

Mitigating GenAI-powered Evidence Pollution for Out-of-Context Multimodal Misinformation Detection

Zehong Yan¹, Peng Qi¹, Wynne Hsu¹ and Mong Li Lee¹

¹NUS Centre for Trusted Internet & Community, National University of Singapore
zyan@u.nus.edu, peng.qi@nus.edu.sg, {whsu, leeml}@comp.nus.edu.sg,

Abstract

While large generative artificial intelligence (GenAI) models have achieved significant success, they also raise growing concerns about online information security due to their potential misuse for generating deceptive content. Out-of-context (OOC) multimodal misinformation detection, which often retrieves Web evidence to identify the repurposing of images in false contexts, faces the issue of reasoning over GenAI-polluted evidence to derive accurate predictions. Existing works simulate GenAI-powered pollution at the claim level with stylistic rewriting to conceal linguistic cues, and ignore evidence-level pollution for such information-seeking applications. In this work, we investigate how polluted evidence affects the performance of existing OOC detectors, revealing a performance degradation of more than 9 percentage points. We propose two strategies, cross-modal evidence reranking and cross-modal claim-evidence reasoning, to address the challenges posed by polluted evidence. Extensive experiments on two benchmark datasets show that these strategies can effectively enhance the robustness of existing out-of-context detectors amidst polluted evidence.

1 Introduction

The rapid development of generative artificial intelligence (GenAI) technologies has led to a surge of synthetic data in the Web [Pan *et al.*, 2023b; Chen and Shu, 2024; Wu *et al.*, 2023]. According to Gartner’s prediction, *by 2025, generative AI will account for 10% of all data produced, up from less than 1% today*¹. While GenAI mitigates the problem of data scarcity to some extent [Babbar and Schölkopf, 2019; Kim *et al.*, 2023; Villalobos *et al.*, 2024], it also facilitates the spread of realistic-looking yet non-factual misinformation [Guo *et al.*, 2022; Zhang *et al.*, 2023b]. Specifically, large language models (LLMs) like GPT-4 [OpenAI, 2023] produce both deliberate disinformation and unintentional hallucinations [Pan *et al.*, 2023b]; the growing use of diffusion

¹<https://www.gartner.com/en/newsroom>



Figure 1: Example of how misinformation detectors are misled by *claim-level* versus *evidence-level* pollution posed by GenAI. Faces of individuals are obscured to reduce privacy risks and mitigate the effects of misinformation exposure.

models for visual manipulation exacerbates these safety issues [Ramesh *et al.*, 2022; Rombach *et al.*, 2022]. Therefore, it is urgent to develop robust methods for information-seeking applications to mitigate pollution in the era of GenAI.

Existing studies has predominantly examined the GenAI-posed threats at the *claim level* [Atanasova *et al.*, 2020; Russo *et al.*, 2023; Wu *et al.*, 2024; Yerukola *et al.*, 2023]. To bypass detectors that rely upon superficial features such as language style for detection [Guo *et al.*, 2022], nefarious users typically transform sensational language into a neutral, formal style [Wu *et al.*, 2024]. For example, Figure 1(a) illustrates the scenario where a sensational claim has been rewritten in the style of the New York Times to elude detection.

On the other hand, *evidence-level* threats primarily target information-seeking systems that retrieve related evidence for inference (such as question answering [Pan *et al.*, 2023a; Pan *et al.*, 2023b] and fact-checking systems [Du *et al.*, 2022; Abdelnabi and Fritz, 2023]), by contaminating the evidence corpus with false information. As shown in Figure 1(b), malicious users exploit GenAI technologies to generate texts and images that support the misinformation about Taylor Swift’s



Figure 2: An illustrated example of claim-conditioned **generated evidence**, accompanied by **clean evidence** retrieved from the Web.

pregnancy, leading to incorrect predictions by the detectors. Existing works on evidence-level threats have focused on textual pollution within fixed, highly structured evidence corpora like Wikipedia pages. However, this narrow focus results in a considerable gap for misinformation detectors in the real-world where evidence retrieved from the web are typically unstructured, noisy and polluted.

Out-of-context (OOC) misinformation, where an authentic image is paired with false narratives to create misleading news, is one of the easiest and most effective ways to mislead audiences and has garnered increasing attention [Luo *et al.*, 2021; Abdelnabi *et al.*, 2022]. To combat OOC misinformation, [Zhang *et al.*, 2023c; Papadopoulos *et al.*, 2023b; Yuan *et al.*, 2023; Qi *et al.*, 2024] retrieve related news from web searches for each modality as a supplement to measure the cross-modal inconsistency. These works assume that the retrieved evidence contains only factual information, making the detectors vulnerable to data pollution caused by GenAI, an issue that remains underexplored.

In this work, we explore how GenAI models contribute to the pollution of evidence affecting the performance of OOC detectors. Figure 2 shows an example of how diverse multimodal evidence that closely resembles the original claim can be generated using GPT-4 [OpenAI, 2023] and Stable Diffusion 2 [Rombach *et al.*, 2022]. The generated evidence is mixed with evidence retrieved from the web before feeding into an OOC misinformation detector. Preliminary experiments reveal that existing OOC detectors are susceptible to this type of pollution, with detection efficacy decreasing by more than 9 percentage points.

We propose two strategies to enhance the robustness of existing OOC detectors: cross-modal evidence reranking and cross-modal claim-evidence reasoning. Cross-modal reranking prioritizes the most contextually relevant retrieved textual evidence based on the claim image, as well as the most relevant retrieved visual evidence based on the claim caption. Cross-modal claim-evidence reasoning provides an additional layer of analysis by identifying inconsistencies between the claim image and the top-ranked textual evidence retrieved. Our main contributions are as follows:

- We construct a large diverse collection of multimodal evidence to simulate the challenges posed by GenAI-based pollution for OOC misinformation detectors.
- We propose cross-modal evidence reranking and cross-modal claim-evidence reasoning to significantly enhance the robustness of OOC detectors against evidence pollution.
- Extensive experiments reveal the susceptibility of OOC detectors in the presence of evidence pollution and the effectiveness of the proposed strategies to mitigate such threats.

2 Related Work

Out-of-Context Misinformation Detection. Early works in OOC misinformation detection [Jaiswal *et al.*, 2017; Luo *et al.*, 2021; Papadopoulos *et al.*, 2023a] focus on verifying claims by analyzing the consistency of the image-caption pairs. These methods employ knowledge-rich pre-trained models, such as VGG-19 [Simonyan and Zisserman, 2015], CLIP [Radford *et al.*, 2021] and VisualBERT [Li *et al.*, 2019] to assess consistency. However, they tend to miss complex misinformation [Guo *et al.*, 2022] as they focus solely on the content of claims without considering external information like metadata [Sabir *et al.*, 2018; Aneja *et al.*, 2021] and web search results [Müller-Budack *et al.*, 2020; Abdelnabi *et al.*, 2022].

