Seeing Through Deception: Uncovering Misleading Creator Intent in Multimodal News with Vision-Language Models

Jiaying Wu[1] **Fanxiao Li**[2] **Min-Yen Kan**[1] **Bryan Hooi**[1]

[1]National University of Singapore [2]Yunnan University

jiayingwu@u.nus.edu, lifanxiao@mail.ynu.edu.cn, {kanmy, bhooi}@comp.nus.edu.sg

Abstract

The real-world impact of misinformation stems from the underlying misleading narratives that creators seek to convey. As such, interpreting misleading creator intent is essential for multimodal misinformation detection (MMD) systems aimed at effective information governance. In this paper, we introduce an automated framework that simulates real-world multimodal news creation by explicitly modeling creator intent through two components: the *desired influence* and the *execution plan*. Using this framework, we construct DECEPTIONDECODED, a large-scale benchmark comprising 12,000 image-caption pairs aligned with trustworthy reference articles. The dataset captures both misleading and non-misleading intents and spans manipulations across visual and textual modalities. We conduct a comprehensive evaluation of 14 state-of-the-art vision-language models (VLMs) on three intent-centric tasks: (1) *misleading intent detection*, (2) *misleading source attribution*, and (3) *creator desire inference*. Despite recent advances, we observe that current VLMs fall short in recognizing misleading intent, often relying on spurious cues such as superficial cross-modal consistency, stylistic signals, and heuristic authenticity hints. Our findings highlight the pressing need for intent-aware modeling in MMD and open new directions for developing systems capable of deeper reasoning about multimodal misinformation. [1]

Content Warning: this paper contains potentially harmful text and images.

Figure 1: **The importance of detecting misleading creation intent in multimodal news beyond surface-level cross-modal consistency.** For instance, a malicious creator can craft captions that are semantically consistent with the image yet deliberately convey a false narrative.

1 Introduction

Multimodal misinformation, which combines persuasive text with compelling visuals, poses significant threats to public understanding and can lead to serious societal harm (Zhou and Zafarani, 2020; Alam et al., 2022; Do Nascimento et al., 2022). A growing body of research in multimodal misinformation detection (MMD) focuses on identifying *cross-modal misalignment*, particularly in two major forms: (1) *out-of-context* (OOC) misinformation (Yuan et al., 2023; Qi et al., 2024), where images and captions from unrelated events or time periods are falsely paired; and (2) *multimodal media manipulation* (Shao et al., 2023; Liu et al., 2025), involving subtle changes such as shifts in facial expressions or caption phrasing that alter the perceived message. However, current MMD benchmarks often simulate such misalignments in artificial ways, either by mismatching image and

[1]Link to DECEPTIONDECODED: [GitHub Repo].

text with CLIP-based (Radford et al., 2021) similarity scores (Luo et al., 2021) or through sentiment word substitutions (Sudhakar et al., 2019; Shao et al., 2023). These arbitrary strategies create a fundamental gap between constructed benchmarks and the complexity of real-world MMD.

Beyond surface-level cross-modal misalignment, ***the key to effective real-world misinformation governance lies in detecting and understanding misleading creator intent*** (Appelman et al., 2022; Jaidka et al., 2025). Many misinformation campaigns are deliberately crafted to advance specific agendas, often without the audience's awareness, yet still manage to substantially influence public opinion (Ecker et al., 2022; Broda and Strömbäck, 2024). For example, as illustrated in Figure 1, a slight modification to a news caption can lead readers to attribute significant casualties to government negligence, ultimately eroding institutional trust. Although preliminary efforts have explored intent interpretation (Da et al., 2021; Gabriel et al., 2022; Wang et al., 2023), they are often limited to unimodal settings and fail to capture the synergy between visual and textual cues. Moreover, these approaches rely on reader inference of creator intent, which are prone to error due to the deceptive framing of such content. These limitations highlight a pressing need for transparent modeling of ground-truth creator intents in multimodal content, along with systematic benchmarks for detecting misleading communicative objectives.

To this end, we establish an *intent-guided framework* for simulating intent-aware multimodal news reporting. Specifically, we define diverse combinations of desired influence and execution plans, two core components that characterize real-world communication strategies (Hallahan et al., 2007; Paul, 2011). Building on this, we introduce DECEPTIONDECODED, *a large-scale benchmark* designed to reveal misleading creator intents in multimodal news. Constructed upon real-world image-caption pairs from the VisualNews (Liu et al., 2021) repository, DECEPTIONDECODED contains 12,000 instances created with misleading and non-misleading intent, spanning a broad spectrum of misleading intent types and framing strategies. Each instance is paired with a trustworthy reference article to serve as contextual grounding, and the quality of the generated data are further validated through human evaluation (Section 3.4). The benchmark supports evaluation of three intent-centric tasks related to MMD: (1) *Misleading Intent Detection*, which determines whether a news piece was intentionally crafted to mislead; (2) *Misleading Source Attribution*, which identifies whether the misleading signal originates from the image or the text; and (3) *Creator Desire Inference*, which infers the creator's intended area of societal impact, such as political polarization or public health disruption.

Using DECEPTIONDECODED, we evaluate 14 representative vision-language models (VLMs). Despite recent advances in multimodal reasoning (Hurst et al., 2024; Anthropic, 2025c), we find that ***state-of-the-art VLMs still struggle with intent understanding in the context of multimodal news***. For instance, even relatively strong models such as GPT-4o-mini (OpenAI, 2024) achieve only 18.8% accuracy when misleading intent is subtly expressed through caption manipulation. More concerningly, our analysis reveals that VLMs tend to rely on spurious features when detecting misleading intent, including superficial image–text consistency, stylistic patterns in the text, and heuristic cues suggesting authenticity. This over-reliance leads to significant performance degradation under simple adversarial manipulations, such as stylistic reframing (Chen and Shu, 2024; Wu et al., 2024) and persuasive prompting attacks (Zeng et al., 2024). Our benchmark and empirical analyses establish a foundation for future MMD research that targets deeper semantic reasoning and intent-aware understanding beyond shallow cross-modal consistency patterns.

