Drifting Away from Truth: GenAI-Driven News Diversity Challenges LVLM-Based Misinformation Detection

Fanxiao Li[1], **Jiaying Wu** [2*], **Tingchao Fu** [1], **Yunyun Dong** [3], **Bingbing Song** [3], **Wei Zhou** [4, 5*]

[1] School of Information Science and Engineering, Yunnan University,
[2] National University of Singapore,
[3] School of Software and AI, Yunnan University,
[4] Yunnan-Malaya Institute (School of Engineering), Yunnan University
[5] State Key Laboratory of Vegetation Structure, Function and Construction (VegLab)
lifanxiao@stu.ynu.edu.cn, jiayingw@nus.edu.sg, zwei@ynu.edu.cn

Abstract

The proliferation of multimodal misinformation poses growing threats to public discourse and societal trust. While Large Vision-Language Models (LVLMs) have enabled recent progress in multimodal misinformation detection (MMD), the rise of generative AI (GenAI) tools introduces a new challenge: *GenAI-driven news diversity*, characterized by highly varied and complex content. We show that this diversity induces *multi-level drift*, comprising (1) *model-level misperception drift*, where stylistic variations disrupt a model's internal reasoning, and (2) *evidence-level drift*, where expression diversity degrades the quality or relevance of retrieved external evidence. These drifts significantly degrade the robustness of current LVLM-based MMD systems. To systematically study this problem, we introduce DRIFTBENCH, a large-scale benchmark comprising 16,000 news instances across six categories of diversification. We design three evaluation tasks: (1) robustness of truth verification under multi-level drift; (2) susceptibility to adversarial evidence contamination generated by GenAI; and (3) analysis of reasoning consistency across diverse inputs. Experiments with six state-of-the-art LVLM-based detectors show substantial performance drops (average F1 $\downarrow$ 14.8%) and increasingly unstable reasoning traces, with even more severe failures under adversarial evidence injection. Our findings uncover fundamental vulnerabilities in existing MMD systems and suggest an urgent need for more resilient approaches in the GenAI era.

Introduction

Multimodal misinformation, often in the form of image-text combinations, poses escalating threats to public discourse, societal trust, and civic stability (Bovet and Makse 2019; Murayama 2021; Wang et al. 2024a; Wu et al. 2025c). To counter this threat, Large Vision-Language Models (LVLMs) (Hurst et al. 2024; Anil et al. 2023; Liu et al. 2024a; Meta 2024) has emerged as a dominant approach for multimodal misinformation detection (MMD), offering strong multimodal reasoning and retrieval-augmented verification capabilities (Kangur et al. 2025; Qi et al. 2024; Braun et al. 2024; Li et al. 2025a; Geng et al. 2024; Wu et al. 2025b).

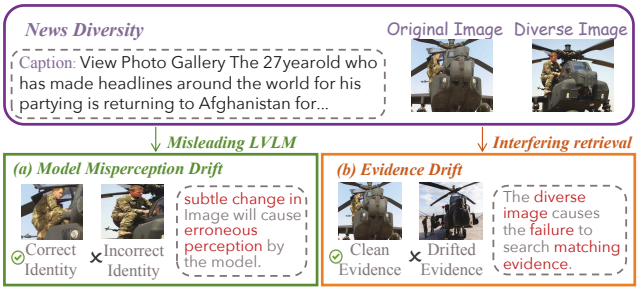

Figure 1: Illustration of **multi-level drift** induced by **GenAI-driven news diversity**, comprising (1) model-level misperception drift and (2) evidence-level drift.

However, the rise of generative AI (GenAI) systems such as GPT-4o (Hurst et al. 2024) is reshaping the information landscape. GenAI enables the creation of highly diverse and stylistically varied content at scale (Kiskola et al. 2025; Davenport and Mittal 2023; Nishal and Diakopoulos 2024; Kieslich, Diakopoulos, and Helberger 2024), fundamentally transforming how news is composed, visualized, and disseminated. We refer to this phenomenon as **GenAI-driven News Diversity**, which manifests in two key forms: (1) *Controlled News Diversity*, where users produce semantically consistent variants of the same news content through rewording, paraphrasing, or shifts in visual framing; and (2) *Open-ended News Diversity*, where GenAI generates entirely novel content via text-to-image synthesis or narrative reconstruction.

While this diversity enhances expressiveness and audience reach, it also introduces substantial challenges for LVLM-based MMD systems. Most existing models either assess image-text coherence using internal knowledge (Wang et al. 2024b) or retrieve external evidence for fact verification (Li et al. 2025b; Qi et al. 2024). Both approaches become fragile under content variation, which gives rise to what we term **multi-level drift**. As illustrated in Figure 1, we identify two key drift phenomena: (1) **Model-Level Mis-**

*Corresponding authors.

perception Drift: Variations in surface form, including lexical, syntactic, or visual aspects, may distort the model's perception, even when semantics remain unchanged. This issue is exacerbated by LVLMs' known hallucination tendencies (Bai et al. 2024; Liu et al. 2024b). (2) **Evidence-Level Drift**: Content diversity compromises evidence retrieval by reducing the likelihood of finding exact matches, leading to loosely relevant or misleading external evidence (Chung et al. 2023; Xiao and Wang 2021; Feng et al. 2020). Beyond natural drift, GenAI also enables adversarial manipulation through **guided retrieval attacks**, wherein malicious actors flood the web with strategically crafted synthetic content. Such contamination can bias retrieval pipelines, poisoning the fact-checking process and degrading LVLM performance (Xu et al. 2024; Dai et al. 2023a). Despite their growing risk, these vulnerabilities remain insufficiently studied.

