

What's Left Unsaid? Detecting and Correcting Misleading Omissions in Multimodal News Previews

Fanxiao Li¹, Jiaying Wu², Tingchao Fu¹, Dayang Li¹, Herun Wan³,
Wei Zhou⁴*, Min-Yen Kan²

¹ School of Information Science and Engineering, Yunnan University,

² National University of Singapore, ³ Xi'an Jiaotong University,

⁴ School of Engineering, Yunnan University

lifanxiao@stu.ynu.edu.cn, jiayingwu@u.nus.edu, zwei@ynu.edu.cn

Abstract

Even when factually correct, social-media news previews (image–headline pairs) can induce **interpretation drift**: by selectively *omitting* crucial context, they lead readers to form judgments that diverge from what the full article supports. This covert harm is subtler than explicit misinformation, yet remains underexplored. To address this gap, we develop a multi-stage pipeline that simulates preview-based and context-based understanding, enabling construction of the MM-MISLEADING benchmark. Using MM-MISLEADING, we systematically evaluate open-source LVLMs and uncover pronounced blind spots in *omission-based misleadingness detection*. We further propose OM-GUARD, which combines (1) **Interpretation-Aware Fine-Tuning** for misleadingness detection and (2) **Rationale-Guided Misleading Content Correction**, where explicit rationales guide headline rewriting to reduce misleading impressions. Experiments show that OM-GUARD lifts an 8B model's detection accuracy to the level of a 235B LVLM while delivering markedly stronger end-to-end correction. Further analysis shows that misleadingness usually arises from local narrative shifts, such as missing background, instead of global frame changes, and identifies image-driven cases where text-only correction fails, underscoring the need for visual interventions.¹

1 Introduction

The governance of online misinformation is evolving beyond the detection of explicit fabrications. While existing efforts primarily target factually incorrect content (Qi et al., 2024; Wan et al., 2025b), a more insidious risk arises from **omission-based deception** (Van Swol and Braun, 2014; Allen et al., 2024), where creators subtly reshape impressions by omitting essential background context. This risk

*Corresponding Author

¹Data and code are available at <https://github.com/fanxiao15/OMGuard>.

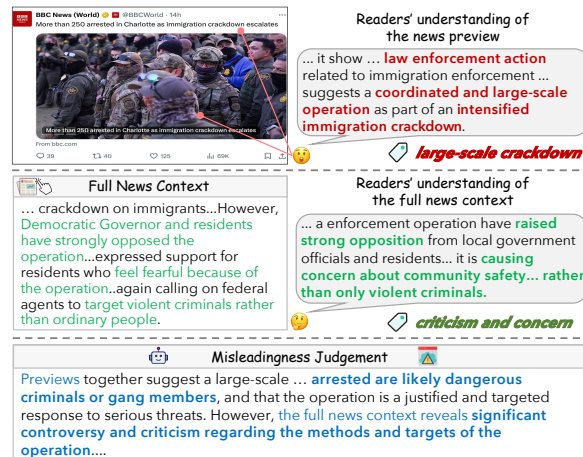


Figure 1: **Illustration of misleading omissions in multimodal news previews.** Social media users often encounter only a *news preview* (image–headline pair), while the *full context* remains unavailable unless they click through. When key information is omitted or selectively presented, the preview can induce interpretations that diverge from those supported by the full article.

is intensified on social media, where users often rely on **news previews** (a brief headline and image), and more than 75% of links are shared without a click (Sundar et al., 2025). When the interpretation formed from the preview significantly diverges from the narrative supported by the full article, it creates what we define as **misleading omissions**. As illustrated in Figure 1, a news preview may frame the event as a coordinated crackdown on illegal actors, even though the full article highlights public concerns about overly aggressive operations affecting ordinary immigrants.

Despite growing interest in misinformation governance, existing methods offer limited support for this challenge. Many approaches emphasize fact verification and treat omission primarily as an auxiliary cue (Tang et al., 2025; Wang et al., 2025e). Other lines of work study creator intent (Wu et al., 2026; Wang et al., 2025d,c), framing strategies (Arora et al., 2025; Moernaut et al.,

2020; Lüking and Pfeiffer, 2012; Ali and Hasan, 2022), or clickbait engagement (Wang et al., 2025b; Yu et al., 2024; Hagen et al., 2022). These methods do not examine whether preview-level omissions create a *systematic mismatch* relative to the full context, and they provide no mechanisms to proactively mitigate this effect through preview correction.

To bridge this gap, we develop MM-MISLEADING, a benchmark of 6,000 multimodal news previews derived from VisualNews (Liu et al., 2021). Constructed through a cognitive simulation pipeline, MM-MISLEADING captures the interpretive divergence between preview and full context, providing omission labels, supporting rationales, and high-quality corrections. Cross-model agreement filtering ensures annotation reliability, and human evaluation confirms substantial alignment with expert judgment. This establishes MM-MISLEADING as a rigorous testbed for both detection and mitigation. However, our evaluation of representative open-source LVLMs on MM-MISLEADING uncovers a fundamental weakness: the models often fail to compare the preview against the full context, leading to systematic failures to detect the subtle omission-induced shifts that define misleadingness.

Motivated by these findings, we propose OMGUARD, a framework for detecting and mitigating misleading omissions. At its core is *interpretation-aware fine-tuning*, which contrasts preview-based and context-based interpretations to strengthen small models’ ability to detect and explain omission-based misleadingness. Beyond detection, OMGUARD uses its generated rationales to support **active mitigation**: the model performs minimally invasive edits to the headline to reduce misleading interpretations relative to the full context. This allows OMGUARD to function as both (1) a platform-side screening tool for warning generation and (2) a publisher-side newsroom copilot for pre-publication self-checks.

Extensive experiments on MM-MISLEADING validate the effectiveness of our approach. On both omission detection and mitigation, OMGUARD based on an 8B model matches or surpasses the performance of a strong 235B LVLm. We identify a *perception–correction gap* in omission correction, where end-to-end correction success critically depends on upstream detection and the quality of generated rationales, and we demonstrate that explicit interpretation guidance produces significant

zero-shot transfer gains. Our analysis further shows that omission-based misleadingness typically manifests as local narrative shifts, such as missing background information, rather than global reframing. For image-driven cases where text-only correction is insufficient, we also show that OMGUARD supports visual prototyping, a necessary strategy for fully multimodal mitigation.

2 Related Work

Evidence-Based Fact Checking. Most existing work (Qi et al., 2024; Li et al., 2025c; Wu et al., 2025b; Wan et al., 2025a) focuses on fact-checking, leveraging retrieved external evidence to assess the factual correctness of image–text pairs. Recent studies (Tang et al., 2025; Wang et al., 2025e) have also examined omissions, but they mainly retrieve missing information to infer implied intent and use it as an auxiliary signal for fact verification. In contrast, we study a different yet underexplored form of deception: content that remains factually correct in the full context but becomes misleading due to critical omissions at the preview level.

News Preview Understanding Beyond Veracity. Recent studies have also examined how the *presentation* of news shapes audience understanding and dissemination outcomes. DeceptionDecoded (Wu et al., 2026) analyzes misleading presentations constructed by creators under different intents, while NINT and InSide (Wang et al., 2025d,c) leverage intent as an auxiliary signal for misinformation detection. Journalism research on framing (De Vreese, 2005; Arora et al., 2025; Moernaut et al., 2020) focuses on identifying and attributing frame categories, and clickbait studies (Wang et al., 2025b; Yu et al., 2024; Hagen et al., 2022) investigate how previews influence clicks and engagement. However, none of these lines of work explicitly characterizes the reader understanding gap induced between the preview and the full context.

A fuller discussion of related work is provided in Appendix A.1.

3 Problem Definition

3.1 Misleading Omission Detection

Let $P_{\text{news}} = (T, I)$ denote a multimodal news preview, where T is a one-sentence headline and I is the accompanying image. The preview links to a full news article C_{news} . The goal is to detect misleading omissions in P_{news} , where selective pre-

sensation creates an impression that deviates from reading C_{news} .

Motivated by prior work on news comprehension (Entman, 1993; Rich and Zaragoza, 2016; Petty and Cacioppo, 1986), we simulate the reader’s understanding under preview-only and full-context exposure by constructing two intermediate interpretations:

$$U_p = h_\theta(P_{\text{news}}), \quad U_c = h_\theta(C_{\text{news}}). \quad (1)$$

These serve as reasoning states that capture preview-based and context-based impressions.

The model then predicts a binary label $y \in \{0, 1\}$ and a rationale r :

$$y, r = h_\theta(P_{\text{news}}, C_{\text{news}}, U_p, U_c). \quad (2)$$

A preview is labeled as misleading ($y = 1$) if discrepancies between U_p and U_c result in a substantial shift in the perceived nature or implications of the event; otherwise it is labeled as non-misleading ($y = 0$). The rationale r explains the source and effect of the omission.

3.2 Misleading Omission Correction

Given a misleading instance and its rationale $((T, I), C_{\text{news}}, r)$, the goal is to correct the preview by eliminating misleading omissions. Since news images serve as documentary evidence and editing them may introduce fabrication or reduce authenticity, we fix the image I as an anchor and focus on *headline revision*: generating a revised headline \hat{T} so that the preview (\hat{T}, I) accurately reflects the full context C_{news} .

To ensure practical usability, we impose a strict length budget, allowing at most three additional words. Under this constraint, we introduce two correction protocols (see Figure 14) that emphasize *style preservation* versus *factual restoration*: **(1) Minimal-Edit Revision**: prioritizes *editorial continuity* by preserving the original style, tone, and syntactic structure as much as possible, simulating a newsroom-copilot setting where the preview must remain close to its published form; **(2) Free-Form Revision**: prioritizes informational fidelity by relaxing stylistic constraints and focusing solely on eliminating misleading impressions, approximating an upper bound of correction when optimized for truthfulness. The correction process is formalized as:

$$\hat{T} = h_\theta((T, I), C_{\text{news}}, r, \mathcal{C}), \quad (3)$$

where \mathcal{C} denotes the chosen correction protocol.

4 The MM-MISLEADING Benchmark

4.1 Data Collection

To rigorously characterize omission-based misleadingness, a benchmark must meet two criteria: **(1)** factual reliability, where the full context provides an authoritative ground truth; and **(2)** narrative complexity, where the story is rich enough for omissions to meaningfully alter interpretation.

To meet these criteria, we build MM-MISLEADING on top of VisualNews (Liu et al., 2021), a trustworthy repository of multimodal news reporting, and apply a strict two-stage filtering process to ensure both relevance and difficulty. First, to ensure **societal relevance**, we stratify sampling across ten high-impact topics (e.g., politics and public health; see Appendix A.2.1), covering domains most vulnerable to misinformation. Second, to enforce **informational density**, we retain instances whose previews can only be properly interpreted with answers to high-level questions such as “*Why did this happen?*” and “*How will it unfold?*” (see Appendix A.2.2), and discard simple factual announcements (e.g., dates or locations).