2 Related Work

2.1 Multimodal Misinformation Detection

With the increasing prevalence of persuasive visual–textual misinformation on online platforms, research has expanded from detecting purely textual misinformation (Rashkin et al., 2017; Chen and Shu, 2024; Wu et al., 2024) to tackling multimodal misinformation detection (MMD). A growing body of work emphasizes misinformation arising from subtle cross-modal misalignments, such as out-of-context (OOC) misinformation (Qi et al., 2024) and multimedia manipulation (Shao et al., 2023; Liu et al., 2025). Recent advances leverage VLMs like LLaVA (Liu et al., 2023) and GPT-4o (Hurst et al., 2024) within retrieval-augmented frameworks, where the model reasons over multimodal inputs using retrieved trustworthy evidence (Khaliq et al., 2024; Xuan et al., 2024; Zhou et al., 2024; Braun et al., 2024). While effective in grounding re-

Benchmark	Task Setup	Content Modality		Creator Intent	News Context
		Textual	Visual		
EMU (Da et al., 2021)	visual manipulation understanding	-	✔	✔ (viewer-perceived)	-
Fakeddit (Nakamura et al., 2020)	mixed-source MMD	✔	✔	-	-
MRF (Gabriel et al., 2022)	reader perception reasoning	✔	-	✔ (reader-perceived)	-
NewsINT (Wang et al., 2023)	news intent interpretation	✔	-	✔ (reader-perceived)	-
PolitiFact (Shu et al., 2020)	MMD on social media	✔	✔	-	✔
GossipCop (Shu et al., 2020)	MMD on social media	✔	✔	-	✔
NewsCLIPpings (Luo et al., 2021)	OOC misinformation detection	✔	✔	-	-
DGM4 (Shao et al., 2023)	multimedia manipulation detection	✔	✔	-	-
MMFakebench (Liu et al., 2025)	mixed-source MMD	✔	✔	-	-
DECEPTIONDECODED (Ours)	misleading intent detection	✔	✔	✔ (creator-produced)	✔

Table 1: Comparison between DECEPTIONDECODED and prior benchmarks on misinformation-related tasks.

sponses, these approaches largely overlook the role of news creation intent, which has been identified as a key driver of misinformation (Sharma et al., 2019). Our work builds upon this evidence-based paradigm by evaluating VLMs on their ability to detect misleading communicative intent in multimodal news. We show that even the most advanced models still fall short in this setting, highlighting the need for intent-aware frameworks in MMD.

2.2 Misinformation Benchmarks

Various benchmarks has been developed to support misinformation-related research (Hanselowski et al., 2019; Yao et al., 2023). Representative datasets such as PolitiFact and GossipCop(Shu et al., 2020) collect naturally occurring misinformation on social media. Other works approach MMD from the perspective of multimedia manipulation; for example, NewsCLIPpings (Shu et al., 2020) and DGM4 (Shao et al., 2023) introduce cross-modal mismatches by substituting images or captions based on visual or textual similarity. Mixed-source datasets such as Fakeddit (Nakamura et al., 2020) and MMFakeBench (Liu et al., 2025) include both naturally occurring and synthetically manipulated content. While preliminary efforts such as EMU (Da et al., 2021), NewsINT (Wang et al., 2023), and MRF (Gabriel et al., 2022) have considered the role of news creator intent, they are limited to unimodal misinformation. Moreover, the notion of creator intent in these datasets is typically annotated from readers' subjective perspective, which may misalign with the actual communicative purpose. In this work, we introduce DECEPTIONDECODED, a benchmark featuring creator-produced multimodal misinformation, where intent is explicitly defined during content generation. A comparison of DECEPTIONDECODED to prior work is shown in Table 1.

3 DECEPTIONDECODED

To systematically investigate the role of misleading creator intent in multimodal news, it is crucial to construct a large-scale benchmark that explicitly defines diverse communicative intents across various news domains and uses these intents to guide content generation. However, this is challenging to achieve with real-world news data, where content is already published and annotators can only infer intent retrospectively from a reader's perspective. To overcome this, we introduce DECEPTIONDECODED (overviewed in Figure 2), a dataset of 12,000 multimodal news instances generated through an intent-guided synthesis pipeline. Each instance is grounded in a predefined creator intent setup, enabling controlled formulation while preserving content realism. Data quality is validated through rigorous human evaluation.

3.1 Source News Collection

We draw from VisualNews (Liu et al., 2021), a large-scale repository of trustworthy multimodal news, to ensure factual grounding. Specifically, we select 10 topic categories that offer broad societal relevance and sample richness. These span diverse domains such as politics, disasters, and public health. The full list of selected topics is provided in Appendix A.1.

To support meaningful evaluation of misinformation detection, we apply rigorous filtering criteria guided by principles from both misinformation governance and professional journalism. From a governance perspective, we prioritize content that serves the **public interest** (Kruger et al., 2024), focusing on events with high societal relevance and potential for wide public impact. In line with journalistic standards, we require that selected content demonstrates **professionalism, neutrality, and clarity** (Maras, 2013; Farley et al., 2014), ensuring that

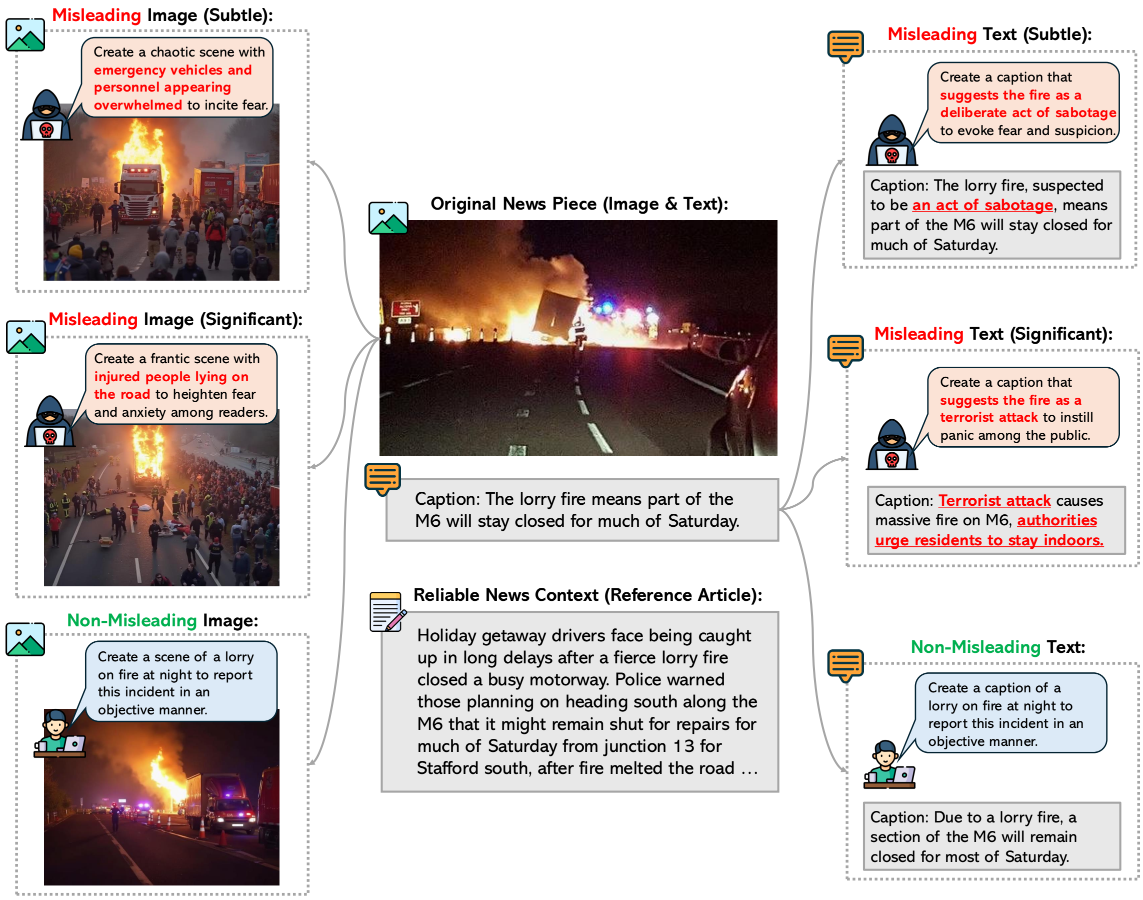

Figure 2: DECEPTIONDECODED: Overview of multimodal news curation guided by diverse simulated news creator intents, including both misleading and non-misleading cases.

language is objective, free from bias, and unambiguous. To further mitigate harm, we exclude instances that reference specific individuals or identifiable public figures (detailed in Appendix A.2).