To bridge this gap, we conduct a systematic investigation into how GenAI-driven news diversity affects LVLM-based misinformation detection. We pose three core research questions: (1) How does GenAI-driven news diversity affect the robustness of current LVLM-based MMD systems? (2) To what extent does adversarial evidence contamination degrade factual verification via guided retrieval of fabricated evidence? (3) How does content diversity influence LVLMs' internal reasoning behavior during fact-checking?

To this end, we introduce DRIFTBENCH, a large-scale benchmark consisting of 16,000 news instances spanning six diversified categories across both image and text modalities, with human validation to ensure semantic fidelity. Specifically, we source high-quality real and fake samples from NewsCLIPpings (Luo and Trevor Darrell 2021) and systematically diversify them using an automated GenAI-based pipeline grounded in our news diversity taxonomy. For real news, we apply *controlled* diversification to the image, text, or both while preserving factual consistency. For fake news, we incorporate both *controlled* and *open-ended* transformations, rewriting text to fabricate false narratives and generating misleading images aligned with the fabricated content. The resulting dataset includes eight structured categories, consisting of six diversified variants and two original classes, thereby enabling rigorous evaluation of LVLM robustness under GenAI-driven content variation.

With DRIFTBENCH, we design three corresponding evaluation tasks: (1) **Performance under Diversity**: Evaluate how SOTA LVLM-based detectors handle controlled and open-ended content variation, focusing on degradation due to model- and evidence-level drift. (2) **Robustness to Malicious Evidence Contamination**: Simulate guided retrieval attacks by injecting GenAI-generated misleading evidence into the verification pipeline. (3) **Reasoning Behavior Analysis**: Employ an LVLM-as-a-Judge framework (Li et al. 2024; Chen et al. 2024; Pu et al. 2025) to assess models' explanation quality and attribution patterns under diverse input conditions.

Our empirical analysis covers both general-purpose LVLMs and task-specific detectors. Results show that GenAI-driven diversity leads to significant performance drops across all models (average F1 $\downarrow$ 14.8%), with malicious evidence causing further degradation. Notably, we observe that real and fake content are differently affected by diversity; vanilla and task-specific models exhibit divergent sensitivities; and diverse inputs distort reasoning trajectories in non-trivial ways. Our findings reveal fundamental limitations in current LVLM-based detection pipelines and lay the groundwork for designing robust, diversity-aware MMD systems to accommodate GenAI-driven news diversity.

Related Work

LVLM-Based MMD Approaches

LVLM-based multimodal fact-checking is the current mainstream method for multimodal misinformation detection. SNIFFER (Qi et al. 2024) performs two-stage fine-tuning based on InstructBLIP (Dai et al. 2023b), aiming to detect misinformation from the perspective of entity matching and evidence verification. CMIE (Li et al. 2025b) and E2LVLM (Wu et al. 2025a) further explore how to discover deeper relationships and optimize the selection and usage of external evidence. LRQ-FACT (Beigi et al. 2024) and LEMMA (Xuan et al. 2024) optimize retrieval strategies to acquire more targeted external evidence. These methods fully investigate the capabilities of LVLMs in retrieval, comprehension, and reasoning, and are optimized around the acquisition and utilization of high-quality evidence. However, most of them assume that the input claim is static and expressed in a single form, lacking systematic investigation and evaluation of such methods' performance under news diversity and the resulting multi-level drift.

Multimodal Misinformation Benchmarks

To evaluate multimodal misinformation detection (MMD), several benchmarks have been introduced. DGM4 (Shao, Wu, and Liu 2023) targets fine-grained visual and textual manipulations. NewsCLIPpings (Luo and Trevor Darrell 2021) constructs out-of-context (OOC) cases by mismatching images and text with external evidence. VERITE (Papadopoulos et al. 2024) collects real-world OOC misinformation from fact-checking platforms to analyze modality bias. MiRAGenNews (Huang et al. 2024) employs GenAI techniques to synthesize misinformation image–text pairs, while MMFakeBench (Liu et al. 2024c) aggregates misinformation from both real and GenAI sources to assess robustness under mixed conditions. MFC-Bench (Wang et al. 2024b) probes LVLMs' internal fact-checking capabilities.

In contrast, DRIFTBENCH is the first benchmark to systematically examine challenges introduced by **GenAI-driven news diversity**, with a focus on two overlooked forms of drift: **model misperception drift** and **evidence drift**. By generating semantically diverse real and fabricated variants through both controlled and open-ended transformations, it addresses a critical gap in existing benchmarks. A comparative overview is provided in Table 1.

DRIFTBENCH

To systematically evaluate how *GenAI-driven news diversity* induces multi-level drift in LVLM-based multimodal misinformation detection, we introduce DRIFTBENCH, a large-scale benchmark of 16,000 news instances across eight cat-

Benchmark	Information Type	Modality		Generated Content	External Evidence	News Diversity	Drift
		Textual	Visual				
DGM4 (Shao, Wu, and Liu 2023)	Multimedia manipulation	✔	✔	-	-	-	-
NewsCLIPpings (Luo and Trevor Darrell 2021)	OOC misinformation	✔	✔	-	✔	-	-
VERITE (Papadopoulos et al. 2024)	Real-world misinformation	✔	✔	-	✔	-	-
MMFakeBench (Liu et al. 2024c)	Multi-source misinformation	✔	✔	✔	-	-	-
MFC-Bench (Wang et al. 2024b)	Multimodal fact-checking	✔	✔	-	-	-	-
MiRAGenNews (Huang et al. 2024)	AI-generated misinformation	✔	✔	✔	-	-	-
DRIFTBENCH (ours)	News diversity & Drift	✔	✔	✔	✔	✔	✔

Table 1: Comparison between DRIFTBENCH and prior benchmarks on multimodal misinformation detection.

egories. These categories comprise two base types (real and fake news with human-written text and authentic images) and six diversification strategies. Built through an automated GenAI-based synthesis pipeline applied to real-world news, DRIFTBENCH enables scalable, controlled, and semantically aligned diversification. Each instance contains an image–text pair—either original or diversified—together with externally retrieved web evidence. An overview of the dataset construction and evaluation setup is shown in Figure 2. Our code, data, and supplementary materials are publicly available[1].