For external evidence reasoning, [Abdelnabi *et al.*, 2022] first collects multimodal evidence from the Web and use a Consistency-Checking Network (CCN) to analyze the consistency between the claim and retrieved evidence. [Papadopoulos *et al.*, 2023b] introduces the RED-DOT model, which ranks and filters evidence based on similarity scores to determine its relevance to the claim before using them for verification. [Yuan *et al.*, 2023] extends this approach by employing stance extraction networks to analyze whether the evidence supports or refutes the claim.

To improve the explainability of the veracity prediction, [Zhang *et al.*, 2023a] integrates multi-clue feature extraction, multi-level reasoning, and a decoder into a unified framework to explain the reasoning behind predictions. [Qi *et al.*, 2024] introduces SNIFFER, an explainable multimodal large language model that uses a two-stage instruction tuning process and three-stage reasoning framework. Despite these advancements, these works assume the factual integrity of retrieved evidence, which might not hold in real-world scenarios where evidence can be tainted with misleading or fabricated content.

Fact Checking with Polluted Evidence While substantial progress has been made in developing automated fact checking systems [Thorne and VLachos, 2021; Chakraborty *et al.*, 2023; Yao *et al.*, 2023] that verify claims based on reference knowledge bases, these systems suffer a marked decrease in performance when faced with compromised evidence. [Du *et al.*, 2022] utilizes language models to generate coherent yet false evidence which is then inserted into the evidence base. Building on this, [Abdelnabi and Fritz, 2023] proposes a taxonomy of pollution strategies targeting evidence, including planting and camouflaging, which expose the susceptibility of current fact-checking systems to manipulation. While these studies provide insights into evidence pollution, they focus on textual pollution in a controlled and highly structured ev-

idence source, such as Wikipedia. Our work considers more complex and realistic scenarios posed by GenAI, examining how such technologies affect fact-checking across a diverse range of evidence sources in an open-domain setting.

3 Methodology

In this section, we first simulate the scenarios where GenAI technologies are used to create realistic multimodal evidence pollution. Then we introduce two strategies, namely cross-modal reranking and cross-modal claim-evidence reasoning, to improve the robustness of OOC detectors against pollution.

3.1 Base OOC Detector

Figure 3 gives an overview of a typical framework of OOC misinformation detector. Given a claim comprising of an image I^q and a caption T^q , we first retrieve visual and textual evidence from the web using Google Vision and Google Custom Search. The claim and the retrieved evidence undergo a framework comprising of three key modules: visual, textual and image-caption consistency reasoning [Abdelnabi *et al.*, 2022]. The visual reasoning module examines the relevance between the claim image I^q and the polluted image evidence $\{I^c, I^g\}$. The textual reasoning module assesses how well the query caption T^q corresponds with the polluted text evidence $\{T^c, T^g\}$. Beyond these individual assessments, the consistency reasoning module checks the consistency between the claim image and the caption. The outputs from these reasoning modules are combined through a fusion module and then passed to a classifier to determine the veracity.

3.2 Evidence Pollution with GenAI

Polluted evidence poses significant challenges for both visual and textual reasoning modules, as they are susceptible to distractions from noisy or conflicting information, leading to inaccurate predictions. Unlike previous works [Abdelnabi *et al.*, 2022; Zhang *et al.*, 2023a; Papadopoulos *et al.*, 2023b; Qi *et al.*, 2024] that assume a clean evidence corpus, we consider the scenario where the evidence on the Web is polluted with highly similar yet potentially false information, thus challenging the robustness of evidence-based detectors.

For textual evidence pollution, we utilize LLMs to obtain realistic textual evidence for pollution at scale. Specifically, we employ GPT-4 [OpenAI, 2023] in a zero-shot manner and prompt it with two types of instructions motivated by real-world scenarios where noisy and conflicting information is prevalent, especially on social media platforms. The first type of instruction is used to generate textual evidence related to the entity mentioned in the caption: “*Write a short text about the main entity mentioned in the caption. Caption: <INPUT>*”. The second type of instruction generates textual evidence that either supports or refutes the claim caption: “*Write a piece of evidence to support or refute the given caption. Caption: <INPUT>*”. Since LLMs are prone to hallucinate [Cao *et al.*, 2022; Ji *et al.*, 2023], the generated text may contain inaccuracies.

Visual evidence also exhibits significant diversity across various domains, particularly in news, where different outlets may display different images of the same event [Chakraborty

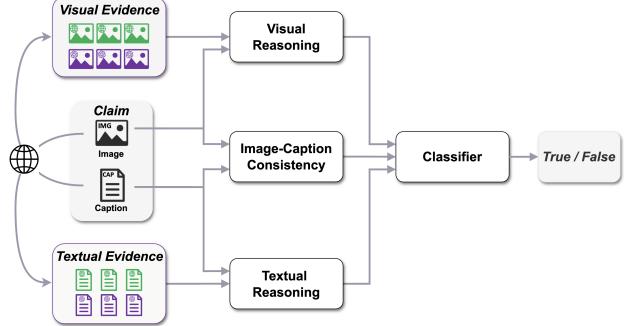


Figure 3: Overview of a typical OOC detection framework.

et al., 2023]. To simulate such diversity in real-world visual information, we employ the entity-preserving capabilities of Depth-Conditional Stable Diffusion [Rombach *et al.*, 2022] to generate visual evidence with varied camera angles and scene compositions, thereby providing a more challenging visual context for evaluating multimodal claims.

Recall the multimodal claim in Figure 2. The generated visual evidence shows variations of the same individual in the image, enriched with contextual details, visual modifications, and different backgrounds. With the claim caption, LLM generates the text based on the main entity, where the description of “British television channel” is factual. However, it also produces hallucinations, such as “BBC3 primarily focused on political debates”, which is incorrect, as BBC3 targets a younger audience and does not specifically focus on political content. Additionally, the generated support and refute textual evidence tends to extend beyond the context of the caption and produce nonfactual statements like “BBC3 won the BAFTA, not the RTS award”.

3.3 Proposed Strategies

OOC detectors assess the information authenticity and the consistency between text and associated images. However, the sophistication of LLMs introduces a new layer of complexity as it generates convincing polluted evidence that is not easily detected as LLM-generated content [Chen and Shu, 2024; Wu *et al.*, 2023; Xiang *et al.*, 2024]. We demonstrate this by evaluating the Vicuna-13B model, an open-source detector, on a dataset comprising of 10,000 pieces of textual evidence, evenly split between human-written and LLM-generated texts. The model achieves only a 41.3% accuracy in identifying LLM-generated content. This motivates us to develop two strategies, cross-modal evidence reranking and cross-modal claim-evidence reasoning, to enhance the robustness of OOC detectors (see Figure 4).