This filtration yields a foundational corpus of high-quality news pairs $(P_{\text{news}}, C_{\text{news}})$, which serve as the input for the multi-stage annotation pipeline detailed in the following subsection.

4.2 Multi-Stage Data Annotation

Motivated by cognitive theories of news comprehension, where readers first form an initial situation model from surface cues before updating it with full text (Entman, 1993), we leverage the social simulation capabilities of LLMs (Liu et al., 2025) to design a multi-stage annotation pipeline (Figure 2). This pipeline explicitly models the interpretative gap between the preview and the full context.

Stage 1: Preview-based Understanding Simulation (U_p). Simulating a user scrolling through a feed, the LLM is presented strictly with the news preview P_{news} (headline + image). It generates an understanding U_p comprising **(1)** a surface description of the visible content and **(2)** a deeper inference of the event’s implications, conditioned solely on these limited cues (see Figure 11).

Stage 2: Context-based Understanding Simulation (U_c). To establish the news ground truth, the LLM processes the full article text C_{news} to generate a comprehensive understanding U_c . This represents the informed interpretation of a reader who has fully consumed the story (see Figure 12).

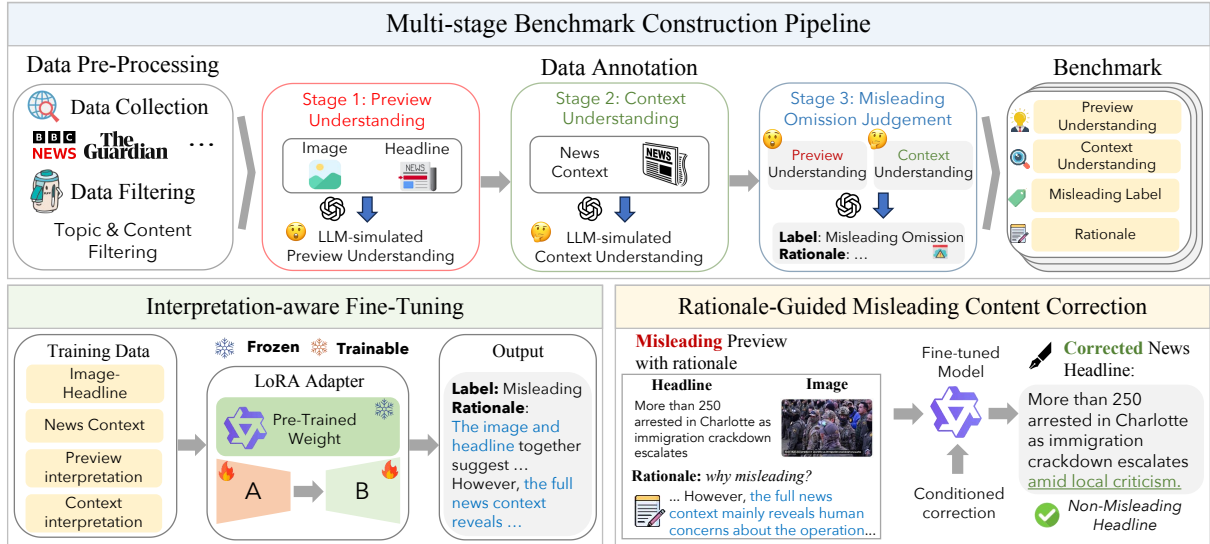


Figure 2: **Overview of OMGUARD.** The upper section illustrates the multi-stage annotation pipeline described in § 4.2. The lower section shows OMGUARD, where the model is first fine-tuned with interpretation-aware supervision using misleadingness rationales and then used for rationale-guided correction of misleading previews.

Stage 3: Misleading Omission Judgment. We detect misleadingness by comparing the semantic divergence between the preview interpretation (U_p) and the context interpretation (U_c). Specifically, the model evaluates whether the preview omits critical information that alters the perceived event nature, causal relationships, or emotional valence (details in Appendix A.2.3).

Quality Control via Cross-Model Agreement. To mitigate single-model bias, we use Gemini-2.5-pro and GPT-4.1 as independent annotators and apply strict cross-model agreement filtering, retaining only instances where both models assign the same misleadingness label (see Appendix A.2.4 for details). This prioritizes label precision over recall, ensuring that our final dataset consists of unambiguous cases. The resulting MM-MISLEADING benchmark contains 6,000 instances (5,000 train, 1,000 test), balanced 1:1 between misleading and non-misleading previews.

4.3 Corrective Headline Generation

To establish a gold-standard reference for active mitigation, we generate corrected headlines \hat{T} corresponding to the two protocols defined in Section 3.2: **Minimal-Edit** and **Free-Form**.

To ensure benchmark solvability, we apply a strict **validity filter**: because not all instances can be corrected successfully, we retain only those corrections that are verified as the ground truth (via the pipeline in Section 4.2) to successfully eliminate misleadingness under *both* protocols (see details in

Appendix A.2.5).

Instances that cannot be resolved through headline revision alone, typically because the misleadingness stems from the visual component, are excluded from the primary benchmark and analyzed separately in Section 6.5.

4.4 Data Quality Assessment

To validate the quality of our automated pipeline, we conduct human evaluation with three independent annotators, assessing detection fidelity on the full test set and correction validity on 100 randomly sampled instances. We evaluate alignment along three dimensions: (1) **Interpretation Consistency**, which measures whether the simulated (U_p, U_c) align with human perception; (2) **Detection Accuracy**, which evaluates the correctness of the predicted misleadingness labels; and (3) **Correction Effectiveness**, which assesses whether the revised headlines successfully remove the misleading impression.

The results show strong agreement between our pipeline and expert annotators across all three dimensions, with Fleiss’ $\kappa \geq 0.81$ and accuracy above 94% (see Appendix A.2.6). These findings indicate that our automated annotations provide reliable ground truth for both training and evaluation.

5 OMGUARD: Mitigating Misleading Omissions in News Previews

To govern omission-based misleadingness, we propose OMGUARD, a framework that integrates

both detection and mitigation. As shown in Figure 2, OMGUARD operates in two stages: **(1) Interpretation-Aware Fine-tuning**, which conditions the model on explicit narrative interpretations to strengthen its detection capability; and **(2) Rationale-Guided Mitigation**, which uses the model’s diagnostic rationales to guide targeted headline correction.

5.1 Interpretation-Aware Fine-Tuning

We formulate omission-based misleadingness detection in multimodal news previews as an instruction-tuning task. Each training instance is represented as a tuple $\{I, T, C_{\text{news}}, U_p, U_c, y, r\}$. The intermediate interpretations (U_p, U_c) are provided as *auxiliary reasoning context*, while loss supervision is applied only to the final judgment y and its rationale r .

We opt *not* to compute loss on U_p and U_c for two methodological reasons. First, narrative interpretation is inherently non-unique. Enforcing exact token-level alignment risks promoting stylistic imitation rather than helping the model learn a robust decision boundary for misleadingness. Second, our objective is *logic extraction*. The model should learn how to leverage interpretive discrepancies to justify its conclusion, rather than learn to reproduce a specific set of reference interpretations.

Formally, we maximize the likelihood of the judgment and rationale conditioned on the full interpretive context:

$$\mathcal{L}_{\text{FT}} = - \sum_{t=1}^{|s|} \log P_{\theta}(s_t | x, s_{<t}),$$

$$\text{where } s = [r; y], x = \{I, T, C_{\text{news}}, U_p, U_c\}. \quad (4)$$

Inference Pipeline. During inference, ground-truth interpretations are not available. We therefore adopt a **generate-then-reason** process. The model first generates its own preview-based and context-based interpretations (U_p and U_c). These generated interpretations are then incorporated into the input to trigger the fine-tuned detection head, which produces the final label y and rationale r .

5.2 Rationale-Guided Mitigation

A key advantage of OMGUARD is that its detection rationales (r) enable **active mitigation**. Unlike standard black-box classifiers, OMGUARD provides a structured diagnosis that identifies the specific omission and clarifies why the preview is misleading relative to the full context.

We use this diagnosis to support zero-shot headline correction. When a preview is identified as misleading, the model receives the generated rationale r together with the correction constraints \mathcal{C} (for example, minimal-edit requirements). Conditioning the rewrite on r directs the model to make targeted adjustments that resolve the omission while preserving the stylistic quality of the original headline, as described in Equation 3.

6 Experiments

We organize our experiments around the following research questions:

RQ1:(§6.2) To what extent can LVLMs reliably detect omission-induced misleadingness in multimodal news previews?

RQ2:(§6.3) What factors affect the effectiveness of misleadingness correction in news previews?

RQ3:(§6.4) What mechanisms give rise to omission-based misleadingness in news previews?

RQ4:(§6.5) What visual cues contribute to omission-induced misleadingness?

6.1 Implementation Details

Models. We evaluate 8 representative open-source LVLMs on MM-MISLEADING, covering diverse parameter scales and major model families: **(1) Non-Reasoning models:** Qwen-VL (Qwen Team, 2025) (Qwen3-VL-8B, Qwen3-VL-235B), Llama3-VL (Grattafiori et al., 2024) (Llama3-VL-11B, Llama3-VL-90B), InternVL3.5-8B (Wang et al., 2025a), and LLaVA-7B (Liu et al., 2023); **(2) Reasoning models:** Qwen-VL (Qwen Team, 2025) (Qwen3-VL-8B-Thinking, Qwen3-VL-235B-Thinking) and GLM-4.1V-9B-Thinking (Hong et al., 2025). We provide detailed model cards in Table 6.

Fine-tuning details. We fine-tuned Qwen3-VL-8B using LoRA (Hu et al., 2022). Specifically, we set the LoRA rank to $r = 64$, the scaling factor to $\alpha = 16$, and the dropout rate to 0.05. The model was trained for 1 epoch using the AdamW optimizer with a learning rate of 2×10^{-5} , and an effective batch size of 4 was achieved via gradient accumulation.