Taxonomy of GenAI-Driven News Diversity

To comprehensively assess the impact of content variation on multimodal misinformation detection, we propose a taxonomy encompassing eight types of image-text instances. These are defined along two orthogonal dimensions: **(1) news authenticity** (*Real* vs. *Fake*) and **(2) diversification strategy** (*Controlled* vs. *Open-Ended*). Controlled diversity involves deliberate and precise modifications to image or text while preserving semantic consistency. Open-ended diversity allows flexible, generative transformations, often introducing false or misleading information, thereby increasing uncertainty.

We begin with two foundational types that serve as the seed for diversification:

- **R_OI_OT**: A real news item sourced from a reputable outlet, composed of an authentic image (*OI*) and a human-written text (*OT*) that are semantically aligned.
- **F_OI_OT**: An out-of-context (OOC) misinformation instance (Luo and Trevor Darrell 2021; Qi et al. 2024), created by mismatching the image and text from two different but individually trustworthy news reports—misleading when paired together.

Building on these base types, we define six variant categories by modifying the image, text, or both using controlled or open-ended GenAI techniques:

- **R_DI_OT**: A real news instance where the image is diversified using GenAI (*DI*), while the original text remains unchanged.
- **R_OI_DT**: A real news instance where the text is diversified (*DT*) while preserving the original image.
- **R_DI_DT**: A real instance where both the image and text are diversified in a semantically faithful manner.
- **F_DI_OT**: An OOC fake instance where the image is replaced by a GenAI-diversified version, while keeping the mismatched original text.
- **F_OI_FT**: A fake instance where the original image is paired with a GenAI-fabricated text (*FT*) that conveys a false narrative.
- **F_FI_FT**: A fully fabricated fake instance where both the image (*FI*) and text (*FT*) are GenAI-synthesized.

This taxonomy enables a systematic evaluation of how controlled and open-ended diversification strategies, applied to both real and fake instances, influence the robustness and reasoning behavior of LVLM-based misinformation detection systems.

Dataset Creation

Based on the proposed taxonomy, we construct DRIFTBENCH, a large-scale benchmark comprising real and fake news instances systematically diversified through GenAI-driven content variation. We begin by sourcing 4,000 high-quality samples (2,000 real and 2,000 fake) from the NewsCLIPpings dataset (Luo and Trevor Darrell 2021). Real instances comprise semantically aligned image-caption pairs published by reputable news outlets. Fake instances are constructed via OOC manipulation, where the image and text are drawn from two different trustworthy reports. While individually accurate, their combination results in a misleading narrative with high surface plausibility.

We apply different transformation strategies based on the type of instance. For real samples, we apply only *controlled* diversification to preserve factual integrity. For fake samples, we introduce both controlled and *open-ended* diversity to simulate more flexible or fabricated misinformation.

Text diversifications and fabrications are generated using GPT-4o (Hurst et al. 2024). For text diversification (*DT*), we prompt the model to rephrase the original text while preserving its semantics. For text fabrication (*FT*), we prompt the model to generate a false narrative by altering the described event. Image diversifications (*DI*) are generated using *FLUX.1 Redux [dev]* (an open-source image variant generation model) (Labs et al. 2025), the model generates visually varied but semantically consistent depictions of the original content. Image fabrications (*FI*) are generated using *FLUX.1 [dev]* (an open-source text-to-image model)