Cross-modal Evidence Reranking. This strategy addresses the issue of OOC detectors inadvertently focus on polluted evidence by giving priority to evidence that best aligns with the claim. Inspired by [Yao *et al.*, 2023], we use CLIP to identify the most contextually relevant textual evidence from a corpus that may contain polluted information, based on the claim image. Similarly, this method is employed to determine the most relevant visual evidence based on the claim caption. Algorithm 1 gives the details. Specifically, we utilize CLIP embeddings to compute cross-modal similarity scores and obtain the re-ranked lists of visual and textual evidence. The

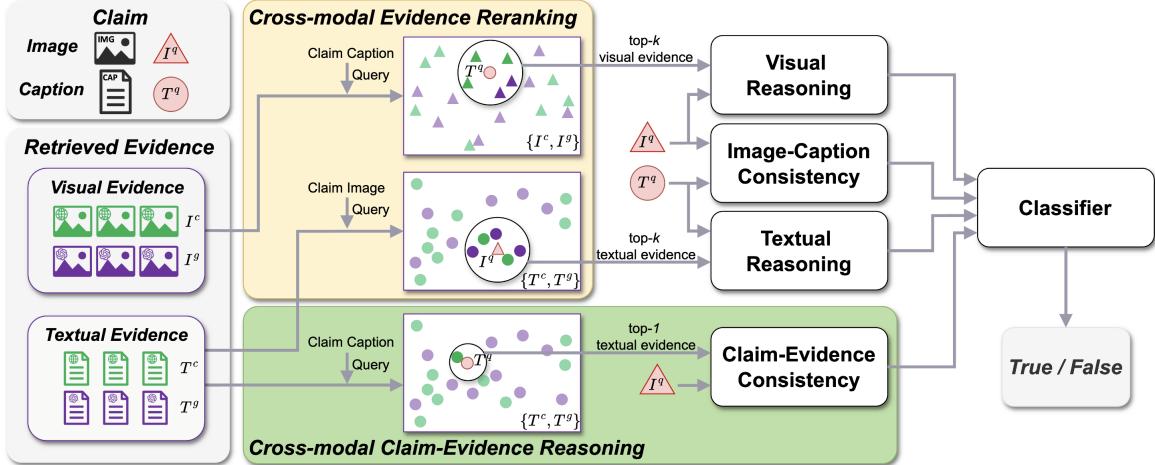


Figure 4: OOC misinformation detection framework in the presence of polluted evidence with proposed cross-modal reranking and cross-modal claim-evidence reasoning strategies.

Algorithm 1 Cross-modal Evidence Reranking

Input: claim $\langle I^q, T^q \rangle$, sets of retrieved textual evidence $\mathcal{T} = \{T^c, T^g\}$ and visual evidence $\mathcal{V} = \{I^c, I^g\}$
Output: sorted textual and visual evidences

```

1: initialize  $S_1 \leftarrow []$ ,  $S_2 \leftarrow []$ 
2: for  $T \in \mathcal{T}$  do
3:   compute cross-modal similarity score
4:    $s \leftarrow \text{cos}(\text{CLIP}(I^q), \text{CLIP}(T))$ 
5:    $S_1.\text{insert}(s)$ 
6: end for
7: for  $V \in \mathcal{V}$  do
8:   compute cross-modal similarity score
9:    $s \leftarrow \text{cos}(\text{CLIP}(T^q), \text{CLIP}(V))$ 
10:   $S_2.\text{insert}(s)$ 
11: end for
12: return  $\text{argsort}(S_1)$ ,  $\text{argsort}(S_2)$  in descending order

```

Algorithm 2 Cross-modal Claim-Evidence Reasoning

Input: claim $\langle I^q, T^q \rangle$, set of retrieved textual evidence \mathcal{T} , claim-evidence consistency reasoning module \mathcal{M}
Output: reasoning-representation

```

1: Initialize  $S \leftarrow []$ .
2: for  $T \in \mathcal{T}$  do
3:   compute intra-modal similarity score
4:    $s \leftarrow \text{cos}(\text{CLIP}(T^q), \text{CLIP}(T))$ 
5:    $S.\text{insert}(s)$ 
6: end for
7: Return  $\mathcal{M}(\mathcal{T}[\text{argmax}(S)], I^q)$ 

```

top- k visual and textual evidence are then passed to the visual reasoning module and textual reasoning module respectively.

Cross-modal Claim-Evidence Reasoning. Cross-modal claim-evidence reasoning *goes beyond traditional caption-image consistency check, which often misses critical contextual details* provided by external evidence. For example, a false caption may correctly describe the visible elements in an image but misrepresent its context, such as attributing a

Dataset	NewsCLIPpings			VERITE
	Train	Validation	Test	Test
Claim	71,072	7,024	7,264	662
Evidence				
Clean Text	689,995	58,388	60,848	1,261
Generated Text	903,067	82,112	67,016	2,002
Clean Image	650,738	64,562	66,772	8,309
Generated Image	655,848	65,082	67,092	8,389

Table 1: Dataset statistics.

news event to the wrong location or time. These discrepancies can only be verified using external information that is most pertinent to the main entity in the caption. As such, we use the most relevant textual evidence related to the caption for a consistency check with the claim image, ensuring the model’s robustness even when confronted with polluted evidence. Algorithm 2 gives the details.

The two proposed strategies can be utilized in a plug-and-play manner, allowing for easier integration into real world applications, without the need for re-training. Further, these strategies are adaptable to various types of pollutions with the emphasis on enhancing semantic-level reasoning rather than the feature distribution of a specific pollution model.

4 Performance Study

4.1 Experimental Setup

Datasets. We use two datasets in our experiments:

- **NewsCLIPpings** [Luo *et al.*, 2021] is the largest synthetic benchmark for OOC misinformation detection. It synthesizes out-of-context samples by replacing the images in the original image-caption pairs with retrieved images that are semantically related but belong to different news events. [Abdelnabi *et al.*, 2022] extends this dataset by supplementing both textual and visual evidence using Google Search APIs.