6.2 Omission Detection Performance

Standard LVLMs under-detect misleading omissions. As shown in Table 1, general open-source models exhibit large disparities in class-wise recall. Although larger models, such as the 90B and 235B

Models	Accuracy	Non-Misleading			Misleading		
		Precision	Recall	F1	Precision	Recall	F1
Llava-7B	0.50	0.50	1.00	0.67	0.00	0.00	0.00
Qwen3-VL-8B	0.68	0.62	0.95	0.75	0.89	0.42	0.57
InternVL3.5-8B	0.64	0.59	0.95	0.73	0.86	0.34	0.49
LLama3-VL-11B	0.51	0.51	0.58	0.54	0.51	0.43	0.47
Qwen3-VL-8B-Thinking	0.70	0.70	0.68	0.69	0.69	0.71	0.70
GLM-4.1V-9B-Thinking	0.70	0.74	0.60	0.66	0.67	0.79	0.72
Qwen3-VL-235B-Thinking	0.79	0.77	0.81	0.79	0.80	0.76	0.78
LLama3-VL-90B	0.72	0.67	0.85	0.75	0.80	0.58	0.67
Qwen3-VL-235B	0.86	0.80	0.96	0.88	0.96	0.76	0.85
FT w U_p/U_c (End-to-End Inference)	0.87	0.83	0.94	0.88	0.93	0.80	0.86
FT w U_p/U_c (Multi-stage Inference)	0.78	0.70	0.98	0.81	0.97	0.57	0.72
FT Label-Only	0.84	0.79	0.93	0.85	0.91	0.75	0.82
OMGUARD (Interpretation-Aware)	0.86	0.87	0.84	0.85	0.84	0.88	0.86

Table 1: Misleading omission detection effectiveness of LVLMs and OMGUARD on MM-MISLEADING.

variants, achieve higher overall accuracy, they remain biased toward the non-misleading class and often fail to identify subtle omission-based deception. This reveals a persistent blind spot in recognizing the *interpretation drift* between the preview and the full context.

Reasoning models raise recall but increase false positives. Reasoning-enhanced models improve sensitivity to misleading cues through multi-step processing, but this gain comes at the cost of precision. As shown in Table 1, these models exhibit *overthinking* (Sui et al., 2025): they often amplify minor benign inconsistencies into misleading signals, increasing recall on positive cases while sharply reducing precision. This suggests that generic reasoning mechanisms still require task-specific calibration to avoid excessive false positives.

Interpretation-aware fine-tuning improves class balance and robustness. We compare OMGUARD with a *Label-Only* ablated variant that applies supervision only to the final label y . Although both methods achieve accuracy comparable to a 235B model, the *Label-Only* variant yields substantially lower recall on misleading cases. This suggests that label-only supervision encourages reliance on superficial patterns. In contrast, OMGUARD uses the interpretive gap as a reasoning anchor, enabling stronger causal discrimination that generalizes better to complex omission scenarios. By explicitly contrasting preview-based and context-based understanding, the model is encouraged to attend to the missing contextual elements that actually drive misleadingness, instead of relying on shallow lexical or visual cues. This leads to a more balanced decision boundary across classes

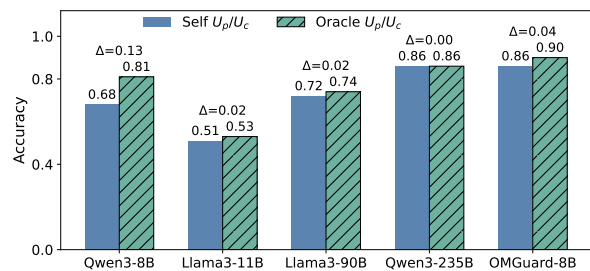


Figure 3: Quantifying error propagation via oracle substitution of U_p/U_c ; $\Delta = \text{Acc}(\text{oracle } U) - \text{Acc}(\text{self } U)$.

and more robust performance on subtle omission cases. Qualitative examples are shown in Figure 8.

OMGUARD reduces error propagation from imperfect interpretations. To assess sensitivity to interpretation quality, we replace the model’s self-generated U_p and U_c with oracle versions produced by GPT-4.1 during annotation and measure the resulting performance gap Δ . As shown in Figure 3, smaller models exhibit larger Δ , indicating stronger error propagation from imperfect intermediate interpretations. By contrast, larger models have Δ close to zero, suggesting that their main bottleneck lies in recognizing misleadingness itself rather than forming intermediate interpretations. Relative to its backbone, OMGUARD reduces Δ from 0.13 to 0.04, showing that supervising only the final decision y and rationale r , without directly supervising U_p and U_c , can effectively mitigate error propagation. This supports our design choice to treat U_p and U_c as reasoning context rather than training targets.

Direct supervision on U_p/U_c introduces conservative bias. As discussed in Section 5.1, OMGUARD does not treat U_p/U_c as direct supervision targets during training. To verify this design choice,

Rewrite Model	Free-Form				Minimal-Edit			
	G1	G2	G3	G4	G1	G2	G3	G4
Qwen3-VL-8B	0.88	0.84	0.84	0.27	0.84	0.78	0.78	0.25
InternVL-3.5-8B	0.82	0.72	0.82	0.20	0.82	0.75	0.79	0.21
GLM-4.1V-9B-Thinking	0.82	0.76	0.84	0.59	0.73	0.66	0.73	0.51
Qwen3-VL-235B	0.90	0.84	0.89	0.56	0.77	0.68	0.75	0.46
OMGuard	0.95	0.90	0.95	0.75	0.94	0.90	0.95	0.75

Table 2: Correction success rate (CSR) comparison for misleading omission correction. **G1**: oracle rationale guided headline correction; **G2**: self-generated rationale guided headline correction, applied only to samples the model predicts as misleading; **G3**: oracle rationale guided headline correction on the same sample subset as G2; **G4**: full end-to-end misleading omission correction pipeline. Detailed setups are provided in Appendix A.3.1.

Rewrite Model	Free-Form			Minimal-Edit		
	BLEU-4	ROUGE-L	Cosine	BLEU-4	ROUGE-L	Cosine
Oracle (GPT-5)	7.36 (-)	23.89 (-)	0.54 (-)	45.51 (-)	61.54 (-)	0.77 (-)
Qwen3-VL-8B	5.43 (4.88)	22.39 (23.20)	0.53 (0.52)	9.57 (10.14)	29.57 (29.89)	0.59 (0.54)
InternVL3.5-8B	4.24 (2.65)	20.98 (18.36)	0.50 (0.43)	4.60 (4.16)	21.30 (20.00)	0.51 (0.44)
GLM-Thinking	8.76 (5.97)	26.54 (25.71)	0.55 (0.55)	28.24 (24.96)	47.40 (44.66)	0.69 (0.65)
Qwen3-VL-235B	3.27 (4.42)	24.03 (24.60)	0.55 (0.56)	27.36 (20.17)	52.16 (45.30)	0.75 (0.71)
OMGuard	7.97 (7.50)	23.73 (22.92)	0.53 (0.53)	13.15 (11.17)	28.70 (25.89)	0.59 (0.55)

Table 3: Similarity comparison between corrected headlines \hat{T} and the original misleading headlines T under the *Free-form* and *Minimal-edit* settings. **Oracle (GPT-5)** denotes the annotated reference correction. Results are reported for two rationale sources: **oracle rationale** and **self-generated rationale**.

we directly supervise U_p/U_c and evaluate both end-to-end and multi-stage inference. As shown in Table 1, the end-to-end setting biases the model toward the non-misleading class, increasing precision to 0.93 but reducing misleading recall to 0.80. This suggests that the model adopts an overly strict threshold for predicting misleadingness and becomes less sensitive to subtle omission-based cases unless its generated reasoning is sufficiently explicit. Under multi-stage inference, direct supervision on intermediate steps further entangles content understanding with misleadingness discrimination, producing a more severe train–test mismatch. As a result, overall accuracy drops to 0.78, and misleading recall declines sharply to 0.57.

Separating explanation from judgment improves deployment suitability. Although end-to-end inference is simpler and performs slightly better on some precision-oriented metrics, separating explanation from final judgment remains the better design for deployment. A multi-stage framework provides greater controllability, intervention, and auditability. End-to-end inference makes it difficult to inspect or correct flawed intermediate understanding before the final decision is made, whereas a multi-stage pipeline explicitly exposes intermediate interpretations and supports targeted intervention, such as adding corrective context or human feedback. This decoupled design also better

supports integration of auxiliary signals, such as user comments, verified background information, and community notes, which is especially valuable for real-world content moderation and governance.

6.3 Omission Mitigation Performance

We evaluate the active mitigation capability of OM-GUARD using **Correction Success Rate (CSR)** across diverse editing modes and setups, and measure headline quality against ground-truth corrections with BLEU, ROUGE, and cosine similarity; see Appendix A.3.1 for details.

Correction of misleading previews reveals a substantial perception–correction gap. As shown in Table 2, a substantial gap emerges when comparing G1 and G4. With oracle rationales (G1), even open-source LVLMs achieve high rewriting success rates (average 0.86 for Free-form and 0.72 for Minimal-edit), demonstrating strong inherent correction capability. In contrast, the full end-to-end pipeline (G4), which requires detection, rationale generation, and correction, reaches much lower success rates (average 0.41 for Free-form and 0.36 for Minimal-edit). This indicates that the primary bottleneck lies in the upstream detection stage. Although fine-tuning improves detection and reduces the gap between G1 and G4, the gap does not fully close, suggesting that rationale quality and other downstream factors also influence final

correction performance.

High-quality rationales are essential for effective correction. Comparing G2 and G3 in Table 2 shows that, among correctly detected misleading instances, higher-quality judgment rationales consistently yield higher CSR (avg. Free-Form: 0.79 to 0.85; Minimal-Edit: 0.72 to 0.76). Consistently, Table 3 indicates that oracle rationales more often produce rewrites that are semantically faithful to the source and better aligned with the ground truth than those using self-generated rationales. We further remove rationales and provide only the *binary misleading label*; CSR drops to 17%, confirming that labels alone do not provide sufficient guidance. Overall, effective correction critically depends on accurately localizing the misleading content and explaining why it is misleading.

Editing constraints expose a trade-off between control and success rate. As shown in Table 3, the stronger constraints in the Minimal-Edit setting make instruction following substantially more difficult. Reasoning-enhanced and large-parameter models remain closer to the oracle references, but their limited room for modification often lowers CSR. Other models achieve higher apparent success only by violating the editing constraints, as reflected in their lower BLEU and ROUGE scores under Minimal-Edit. Under Free-Form editing, relaxing these constraints reduces CSR differences across models. Even so, models with stronger instruction-following ability remain closer to the oracle references, producing more controlled rewrites with smaller deviations. Overall, OMGUARD achieves a better balance between adherence to editing constraints and correction effectiveness, resulting in more stable performance across both settings.

OMGUARD transfers from diagnosis to correction without rewrite supervision. Although OMGUARD is fine-tuned only for detection and rationale generation, and receives no direct supervision on rewriting, it still delivers substantial gains in end-to-end correction. This supports our central hypothesis: improving the model’s ability to *diagnose* subtle interpretive discrepancies through Interpretation-Aware Fine-Tuning naturally strengthens its ability to *correct* them. As a result, OMGUARD enables effective zero-shot mitigation without explicit training on headline rewriting.