This process yields 2,000 high-quality samples, evenly distributed across the 10 selected topics. Each instance is structured as $N = \{I, T, A\}$, where I is the news image, T is the accompanying caption, and A is a trusted reference article.

3.2 Creator Intent Establishment

Given the context provided by each original news instance $N = \{I, T, A\}$, we simulate both malicious and trustworthy content creators using GPT-4o (Hurst et al., 2024), leveraging its demonstrated capacity for social simulation (Anthis et al., 2025).

Inspired by strategic communication theories (Hallahan et al., 2007; Paul, 2011) and prior work on textual intent inference (Wang et al., 2023), we conceptualize creator intent (C_{int}) along two core dimensions: **(1) desired influence** and **(2) execution plan**. The desired influence refers to the societal sector(s) the creator aims to affect, selected from a predefined list of options (up to three per instance; see Appendix A.3). The execution plan is generated as open-ended text, detailing how the creator intends to achieve the intended influence.

3.3 Intent-Guided Multimodal News Creation

Guided by C_{int}, we generate misleading and non-misleading variants of each news piece obtained in Section 3.1, by either modifying the image or text. **Misleading** intents are categorized into two levels: **(1) subtle** (e.g., minor distortions to framing or background details to nudge reader interpretation) and **(2) significant** (e.g., major alterations that substantially change the perceived message). For **non-misleading** cases, we prompt the creator to faithfully paraphrase the caption or reconstruct the image while retaining alignment with the article.

For **textual modification**, we use GPT-4o (Hurst et al., 2024) to generate captions aligned with the specified intent, as defined in Section 3.2, while keeping the original image unchanged. This produces a sample of the form $N_{\text{text}} = \{I, T', A\}$, where I is the original image, T' is the modified caption, and A is the associated reference article.

For **visual modification**, we prompt GPT-4o to produce textual descriptions reflecting the intended visual manipulation. These descriptions are then used to generate images via the open-source FLUX.1 [dev] model (Black Forest Labs, 2024). The resulting sample is of the form $N_{\text{image}} = \{I', T, A\}$, where I' is the modified image and T is the original caption.

Dataset Statistics. For each of the 2,000 filtered news samples from Section 3.1, we generate misleading and non-misleading variants for each of the six creator intent categories outlined in Figure 2, resulting in a total of 12,000 instances. All samples are crafted to simulate the professional tone of real-world news reporting. Data examples are provided in Appendix A.4.

3.4 Human Verification of Data Quality

To assess the quality of news samples created in DECEPTIONDECODED, we conduct a human evaluation by randomly sampling 2% of the dataset, comprising 120 text and 120 image instances covering all six fine-grained intent categories. The goal is to determine whether human annotators perceive the misleading or non-misleading nature of each sample in alignment with the intent simulated by the news creators.

We recruit three graduate students as annotators. Each annotator is presented with 240 label-shuffled pairs of (original, DECEPTIONDECODED-generated) news pieces, where the generated version may be either misleading (M) or non-misleading (NM). Annotators are instructed to provide a binary label indicating whether the generated version sufficiently deviates from the original such that it could plausibly mislead an average reader. Additional details on the annotation protocol are provided in Appendix B.

We assess data quality using two metrics: **(1) accuracy** of labels in DECEPTIONDECODED (misleading/non-misleading) relative to the human-assigned label, based on aggregated human annotations, and **(2) inter-annotator agreement**, measured by Fleiss' κ (Fleiss, 1971). The results indicate high reliability and clarity, with 99.2% Acc. and $\kappa = 0.877$ for text samples, and 89.2% Acc. with $\kappa = 0.703$ for image samples.

3.5 Evaluation Tasks and Metrics

DECEPTIONDECODED facilitates the evaluation of the following three intent-centric tasks:

- **Task 1: Misleading Intent Detection**. Given an input triplet of either $\{I', T, A\}$ or $\{I, T', A\}$, the goal is to predict whether the instance is misleading (M) or non-misleading (NM).

- **Task 2: Misleading Source Attribution**. A 3-way classification task in which the model must attribute the source of misleading intent to either the image or the text modality, or output "NA" if the instance is non-misleading.

- **Task 3: Creator Desire Inference**. A multi-label classification task where the model must identify the creator's intended societal influence, selecting up to three options from a predefined list (see Appendix A.3).

We evaluate Tasks 1 and 2 using classification accuracy, and Task 3 using F1 score.

4 Experiments

4.1 Experimental Details

With DECEPTIONDECODED, we evaluate 14 representative vision-language models (VLMs) spanning a range of sizes, model families, and access levels. These include **(1) multimodal Large Reasoning Models (MLRMs)**, such as Claude-3.7-Sonnet (Anthropic, 2025c); **(2) proprietary multimodal Large Language Models (MLLMs)**, such as GPT-4o (Hurst et al., 2024); and (3) **open-source MLLMs**, such as the Qwen-2.5-VL series (Bai et al., 2025). The complete list of models used in our experiments is provided in Table 7.

We evaluate two representative paradigms in misinformation detection, distinguished by their reasoning focus: **(1)** The **implication-oriented (I)** approach, which focuses on the inferred implications of the news sample; and **(2)** The **consistency-oriented (C)** approach, which assesses consistency both across modalities and between the image–caption pair and the reference article. ***Unless otherwise specified, all experiments on* DECEPTIONDECODED *default to the consistency-oriented (C) approach.*** Experimental setup and evaluation prompt details are shown in Appendix C.1.