- **VERITE** [Papadopoulos *et al.*, 2024] is a real-world benchmark for evaluating multimodal misinformation detection. It consists of real and out-of-context pairs from fact-

Evidence	NewsCLIPpings			VERITE			
	Acc.	F1-True	F1-False	Acc.	F1-True	F1-False	
CCN	Clean	84.28	84.29	84.27	67.25	71.52	61.48
	Polluted Text	75.12 ($\downarrow 9.16$)	78.10 ($\downarrow 6.19$)	71.22 ($\downarrow 13.05$)	59.06 ($\downarrow 8.19$)	69.91 ($\downarrow 1.61$)	35.97 ($\downarrow 25.51$)
	Polluted Image	82.11 ($\downarrow 2.17$)	82.85 ($\downarrow 1.44$)	81.30 ($\downarrow 2.97$)	63.41 ($\downarrow 3.84$)	68.93 ($\downarrow 2.59$)	55.51 ($\downarrow 5.97$)
	Polluted Text + Image	71.78 ($\downarrow 12.50$)	76.48 ($\downarrow 7.81$)	64.72 ($\downarrow 19.55$)	55.92 ($\downarrow 11.33$)	68.65 ($\downarrow 2.87$)	25.81 ($\downarrow 35.67$)
RED-DOT	Clean	84.98	84.62	85.32	64.29	62.39	66.00
	Polluted Text	75.56 ($\downarrow 9.42$)	70.62 ($\downarrow 14.00$)	79.09 ($\downarrow 6.23$)	52.64 ($\downarrow 11.65$)	50.73 ($\downarrow 11.66$)	54.24 ($\downarrow 11.76$)
	Polluted Image	79.85 ($\downarrow 5.13$)	76.81 ($\downarrow 7.81$)	82.19 ($\downarrow 3.13$)	57.49 ($\downarrow 6.80$)	57.93 ($\downarrow 4.46$)	57.04 ($\downarrow 8.96$)
	Polluted Text + Image	73.75 ($\downarrow 11.23$)	67.19 ($\downarrow 17.43$)	78.12 ($\downarrow 7.20$)	48.75 ($\downarrow 15.54$)	48.65 ($\downarrow 13.74$)	48.85 ($\downarrow 17.15$)
SNIFFER	Clean	88.85	88.92	88.78	73.69	76.15	70.68
	Polluted Text	78.55 ($\downarrow 10.30$)	80.08 ($\downarrow 8.84$)	77.21 ($\downarrow 11.57$)	65.16 ($\downarrow 8.53$)	68.75 ($\downarrow 7.40$)	60.99 ($\downarrow 9.69$)
	Polluted Image	82.25 ($\downarrow 6.60$)	82.19 ($\downarrow 6.73$)	82.50 ($\downarrow 6.28$)	67.94 ($\downarrow 5.75$)	71.13 ($\downarrow 5.02$)	64.48 ($\downarrow 6.20$)
	Polluted Text + Image	76.42 ($\downarrow 12.43$)	77.47 ($\downarrow 11.45$)	75.31 ($\downarrow 13.47$)	59.41 ($\downarrow 14.28$)	64.71 ($\downarrow 11.44$)	53.04 ($\downarrow 17.64$)
GPT-4o	Clean	87.27	86.58	87.89	77.53	76.76	78.25
	Polluted Text	79.02 ($\downarrow 8.25$)	75.88 ($\downarrow 10.70$)	81.44 ($\downarrow 6.45$)	67.42 ($\downarrow 10.11$)	63.12 ($\downarrow 13.64$)	70.83 ($\downarrow 7.42$)
	Polluted Image	82.48 ($\downarrow 4.79$)	81.44 ($\downarrow 5.14$)	83.41 ($\downarrow 4.48$)	68.64 ($\downarrow 8.89$)	65.91 ($\downarrow 10.85$)	70.97 ($\downarrow 7.28$)
	Polluted Text + Image	77.72 ($\downarrow 9.55$)	74.69 ($\downarrow 11.89$)	80.11 ($\downarrow 7.78$)	64.29 ($\downarrow 13.24$)	56.29 ($\downarrow 20.47$)	69.81 ($\downarrow 8.44$)

Table 2: OOC detection performance (%) under evidence pollution of different modalities. The first row (Clean) refers to the original performance without any pollution introduced. The absolute change compared to the Clean setting is highlighted in red.

checking websites. We use the corresponding multimodal evidence from [Papadopoulos *et al.*, 2023b].

For each piece of textual evidence, we randomly apply one of the LLM instruction to create the corresponding polluted entity-based, supporting or refuting evidence. For each piece of visual evidence, we use Depth-conditioned Stable Diffusion to generate the corresponding images. These generated evidence are added to the original clean evidence corpus. Table 1 shows the statistics for the two datasets.

Baselines. We use the following OOC misinformation detectors in our experiments:

- **CCN** [Abdelnabi *et al.*, 2022]. This employs attention-based memory networks for visual and textual reasoning between the claim and evidence, and a fine-tuned CLIP component to check the claim image and caption consistency.

- **RED-DOT** [Papadopoulos *et al.*, 2023b]. This leverages the pre-trained CLIP as the backbone to extract visual and textual features. Transformer-based fusion module is used to facilitate interaction and reasoning among these features.

- **SNIFFER** [Qi *et al.*, 2024]. This is the state-of-the-art multimodal large language model designed for OOC misinformation detection. It employs a two-stage instruction tuning on InstructBLIP for the cross-modal consistency checks.

- **GPT-4o** [OpenAI, 2024]. This is currently one of the most powerful multimodal large language models. We utilize GPT-4o in a zero-shot manner with step-by-step instructions for OOC detection. Details are provided in Appendix.

4.2 Effect of Evidence Pollution on OOC Detectors

Table 2 shows the OOC detection performance across different evidence modalities. We observe that: 1) The combination of polluted text and image poses a significant threat to OOC detectors. Specifically, the accuracy of all detectors drop by more than 9 percentage points, revealing the vulnerabilities of existing OOC detectors against generated multi-

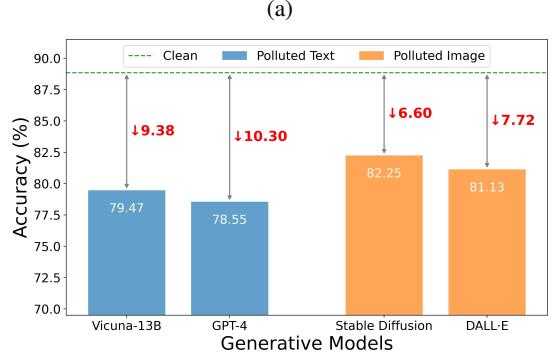
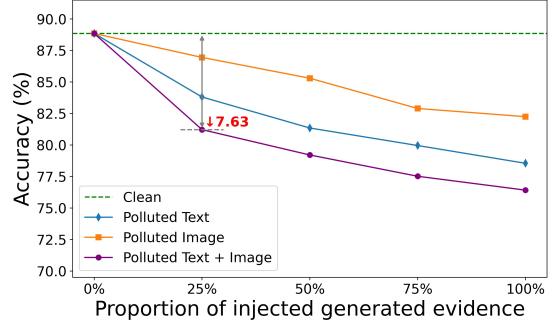


Figure 5: SNIFFER’s performance across varying proportion of polluted evidence and GenAI models on NewsCLIPpings.

modal pollution. 2) Textual pollution has a greater impact than visual pollution, indicating that existing OOC detectors are more dependent on textual information. This modality bias may stem from the fact that textual evidence often provides more semantics such as relationships between entities compared to images. 3) Detection of false claims in the presence of with polluted evidence proves to be more challeng-