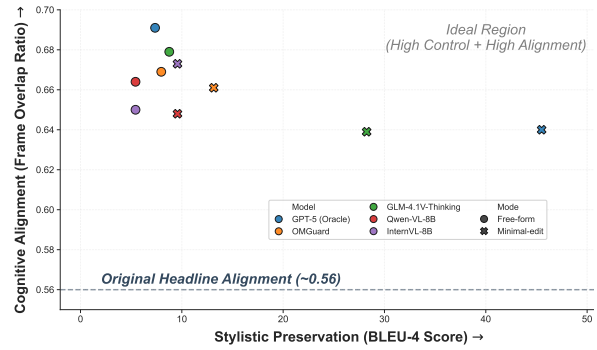


Figure 4: Frame-shift analysis comparing stylistic preservation and frame alignment in misleading headline correction.

6.4 Attributions of Misleading Omissions

We analyze how previews induce misleadingness from two complementary perspectives: **frame shift** and **fine-grained attribution**. First, following (Arora et al., 2025), we extract the top-3 frames from both the preview and the full context and compute their overlap ratio (see Appendix A.4.1). Second, we identify major categories of misleadingness through expert annotation: two human experts first summarize five recurring causes, and an LLM judge then assigns an attribution label to each sample (see Appendix A.4.2).

Misleading omissions usually arise from local narrative shifts. As shown in Figure 4, the frame overlap ratio between misleading headlines and full contexts remains relatively high (≈ 0.56), suggesting that omission-based misleadingness rarely changes the topic itself, but instead distorts its perspective or implication. Although correction improves frame alignment, it does not fully close the gap, indicating that *truthfulness* and *framing* are related but distinct dimensions. As shown in Figure 4, *Minimal-Edit* better preserves stylistic continuity but shifts the frame less effectively, whereas *Free-Form* achieves stronger alignment at the cost of the original stylistic voice.

Missing background is the dominant source of misleadingness. We further examine fine-grained causes beyond frame shift. For misleading samples in the test set, we first ask human experts to summarize five major causes, and then use an LLM-as-a-judge approach (Li et al., 2024; Pu et al., 2025) to assign an attribution label to each sample (see Appendix A.4.2).

The resulting distribution follows a long-tail pattern. As shown in Figure 5, *missing background and conditions* accounts for 67% of all

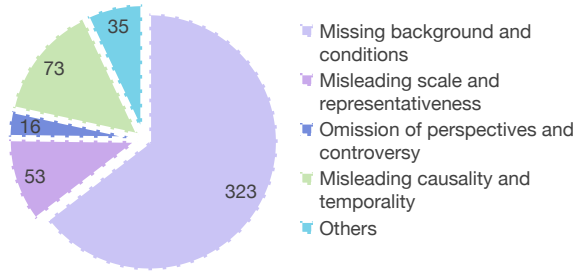


Figure 5: Distribution of misleadingness types.

Data type	Precision	Recall	F1	Accuracy
Non-Misleading	0.84	0.91	0.87	0.87
Misleading	0.90	0.82	0.86	

Table 4: Ablation study using only the headline (Text-Only) for misleadingness detection, showing the contribution of visual signals.

misleading cases, far exceeding the other categories. This distribution also helps explain why effective correction often depends on *recovering omitted context* instead of extensive rewriting. For the dominant omission-driven cases, minimally adding the missing background or conditions is often sufficient to prevent incorrect inferences. By contrast, long-tail categories such as **misleading causality/temporality** and **misleading scale/representativeness** often require the rewrite to make implicit relations or scope conditions explicit, and, when possible, to ground the revision more clearly in supporting evidence. Overall, these findings reinforce our central claim that omission-based misleadingness is driven primarily by *insufficient information and selective presentation*: seemingly factual previews can still induce biased interpretations by withholding critical conditions, even without introducing explicit falsehoods.

6.5 Discussion: Multimodal Mitigation

Visual cues are essential for robust misleadingness detection. To assess the need for multimodal modeling, we ablate the image channel and evaluate a headline-only variant while keeping the rest of the pipeline unchanged. As shown in Table 4, removing visual information degrades overall performance and introduces a clear asymmetry across classes: the model retains relatively high recall on non-misleading instances but struggles more on misleading ones. This suggests that many failures arise from *text-image dissonance*, and that reliable detection requires multimodal evidence.

Visual semantics limit the effectiveness of text-only correction. As noted in Section 4.3, not all misleading instances can be resolved during ground-truth construction. To better characterize this bottleneck, we partition misleading samples into *Text-Fixable* and *Image-Driven* subsets via modality attribution (see Appendix A.5.1). Even before attribution, the Free-Form rewriting success rate reaches only 0.78 (Appendix A.2.5); after attribution, performance rises to 0.88 on the text-fixable subset but falls to 0.46 on image-driven cases. This sharp gap highlights the limiting role of visual semantics in omission correction.

Image intervention can mitigate otherwise uncorrectable cases. To test whether *image-driven misleading instances* can be resolved through visual intervention, we conduct a pilot study on 20 cases that text modification alone cannot correct. We use a generative model to synthesize alternative images better aligned with the full context (Figure 9), increasing the correction success rate to 90%. Although we do not advocate replacing authentic news photos with generated images in practice, this result highlights *image selection* as a critical variable in governing misleading previews. More practically, it suggests that future newsroom tools should help editors select more representative archival images, instead of relying solely on headline edits.

7 Conclusion

We present a formal study of **misleading omissions** in multimodal news previews and introduce MM-MISLEADING, a 6,000-instance benchmark for evaluating both detection and correction. We further propose OMGUARD, which uses interpretation-aware fine-tuning to lift an 8B model to the detection accuracy of 235B-scale LLMs while enabling zero-shot headline correction through diagnostic rationales. Our analyses show that most misleadingness arises from local narrative shifts that can often be resolved through targeted text editing, while a smaller set of image-driven cases exposes the limits of text-only correction and motivates multimodal intervention. These findings position omission-based misleadingness as a distinct governance challenge beyond fact verification alone. We hope MM-MISLEADING and OMGUARD provide a foundation for future work on multimodal moderation, newsroom support, and more context-faithful communication online.

Limitations

Our work also has the following limitations: (1) We primarily mitigate preview-level misleadingness by rewriting news headlines. While this text-based correction can effectively recover critical missing information and reduce erroneous inferences, rewriting itself may introduce new biases, such as shifts in emphasis or unintended formulations driven by the model’s inherent biases. Ensuring mitigation without introducing new distortions is an important and promising direction, but it is beyond the scope of this paper. (2) We conduct image-side visual prototyping. However, our current exploration is largely limited to generating semantically aligned images to validate feasibility, and we do not investigate more practical solutions, such as retrieving candidate images from newsroom archives or related media repositories. Retrieval-based real-image replacement and joint text–image adjustment frameworks remain promising avenues for future work.

Ethical Considerations

Our work aims to model and identify reader-understanding deviations induced by a news preview relative to the full news context, and to mitigate misleading impressions caused by information omission and selective presentation through headline rewriting. Overall, this direction has the potential to reduce erroneous inferences in low-information preview settings and to lessen the societal harms associated with the spread of misleading content, thereby supporting content governance and healthier information consumption.

We also recognize potential misuse risks. First, the detection and correction capabilities themselves could be exploited adversarially. Attackers may treat misleadingness detection as an optimization signal to iteratively craft omission-based misleading content that is harder to detect, or to perform evasion-oriented optimization against our evaluation setup. Second, our exploration of visual prototyping is intended to highlight the importance of image selection in governance, but similar techniques could be misused to generate or curate more suggestive images that amplify misleading impressions. Our goal is to promote responsible governance practices, not to provide tools for generating deceptive content.

Acknowledgments

This work is supported by the Yunnan Research Project (Grant Nos. 202503AG380006, 202401AT070474, 202501AU070059, and 202403AP140021), the National Natural Science Foundation of China (Grant Nos. 62562061, 62502422, and 62462067), the Yunnan Provincial Department of Education Science Research Project (Grant Nos. 2025J0006, 2024J0010, and 2025J0007) and China Scholarship Council (CSC) program. This research is also supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 1 (T1 251RES2508) and MOE AcRF TIER 3 Grant (MOE-MOET32022-0001).

References

- Mohammad Ali and Naeemul Hassan. 2022. A survey of computational framing analysis approaches. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 9335–9348.
- Jennifer Allen, Duncan J Watts, and David G Rand. 2024. Quantifying the impact of misinformation and vaccine-skeptical content on facebook. *Science*, 384(6699):eadk3451.
- Arnav Arora, Srishti Yadav, Maria Antoniak, Serge Belongie, and Isabelle Augenstein. 2025. Multimodal framing analysis of news. *arXiv preprint arXiv:2503.20960*.
- Alimohammad Beigi, Bohan Jiang, Dawei Li, Zhen Tan, Pouya Shaeri, Tharindu Kumarage, Amrita Bhat-tacharjee, and Huan Liu. 2024. Can llms improve multimodal fact-checking by asking relevant questions? *arXiv preprint arXiv:2410.04616*.
- Claes H De Vreese. 2005. News framing: Theory and typology. *Information design journal+ document design*, 13(1):51–62.
- Robert M Entman. 1993. Framing: Towards clarification of a fractured paradigm. *McQuail’s reader in mass communication theory*, 390:397.
- Google. 2025. Gemini 2.5 pro. <https://deepmind.google/technologies/gemini/pro/>.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Matthias Hagen, Maik Fröbe, Artur Jurk, and Martin Potthast. 2022. Clickbait spoiling via question answering and passage retrieval. *arXiv preprint arXiv:2203.10282*.