Model		Text M-Sub	Text M-Sig	Text NM	Text Avg.	Image M-Sub	Image M-Sig	Image NM	Image Avg.
o4-mini	I	64.8 (63.9)	91.2 (87.9)	95.4 (95.4)	83.8 (82.4)	25.7 (20.3)	41.8 (33.8)	91.6 (91.6)	53.0 (48.5)
	C	72.4 (70.3)	94.0 (92.4)	88.7 (88.7)	85.0 (83.8)	46.3 (33.7)	66.5 (50.3)	78.7 (78.7)	63.8 (54.2)
Claude-3.7-Sonnet	I	64.5 (55.0)	90.0 (82.0)	92.8 (92.8)	82.4 (76.6)	46.9 (43.8)	67.2 (64.0)	84.5 (84.5)	66.2 (64.1)
	C	67.4 (66.1)	90.2 (89.7)	90.3 (90.3)	82.6 (82.0)	50.3 (31.3)	71.1 (50.6)	81.9 (81.9)	67.8 (54.6)
Gemini-2.5-Pro	I	71.0 (70.7)	92.9 (92.5)	93.5 (93.5)	85.8 (85.6)	43.9 (31.1)	66.4 (50.2)	86.4 (86.4)	65.6 (55.9)
	C	72.0 (71.7)	93.0 (93.0)	92.2 (92.2)	85.7 (85.6)	53.3 (35.5)	73.9 (51.9)	80.3 (80.3)	**69.2 (55.9)**
GPT-4o	I	67.2 (60.9)	93.1 (85.8)	91.6 (91.6)	84.0 (79.4)	42.6 (39.8)	61.9 (59.9)	85.8 (85.8)	63.4 (61.8)
	C	70.2 (63.6)	93.3 (85.1)	86.6 (86.6)	83.4 (78.4)	51.7 (46.9)	69.3 (66.2)	78.0 (78.0)	66.3 (63.7)
Claude-3.5-Sonnet	I	58.0 (53.2)	88.5 (84.0)	94.8 (94.8)	80.4 (77.3)	25.3 (18.8)	42.7 (34.0)	93.8 (93.8)	53.9 (48.9)
	C	66.8 (66.0)	92.1 (91.6)	90.3 (90.3)	83.1 (82.6)	41.9 (23.7)	63.5 (37.4)	87.0 (87.0)	64.1 (49.4)
Gemini-1.5-Pro	I	**84.7 (83.6)**	**97.0 (96.3)**	76.5 (76.5)	**86.1 (85.5)**	54.6 (36.1)	72.4 (51.6)	74.8 (74.8)	67.3 (54.2)
	C	80.5 (79.0)	94.6 (94.0)	78.0 (78.0)	84.4 (83.7)	**56.9 (34.6)**	**74.5 (50.1)**	72.2 (72.2)	67.9 (52.3)
GPT-4o-mini	I	7.3 (5.1)	25.3 (19.8)	99.3 (99.3)	44.0 (41.4)	5.2 (4.6)	9.7 (9.1)	99.2 (99.2)	38.0 (37.6)
	C	18.8 (17.2)	45.9 (43.2)	95.7 (95.7)	53.5 (52.0)	21.3 (18.1)	35.8 (33.2)	94.6 (94.6)	50.6 (48.6)
Claude-3.5-Haiku	I	3.9 (2.7)	18.2 (12.6)	99.9 (99.9)	40.7 (38.4)	1.1 (0.9)	1.3 (1.2)	99.9 (99.9)	34.1 (34.0)
	C	4.2 (3.5)	15.7 (14.1)	99.7 (99.7)	39.9 (39.1)	3.0 (2.7)	5.6 (5.2)	99.7 (99.7)	36.1 (35.9)
Qwen2.5-VL-72B	I	47.8 (44.1)	81.0 (75.0)	97.0 (97.0)	75.3 (72.0)	19.5 (17.0)	30.2 (26.9)	92.6 (92.6)	47.4 (45.5)
	C	49.6 (47.4)	82.5 (79.9)	94.4 (94.4)	75.5 (73.9)	28.8 (25.3)	42.2 (38.8)	88.6 (88.6)	53.2 (50.9)
Qwen2.5-VL-32B	I	18.3 (17.3)	45.4 (44.0)	98.4 (98.4)	54.0 (53.2)	16.0 (15.3)	25.9 (25.2)	96.7 (96.7)	46.2 (45.7)
	C	27.5 (27.1)	58.6 (58.1)	97.3 (97.3)	61.1 (60.8)	21.8 (17.9)	33.6 (28.3)	94.6 (94.6)	50.0 (46.9)
Llama-3.2-11B	I	8.5 (7.9)	18.1 (13.7)	97.0 (92.7)	41.2 (38.1)	7.0 (4.7)	8.4 (7.1)	96.6 (93.9)	37.3 (35.2)
	C	9.1 (5.7)	19.4 (9.2)	94.7 (88.5)	41.1 (34.5)	11.9 (12.2)	15.5 (13.9)	94.2 (88.0)	40.5 (38.0)
InternVL-8B	I	3.4 (6.9)	3.9 (6.9)	97.6 (89.4)	35.0 (34.4)	2.3 (7.7)	2.1 (8.8)	98.0 (88.1)	34.1 (34.9)
	C	5.5 (2.2)	6.2 (2.4)	96.3 (89.5)	36.0 (31.4)	6.0 (11.4)	7.1 (12.1)	96.3 (89.4)	36.5 (37.6)
LLaVA-v1.6-7B	I	0.0 (0.1)	0.0 (0.1)	100.0 (100.0)	33.3 (33.4)	0.0 (0.0)	0.0 (0.0)	100.0 (100.0)	33.3 (33.3)
	C	0.0 (0.0)	0.0 (0.0)	100.0 (100.0)	33.3 (33.3)	0.0 (0.0)	0.0 (0.0)	100.0 (100.0)	33.3 (33.3)
Qwen-VL-7B	I	0.0 (0.0)	0.0 (0.0)	100.0 (100.0)	33.3 (33.3)	0.0 (0.0)	0.0 (0.0)	100.0 (100.0)	33.3 (33.3)
	C	0.0 (0.0)	0.0 (0.0)	100.0 (100.0)	33.3 (33.3)	0.0 (0.0)	0.0 (0.0)	100.0 (100.0)	33.3 (33.3)

Table 2: **VLM accuracy (%) on misleading intent detection (black) and misleading source attribution (colored).** **I** refers to the implication-oriented approach, and **C** refers to the consistency-oriented approach (see Section 4.1). Green cells indicate the top-2 best-performing models. Red cells indicate hallucinations observed in smaller VLMs, where the model attributes a misleading source (M) despite predicting the instance as non-misleading (NM).

Model	Text M-Sub	Text M-Sig	Text Avg.	Image M-Sub	Image M-Sig	Image Avg.
o4-mini	60.9	79.7	70.3	37.7	56.3	47.0
Claude-3.7-Sonnet	61.1	82.6	71.9	44.7	**65.4**	**55.1**
GPT-4o	57.0	75.9	66.5	39.9	55.4	47.7
Claude-3.5-Sonnet	59.4	**82.8**	71.1	35.0	55.0	45.0
Gemini-1.5-Pro	**68.5**	81.9	**75.2**	**46.6**	63.6	**55.1**
GPT-4o-mini	0.8	3.2	2.0	1.7	2.6	2.2
Claude-3.5-Haiku	0.0	0.0	0.0	0.0	0.0	0.0
Qwen2.5-VL-72B	44.1	74.0	59.1	24.6	37.6	31.1
Qwen2.5-VL-32B	0.0	0.0	0.0	0.4	0.0	0.2

Table 3: **Performance (F1%) of VLMs on creator desire inference.** Green cells indicate the best-performing models, while red cells indicate models that failed to make a single correct prediction.