Strategy	NewsCLIPpings			VERITE		
	Acc.	F1-True	F1-False	Acc.	F1-True	F1-False
CCN	None	71.78	76.48	64.72	55.92	68.65
	Cross-modal Reranking	79.70 ($\uparrow 7.92$)	79.88 ($\uparrow 3.40$)	79.51 ($\uparrow 14.79$)	61.67 ($\uparrow 5.75$)	65.08 ($\downarrow 3.57$)
	Cross-modal Reasoning	75.17 ($\uparrow 3.39$)	78.38 ($\uparrow 1.90$)	70.83 ($\uparrow 6.11$)	59.76 ($\uparrow 3.84$)	63.39 ($\downarrow 5.26$)
	Both	80.21 ($\uparrow 8.43$)	80.86 ($\uparrow 4.38$)	79.52 ($\uparrow 14.80$)	65.51 ($\uparrow 9.59$)	70.54 ($\uparrow 1.89$)
RED-DOT	None	73.75	67.19	78.12	48.75	48.65
	Cross-modal Reranking	82.92 ($\uparrow 9.17$)	81.33 ($\uparrow 14.14$)	84.26 ($\uparrow 6.14$)	62.54 ($\uparrow 13.79$)	60.98 ($\uparrow 12.33$)
	Cross-modal Reasoning	83.41 ($\uparrow 9.66$)	82.17 ($\uparrow 14.98$)	84.49 ($\uparrow 6.37$)	62.02 ($\uparrow 13.27$)	62.41 ($\uparrow 13.76$)
	Both	84.69 ($\uparrow 10.94$)	84.11 ($\uparrow 16.92$)	85.24 ($\uparrow 7.12$)	63.24 ($\uparrow 14.49$)	63.93 ($\uparrow 15.28$)
SNIFFER	None	76.42	77.47	75.31	59.41	64.71
	Cross-modal Reranking	87.68 ($\uparrow 11.26$)	87.74 ($\uparrow 10.27$)	87.62 ($\uparrow 12.31$)	71.78 ($\uparrow 12.37$)	74.77 ($\uparrow 10.06$)
	Cross-modal Reasoning	87.51 ($\uparrow 11.09$)	87.95 ($\uparrow 10.48$)	87.05 ($\uparrow 11.74$)	70.21 ($\uparrow 10.80$)	73.89 ($\uparrow 9.18$)
	Both	88.82 ($\uparrow 12.40$)	89.15 ($\uparrow 11.68$)	88.48 ($\uparrow 13.17$)	72.82 ($\uparrow 13.41$)	76.00 ($\uparrow 11.29$)
GPT-4o	None	77.72	74.69	80.11	64.29	56.29
	Cross-modal Reranking	87.07 ($\uparrow 9.35$)	85.82 ($\uparrow 11.13$)	88.11 ($\uparrow 8.00$)	73.17 ($\uparrow 8.88$)	72.20 ($\uparrow 15.91$)
	Cross-modal Reasoning	86.87 ($\uparrow 9.15$)	86.10 ($\uparrow 11.41$)	87.66 ($\uparrow 7.55$)	74.39 ($\uparrow 10.10$)	74.79 ($\uparrow 18.50$)
	Both	88.00 ($\uparrow 10.28$)	87.51 ($\uparrow 12.82$)	88.53 ($\uparrow 8.42$)	75.44 ($\uparrow 11.15$)	76.30 ($\uparrow 20.01$)

Table 3: OOC detection performance (%) with the proposed strategies under the evidence pollution. The first row (None) refers to the original performance under multimodal pollution. The absolute change to the original one is highlighted in blue.

ing than true claims. Specifically, CCN experiences a significant drop of 35.67 points in the F1 score for false claims on the VERITE dataset, highlighting the difficulties in reasoning with contradictory evidence.

Quantitative Analysis. Figure 5a shows the performance of SNIFFER when we vary the proportion of polluted evidence. We see that the accuracy of the model drops as the proportion of pollution increases. Even a small amount of pollution can significantly affect the model’s detection capabilities where introducing 25% of polluted evidence results in a decrease of 7.63 points. The impact of varying pollution ratios on different models such as CCN and different types of textual evidence pollution are given in the Appendix.

Generalization Analysis. Figure 5b shows the impact of pollution in textual and visual modalities under different generative models. Notably, for visual pollution, advanced models like DALL-E, which significantly improves image quality and resolution, further amplify the effects of visual pollution.

Human Evaluation. We conduct a human evaluation on ten randomly selected misinformation samples with polluted evidence. Twenty participants were asked to judge each piece of evidence’s authenticity and each claim’s veracity before and after reading the polluted evidence. The results show that (a) only 49.39% of the generated evidence was correctly identified as AI-generated; (b) 41.84% of the initially correct veracity judgments for misinformation samples were reversed to wrong predictions after reading the polluted evidence.

4.3 Effect of Proposed Strategies

Table 3 shows the performance of the various OOC misinformation detectors when we incorporate the proposed defense strategies. We see that: 1) The combination of both strategies yields the best results, increasing the overall accuracy to 88.82% (+12.40) and 75.44% (+11.15) for SNIFFER on the NewsCLIPpings and VERITE dataset respec-

tively. This indicates that the two strategies complement each other, enhancing the model’s robustness against multimodal pollution. Further, the strategies can be generalized to the real-world VERITE dataset. 2) Incorporating cross-modal evidence re-ranking significantly boosts performance. The overall accuracy of SNIFFER increases to 87.68%, marking an improvement of 11.26%, on the NewsCLIPpings dataset. This strategy also enhances the detection of true and false claims to 87.74% (+10.27) and 87.62% (+12.31), respectively. The results suggest that re-ranking evidence and focusing on the top relevant evidence greatly aids in reconciling discrepancies introduced by multimodal pollution. 3) Similar to cross-modal reranking, cross-modal claim-evidence reasoning module also shows substantial gains, particularly in the detection of true claims.

Table 4 further compares the performance of LLM-based detectors with three general approaches under evidence pollution. The extra detector approach involves adding an auxiliary classifier to filter out the generated evidence, the vigilant prompting approach introduces hints at the presence of false evidence in the prompt, and the reader ensemble approach combines multiple judgments based on different evidence by voting [Pan *et al.*, 2023b]. SNIFFER, equipped with our proposed solution, achieves the highest performance across two datasets, with significant improvements of 12.40% on NewsCLIPpings and 13.41% on VERITE, demonstrating its superiority in the presence of polluted evidence. Notably, our approaches can be easily integrated into existing OOC detection frameworks, whereas the prompting-based and voting-based approaches are restricted to LLM-based detectors.

