样性。

可训练的检测器 代表传统的多模态虚假信息检测器，这些检测器需要学习参数，包括：

- CLIP (Radford et al. 2021) 是一种广泛使用的骨干网络，用于编码文本和视觉信息。在获得表示后，我们采用 MLP 层来分类虚假信息。
- CAFE (Chen et al. 2022) 是一种模糊感知的多模态假新闻检测方法，包括跨模态对齐模块、跨模态模糊学习模块和跨模态融合模块。
- COOLANT (Wang et al. 2023) 是一个用于多模态假新闻检测的跨模态对比学习框架，包括一个辅助任务、一个跨模态融合模块以及带有注意力引导模块的注意力机制。

LVLM **作为增强器** 检测器包含两个具有这种范式的代表性基线，包括：

- EARAM (Zheng et al. 2025) 可以利用多模态小语言模型，从 LVLMs 的多视角分析中提取有用的理由。LVLMs 分析新闻文章的常识性和互补性。
- EFND (Wang et al. 2024) 设计了一个基于提示的模块，该模块利用大型语言模型通过推理两种可能的真实性的原因来生成理由。它还提出了一个特定的可训练模块以从两个角度捕捉信号。

我们认为这四类多模态虚假信息检测器涵盖了大多数现有的方法，并代表了该领域的进展。

B 设置

LVLMs **的推断**。 我们将温度设置为 0，或将 do sample 设置为 false，以确保结果的可重现性。对于 GPT-4o，我们使用官方 API。对于 InternVL V3，我们使用两块 24GB 显存的 RTX4090 GPU 进行操作。

可训练检测器 为了获得公平的比较，我们在每个折叠中为每个检测器设置相同的超参数。每个基准测试可以在一块 24GB 显存的 RTX4090 GPU 上运行。同时，我们对每个基准测试运行五次，并报告每个折叠中微 f1-score 最好的结果。表 7 展示了每个基准测试的超参数。我们还在补充材料中提供了相关代码。

C 消融研究

我们在表格 6 中展示了完整的消融研究。

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

Table 6: DiFaR 的消融研究。这表明 DiFaR 的每个模块都可以提升检测性能。

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: 基线的超参数。

D 人工评估

评估指南文档。 我们首先为每位专家提供一个简要的指导文件。指导内容如下：

大型视觉语言模型在增强多模态虚假信息检测方面被证明是有效的。

一种广泛使用的范式，LVLM-as-enhancer，提出生成外部解释/理由来增强可训练检测器的性能。

然而，由 LVLM 生成的解释/理由面临三个挑战：

- 缺乏多样性：这些理由未能为分析新闻文章提供多角度的视角
- 缺乏事实性：理由可能包含事实错误
- 缺乏相关性：理由可能包含无助于判断的噪声信息。

因此，该评估旨在评估哪个理由更有助于判断特定新闻文章的真实性。

在“news_articles”中的文件包含特定新闻文章的文本内容和两个相应的理由。

“images”中的文件包含新闻文章的视觉图像。

您需要在每篇新闻文章的'answer.csv' 中输入您首选的依据（基于多样性、准确性和相关性）。

- 1 对解释 1 的偏好
- 选项解释 2 的偏好为 2
- 3 表示没有明确的偏好

请遵循你的主观感受。

多数投票 每位人工评估员需要评估 200 对理由对，所选择的对对于所有评估员来说都是相同的。对于每一对，我们采用少数服从多数来获得最终的结果。值得注意的是，如果三位专家的答案各不相同，我们将最终结果视为“无法区分”。

图

E 多样性分析

7 显示了标记频率和相似度的分布。