4.2 Main Results

Consistency-Oriented Reasoning Generally Outperforms Implication-Oriented Reasoning. As shown in Table 2, a possible reason is that misleading news is often constructed through deliberate content manipulation. Detecting unsubstantiated inconsistencies, either between the image and its caption or between the image-caption pair and the reference article, helps models better identify misleading intent.

VLMs Struggle to Reason About Misleading Creator Intent. As shown in Table 2, even state-of-the-art MLRMs such as Claude-3.7 demonstrate limited performance in detecting and attributing misleading creator intent. Performance is even lower on the more complex task of creator desire inference (Table 3), suggesting a deeper challenge in reasoning about communicative goals.

These findings raise alarming concerns. Our setup reflects a realistic deployment scenario, where readers are exposed to subtle yet misleading representations of otherwise trustworthy multimodal news and must rely on automated systems for guidance. The consistent underperformance of VLMs under such conditions suggests an urgent need to investigate the underlying patterns these models exploit when assessing content authenticity, inspiring our in-depth analysis in Section 5.

Model	All Misleading (Acc. %)			All Non-Misleading (Acc. %)		
	Original	w/ Helpful Hint	w/ Adversarial Hint	Original	w/ Helpful Hint	w/ Adversarial Hint
o4-mini	69.8	81.0 (+11.2)	46.8 (-23.0)	83.7	90.2 (+6.5)	68.4 (-15.3)
Claude-3.7-Sonnet	69.7	88.5 (+18.8)	49.3 (-20.4)	86.1	93.9 (+7.8)	58.5 (-27.6)
Gemini-2.5-Pro	73.0	84.6 (+11.6)	43.2 (-29.8)	86.3	99.8 (+13.5)	72.8 (-13.5)
GPT-4o	71.1	87.6 (+16.5)	29.3 (-41.8)	82.3	97.4 (+15.1)	61.1 (-21.2)
Claude-3.5-Sonnet	66.1	74.6 (+8.5)	47.5 (-18.6)	88.6	95.6 (+7.0)	82.4 (-6.2)
Gemini-1.5-Pro	76.6	96.3 (+19.7)	54.3 (-22.3)	75.1	97.3 (+22.2)	51.0 (-24.1)
GPT-4o-mini	30.4	84.9 (+54.5)	6.4 (-24.0)	95.1	99.8 (+4.7)	50.4 (-44.7)
Claude-3.5-Haiku	7.1	34.5 (+27.4)	0.8 (-6.3)	99.7	100.0 (+0.3)	95.7 (-4.0)
Qwen2.5-VL-72B	50.7	81.9 (+31.2)	30.2 (-20.5)	91.5	98.2 (+6.7)	61.6 (-29.9)
Qwen2.5-VL-32B	35.3	99.5 (+64.2)	1.2 (-34.1)	95.9	99.6 (+3.7)	3.4 (-92.5)

Table 4: **Misleading intent detection performance of VLMs when exposed to spurious authenticity-related hints**, averaged over both misleading and non-misleading cases. Hint construction details are provided in Section 5.1.

Model	M-Sub (Text)			M-Sig (Text)		
	I+T	T+A	Full	I+T	T+A	Full
GPT-4o	21.8	63.4	**70.2**	51.6	88.4	**93.3**
GPT-4o-mini	3.6	**36.8**	18.8	10.2	**64.4**	45.9
Qwen2.5-72B	10.6	**53.6**	49.6	24.6	**83.8**	82.5
Qwen2.5-32B	6.2	**35.8**	27.5	18.4	**67.2**	58.6

Table 5: **VLM performance (Acc. %) on misleading intent prediction under partial modality settings**, suggesting that smaller VLMs may overly rely on image-text consistency for misleading intent detection. (**I+T**: image & caption only; **T+A**: text & reference article only; **Full**: all modalities provided.)

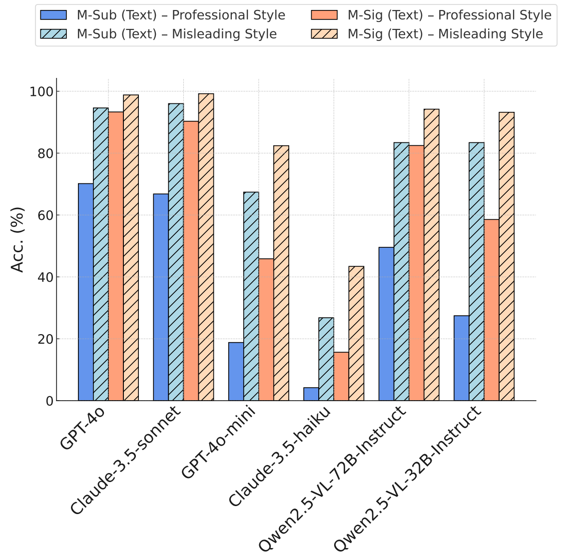

Figure 3: Misleading intent detection performance of VLMs on misleading text presented either in a **professional** tone (as used in the original DECEPTIONDECODED setting) or in an **explicitly misleading** style.

5 Analysis

5.1 Probing the Limitations of VLMs

VLMs Can Be Misled by Surface-Level Image–Text Consistency. Although misleading intent is often not substantiated when compared with the trustworthy reference article, the manipulated image and caption typically appear internally consistent, as illustrated in Figure 1. Ideally, models should integrate information across all modalities (i.e., image, text, and article) to detect such deceptive intent. To investigate this, we evaluate VLMs under two partial input settings: (1) Image + Text, and (2) Text + Article.

However, results in Table 5 reveal a contradiction to this ideal. For instance, GPT-4o-mini performs better when detecting misleading text by comparing it with the article, but its performance drops notably when the misleading image is added. This suggests that consistency between image and text can mask deception, misleading models that lack deeper reasoning about cross-source veracity.

VLMs Fail to Detect Misleading Text Framed in Credible-Sounding Styles. Prior work in textual misinformation detection (Chen and Shu, 2024; Wu et al., 2024) has shown that large language models (LLMs) often rely on stylistic cues rather than factual correctness when making veracity judgments. To investigate whether this stylistic bias extends to the multimodal setting, we reframe misleading captions in DECEPTIONDECODED from professional-sounding styles to more overtly deceptive or sensational tones (see Appendix C.2). As illustrated in Figure 3, we observe a similar trend: VLMs are more likely to misclassify misleading content when it is framed in a polished, authoritative style. This suggests that current models may be overly influenced by surface-level linguistic cues, posing a serious risk when misinformation is intentionally crafted to appear credible.

Models React Strongly to Spurious Authenticity Cues in Prompts. To examine whether VLMs are influenced by high-level framing cues that are not grounded in the input content, we inject two types of authenticity-related hints into the prompt (see Appendix C.3): **(1) Skeptical**, which presumes the news piece may contain intentional distortions. This hint serves as a *helpful* signal for misleading samples but becomes *adversarial* for non-misleading samples; and **(2) Trusting**, which assumes the content is reliable, reversing the helpful and adversarial roles.

As shown in Table 4, these prompts produce significant and consistent performance changes: they improve accuracy in helpful cases and degrade performance in adversarial ones. This reveals that VLMs are susceptible to spurious prior assumptions about content authenticity, even when such cues are not supported by the input.