[1]https://github.com/fanxiao15/DriftBench

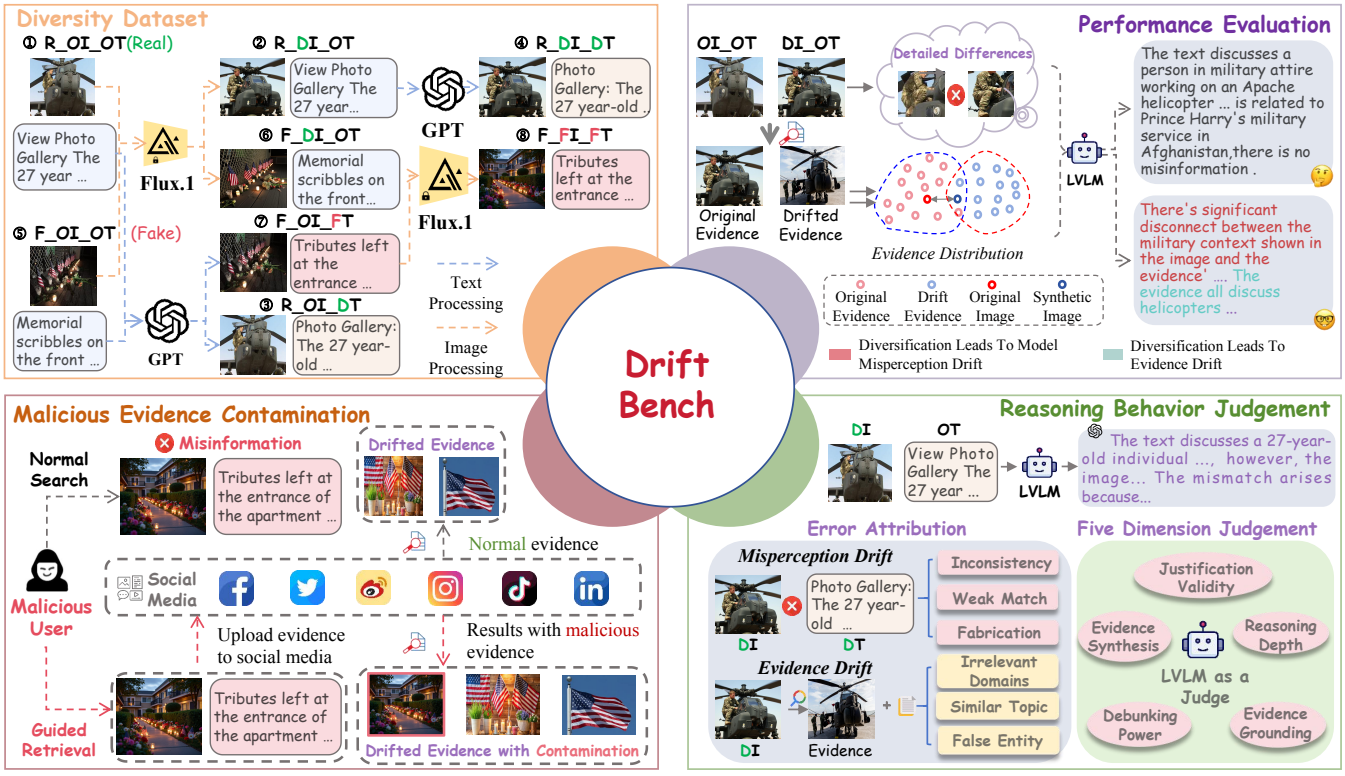

Figure 2: Overview of DRIFTBENCH, illustrating the construction of diversified news variants and the three evaluation tasks.

(Labs 2024), we condition the image generation on the fabricated narrative *FT*, yielding fully synthetic misinformation instances.

Implementation details and prompt templates are provided in Appendix A.1.

Evidence Retrieval. For each image-text pair, we retrieve external evidence to support LVLM-based multimodal fact-checking. For image-based retrieval, we use a web crawler to identify webpages containing the image and extract their page titles as evidence. For text-based retrieval, we adopt the query generation strategy from LEMMA (Xuan et al. 2024), using the Serper API to retrieve relevant web documents.

Data Quality Assessment

Preserving semantic meaning is essential for high-quality news diversification. To assess the quality of the diversified instances in DRIFTBENCH, we conducted a human evaluation to determine whether the core meaning of the original content is retained after diversification. Since the *OI_FT* and *FI_FT* types involve fake news with fabricated content, they represent open-ended generation and lack a verifiable semantic reference. Therefore, we restrict our evaluation to the four types of diversified *real* instances, where semantic fidelity can be reliably assessed.

We randomly sampled 40 image-text pairs from each diversification type, yielding a total of 160 diversified instances. Three graduate students served as annotators. Each annotator independently reviewed all 160 original-diversified pairs and provided a binary judgment on whether the diversified version preserved the core meaning of the original. Annotation guidelines and examples are detailed in Appendix A.2.

We report two metrics to assess data quality: (1) semantic preservation accuracy, computed from aggregated human annotations, and (2) inter-annotator agreement using Fleiss' κ (Fleiss 1971). Results show that DRIFTBENCH exhibits high semantic consistency and labeling reliability, with 98.6% accuracy and a Fleiss' κ of 0.872 for real-instance variants, and 96.6% accuracy with a Fleiss' κ of 0.741 for controlled fake variants.

Evaluation Tasks and Metrics

We design three evaluation tasks to assess: (1) performance degradation under *model-level misperception drift* and *evidence-level drift* introduced by news diversity; (2) robustness to *malicious evidence contamination*; and (3) the influence of news diversity on *reasoning behavior and explanation quality*.

Task 1: Performance Analysis under News Diversity. We evaluate how both Vanilla LVLMs and task-specific MMD models perform on diversified instances in DRIFTBENCH, measuring the impact of semantic-preserving (controlled) and misleading (open-ended) variations on the models' MMD performance.

Infer Type	Data Type	Accuracy	Real			Fake		
			Precision	Recall	F1	Precision	Recall	F1
GPT-4o-mini	Realistic	83.3	89.4	75.6	81.9	78.9	91.1	84.5
	Diversified	64.4 (-18.9)	83.7 (-6.0)	35.7 (-39.9)	50.0 (-31.9)	59.1 (-19.8)	93.1 (+2.0)	72.3 (-12.2)
Claude-3.7-Sonnet	Realistic	88.3	87.7	89.2	88.4	89.0	87.5	88.2
	Diversified	72.4 (-15.9)	86.1 (-1.6)	53.5 (-35.7)	66.0 (-22.4)	66.3 (-22.7)	91.4 (+3.9)	76.8 (-11.4)
Qwen-VL	Realistic	73.7	67.6	91.1	77.6	86.3	56.3	68.2
	Diversified	63.7 (-10.0)	63.9 (-3.7)	62.7 (-28.4)	63.3 (-14.3)	63.5 (-22.8)	64.7 (+8.4)	64.1 (-4.1)
CMIE	Realistic	90.9	88.8	93.6	91.1	93.2	88.2	90.6
	Diversified	72.4 (-18.5)	72.2(-16.6)	72.8(-20.8)	72.5(-18.6)	72.6(-20.6)	72.1(-16.1)	72.3(-18.3)
SNIFFER	Realistic	84.3	78.4	93.2	85.1	92.3	76.2	83.5
	Diversified	73.5 (-10.8)	71.2 (-7.2)	76.0 (-17.2)	73.5 (-11.6)	75.8 (-16.5)	71.1 (-5.1)	73.4 (-10.1)
LEMMA	Realistic	79.6	73.4	92.8	82.0	90.2	66.5	76.5
	Diversified	68.9 (-10.7)	64.7 (-8.7)	83.1 (-9.7)	72.7 (-9.3)	76.3 (-13.9)	54.7 (-11.8)	63.7 (-12.8)

Table 2: Performance comparison of LVLM-based multimodal misinformation detection methods under the DI_OT setting within Controlled News Diversity. Since DI_OT applies to both real and fake instances, we report results (in percentage) for this category as a representative example.