- Jing He, Han Zhang, Yuanhui Xiao, Wei Guo, Shaowen Yao, and Renyang Liu. 2025. Factguard: Event-centric and commonsense-guided fake news detection. *arXiv preprint arXiv:2511.10281*.
- Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, and 1 others. 2025. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Dayang Li, Fanxiao Li, BingBing Song, Li Tang, and Wei Zhou. 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)*, pages 9127–9142.
- Fanxiao Li, Jiaying Wu, Tingchao Fu, Yunyun Dong, Bingbing Song, and Wei Zhou. 2025b. Drifting away from truth: Genai-driven news diversity challenges lvlm-based misinformation detection. *arXiv preprint arXiv:2508.12711*.
- Fanxiao Li, Jiaying Wu, Canyuan He, and Wei Zhou. 2025c. CMIE: Combining MLLM insights with external evidence for explainable out-of-context misinformation detection. In *Findings of the Association for Computational Linguistics: ACL 2025*.
- Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.
- Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2021. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 6761–6771.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Yuhan Liu, Zirui Song, Juntian Zhang, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2025. The stepwise deception: Simulating the evolution from true news to fake news with llm agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26187–26203.
- Andy Lüking and Thies Pfeiffer. 2012. Framing multimodal technical communication. *Handbook of technical communication*, pages 591–644.
- Xueqi Ma, Shuo Yang, Yanbei Jiang, Shu Liu, Zhenzhen Liu, Jiayang Ao, Xingjun Ma, Sarah Monazam Erfani, and James Bailey. 2026. Attention in space: Functional roles of vlm heads for spatial reasoning. *arXiv preprint arXiv:2603.20662*.
- Zihan Ma, Minnan Luo, Zhi Zeng, Herun Wan, Yifei Li, and Xiang Zhao. 2025. Graphing the truth: Harnessing causal insights for advanced multimodal fake news detection. *IEEE Transactions on Information Forensics and Security*, 20:12934–12949.
- Renée Moernaut, Jelle Mast, and Luc Pauwels. 2020. Visual and multimodal framing analysis. *The SAGE handbook of visual research methods*, pages 484–499.
- OpenAI. 2025a. Introducing gpt-4.1 in the api. <https://openai.com/index/gpt-4-1/>.
- OpenAI. 2025b. Introducing gpt-5. <https://openai.com/index/introducing-gpt-5/>.
- Richard E Petty and John T Cacioppo. 1986. The elaboration likelihood model of persuasion. In *Advances in experimental social psychology*, volume 19, pages 123–205. Elsevier.
- Shu Pu, Yaochen Wang, Dongping Chen, Yuhang Chen, Guohao Wang, Qi Qin, Zhongyi Zhang, Zhiyuan Zhang, Zetong Zhou, Shuang Gong, and 1 others. 2025. Judge anything: Mllm as a judge across any modality. *arXiv preprint arXiv:2503.17489*.
- Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. 2024. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13052–13062.
- Qwen Team. 2025. *Qwen3 technical report*. Preprint, arXiv:2505.09388.
- Patrick R Rich and Maria S Zaragoza. 2016. The continued influence of implied and explicitly stated misinformation in news reports. *Journal of experimental psychology: learning, memory, and cognition*, 42(1):62.
- Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, and 1 others. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*.
- S Shyam Sundar, Eugene Cho Snyder, Mengqi Liao, Junjun Yin, Jinping Wang, and Guangqing Chi. 2025. Sharing without clicking on news in social media. *Nature Human Behaviour*, 9(1):156–168.
- Yixuan Tang, Jincheng Wang, and Anthony Kum Hoe Tung. 2025. The missing parts: Augmenting fact verification with half truth detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33967–33984.

- Lyn M Van Swol and Michael T Braun. 2014. Communicating deception: Differences in language use, justifications, and questions for lies, omissions, and truths. *Group Decision and Negotiation*, 23(6):1343–1367.
- Herun Wan, Jiaying Wu, Minnan Luo, Xiangzheng Kong, Zihan Ma, and Zhi Zeng. 2025a. Difar: Enhancing multimodal misinformation detection with diverse, factual, and relevant rationales. *arXiv preprint arXiv:2508.10444*.
- Herun Wan, Jiaying Wu, Minnan Luo, Zhi Zeng, and Zhixiong Su. 2025b. Truth over tricks: Measuring and mitigating shortcut learning in misinformation detection. *arXiv preprint arXiv:2506.02350*.
- Eileen Wang, Hiba Arnaout, Dhita Pratama, Shuo Yang, Danyang Liu, Jie Yang, Josiah Poon, Jeff Pan, and Caren Han. 2026. Mmcomet: A large-scale multimodal commonsense knowledge graph for contextual reasoning. *arXiv preprint arXiv:2603.01055*.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025a. Internv13.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.
- Ye Wang, Yi Zhu, Yun Li, Liting Wei, Yunhao Yuan, and Jipeng Qiang. 2025b. Multi-modal soft prompt-tuning for chinese clickbait detection. *Neurocomputing*, 614:128829.
- Zhengjia Wang, Qiang Sheng, Danding Wang, Beizhe Hu, and Juan Cao. 2025c. Bridging thoughts and words: Graph-based intent-semantic joint learning for fake news detection. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 3250–3260.
- Zhengjia Wang, Danding Wang, Qiang Sheng, Juan Cao, Siyuan Ma, and Haonan Cheng. 2025d. Exploring news intent and its application: A theory-driven approach. *Information Processing & Management*, 62(6):104229.
- Zhengjia Wang, Danding Wang, Qiang Sheng, Jiaying Wu, and Juan Cao. 2025e. Reasoning about the unsaid: Misinformation detection with omission-aware graph inference. *arXiv preprint arXiv:2512.01728*.
- Jiaying Wu, Zihang Fu, Haonan Wang, Fanxiao Li, and Min-Yen Kan. 2025a. Beyond the crowd: Llm-augmented community notes for governing health misinformation. *arXiv preprint arXiv:2510.11423*.
- Jiaying Wu, Fanxiao Li, Zihang Fu, Min-Yen Kan, and Bryan Hooi. 2026. Seeing through deception: Uncovering misleading creator intent in multimodal news with vision-language models. In *The Fourteenth International Conference on Learning Representations*.
- Junjie Wu, Yumeng Fu, Nan Yu, and Guohong Fu. 2025b. E2lvm: Evidence-enhanced large vision-language model for multimodal out-of-context misinformation detection. *arXiv preprint arXiv:2502.10455*.
- Keyang Xuan, Li Yi, Fan Yang, Ruochen Wu, Yi R Fung, and Heng Ji. 2024. Lemma: towards lvm-enhanced multimodal misinformation detection with external knowledge augmentation. *arXiv preprint arXiv:2402.11943*.
- Jianxing Yu, Shiqi Wang, Han Yin, Zhenlong Sun, Ruobing Xie, Bo Zhang, and Yanghui Rao. 2024. Multimodal clickbait detection by de-confounding biases using causal representation inference. *arXiv preprint arXiv:2410.07673*.
- Zhi Zeng, Yifei Yang, Jiaying Wu, Xulang Zhang, Xiangzheng Kong, Herun Wan, Zihan Ma, and Minnan Luo. 2026. From manipulation to mistrust: Explaining diverse micro-video misinformation for robust debunking in the wild. *arXiv preprint arXiv:2603.25423*.
- Qianchi Zhang, Hainan Zhang, Liang Pang, Yongxin Tong, Hongwei Zheng, and Zhiming Zheng. 2026a. Less is more: Compact clue selection for efficient retrieval-augmented generation reasoning. In *Proceedings of the ACM Web Conference 2026*, pages 1971–1982.
- Qianchi Zhang, Hainan Zhang, Liang Pang, Hongwei Zheng, and Zhiming Zheng. 2026b. Stable-rag: Mitigating retrieval-permutation-induced hallucinations in retrieval-augmented generation. *arXiv preprint arXiv:2601.02993*.

A Appendix

A.1 Related Work

Evidence-based Fact Checking. Capitalizing on the powerful capabilities of large language models (LLMs) across various tasks (Wang et al., 2026; Ma et al., 2026; Zhang et al., 2026a,b), recent studies (Qi et al., 2024; Li et al., 2025c; He et al., 2025; Wu et al., 2025a; Li et al., 2025a,b; Zeng et al., 2026; Ma et al., 2025) predominantly rely on LLMs augmented with retrieved external evidence to evaluate whether image–text pairs align with real-world facts. SNIFFER (Qi et al., 2024) fine-tunes LLMs and incorporates external cues to verify authenticity. CMIE (Li et al., 2025c), E2LVLM (Wu et al., 2025b), and DiFaR (Wan et al., 2025a) improve cross-modal alignment and optimize evidence utilization. LRQ-FACT (Beigi et al., 2024) and LEMMA (Xuan et al., 2024) focus on enhancing retrieval to obtain more targeted evidence, thereby improving verification performance. Recent studies have also examined omissions, but primarily in the form of contextual omission. TRACER (Tang et al., 2025) and OmiGraph (Wang et al., 2025e) retrieve missing information to infer implied intent, yet they largely treat such omissions as auxiliary cues for fact-checking and are mostly confined to plain-text settings.

Rather than verifying the factual veracity of the news preview, we focus on scenarios in which it is factually correct yet omits key information relative to the full news context, thereby inducing a shift in interpretation.

News Preview Understanding Beyond Veracity. Recent studies have also examined how the *presentation* of news itself shapes audience understanding and dissemination outcomes. DeceptionDecoded (Wu et al., 2026) adopts a generative perspective to analyze how news creators construct misleading presentations under different intents, while NINT (Wang et al., 2025d) and In-Side (Wang et al., 2025c) leverage intent as an auxiliary signal to strengthen misinformation detection. Meanwhile, journalism research commonly regard frames (De Vreese, 2005) as a concrete manifestation of selective emphasis and omission, and existing work (Arora et al., 2025; Moernaut et al., 2020; Lücking and Pfeiffer, 2012; Ali and Hassan, 2022) primarily focuses on identifying and attributing frame categories. In addition, studies on clickbait and curiosity-driven headlines (Wang et al., 2025b;

Yu et al., 2024; Hagen et al., 2022), analyze the effects of news presentation by examining how exaggerated, suspenseful, and information-incomplete previews attract clicks and influence click-through rates and user engagement.

Different from our work, these studies typically do not characterize the interpretation shift induced by a preview relative to its underlying full context. In contrast, we formalize misleadingness as a reader-side interpretation shift and further introduce a correction mechanism to mitigate the potential misleadingness caused by news previews.

A.2 Data Annotation Details

A.2.1 Topic Selection

VisualNews (Liu et al., 2021) provides fine-grained topic annotations for each news article. Leveraging these annotations, we stratify sampling across ten high-impact topic categories to improve **societal relevance** and ensure the benchmark emphasizes domains most susceptible to misinformation.

The selected topics are:

```
"world",  
"international_relations",  
"politics_elections",  
"politics",  
"law_crime",  
"business_economy",  
"environment",  
"science_technology",  
"technology",  
"conflict_attack"
```

A.2.2 Content Filtering

In Section 4.1, we discussed to enforce **information density** via content filtering. We filter out low-quality instances (e.g., cases where the headline is simply a direct caption of the image) and focus the data on semantically more complex news reports. We retain instances where the full story raises and answers complex high-level inquiries, such as “*what happened*”, “*why it happened*”, and “*how will it unfold*”. The detailed filtering strategy is outlined in Figure 10.

A.2.3 Misleading Omission Judgment

Section 4.2 details our procedure for determining whether a news preview P_{news} is misleading based on U_p and U_c . Concretely, we compare the initial impression formed from the image–headline pair

alone (U_p), including the event’s nature and status, causal relations, responsibility attribution, and perceived severity, with the judgment after reading the full context (U_c). We label a preview as *misleading* when the full context substantially revises or overturns these core judgments, rather than merely refining or supplementing them. Otherwise, when the context mainly adds details while preserving the overall direction of the core judgment, we label it as *Non-misleading* (Figure 13).