5.2 Impact of Image Generation Advances

DECEPTIONDECODED includes reasonably high-quality images generated using the FLUX model, as supported by human validation results in Section 3.4. As image generation capabilities continue to advance, we conduct a pilot study to assess whether VLM performance improves when exposed to higher-fidelity images generated by `GPT-image-1` (OpenAI, 2025a), using a subset of 200 sampled misleading instances.

As shown in Table 6 and Figure 4, we observe modest performance gains, likely due to the more vivid visual portrayal of creator intent. However, despite these improvements, understanding the underlying communicative intent behind multimodal news remains a significant challenge even when image realism is enhanced. These findings highlight that surface-level fidelity alone is insufficient, reinforcing the need for models that reason explicitly about creator objectives.

The growing realism of deceptive images as reflected by GPT-generated instances introduces a significant risk. When high-fidelity images convincingly align with misleading textual narratives, the deception becomes more subtle and harder to detect. These advances further underscore the importance of reasoning about the creator's underlying intent, rather than relying solely on content realism or modality alignment.

Figure 4: **Case Study:** a VLM (GPT-4o-mini) fails to detect misleading creator intent in a FLUX-generated image due to its inability to recognize the unsubstantiated presence of a crowd with fireworks, given the news context. The state-of-the-art `GPT-image-1` model produces a clearer depiction, leading to a correct prediction.

Model	M-Sub (Image)		M-Sig (Image)	
	Flux	GPT-image	Flux	GPT-image
GPT-4o	51.0	66.0	69.0	84.0
Claude-3.5-Sonnet	42.5	64.0	65.0	80.0
GPT-4o-mini	22.5	47.5	37.5	61.5
Claude-3.5-Haiku	3.0	9.0	6.0	18.0
Qwen2.5-72B	22.5	53.5	33.0	57.0
Qwen2.5-32B	30.0	37.0	46.5	46.0

Table 6: **Misleading intent detection performance of VLMs on images created using different image generation models**, evaluated on a pilot test set of 200 samples. While performance improves with higher-fidelity GPT images, a substantial gap remains.

6 Conclusion

In this work, we establish the important yet underexplored task of understanding creator intent in multimodal misinformation. We introduce DECEPTIONDECODED, a large-scale benchmark designed to evaluate VLMs across intent-centric tasks. Through extensive experiments, we uncover alarming vulnerabilities in state-of-the-art VLMs, which over-rely on spurious shortcuts such as cross-modal consistency, stylistic cues, and heuristic indicators of authenticity to detect misleading intent. Our findings pave the way for future research on MMD systems capable of intent-aware, authenticity-centric semantic reasoning toward effective misinformation governance.

Limitations

Through DECEPTIONDECODED, we establish a framework for understanding misleading creator intent in multimodal news, providing a meaningful first step toward modeling and detecting misleading creator intent. While DECEPTIONDECODED presents a structured and scalable framework, it also opens up several avenues for further research in advancing intent-aware misinformation governance across more realistic and socially grounded scenarios.

First, although our framework supports the evaluation of creator intent interpretation through tasks such as misleading intent detection and creator desire inference, **fully open-ended intent articulation and corresponding evaluation** remains an open challenge. Future work could explore adaptive prompting strategies and leverage the emerging LLM-as-a-Judge paradigm (Gu et al., 2024) to assess generated intent representations in more flexible and scalable ways.

Second, our current benchmark focuses on intent detection and interpretation, but does not directly measure downstream societal influence. In real-world scenarios, the effectiveness of misinformation depends not only on creator intent but also on how audiences interpret and react to the content. **Incorporating social grounding**, such as user feedback signals, comment chains, or engagement patterns on social media, would provide richer context for evaluating the influence of misleading content and calibrating model responses accordingly.

Despite these limitations, DECEPTIONDECODED represents an important contribution toward systematic, intent-aware evaluation of VLMs in real-world misinformation contexts. We hope our benchmark and insights will serve as a solid foundation for future work.

Ethics Statement

The intent-guided simulation of multimodal misinformation in DECEPTIONDECODED involves strategies that could potentially be misused to generate misleading content using VLMs and open-source image generation models (e.g., FLUX). Nonetheless, given the realistic and challenging nature of detecting misleading creator intent in multimodal news, we believe it is important to transparently acknowledge these risks while reporting empirical findings on the performance and limitations of current state-of-the-art VLMs.

Our benchmark is developed solely to highlight the important yet underexplored aspect of modeling malicious creator intent in MMD, and to advance the understanding of current system limitations.To minimize harm, we explicitly avoid manipulating news content involving specific, identifiable individuals; all simulated content is based on anonymized or synthetic entities. To further mitigate misuse, we will open-source the dataset and evaluation scripts, but refrain from releasing the specific generation prompts that could be repurposed for deceptive use. Data access will be granted only to verified researchers under a binding usage agreement to ensure responsible use. All data collection and usage comply with the terms of service of the underlying models and platforms.

References

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643.

Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. 2025. Llm social simulations are a promising research method. *arXiv preprint arXiv:2504.02234*.

Anthropic. 2025a. Claude 3.5 haiku. https://www.anthropic.com/claude/haiku.

Anthropic. 2025b. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet.

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

Naomi Appelman, Stephan Dreyer, Pranav Manjesh Bidare, and Keno C Potthast. 2022. Truth, intention and harm: Conceptual challenges for disinformation-targeted governance. *Internet Policy Review*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Black Forest Labs. 2024. Flux. https://github.com/black-forest-labs/flux.

Tobias Braun, Mark Rothermel, Marcus Rohrbach, and Anna Rohrbach. 2024. Defame: Dynamic evidence-based fact-checking with multimodal experts. *arXiv preprint arXiv:2412.10510*.

Elena Broda and Jesper Strömbäck. 2024. Misinformation, disinformation, and fake news: lessons from an interdisciplinary, systematic literature review. *Annals of the International Communication Association*, 48(2):139–166.

Canyu Chen and Kai Shu. 2024. Can LLM-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Jeff Da, Maxwell Forbes, Rowan Zellers, Anthony Zheng, Jena D. Hwang, Antoine Bosselut, and Yejin Choi. 2021. Edited media understanding frames: Reasoning about the intent and implications of visual misinformation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2026–2039.

Israel Junior Borges Do Nascimento, Ana Beatriz Pizarro, Jussara M Almeida, Natasha Azzopardi-Muscat, Marcos André Gonçalves, Maria Björklund, and David Novillo-Ortiz. 2022. Infodemics and health misinformation: a systematic review of reviews. *Bulletin of the World Health Organization*, 100(9):544.

Ullrich KH Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K Fazio, Nadia Brashier, Panayiota Kendeou, Emily K Vraga, and Michelle A Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29.

Elizabeth Farley, Fiona Grady, Dean S Miller, Rory O'Connor, Howard Schneider, Michael Spikes, Constantia Constantinou, et al. 2014. Spj code of ethics. *The Power of Images*.