Task 2: Robustness to Malicious Evidence Contamination. To simulate guided retrieval attacks, we inject adversarially generated evidence (i.e., crafted to falsely support or refute the claim) into the evidence set. This mimics real-world scenarios where search engines are manipulated by fabricated content. We evaluate models' robustness by measuring performance drop before and after evidence poisoning, isolating the sensitivity of each model to retrieval contamination. The generation procedure and insertion strategy are detailed in Appendix A.3.

Task 3: Reasoning Behavior Analysis. We adopt an LVLM-as-Judge evaluation framework to analyze model reasoning under drift. This includes:

- **Error Attribution:** We manually label failure cases according to two error types: (1) model-level misperception drift and (2) evidence-level drift. A detailed taxonomy is provided in Appendix A.4.
- **Explanation Evaluation:** We assess the quality of model-generated explanations across five dimensions: justification validity, evidence grounding, evidence synthesis, reasoning depth, and debunking power. Full guidelines and scoring rubrics are included in Appendix A.5.

Following established practices in LVLM-based MMD evaluation (Li et al. 2025b; Xuan et al. 2024; Qi et al. 2024), we report Accuracy and F1 as the primary metrics for detection performance.

Experiments

Experimental Setup

Using DRIFTBENCH, we conduct a comprehensive evaluation of both representative Vanilla LVLMs and state-of-the-art LVLM-based MMD approaches.

Vanilla LVLMs. We assess three widely used models: (1) **GPT-4o-mini** (Hurst et al. 2024): a proprietary large vision-language model. (2) **Claude-3.7-Sonnet** (Anthropic 2025): a large reasoning model with multimodal capabilities. (3) **Qwen-VL-Plus** (Bai et al. 2023): a strong open-source vision-language model. Prompting strategies and instruction templates used for Vanilla LVLMs are detailed in Appendix (Figure 11).

LVLM-Based MMD Methods. We also evaluate three state-of-the-art MMD approaches. (1) **SNIFFER** (Qi et al. 2024), which fine-tunes LVLMs and retrieves external evidence directly through image-based querying. (2) **CMIE** (Li et al. 2025b), which models deep semantic alignment between image and text while mitigating noise in retrieved evidence. and (3) **LEMMA** (Xuan et al. 2024), which employs a modular pipeline that first generates optimized queries and then retrieves high-quality evidence.

Performance under GenAI-Driven Diversity

Multimodal LVLMs exhibit limited generalization under news diversity, with asymmetric performance between real and fake Instance. As shown in Table 2, GenAI-driven news diversity causes significant performance degradation across all evaluated models. This demonstrates that even controlled, semantically consistent perturbations can severely compromise the reliability of current LVLM-based detectors. Multi-level drift induces a bias toward overpredicting "fake". For real instances, both precision and recall drop substantially due to a combination of model-level misperception drift (e.g., hallucinated misalignment between modalities) and evidence-level drift (e.g., irrelevant or misleading retrieval). For fake instances, while recall improves, precision declines. This suggests that the models become overconfident in detecting misinformation, even when uncertain.

Image diversity has a stronger impact than text diversity, and open-ended diversity is more disruptive than controlled diversity. As shown in Figure 3, text diversification (*OI_DT*) has minimal impact, while performance declines are primarily driven by image diversification (*DI_OT*), detailed results in Appendix A.6. Notably, open-ended diversity leads to greater performance drops than controlled diversity. In these cases, the generated images may be perceived as too well-aligned with the fabricated text, mis-

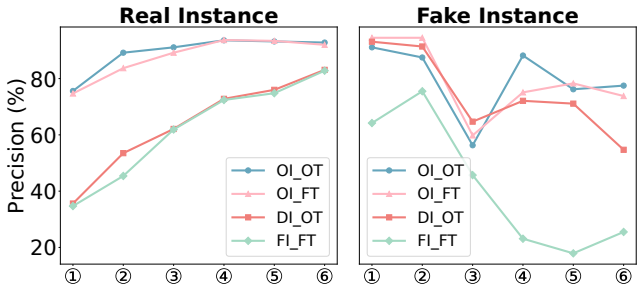

Figure 3: Precision (%) of representative LVLM-based MMD approaches (①: *GPT-4o-mini*; ②: *Claude-3.7*; ③: *Qwen-VL*; ④: *CMIE*; ⑤: *Sniffer*; ⑥: *Lemma*) under all types of *controlled* and *open-ended* news diversity.

leading the model into incorrectly classifying the instance as real. This highlights a fundamental challenge posed by GenAI-generated misinformation, where visual-textual coherence can be artificially inflated.

Model performance degrades more sharply with increasing evidence drift severity. Figure 4 illustrates how detection accuracy declines as the degree of evidence drift increases, ranging from mild (Degree 1) to severe (Degree 5). While the degradation is not strictly linear, the trend is clear: performance steadily worsens from low to moderate to high drift levels. As drift severity grows, the retrieved evidence becomes increasingly misaligned or ambiguous, making it harder for models to ground their predictions accurately. These findings highlight the urgent need for drift-aware verification mechanisms that can remain robust in the face of evidence drifts induced by GenAI-driven diversity.

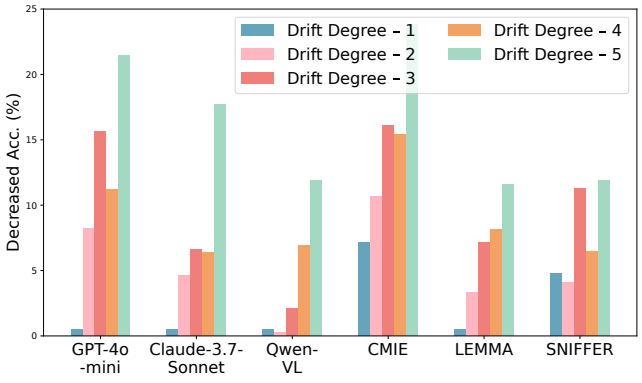

Figure 4: Accuracy degradation of MMD methods under increasing evidence drift severity in the *DI_OT* setting. Higher drift degrees lead to more pronounced performance drops.