A.2.4 Details of Cross-Model Agreement

In Section 4.2, we describe our cross-model agreement filtering strategy to mitigate single-model bias and improve annotation reliability. Based on our analysis, rationales r produced by the GPT family better match human writing style, so we first use GPT-4.1 (OpenAI, 2025a) for multi-stage annotation. We then have Gemini-2.5-pro (Google, 2025) annotate the same instances independently and apply agreement filtering. The two models produce consistent labels for over 84% of the samples. Because our goal is to identify omission-based misleading content, we retain only high-confidence instances where both models agree, prioritizing label precision over coverage of borderline cases.

A.2.5 Details of Misleadingness Correction

In Section 4.3, we describe how we construct a gold-standard reference for actively mitigating misleading instances. Specifically, for each misleading sample, we use the stronger GPT-5 to generate two corrected versions under the two settings of \mathcal{C} , namely MINIMAL-EDIT and FREE-FORM, and Figure 14 illustrates the detailed procedure. We then follow the annotation protocol in Section 4.2 to re-assess whether $\hat{P}_{news} = (\hat{T}, I)$ remains misleading. Finally, to obtain a high-confidence reference set that balances correctability and structural controllability, we retain only instances that are successfully corrected under both settings. In the test split, which contains 500 misleading instances, 78% can be corrected under FREE-FORM, whereas only 54% can be corrected under MINIMAL-EDIT. Ultimately, we select the 265 instances that are successfully corrected under both \mathcal{C} settings as the gold-standard reference.

A.2.6 Human Evaluation Details

In Section 4.4, we describe our human evaluation. We assess annotation quality using accuracy and inter-annotator agreement. For the misleadingness

detection task, we evaluate both (i) understanding consistency and (ii) judgment consistency of the model outputs. Accuracy is defined as whether the model-simulated understanding or the final misleadingness judgment matches the human annotators’ decisions, and the evaluation protocol is shown in Figure 6. For the headline correction task, accuracy is defined as whether human annotators judge the rewritten headline to be no longer misleading, and the evaluation protocol is shown in Figure 7.

As shown in Table 5, we report detailed validation results. Here, (U_p, U_c) corresponds to the evaluation of understanding consistency, *Detection* corresponds to judgment consistency, and *Correction* corresponds to the headline correction task. The results indicate that our automated annotations achieve high agreement with expert judgments.

Metrics	U_p	U_c	Detection	Correction
Accuracy	0.95	0.96	0.94	0.95
Fleiss’ κ	0.83	0.86	0.81	0.82

Table 5: Automated annotations exhibit high agreement with expert judgments across interpretation consistency, detection accuracy, and correction effectiveness.

In addition, we analyze cases with annotator disagreement. For misleading omission detection, disagreements primarily stem from annotators adopting different event frames or inferred implications when interpreting the same image–headline preview, which in turn yields different judgments after reading the full context, as illustrated in Figure 6. For headline correction, while rewrites typically add or amend key information, some annotators may not view the added content as salient, or may apply different thresholds for whether the rewrite sufficiently eliminates misleadingness, leading to inconsistent judgments, as shown in Figure 7.

A.3 Experimental Details

A.3.1 Experimental Setup

We provide detailed definitions of G1–G4 in Table 2 as follows:

- **G1 (Oracle Upper Bound):** For all samples with gold-standard references, we prompt open-source LLMs to rewrite the headline, using the *Rationale* produced during annotation as guidance for correction. This setting estimates whether the open-source LLM can

perform effective correction when provided with oracle rationales.

- **G2 (Self-Rationale Guidance):** For all samples with gold-standard references, we first require open-source LVLMs to detect misleading omissions. For the subset predicted as *misleading*, we further evaluate the model’s ability to perform headline correction using its self-generated rationale.
- **G3 (Oracle-Rationale Guidance):** Using the same samples as G2, we replace the guidance with the gold-standard *Rationale* and perform headline rewriting, thereby quantifying the effect of rationale quality on correction performance.
- **G4 (End-to-End):** The full pipeline from detection to rationale generation to rewriting. For all samples with gold-standard references, the LVLM first predicts whether the preview is *misleading*. If it is deemed *misleading*, the model then performs headline correction conditioned on its self-generated *rationale*. This setting reflects real-world governance performance.

For misleading content correction, we evaluate performance using the correction success rate, defined as the proportion of misleading instances that are relabeled as non-misleading after correction. Given the corrected headlines, we re-assess misleadingness using the same evaluation model and multi-stage pipeline as in data annotation, ensuring a fair comparison of success rates across different rewrite models. In addition, we report BLEU, ROUGE, and cosine similarity to quantify the gap between constraint-based rewrites and the oracle headline.

A.4 Misleading Attribution Analysis

A.4.1 Frame Shift Analysis

In Section 6.4, we examine frame shifts between misleading and non-misleading previews. We adopt the set of generic news frames from (Arora et al., 2025), which we list below, and Figure 15 outlines the frame identification procedure. Specifically, we extract three semantic frames from the original preview, the rewritten preview, and the corresponding full news context. We then quantify frame shift by counting the number of overlapping frames between each preview (before and

after rewriting) and the full context. The generic news frames:

“Economic”,
“Capacity and Resources”,
“Morality”,
“Fairness and Equality”,
“Legality”,
“Policy”,
“Crime and Punishment”,
“Security and Defense”,
“Health and Safety”,
“Quality of Life”,
“Cultural Identity”,
“Public Opinion”,
“Political”,
“External Regulation”,
“Other”

A.4.2 Fine-grained Misleadingness Attribution

In Section 6.4, we further analyze fine-grained factors beyond frame shift that can induce misleadingness. For misleading samples in the test set, we first ask human experts to summarize five major causes, and then adopt an LLM-as-a-judge approach (Li et al., 2024; Pu et al., 2025) to assign an attribution label to each sample. Figure 16 provides definitions and explanations of these causes, along with the full attribution pipeline. As shown in Figure 5, **Missing background and conditions** dominates the distribution (about 65%, 323/500), while the remaining causes exhibit a clear long-tail pattern. This suggests that most misleading previews do not negate the core facts. Instead, they exploit *missing qualifiers*, such as temporal scope, applicability constraints, comparison baselines, subsequent developments, entity identities, or causal premises. When such prerequisites are omitted, readers are prone to extrapolate a local description into a global conclusion, leading to systematic bias.

A.5 Multimodal Mitigation

A.5.1 Image-Modal Analysis for Correction

Modality Attribution for Correction Failure:

In Section 6.5, to examine whether correction performance is significantly affected by the image modality, we first identify a news preview as misleading and then, based on its misleading rationale, use an LLM judge to categorize it as either *Text-Fixable* (misleadingness can be corrected by rewrit-

ing the headline) or *Image-Driven* (misleadingness is primarily driven by the image and cannot be corrected by headline rewriting alone). We then compare correction outcomes across these two subsets to quantify the impact of the image modality. The detailed prompts are provided in Figure 17 and definitions of *Text-Fixable* and *Image-Driven* are as follows:

Text-Fixable:

- If the misleading effect mainly stems from information omission, missing outcome, or omitted controversy in the headline, and the image itself merely serves as scene or atmosphere rendering—without anchoring a narrative, identity, event type, or timeline that is fundamentally inconsistent with the main theme of the article—then the case is considered “headline amendable.” In such cases, the misleading impression can be eliminated by rewriting the headline.

Image-Driven:


- If the image content strongly dominates the reader’s interpretation, anchoring an event type, emotion, identity, causality, or historical timeline that is seriously inconsistent with the true news context—even when the headline is maximally revised—the misleading effect cannot be corrected. Such cases are considered “not amendable,” and require image replacement or other multimodal interventions.

We note that the generated images in this experiment may not be fully factual, and our misleadingness assessment does not additionally verify their factual correctness. Therefore, our goal is not to produce factually accurate images, but to validate a key point: pairing news previews with semantically appropriate images can substantially reduce potential misleadingness. In practice, before publishing a news preview, creators can use similar semantic guidance to retrieve and select semantically matched real images from a repository. Accordingly, newsroom tools should assist editors in choosing more representative archival images to reduce misleading risk, rather than relying solely on headline edits.

Visual Prototyping: In Section 6.5, we conduct a pilot study on 20 “uncorrectable” instances and perform visual prototyping by using a generative model to synthesize alternative images that are more semantically consistent with the full context. Concretely, we know the reason why each preview is misleading, and we also have the failed corrected headline along with the reason it remains misleading. This allows us to diagnose why headline-only edits are insufficient. More importantly, it enables us to infer what image semantics would mitigate the misleading impression when paired with the preview, and to derive prompts that would produce such images. We then use GPT to generate the corresponding images from these prompts. The full generation pipeline is shown in Figure 0, and the generated images together with their mitigation effects are presented in Figure 9.

NEWS

News Headline:
Members of the Turkish military stand guard in Istanbul's Taksim Square.



News Context

After an attempted military coup in Turkey, the government detained thousands of suspected participants ... The failed coup resulted in over 265 deaths and was partly thwarted by mass public opposition after President Erdogan called citizens to the streets. The coup attempt disrupted military operations, strained Turkey's international standing, and left the country facing deep political uncertainty.

Model Annotated data:

Preview Understanding: ...it seems that the Turkish military ... possibly to maintain order ... The soldiers' stance and the presence of civilians suggest a tense or significant event, such as a security operation, protest...

Context Understanding: ...a major internal conflict in Turkey... a portion of the military tried to seize power, but failed ... The aftermath seems to be a crackdown by the government on potential opposition, raising fears...

Misleading Judgment: The image and headline together suggest that the Turkish military is maintaining order in Istanbul's Taksim Square, implying a situation where the military is in control ... However, the full news context reveals the military presence was actually part of a failed coup attempt against the government, not a routine or government-sanctioned security operation. The government detained thousands of military personnel ... Thus, the image-headline pair is misleading.

Human Annotation Instruction:
You will be presented with a multimodal news preview (consisting of an image and a headline) and its corresponding full news content (news context), along with two model-generated interpretations: (1) a preview-based understanding, and (2) a context-based understanding. In addition, you will see the model's final judgment on whether the news is misleading (misleading judgment).

Question 1 (Understanding Consistency):
Please assess whether the model-generated understanding is consistent with your own understanding at the event level, including core aspects such as the nature and status of the event, causal relations, attribution of responsibility, or severity.
If the overall interpretation is consistent, annotate it as "1"; if there is a substantive discrepancy, annotate it as "0".

Question 2 (Misleading Judgment Consistency):
Based on the model's final decision, please determine whether its misleading / non-misleading judgment is consistent with your own judgment.
If consistent, annotate it as "1"; otherwise, annotate it as "0".