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

Saadia Gabriel, Skyler Hallinan, Maarten Sap, Pemi Nguyen, Franziska Roesner, Eunsol Choi, and Yejin Choi. 2022. Misinfo reaction frames: Reasoning about readers' reactions to news headlines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3108–3127.

Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Google. 2025. Genimi 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, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Kirk Hallahan, Derina Holtzhausen, Betteke Van Ruler, Dejan Verčič, and Krishnamurthy Sriramesh. 2007. Defining strategic communication. *International journal of strategic communication*, 1(1):3–35.

Andreas Hanselowski, Christian Stab, Claudia Schulz, Zile Li, and Iryna Gurevych. 2019. A richly annotated corpus for different tasks in automated fact-checking. *arXiv preprint arXiv:1911.01214*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Kokil Jaidka, Tsuhan Chen, Simon Chesterman, Wynne Hsu, Min-Yen Kan, Mohan Kankanhalli, Mong Li Lee, Gyula Seres, Terence Sim, Araz Taeihagh, et al. 2025. Misinformation, disinformation, and generative ai: Implications for perception and policy. *Digital Government: Research and Practice*, 6(1):1–15.

Mohammed Abdul Khaliq, Paul Yu-Chun Chang, Mingyang Ma, Bernhard Pflugfelder, and Filip Miletić. 2024. RAGAR, your falsehood radar: RAG-augmented reasoning for political fact-checking using multimodal large language models. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 280–296.

Ariel Kruger, Morgan Saletta, Atif Ahmad, and Piers Howe. 2024. Structured expert elicitation on disinformation, misinformation, and malign influence: Barriers, strategies, and opportunities. *Harvard Kennedy School Misinformation Review*.

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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Xuannan Liu, Zekun Li, Pei Pei Li, Huaibo Huang, Shuhan Xia, Xing Cui, Linzhi Huang, Weihong Deng, and Zhaofeng He. 2025. MMFakebench: A mixed-source multimodal misinformation detection benchmark for LVLMs. In *The Thirteenth International Conference on Learning Representations*.

Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6817.

Steven Maras. 2013. *Objectivity in journalism*. John Wiley & Sons.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6149–6157.

OpenAI. 2024. GPT-4o mini: Advancing Cost-Efficient Intelligence.

OpenAI. 2025a. Introducing 4o image generation. https://openai.com/index/introducing-4o-image-generation/.

OpenAI. 2025b. Introducing openai o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/.

Christopher Paul. 2011. *Strategic communication: Origins, concepts, and current debates*. Bloomsbury Publishing USA.

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. Qwen2.5-vl.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937.

Rui Shao, Tianxing Wu, and Ziwei Liu. 2023. Detecting and grounding multi-modal media manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating fake news: A survey on identification and mitigation techniques. *ACM transactions on intelligent systems and technology (TIST)*, 10(3):1–42.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3):171–188.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. Transforming delete, retrieve, generate approach for controlled text style transfer. *arXiv preprint arXiv:1908.09368*.

Zhengjia Wang, Danding Wang, Qiang Sheng, Juan Cao, Silong Su, Yifan Sun, Beizhe Hu, and Siyuan Ma. 2023. Understanding news creation intents: Frame, dataset, and method. *arXiv preprint arXiv:2312.16490*.

Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 2024. Fake news in sheep's clothing: Robust fake news detection against llm-empowered style attacks. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 3367–3378.

Keyang Xuan, Li Yi, Fan Yang, Ruochen Wu, Yi R Fung, and Heng Ji. 2024. Lemma: towards lvlm-enhanced multimodal misinformation detection with external knowledge augmentation. *arXiv preprint arXiv:2402.11943*.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2733–2743.

Xin Yuan, Jie Guo, Weidong Qiu, Zheng Huang, and Shujun Li. 2023. Support or refute: Analyzing the stance of evidence to detect out-of-context mis- and disinformation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4268–4280.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade LLMs to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14322–14350.

Xinyi Zhou, Ashish Sharma, Amy X Zhang, and Tim Althoff. 2024. Correcting misinformation on social media with a large language model. *arXiv preprint arXiv:2403.11169*.

Xinyi Zhou and Reza Zafarani. 2020. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40.

A DECEPTIONDECODED Details

A.1 News Topics

VisualNews (Liu et al., 2021) provides fine-grained topic annotations for each news article. For DECEPTIONDECODED, we select 10 major topic categories based on sample abundance, topical diversity, and real-world relevance.

The selected categories are:

'world',
'science_technology',
'politics_elections',
'environment',
'business_economy',
'technology',
'disaster_accident',
'conflict_attack',
'us-news',
'health_medicine_environment'

These topics form the foundation for generating diverse and realistic multimodal misinformation instances in DECEPTIONDECODED.

A.2 Source Data Acquisition

In Section 3.1, we obtain high-quality source data for DECEPTIONDECODED with the prompt outlined in Figure 8.

A.3 Aspects of Desired Influence

We define a set of major desired aspects of societal influence to facilitate creator intent establishment in Section 3.2. The aspects are listed as follows:

Political Polarization,
Social Polarization,
Cultural and Religious Polarization,
Economic Misinformation,
Public Health and Safety,
Environmental and Scientific Polarization,
Geopolitical and International Relations,
Psychological and Emotional Manipulation

A.4 DECEPTIONDECODED Examples

We present two DECEPTIONDECODED data examples in Figure 5 (textual modification) and Figure 6 (visual modification).

To prevent misuse, we refrain from disclosing the data generation prompts for Section 3.3 that could be repurposed for deceptive use.

B Human Evaluation Details

We illustrate the annotation process for human evaluation in Appendix B in Figure 7.

C VLM Evaluation Details

C.1 Experimental Setup

With DECEPTIONDECODED, we systematically evaluate 14 representative VLMs. Detailed model names and versions are provided in Table 7.

For all evaluations, we set the model temperature to 0 to ensure stable outputs, and run each experiment once. Detailed evaluation prompts for the implication- and consistency-based approaches are shown in Appendix C.4.

C.2 Caption Style Reframing Formulation

We adopt the following prompt for the stylistic reframing experiments in Section 5.1 and Table 3.

You are given a factual news caption. Your task is to rewrite it in an unreliable style: that is, make it sound sensational, biased, emotional, or manipulative, while preserving the original content and factual meaning. Do not add or remove any factual information. Only change the tone and style to reflect how misinformation might manipulate readers through language.
Original Caption: {caption}
Unreliable-style Caption:

C.3 Authenticity-Related Hint Formulation

To investigate how VLMs respond to subtle authenticity cues, we design two types of prompt-level hints aligned with the analysis in Section 5.1. These hints are inserted into the input to simulate prior assumptions about content reliability.

Hint (Skeptical):
"Hint: The news piece is likely to contain intentional distortions. Carefully analyze the image and caption to identify any misleading or manipulative elements."