Robustness to Evidence Contamination

Contaminated evidence further degrades model performance and can steer predictions in misleading directions. As discussed in Section , malicious actors can manipulate retrieval pipelines by injecting fabricated or misleading content into the evidence set. Table 3 shows that introducing such adversarial evidence causes a clear performance drop across all evaluated methods. Detailed results in Appendix A.6 indicate that real instances are increasingly misclassified as fake when contaminated evidence is present. For fake instances, the misleading evidence often supports the false narrative, leading models to mistakenly classify them as real and thus producing the outcome intended by the adversary.

Infer Type	Data Type	Accuracy
GPT-4o-mini	DI_OT	64.4
	Polluted	44.6 (-19.8)
Claude-3.7-Sonnet	DI_OT	72.4
	Polluted	66.7 (-5.7)
Qwen-VL	DI_OT	63.7
	Polluted	37.2 (-26.5)
CMIE	DI_OT	72.4
	Polluted	66.8 (-5.6)
SNIFFER	DI_OT	73.5
	Polluted	70.3 (-3.2)
LEMMA	DI_OT	68.9
	Polluted	41.9 (-27.0)

Table 3: Comparison of LVLM-based multimodal misinformation detection under malicious evidence contamination, evaluated on the DI_OT setting with controlled news diversity.

Notably, Vanilla LVLMs show greater performance fluctuation under contamination, suggesting they are more vulnerable to evidence interference in noisy environments. In comparison, task-specific MMD approaches such as CMIE and LEMMA maintain more stable F1 scores, due to their dedicated mechanisms for filtering and interpreting evidence. However, their performance also declines when exposed to adversarial content. This highlights the critical need for retrieval-aware defenses that can detect and mitigate the effects of evidence manipulation in LVLM-based misinformation detection systems.

Model Reasoning Behavior Analysis

Visual-textual consistency influences model misclassification tendencies across real and fake instances. We analyze error attribution across diversified instances, as illustrated in Figure 6. While final misjudgments often result from both model misperception and evidence drift, their relative contributions differ between real and fake news. For real instances, errors are more frequently caused by model-level misperception drift, suggesting that when image-text consistency remains strong, models tend to rely on this internal alignment for prediction. In contrast, for fake instances, errors are more often attributed to evidence-level drift, indicating that once image-text alignment is questioned, LVLMs tend to shift their focus to external evidence for verification.

Vanilla LVLMs and MMD-specific models show distinct reasoning behaviors under news diversity. As shown in Figure 3, Vanilla LVLMs are more sensitive to variation in real instances, while MMD-specific models exhibit greater ro-

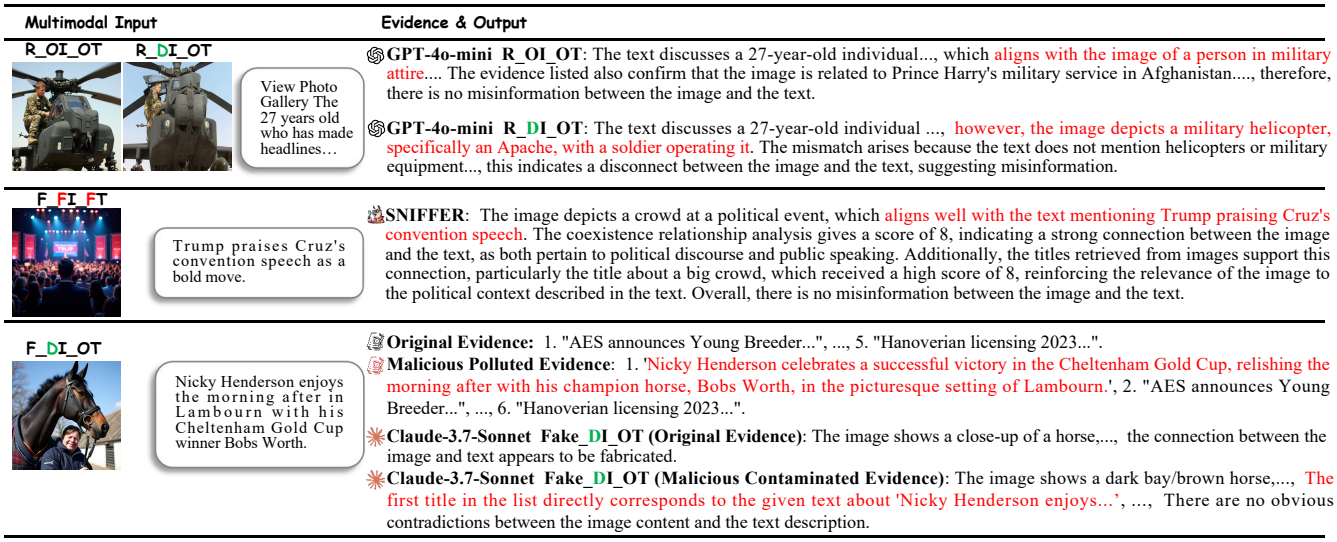

Figure 5: Reasoning explanations generated by LVLMs under controlled settings (R_OI_OT, R_DI_OT), open ended scenarios (F_FI_FT), and cases involving malicious evidence contamination (F_DI_OT).