4.4 Case Study

Figure 6 presents a case study under evidence pollution. Initially, in the the clean setting, the model correctly identifies that the image, depicting Tim Henman, is irrelevant to the po-

Strategy	NewsCLIPpings			VERITE			
	Acc.	F1-True	F1-False	Acc.	F1-True	F1-False	
SNIFFER	None	76.42	77.47	75.31	59.41	64.71	53.04
	Extra Detector	79.00 ($\uparrow 2.58$)	80.73 ($\uparrow 3.26$)	76.92 ($\uparrow 1.61$)	68.99 ($\uparrow 9.58$)	72.01 ($\uparrow 7.30$)	65.23 ($\uparrow 12.19$)
	Vigilant Prompting	79.49 ($\uparrow 3.07$)	80.93 ($\uparrow 3.46$)	77.84 ($\uparrow 2.53$)	69.51 ($\uparrow 10.10$)	72.18 ($\uparrow 7.47$)	66.28 ($\uparrow 13.24$)
	Reader Ensemble	68.51 ($\downarrow 7.91$)	70.70 ($\downarrow 6.77$)	65.94 ($\downarrow 9.37$)	64.81 ($\uparrow 5.40$)	63.54 ($\downarrow 1.17$)	65.99 ($\uparrow 12.95$)
	Ours	88.82 ($\uparrow 12.40$)	89.15 ($\uparrow 11.68$)	88.48 ($\uparrow 13.17$)	72.82 ($\uparrow 13.41$)	76.00 ($\uparrow 11.29$)	68.67 ($\uparrow 15.63$)
GPT-4o	None	77.72	74.69	80.11	64.29	56.29	69.81
	Extra Detector	81.69 ($\uparrow 3.97$)	79.20 ($\uparrow 4.51$)	83.85 ($\uparrow 3.74$)	72.13 ($\uparrow 7.84$)	70.15 ($\uparrow 13.86$)	74.10 ($\uparrow 4.29$)
	Vigilant Prompting	83.50 ($\uparrow 5.78$)	82.50 ($\uparrow 7.81$)	84.84 ($\uparrow 4.73$)	66.03 ($\uparrow 1.74$)	62.14 ($\uparrow 5.85$)	69.41 ($\downarrow 0.40$)
	Reader Ensemble	72.33 ($\downarrow 5.39$)	68.53 ($\downarrow 6.16$)	77.29 ($\downarrow 2.82$)	64.98 ($\uparrow 0.69$)	62.20 ($\uparrow 5.91$)	68.80 ($\downarrow 1.01$)
	Ours	88.00 ($\uparrow 10.28$)	87.51 ($\uparrow 12.82$)	88.53 ($\uparrow 8.42$)	75.44 ($\uparrow 11.15$)	76.30 ($\uparrow 20.01$)	74.50 ($\uparrow 4.69$)

Table 4: Performance comparison of different strategies. The first row (None) refers to the original performance under multimodal pollution.

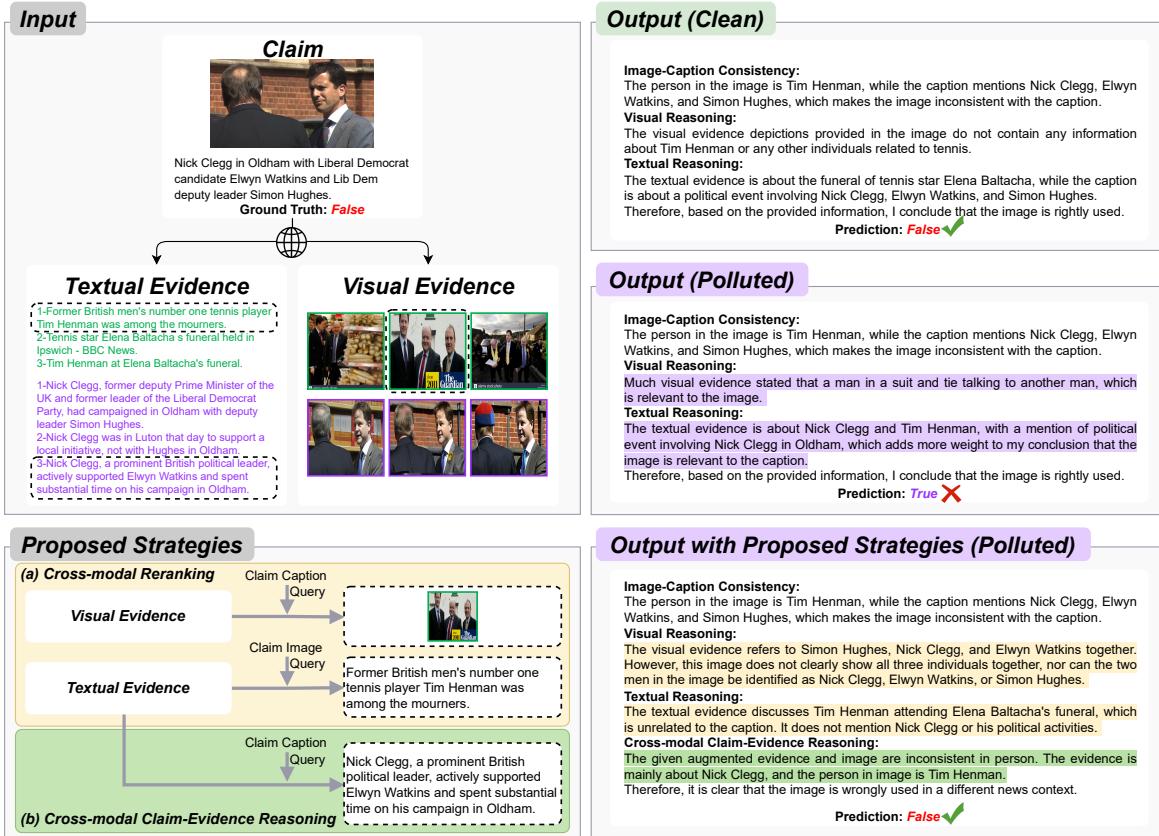


Figure 6: Case study of SNIFFER’s justification outputs under clean and polluted settings. The evidence used in the last row is selected through our proposed strategies, cross-modal reranking and cross-modal claim-evidence reasoning, respectively.

litical figures mentioned in the caption (Nick Clegg, Elwyn Watkins, Simon Hughes). However, after exposure to pollution, SNIFFER erroneously asserts that the image is relevant, citing visual evidence of a man in a suit speaking to another man and textual evidence mentioning both Nick Clegg and Tim Henman. Additionally, it also incorrectly emphasizes a weak connection between the image and textual evidence, leading to an incorrect prediction.

Incorporating the two proposed strategies enables SNIFFER to recognize the inconsistency between the image and the caption, and confirm that the image indeed features Tim Henman which does not match the caption’s context. This leads to the correction prediction.