Supported annotators
Judgment: *Misleading*
Judgment Rationale: Both the annotators and the model agree that the preview conveys routine actions, such as maintaining order, and does not allow one to infer the presence of serious military operations.

Dissenting annotators
Judgment: *Non-Misleading*
Judgment Rationale: The annotator associates the preview with major military actions and therefore considers it part of a failed coup attempt against the government.

Figure 6: Human annotation guidelines for misleading content detection and an analysis of annotation disagreements.

NEWS

Original News Headline:
Democratic presidential candidate Bernie Sanders campaigns in Columbia SC on Monday.



News Context

Anger toward Wall Street remains a rare unifying force among both Democrats and Republicans, years after the financial crisis and government bailouts ... candidates from both parties distance themselves from Wall Street ... candidates like Ted Cruz and Donald Trump use populist rhetoric to criticize Wall Street influence, despite having their own financial industry connections...

Model Annotated data:

Preview Understanding: ...The image suggests he is attending a campaign event or making a public appearance, possibly giving a speech or addressing supporters...

Context Understanding: ...The deeper implication is that criticism of Wall Street has become a common and powerful tool for politicians seeking public support ... anti-Wall Street sentiment are influencing the tactics and messaging...

Misleading Judgment: The image and headline ... leads the reader to interpret the event as a routine campaign appearance by a single Democratic candidate ... However, the full news context reveals that ... bipartisan political trend ... both Democratic and Republican candidates are leveraging anti-Wall Street sentiment to gain public support. he article is not centered on Sanders alone, but rather on how distrust of Wall Street is shaping the strategies of multiple candidates across party lines...

Corrected News Headline:
Democratic presidential candidate Bernie Sanders campaigns in Columbia SC amid bipartisan anti-Wall Street sentiment.

Human Annotation Instruction:
You will be provided with a news preview, the corresponding full news content (news context), and the reason why the preview is identified as misleading (misleading reason). In addition, you will see a rewritten news headline generated based on this reason.

Please determine whether the rewritten headline, when considered together with the full news context, may still mislead readers.
If you believe that the rewritten headline is no longer misleading, annotate it as "1";
if you believe that it remains misleading, annotate it as "0".

Supported annotators
(Considers the corrected headline to be non-misleading)
The annotator believes that the added content (highlighted in yellow) links the preview to the full contextual background (anti-Wall Street sentiment), allowing the full article to serve as a plausible interpretation of the preview.

Dissenting annotators
(Still considers the corrected headline to be misleading)
The annotator thinks that, even after adding the "anti-Wall Street" phrasing, the headline still centers on the candidate's specific campaign stop, leading readers to assume the report primarily focuses on the campaign itself. Therefore, they still considered it misleading..

Figure 7: Human annotation guidelines for misleading content correction and an analysis of annotation disagreements.

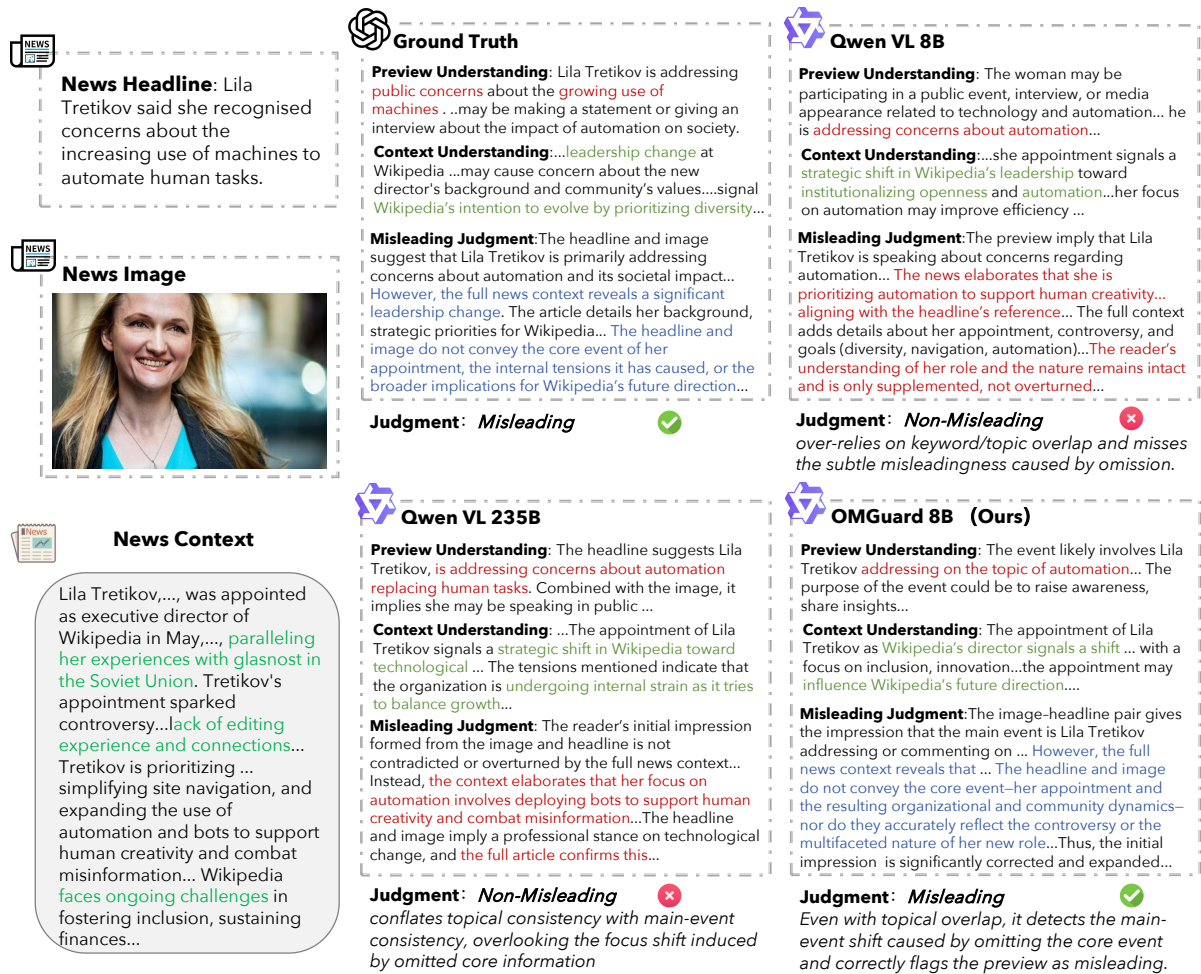


Figure 8: A case study of misleading content detection across different models.

Model	Model Card
GPT-4.1 (OpenAI, 2025a)	gpt-4.1-2025-04-14
GPT-5 (OpenAI, 2025b)	gpt-5-2025-08-07
Gemini-2.5-Pro (Google, 2025)	gemini-2.5-pro-preview-03-25
LlaVA-1.5-7B (Liu et al., 2023)	llava-1.5-7b-hf
Qwen3-VL-8B (Qwen Team, 2025)	Qwen3-VL-8B-Instruct
InternVL3.5-8B (Wang et al., 2025a)	InternVL3_5-8B
Llama3-VL-11B (Grattafiori et al., 2024)	Llama-3.2-11B-Vision-Instruct
GLM-4.1V-9B-Thinking (Hong et al., 2025)	GLM-4.1V-9B-Thinking
Qwen3-VL-8B-Thinking (Qwen Team, 2025)	Qwen3-VL-8B-Thinking
Qwen3-VL-235B-Thinking (Qwen Team, 2025)	Qwen3-VL-235B-Thinking
Llama3-VL-90B (Grattafiori et al., 2024)	Llama-3.2-90B-Vision-Instruct
Qwen3-VL-235B (Qwen Team, 2025)	qwen3-vl-235b-a22b-instruct

Table 6: Model cards for the VLMs used in our work.

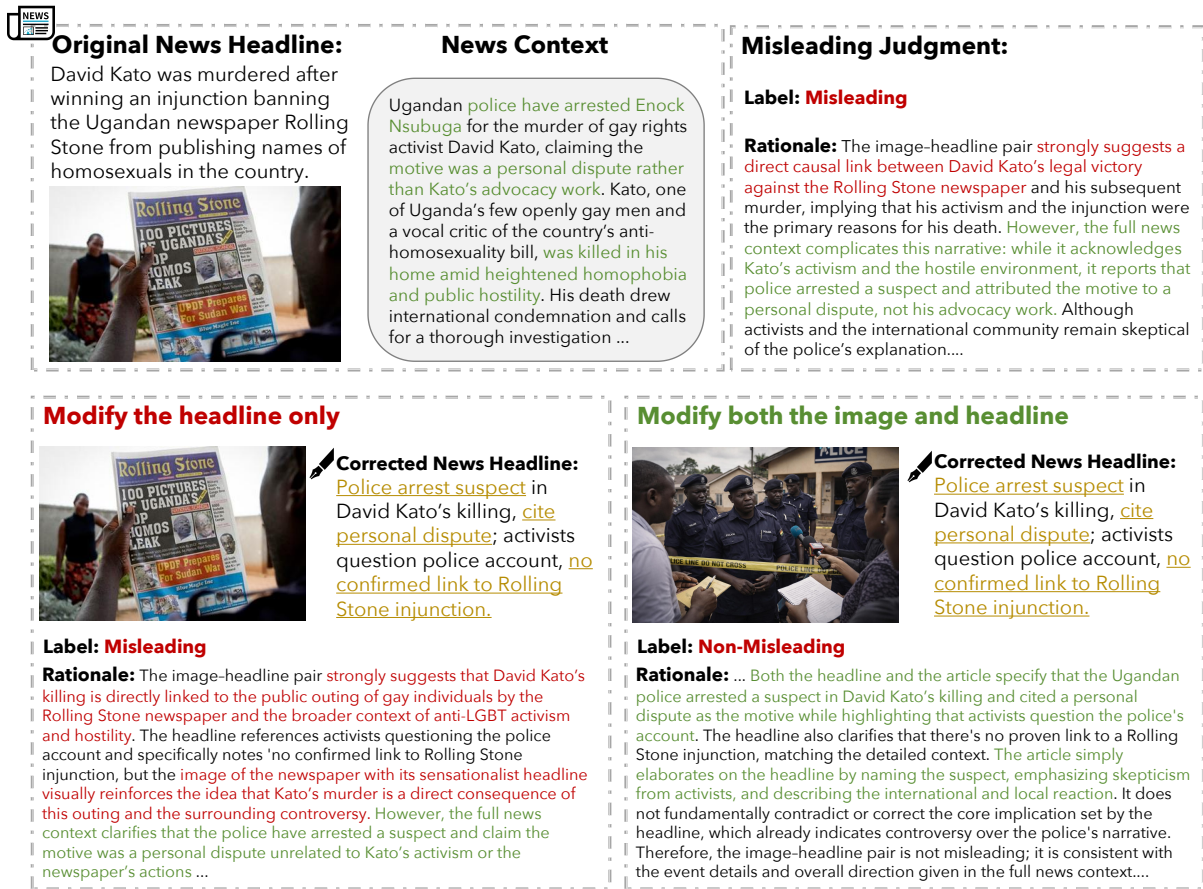


Figure 9: Examples of correction via visual image replacement. We show that replacing the original image with a generated suggestion image that better matches the article context can effectively mitigate misleadingness in news previews.