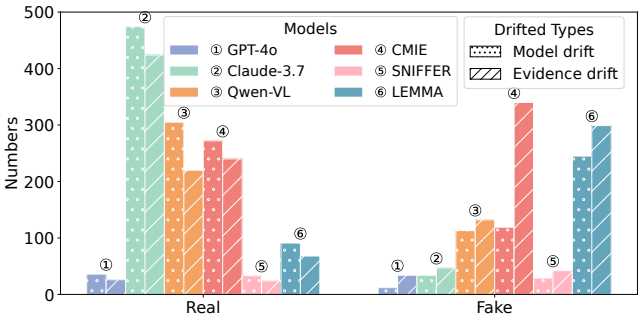

Figure 6: Error attribution showing the distribution of error types across different methods for real and fake instances.

bustness due to their dedicated vision-text alignment modules. However, for fake instances, MMD-specific models experience larger performance drops, primarily due to their reliance on external evidence, which is more vulnerable to drift. This highlights the dual challenge: LVLMs are perceptually fragile, while retrieval-based models are susceptible to evidence contamination.

Different news diversity types affect explanation quality and reasoning dynamics in distinct ways. We report explanation evaluation results across five dimensions in Appendix A.5. Text-based diversification typically yields better explanation quality, potentially because LLMs favor content that resembles their generative style. Open-source LVLMs show more instability, while reasoning-focused models maintain consistent explanation quality but still fall short in accuracy due to lacking specialized modules.

Under diversification, Vanilla LVLMs exhibit weakened evidence integration yet maintain relatively stable reasoning depth, often misusing evidence without full synthesis. In contrast, task-specific MMD baselines show a correlated decline in both evidence usage and reasoning depth, revealing their dependency on structured reasoning components. These findings underscore the need for more adaptive and drift-resilient reasoning mechanisms in the face of GenAI-driven variation.

Case Study

In Figure 5, we demonstrate the impact of different types of news diversity and malicious evidence contamination on LVLMs' ability to detect misinformation. Controlled news diversity leads to the model's misperception drift, while open-ended diversity causes the model to perceive the image-text pair as well-aligned, resulting in misclassification. As for evidence contamination, the model's judgment is influenced by malicious evidence, steering predictions in misleading directions.

Conclusion

We present DRIFTBENCH, the first benchmark explicitly designed to evaluate the vulnerabilities of LVLM-based misinformation detection in the face of GenAI-driven news diversity. Our diversified benchmark demonstrates that both controlled and open-ended content variation, along with adversarial evidence contamination, can substantially weaken the performance of current LVLM-based MMD systems. Beyond accuracy degradation, our analysis reveals distinct reasoning behaviors and error patterns among different model classes, reflecting varied reliance on visual-textual coherence and external evidence. These findings emphasize the need for robust, diversity-aware strategies to ensure reliable misinformation detection. DRIFTBENCH provides a foundation for advancing trustworthy verification methods in an increasingly complex and manipulated information landscape.

Acknowledgments

This work is supported by the Yunnan Province expert workstations (Grant No. 202305AF150078), National Natural Science Foundation of China (Grant Nos. 62162067, 62562061, 62502422 and 62462067), Yunnan Fundamental Research Project (Grant Nos. 202401AT070474, 202501AU070059), Yunnan Province Special Project (Grant No.202403AP140021), Yunnan Provincial Department of Education Science Research Project (Grant Nos. 2025J0006, 2024J0010 and 2025J0007), Scientific Research and Innovation Project of Postgraduate Students in the Academic Degree of Yunnan University (KC-4248590) and China Scholarship Council (CSC) program. This research is also supported by the Ministry of Education, Singapore, under its MOE AcRF TIER 3 Grant (MOE-MOET32022-0001).

References

Anil, R.; Borgeaud, S.; Wu, Y.; Alayrac, J.; Yu, J.; and et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *CoRR*, abs/2312.11805.

Anthropic. 2025. Claude 3.7 Sonnet and Claude Code. https://www.anthropic.com/news/claude-3-7-sonnet.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2): 3.

Bai, Z.; Wang, P.; Xiao, T.; He, T.; Han, Z.; Zhang, Z.; and Shou, M. Z. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Beigi, A.; Jiang, B.; Li, D.; Tan, Z.; Shaeri, P.; Kumarage, T.; Bhattacharjee, A.; and Liu, H. 2024. Can LLMs Improve Multimodal Fact-Checking by Asking Relevant Questions? *arXiv preprint arXiv:2410.04616*.

Bovet, A.; and Makse, H. A. 2019. Influence of fake news in Twitter during the 2016 US presidential election. *Nature communications*, 10(1): 7.

Braun, T.; Rothermel, M.; Rohrbach, M.; and Rohrbach, A. 2024. Defame: Dynamic evidence-based fact-checking with multimodal experts. *arXiv preprint arXiv:2412.10510*.

Chen, D.; Chen, R.; Zhang, S.; Wang, Y.; Liu, Y.; Zhou, H.; Zhang, Q.; Wan, Y.; Zhou, P.; and Sun, L. 2024. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

Chung, H.; Lee, N.; Lee, H.; Cho, Y.; and Woo, J. 2023. GuaRD: Guaranteed robustness of image retrieval system under data distortion turbulence. *Plos one*, 18(9): e0288432.

Dai, S.; Zhou, Y.; Pang, L.; Liu, W.; Hu, X.; Liu, Y.; Zhang, X.; and Xu, J. 2023a. Llms may dominate information access: Neural retrievers are biased towards llm-generated texts. *arXiv preprint arXiv:2310.20501*.

Dai, W.; Li, J.; Li, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P. N.; and Hoi, S. 2023b. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36: 49250–49267.

Davenport, T. H.; and Mittal, N. 2023. How Generative AI Is Changing Creative Work. 2022. *Dostupno na: https://hbr. org/2022/11/how-generative-ai-is-changing-creativework [Pristupljeno: srpanj 2023.]*.