Hint (Trusting):
"Hint: The news piece is sourced from reputable outlets and is therefore unlikely to contain intentional distortions. "

Model	Model Card
o4-mini (OpenAI, 2025b)	`o4-mini-2025-04-16`
Claude-3.7-Sonnet (Anthropic, 2025c)	`claude-3-7-sonnet-20250219`
Gemini-2.5-Pro (Google, 2025)	`gemini-2.5-pro-preview-03-25`
GPT-4o (Hurst et al., 2024)	`gpt-4o-2024-08-06`
Claude-3.5-Sonnet (Anthropic, 2025b)	`claude-3-5-sonnet-20241022`
Gemini-1.5-Pro (Georgiev et al., 2024)	`gemini-pro-1.5`
GPT-4o-mini (OpenAI, 2024)	`gpt-4o-mini-2024-07-18`
Claude-3.5-Haiku (Anthropic, 2025a)	`claude-3-5-haiku-20241022`
Qwen2.5-VL-72B-Instruct (Qwen Team, 2025)	`qwen2.5-vl-72b-instruct`
Qwen2.5-VL-32B-Instruct (Qwen Team, 2025)	`qwen2.5-vl-32b-instruct`
LLama-3.2-Vision-Instruct-11B (Grattafiori et al., 2024)	`llama-3.2-11b-vision-instruct`
InternVL2-8B (Chen et al., 2024)	`InternVL2-8B`
LLaVA-1.6-7B (Liu et al., 2024)	`llava-v1.6-mistral-7b`
Qwen2.5-VL-7B-Instruct (Qwen Team, 2025)	`Qwen2.5-VL-7B-Instruct`

Table 7: Model cards for the VLMs evaluated throughout our work.

Figure 5: An example in DECEPTIONDECODED, showcasing a scenario where the creator intentionally misrepresents the news content with a distorted caption to mislead readers.

C.4 Evaluation Prompts

We present our prompts for evaluating VLMs in Figure 9 (implication-oriented) and Figure 10 (consistency-oriented).

Caption: The letters of the new acronym of Mexico City are seen at a square of the city.

Reference Article:
Mexico has rechristened its capital city, embracing the name by which it is known worldwide, but causing a conundrum for residents who for decades have referred to the sprawling megalopolis as as the Federal District – or “DF” …

Misleading (Subtle)

Misleading (Significant)

Non-Misleading

Figure 6: An example in DECEPTIONDECODED, showcasing a scenario where the creator intentionally misrepresents the news content with a distorted image to mislead readers.

Original News

Caption: The lorry fire means part of the M6 will stay closed for much of Saturday.

Modified News

Caption: The lorry fire means part of the M6 will stay closed for much of Saturday.

Instruction:
You are provided with two multimodal news pieces reporting the same event, each consisting of an image and its corresponding caption (Left: original, Right: modified). Does the modified news piece distort the meaning or implication of the original news piece in a way that could mislead readers? Annotate with “0” if the modified image is Not Misleading, or “1” if it is Misleading.

Important Note: Focus only on the multimodal news content and implications. The presence of AI-generated artifacts should not be considered misleading unless they significantly change the implication or meaning conveyed by the image.

Figure 7: Illustration of human evaluation instructions for validating the quality of DECEPTIONDECODED.

Data filtering in DECEPTIONDECODED

You are a senior journalist tasked with evaluating a news caption based on five key criteria:

- Public Interest: Does the caption describe a significant event likely to capture public attention?
- Professionalism: Does it maintain a professional, objective tone?
- Neutrality: Is it free from political, ethnic, gender, or religious bias?
- Anonymity: Does it avoid mentioning any specific person's name or national/world leaders?
- Clarity: Is the message clearly conveyed without ambiguity, misunderstanding, or controversy?

Conclude with a single word "Yes" if the caption meets all five criteria or a single word "No" if it fails any.
Caption: {caption}
Response:

Figure 8: Prompt for selecting VisualNews (Liu et al., 2021) samples as source data for DECEPTIONDECODED.

Implication-Oriented Approach (I)

Your task is to determine whether a news piece has been intentionally distorted to mislead readers. The news piece consists of an image with its caption: {caption}. You are also provided a trustworthy reference article related to the same event: {article}.
Follow the structured steps below and strictly adhere to the output format.
Important Note: Focus only on the content of the news piece. The use of AI tools to assist image creation does not, by itself, indicate distortion.

Answer the following questions:
Q1. Has the news piece been intentionally distorted by its creator to mislead readers, such that **it conveys a biased, controversial, or manipulated message?**
- Respond with one of the following: [Yes, No]. Then provide a brief explanation.
Q2. What is the source of the manipulation—text or image?
- If answer to Q1 is "Yes", answer with one of the following: [Text, Image].
- If answer to Q1 is "No", answer: "NA".
Q3. What does the distortion suggest about the news publisher's intent?
- If answer to Q1 is "Yes" :
Primary Influence (Desire): Select up to three from the following list, ranked by relevance: [Political Polarization, Social Polarization, Cultural and Religious Polarization, Economic Misinformation, Public Health and Safety, Environmental and Scientific Polarization, Geopolitical and International Relations, Psychological and Emotional Manipulation].
- If answer to Q1 is "No", respond with "NA" for Q3.

Output Format: Respond in the following exact JSON format, preserving the structure, punctuation, and delimiters. Do not add any extra line breaks or formatting outside this structure:
{ "pred": "Answer to Q1, followed by a brief explanation",
"source": "Answer to Q2",
"desire": "Answer to Q3" }

Figure 9: Prompt for the **implication-oriented approach (I)** for detecting misleading news creator intent.

Consistency-Oriented Approach (C)

Your task is to determine whether a news piece has been intentionally distorted to mislead readers. The news piece consists of an image with its caption: {caption}. You are also provided a trustworthy reference article related to the same event: {article}.
Follow the structured steps below and strictly adhere to the output format.
Important Note: Focus only on the content of the news piece. The use of AI tools to assist image creation does not, by itself, indicate distortion.

Answer the following questions:
Q1. Has the news piece been intentionally distorted by its creator to mislead readers, such that **there exists inconsistencies between the image and caption, or between the news piece and the reference article?**
- Respond with one of the following: [Yes, No]. Then provide a brief explanation.
Q2. What is the source of the manipulation—text or image?
- If answer to Q1 is "Yes", answer with one of the following: [Text, Image].
- If answer to Q1 is "No", answer: "NA".
Q3. What does the distortion suggest about the news publisher's intent?
- If answer to Q1 is "Yes" :
Primary Influence (Desire): Select up to three from the following list, ranked by relevance: [Political Polarization, Social Polarization, Cultural and Religious Polarization, Economic Misinformation, Public Health and Safety, Environmental and Scientific Polarization, Geopolitical and International Relations, Psychological and Emotional Manipulation].
- If answer to Q1 is "No", respond with "NA" for Q3.

Output Format: Respond in the following exact JSON format, preserving the structure, punctuation, and delimiters. Do not add any extra line breaks or formatting outside this structure:
{ "pred": "Answer to Q1, followed by a brief explanation",
"source": "Answer to Q2",
"desire": "Answer to Q3" }

Figure 10: Prompt for the **consistency-oriented approach (C)** for detecting misleading news creator intent.