5 Conclusion

In this paper, we reveal the critical vulnerabilities of existing out-of-context multimodal misinformation detectors when confronted with evidence polluted by large generative models. To counteract this, we introduced and evaluated two innovative strategies: cross-modal evidence reranking and cross-modal claim-evidence reasoning. Our comprehensive experiments across multiple detectors and two benchmarks have shown that these strategies significantly enhance the detectors’ resilience against multimodal evidence pollution. We believe this study paves the way for further research into robust misinformation detection in the era of GenAI.

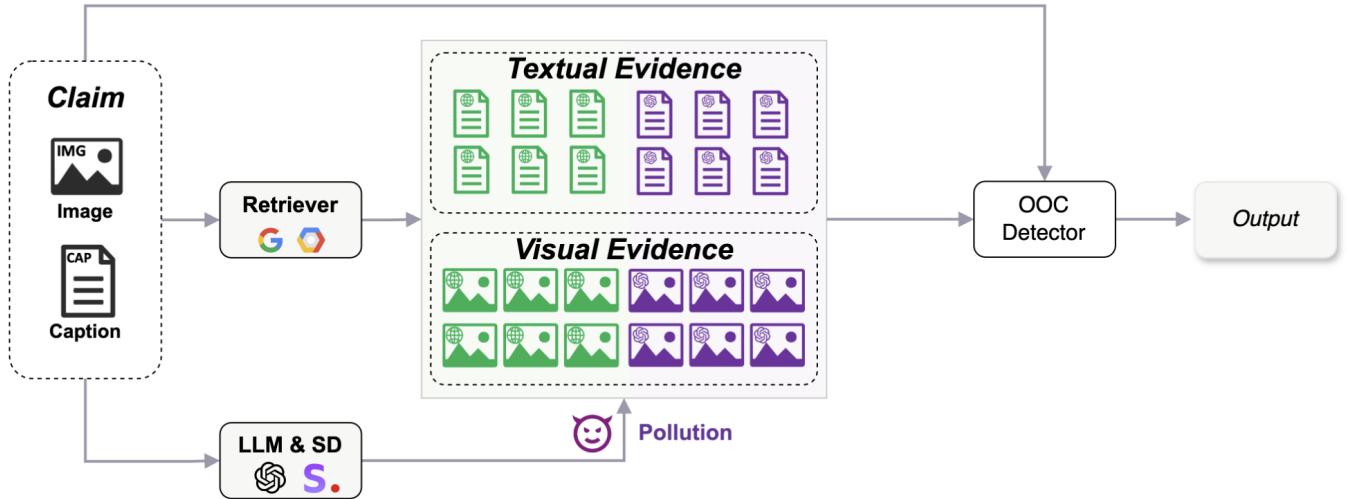


Figure 7: An overview of out-of-context detection system under evidence pollution. A claim image and its caption are processed by retrievers to gather textual and visual evidence from the web (green). Conditioned on the claim, we employ large language models (LLMs) and stable diffusion (SD) models to generate pollution, which is then injected into original evidence corpus (purple). Finally, the claim, along with the textual and visual evidence, is fed into an OOC detector to determine its veracity.

A Task Formulation

Figure 7 provides an overview of an out-of-context (OOC) detection system in the era of GenAI. The input claim is processed through a retriever module to gather relevant textual and visual evidence from the Web. LLMs and Stable Diffusion models play a role in generating and simulating pollution. The claim and evidence are then passed to the OOC detector, which evaluates the claim’s veracity. Here, we further summarize the task components and evidence pollution posed by large generative models as follows:

Model.

- An out-of-context detection model \mathcal{M}
- A retrieval model \mathcal{R}
- A generative model \mathcal{G}

Claim.

- A claim image-caption pair $\{I^q, T^q\}$

Evidence.

• Clean evidence \mathcal{E}^c :

- Text evidence: A list of texts retrieved by $\mathcal{R}(T^c|I^q)$:
 $T^c = [T_1^c, \dots, T_M^c]$
- Image evidence: A list of images retrieved by $\mathcal{R}(I^c|T^q)$:
 $I^c = [I_1^c, \dots, I_K^c]$

• Generated evidence \mathcal{E}^g :

- Text evidence: A list of texts generated by $\mathcal{G}(T^g|T^q)$:
 $T^g = [T_1^g, \dots, T_M^g]$
- Image evidence: A list of images generated by $\mathcal{G}(I^g|I^q)$:
 $I^g = [I_1^g, \dots, I_K^g]$

Task.

- **Clean:** Leverage \mathcal{M} to classify the claim as *true* or *false* using \mathcal{E}^c
- **Polluted:** Leverage \mathcal{M} to classify the claim as *true* or *false* using polluted evidence $\{\mathcal{E}^c, \mathcal{E}^g\}$

B Implementation Details

We use CCN [Abdelnabi *et al.*, 2022] and SNIFFER’s [Qi *et al.*, 2024] public model checkpoints fine-tuned on the NewsCLIPpings training set. We use the InstructBLIP [Dai *et al.*, 2023] as our captioner for visual reasoning path without fine-tuning. We leverage the CLIP (ViT-L/14) as the cross-modal reranking module and select the top-1 sentence and top-5 images for textual and visual evidence. For augmented reasoning, we reuse the original CLIP component from CCN and internal checking from SNIFFER. All models are trained and evaluated on 8 Nvidia H100 (80G) GPUs. We generate textual pollution with GPT-4 (gpt-4) [OpenAI, 2023], which is configured with a temperature of 1.2, a maximum token length of 64, and a top-P setting of 0.95. We employ the variant of Stable Diffusion v2 models (stabilityai/stable-diffusion-2-depth) to generate visual pollution. We report accuracy over all samples, and F1 score for the true and false samples, respectively.

C Visualization of Similarity Distribution

To assess the similarity between the generated evidence and the original clean evidence, we conducted an analysis of similarity for both textual and visual evidence. We then examined the distribution between the clean and generated evidence. For clearer visualization, We randomly select a evidence subset of 500 claims from the test set. As shown in Figure 8a and Figure 8b, the distribution is centered around zero, indicating that the generated evidence closely resembles the original clean evidence. Additionally, we applied t-SNE to visualize the latent spaces. The results prove that our approach is able to generate evidence that not only closely mirrors the original clean evidence but also exhibits greater similarity to the input claim, thereby effectively contaminating the clean evidence while preserving high semantic similarity. This demonstrates the effectiveness of our approach in generating evidence that can blend seamlessly into the original clean evidence set.