Feng, Y.; Chen, B.; Dai, T.; and Xia, S.-T. 2020. Adversarial attack on deep product quantization network for image retrieval. In *Proceedings of the AAAI conference on Artificial Intelligence*, volume 34, 10786–10793.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5): 378.

Geng, J.; Kementchedjhieva, Y.; Nakov, P.; and Gurevych, I. 2024. Multimodal large language models to support real-world fact-checking. *arXiv preprint arXiv:2403.03627*.

Huang, R.; Dugan, L.; Yang, Y.; and Callison-Burch, C. 2024. MiRAGeNews: Multimodal Realistic AI-Generated News Detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, 16436–16448. Association for Computational Linguistics.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; and et al. 2024. GPT-4o System Card. *CoRR*, abs/2410.21276.

Kangur, U.; Agrawal, K.; Singh, Y.; Sabir, A.; and Sharma, R. 2025. MultiReflect: Multimodal Self-Reflective RAG-based Automated Fact-Checking. In *Proceedings of the 1st Workshop on Multimodal Augmented Generation via Multimodal Retrieval (MAGMaR 2025)*, 1–17.

Kieslich, K.; Diakopoulos, N.; and Helberger, N. 2024. Anticipating impacts: using large-scale scenario-writing to explore diverse implications of generative AI in the news environment. *AI and Ethics*, 1–23.

Kiskola, J.; Rydenfelt, H.; Olsson, T.; Haapanen, L.; Vänttinen, N.; Nelimarkka, M.; Vigren, M.; Laaksonen, S.-M.; and Lehtiniemi, T. 2025. Generative AI and News Consumption: Design Fictions and Critical Analysis. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–18.

Labs, B. F. 2024. FLUX. https://github.com/black-forest-labs/flux.

Labs, B. F.; Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; Kulal, S.; Lacey, K.; Levi, Y.; Li, C.; Lorenz, D.; Müller, J.; Podell, D.; Rombach, R.; Saini, H.; Sauer, A.; and Smith, L. 2025. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. arXiv:2506.15742.

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. *arXiv preprint arXiv:2505.23449*.

Li, H.; Dong, Q.; Chen, J.; Su, H.; Zhou, Y.; Ai, Q.; Ye, Z.; and Liu, Y. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 26296–26306.

Liu, H.; Xue, W.; Chen, Y.; Chen, D.; Zhao, X.; Wang, K.; Hou, L.; Li, R.; and Peng, W. 2024b. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Liu, X.; Li, Z.; Li, P.; Huang, H.; Xia, S.; Cui, X.; Huang, L.; Deng, W.; and He, Z. 2024c. Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms. *arXiv preprint arXiv:2406.08772*.

Luo, G.; and Trevor Darrell, A. R. 2021. NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, 6801–6817. Association for Computational Linguistics.

Meta, A. 2024. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI Blog. Retrieved December*, 20: 2024.

Murayama, T. 2021. Dataset of fake news detection and fact verification: a survey. *arXiv preprint arXiv:2111.03299*.

Nishal, S.; and Diakopoulos, N. 2024. Envisioning the applications and implications of generative AI for news media. *arXiv preprint arXiv:2402.18835*.

Papadopoulos, S.-I.; Koutlis, C.; Papadopoulos, S.; and Petrantonakis, P. C. 2024. Verite: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. *International Journal of Multimedia Information Retrieval*, 13(1): 4.

Pu, S.; Wang, Y.; Chen, D.; Chen, Y.; Wang, G.; Qin, Q.; Zhang, Z.; Zhang, Z.; Zhou, Z.; Gong, S.; et al. 2025. Judge Anything: MLLM as a Judge Across Any Modality. *arXiv preprint arXiv:2503.17489*.

Qi, P.; Yan, Z.; Hsu, W.; and Lee, M. L. 2024. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13052–13062.

Shao, R.; Wu, T.; and Liu, Z. 2023. Detecting and Grounding Multi-Modal Media Manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, B.; Wang, S.; Li, C.; Guan, R.; and Li, X. 2024a. Harmfully manipulated images matter in multimodal misinformation detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2262–2271.

Wang, S.; Lin, H.; Luo, Z.; Ye, Z.; Chen, G.; and Ma, J. 2024b. Mfc-bench: Benchmarking multimodal fact-checking with large vision-language models. *arXiv preprint arXiv:2406.11288*.

Wu, J.; Fu, Y.; Yu, N.; and Fu, G. 2025a. E2lvlm: Evidence-enhanced large vision-language model for multimodal out-of-context misinformation detection. *arXiv preprint arXiv:2502.10455*.

Wu, J.; Fu, Z.; Wang, H.; Li, F.; and Kan, M.-Y. 2025b. Beyond the Crowd: LLM-Augmented Community Notes for Governing Health Misinformation. *arXiv preprint arXiv:2510.11423*.

Wu, J.; Li, F.; Fu, Z.; Kan, M.-Y.; and Hooi, B. 2025c. Seeing Through Deception: Uncovering Misleading Creator Intent in Multimodal News with Vision-Language Models. *arXiv preprint arXiv:2505.15489*.

Xiao, Y.; and Wang, C. 2021. You see what I want you to see: Exploring targeted black-box transferability attack for hash-based image retrieval systems. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1934–1943.

Xu, S.; Hou, D.; Pang, L.; Deng, J.; Xu, J.; Shen, H.; and Cheng, X. 2024. Invisible relevance bias: Text-image retrieval models prefer ai-generated images. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, 208–217.

Xuan, K.; Yi, L.; Yang, F.; Wu, R.; Fung, Y. R.; and Ji, H. 2024. LEMMA: towards LVLM-enhanced multimodal misinformation detection with external knowledge augmentation. *arXiv preprint arXiv:2402.11943*.