Context Filtering

You task is to annotate image-text pairs for news-signal screening.

Definitions:

- ld (Literal-Descriptive): venue/object/scene or general information (e.g., opening, decor, menu, discount). Does not convey any event information worth further inquiry; readers would not ask "what happened/why/what's next."
- ms (Message-Suggestive): conveys or implies a real-world event/impact (e.g., conflict, explosion, bombing, disaster, accident, evacuation, casualties, law enforcement, arrest, protest, policy, curfew, sanctions), or uses causal/temporal language that invites "what happened/why/what's next."

Input
IMAGE: You will be provided with the image.
TEXT: {News Headline}

Output Requirements Outputs must follow the JSON format below, consisting of three keys:

"label": "ld" | "ms",
"reason": "rationale citing the main textual cues"

Figure 10: Prompt for selecting high-quality news instances.

LLM-based Preview Understanding Simulation

Task Description

You are an average news reader. you will be provided with a piece of news that includes an image and a news headline. From a reader's perspective, describe your immediate impression of the news and make reasonable inferences at the detail level.

You need to complete the following parts:

- Analyze ****only**** based on the image and the news headline.
- Describe what you see (surface interpretation).
- Infer what event might be happening based on visual cues and the headline (event implication).

Input

News Headline: {NEWS_HEADLINE}

Image: (will be provided)

Output Format (JSON)

“Image-Headline”: {

“Surface_Interpretation”: “What is the surface interpretation?”,

“Event_Implication”: “What is the deep meaning, and what is the purpose?” }

Figure 11: Prompt for LLM-based preview understanding simulation.

LLM-based News Context Understanding Simulation

Task Description

You are an average news reader. you will be provided with a full news article. From a reader's perspective, describe your immediate impression of the news and make reasonable inferences at the detail level.

You need to complete the following parts:

- Analyze based on the news context.
- Describe what you see (surface interpretation).
- Infer what event might be happening based on the news context (event implication).

Input

News Context: {NEWS_CONTEXT}

Output Format (JSON)

“News_Context”: {

“Surface_Interpretation”: “What is the surface interpretation?”,

“Event_Implication”: “What is the deep meaning, and what is the purpose?” }

Figure 12: Prompt for LLM-based news context understanding simulation.

Misleading Omission Judgment

Task Description

You will receive an image, a news headline, a full news context, a reader's surface interpretation and event implication for the image-headline pair, a reader's surface interpretation and event implication for the full news context.

You need to complete the following parts:

- If a reader forms an impression about the nature, status, cause and effect, the responsible party, or severity of a news event when only exposed to images and titles, and this impression is significantly corrected, restricted, or overturned after reading the full news, it is considered misleading.
- On the contrary, if the full news only elaborates, extends, or supplements the content implied by the title (for example, by providing more details, reactions, or outcomes), without altering the reader's understanding of the basic direction or core judgment of the event, it is considered non-misleading.

Input

Image: (will be provided)

News Headline: {NEWS_HEADLINE}

Full News Context: {CONTEXT}

Reader Interpretations based on image-headline and context: {READER_INFER}

Output Format (JSON)

```
{ "Misleading": "Yes/No", "Reason": "Not less than 100 words, focus on the event level." }
```

Figure 13: Prompt for misleading omission judgment.

News Headline Correction

Task Description

You are a news rewriting expert. You will receive an news image, an news headline, and the full news context. Compared with the news context, the image-headline pair is considered misleading. You will also be provided with the corresponding reason why it is misleading.

Please follow the steps below to generate a non-misleading headline:

1. Analyze the Misleading Cause - Based on the provided data, identify the main reasons why the original headline is misleading, including any factual, contextual, or expressive distortions.
2. Suggestions on Improvement - Consider what kinds of information or phrasing should be included in the headline to prevent misleading readers and accurately convey the core message of the news.
3. Generate the Headline - Based on the above analysis, produce a non-misleading headline that is factually accurate, semantically clear, and maintains a neutral tone.

Rewriting requirements:

[This can be replaced according to different rewritten types]

Minimal-Edit: - The rewritten news headline may contain at most {limit_words} additional words compared to the original headline. - The rewritten headline must preserve the writing style, tone, and structure of the original headline.

Free-Form: - The rewritten news headline may contain at most {limit_words} additional words compared to the original headline.

Input: Image: You will be provided.

News Headline: {NEWS_HEADLINE}

Full News Context: {NEWS_CONTEXT}

Misleading reason of image-headline pair: {MISLEADING_REASON}

Output(json):

```
{ "Misleading_Cause": xxx, "Suggested_Improvement": xxx, "Rewritten_Caption": xxx }
```

Figure 14: Prompt for misleading headline correction.

News Frame Identification

Task Definition

You are an expert media analyst. Your task is to identify the relevant generic news frames presented by the combination of the provided News Image and Headline.

Instruction:

- Analyze how the image and headline interact. A news item often contains multiple angles (e.g., both "Political" and "Policy").
- Select the ****Top-3 most relevant frames**** that represent the dominant perspectives from the taxonomy: {taxonomy}

Input

IMAGE: You will be provided with the image.

TEXT: {News Headline}

Output

Output strictly in JSON format with two keys:

- { - "reasoning": Brief explanation of why these frames apply.
- "frames": A list of strings containing the exact names of the selected frames (e.g., ["Economic", "Political", "Policy"]).

Figure 15: Prompt for frame analysis.

Fine-grained Misleading Attribution

Task You are a misleading attribution classifier, designed to evaluate the reasons why an image–headline pair may be misleading compared to the full news context. Your task is to determine which category of misleading type the given reason belongs to.

Input

- Image: You will be provided.
- News Headline: {NEWS_HEADLINE}
- Full NEWS Context: {NEWS_CONTEXT}
- Reason why an image–headline pair may be misleading compared to the full news context: {REASON}

Categories

Choose exactly one of the following categories:

1. Missing background and conditions:

- The reason mainly points out that the image–headline pair omits essential background or conditions needed to correctly understand the event (for example, prior context, policy constraints, key actors, follow-up developments, or outcomes). Because this context is missing, readers are likely to form an incomplete or distorted overall impression.

2. Misleading scale and representativeness:

- The reason mainly emphasizes that the image–headline pair misleads about how large, frequent, or systemic the event is. It only shows isolated or local cases, or uses extreme examples in a way that underplays or exaggerates the true scale, prevalence, or impact described in the full news context.

3. Omission of perspectives and controversy:

- The reason mainly highlights that the image–headline pair hides important viewpoints or controversy. It presents only one side (for example, an official or dominant narrative) while omitting affected groups, opposition voices, counter-arguments, or social conflict that are present in the full news context, leading to a one-sided understanding.

4. Misleading causality and temporality:

- The reason mainly concerns incorrect or misleading suggestions about cause–effect relations, event sequence, or current status. The image–headline pair may imply that one action directly caused an outcome, that an event is still ongoing, or that a past event is current, in ways that are not supported by the full news context.

5. Others:

- Use this category if the reason does not clearly fall into any of the four types above, or if you are not confident which category is most appropriate.

Output

Return the output in standard JSON format with the following fields:

- { "attribution_class": "Only the most possible class",
- "attribution_reason": "Explain in detail why it belongs to this category, referring to the given text for analysis" }

Figure 16: Prompt for fine-grained misleading attribution.

Modality Attribution

You will be provided with an image, the corresponding headline, the full news article, a reader interpretation based solely on the image–headline pair, a reader interpretation based on the full news article, and an explanation of why the image–headline pair is considered misleading compared to the complete news content.

Task

In multimodal news data, there exist a large number of samples in which the image–headline pair does not align with the main theme of the article context, easily misleading readers. In practice, simply rewriting the headline (text) does not always eliminate this misleading effect. In some cases, the image strongly dominates the narrative focus, emotion, or scene, so even after the headline is maximally revised, readers may still form an understanding that does not match the true news context. Therefore, the goal of this task is to automatically identify and annotate which misleading samples are likely to become non-misleading solely through headline rewriting.

Judgment Criteria

Text-Fixable:

- If the misleading effect mainly stems from information omission, missing outcome, or omitted controversy in the headline, and the image itself merely serves as scene or atmosphere rendering—without anchoring a narrative, identity, event type, or timeline that is fundamentally inconsistent with the main theme of the article—then the case is considered “headline amendable.” In such cases, the misleading impression can be eliminated by rewriting the headline.

Image-Driven:

- If the image content strongly dominates the reader’s interpretation, anchoring an event type, emotion, identity, causality, or historical timeline that is seriously inconsistent with the true news context—even when the headline is maximally revised—the misleading effect cannot be corrected. Such cases are considered “not amendable,” and require image replacement or other multimodal interventions.

Input:

Image: You will be provided.

News Headline: {NEWS_HEADLINE}

Full News Context: {NEWS_CONTEXT}

reader interpretation based only on the image–headline pair: {Reader_Preview}

a reader interpretation based on the full news article, and an explanation of why the image–headline pair is considered misleading compared to the complete news context: Reader_context

Misleading reason of image-headline pair: {MISLEADING_REASON}

Output(json):

```
{ "label": Text-Fixable or Image-Driven, "reason": xxx }
```

Figure 17: Prompt for modality attribution.

Visual Prototyping

You will receive a news preview (including an image and a headline) and the corresponding news context. It is known that this news preview is misleading compared to the news context.

We have rewritten the headline based on the identified original misleading rationale. However, the rewritten headline is still misleading. We believe this is mainly because the image introduces misleading cues.

I will provide you with:

Image: (will be provided)

Headline: {HEADLINE}

Context: {CONTEXT}

Original Misleading Rationale: {Original Misleading Rationale}

Rewritten Headline: {Rewritten Headline}

Rewritten Misleading Rationale: {Rewritten Misleading Rationale}

You need to perform visual prototyping: analyze what kind of contextual image the rewritten headline should be integrated with so that the new preview (New Image + Rewritten Headline) is no longer misleading. You should output a description of the recommended image and an image prompt for generating it.

Output (JSON):

```
{ "Image description": "xxx", "Image Prompt": "xxx" }
```

Figure 18: Prompt for visual prototyping.