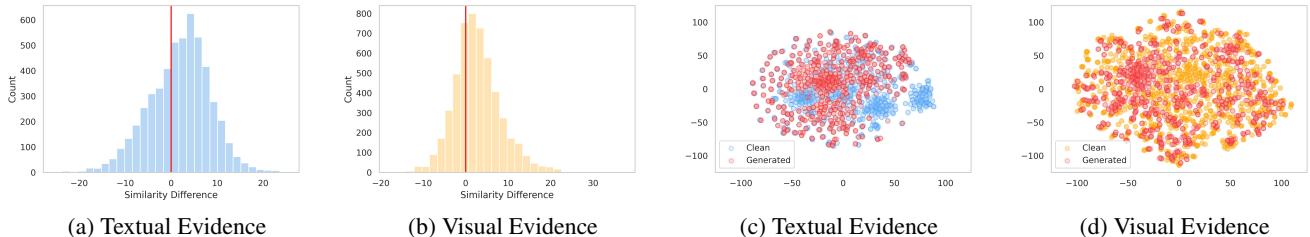


Figure 8: (a): Distribution of differences in CLIP scores between input image and textual evidence. The X-axis represents the difference calculated as the CLIP score of the image-evidence (generated) minus the CLIP score of the image-evidence (clean), while the Y-axis shows the count of these occurrences. (b): Distribution of differences in CLIP scores between input caption and visual evidence. (c)-(d): t-SNE visualization of latent space of clean and generated evidence.

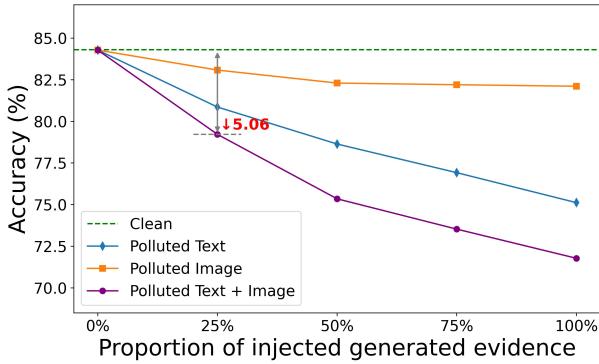


Figure 9: CCN’s performance across varying proportion of polluted evidence on NewsCLIPpings dataset.

D Performance Analysis of Varying Proportion of Polluted Evidence

In addition to SNIFFER, we present the results of the CCN model [Abdelnabi *et al.*, 2022] under different proportions of polluted evidence, as illustrated in Figure 9. The accuracy of CCN demonstrates a marked decline as the level of evidence pollution increases. Furthermore, the results highlight CCN’s heavy reliance on the text modality for misinformation identification, making it particularly vulnerable to pollution introduced by LLMs.

E Comparative Analysis of Types of Textual Pollution

In this section, we study the effects of different ways when generating textual evidence pollution. Figure 10 shows the impact of different types of textual evidence pollution on the performance of CCN and SNIFFER. We see that CCN is more affected by the generated entity based text, while SNIFFER shows the largest decline in the presence of generated supporting and refuting evidence.

F Performance of Cross-modal Reranking

Table 5 shows the percentage of clean evidence within the top-k results after applying the cross-modal re-ranking. By leveraging the capabilities of pre-trained encoder CLIP to facilitate cross-modal semantic matching between textual and visual modalities, we have effectively increase the probability of utilizing clean evidence for misinformation detection.

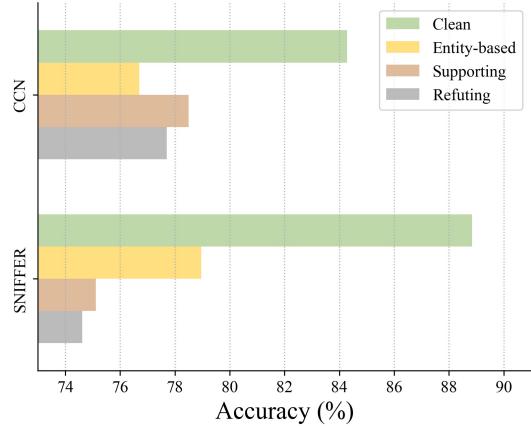


Figure 10: OOC detection performance (%) comparison among different types of textual pollution.

G Comparison of Related Works

Table 6 presents a comparison of related work , each evaluated across different criteria: Textual Modality, Visual Modality, Use of Large Language Models, Targeted Evidence Source, and Stance Diversity. Our work distinctly integrates all these aspects in an open-domain OOC misinformation detection task, which requires reasoning over evidence retrieved from the Web with various sources. We simulate a more realistic pollution posed by the GenAI, calling for an early evaluation. Furthermore, unlike previous efforts that focus solely on textual pollution, our proposed pollution pipeline is the first work to introduce multimodal pollution.

H Detecting Polluted Evidence

Along with the rapid development of LLMs, the issue of data pollution has become increasingly important and observed in the research community [Pan *et al.*, 2023b; Xiang *et al.*, 2024]. There has been increasing attention on detecting LLM-generated data in recent studies [Chen and Shu, 2024]. Following [Chen and Shu, 2024], we adopt the prompt for detection. We randomly select a set of 10,000 pieces of textual evidence samples as the test set, equally divided into human-written clean samples and LLM-generated samples, and use open-source Vicuna-13B model to detect LLM-generated content. The results show that LLM detector can hardly identify LLM-generated text with an overall accuracy

Reranker	Evidence	Query	R@1	R@3	R@5	R@10
CLIP (ViT-B/32)	Polluted Text	Image	70.56%	64.14%	59.98%	55.57%
CLIP (ViT-B/32)	Polluted Image	Caption	64.88%	61.05%	57.00%	49.74%
CLIP (ViT-L/14)	Polluted Text	Image	72.78%	66.73%	62.67%	57.89%
CLIP (ViT-L/14)	Polluted Image	Caption	76.38%	72.23%	67.83%	56.73%

Table 5: Performance evaluation of CLIP-based re-rankers in NewsCLIPpings dataset. The retrieval effectiveness is measured at multiple cutoff points. R@k indicates the percentage of clean evidence is found within the top-k retrieved results.

Targeted Task	Textual Modality	Visual Modality	Use LLM	Targeted Evidence	Stance Diversity
News Veracity Classification [Du <i>et al.</i> , 2022]	✓	✗	✗	Wikipedia, S2ORC, Reddit	✗
News Veracity Classification [Abdelnabi and Fritz, 2023]	✓	✗	✗	Wikipedia	Supporting
Question Answering [Pan <i>et al.</i> , 2023a]	✓	✗	✗	Wikipedia	✗
Question Answering [Pan <i>et al.</i> , 2023b]	✓	✗	✓	Wikipedia, WMT News	Supporting
OOC Misinformation Detection (<i>Ours</i>)	✓	✓	✓	Web	Supporting, Refuting

Table 6: Comparison of related work on evidence pollution.

```
# system message
Task description: some rumormongers use images from other events as illustrations of
the current news event to make multimodal misinformation. Given a news caption and
a news image, you are responsible for judging whether the given image is wrongly
used in a different news context. You will be presented with a caption, an image,
visual evidence, and textual evidence. You should use the following step-by-step
instructions to derive your judgment:

Step 1 - Make a decision based on inconsistency between the caption and the image.
Step 2 - Make a judgement according to the inconsistency between the image and the
visual evidence.
Step 3 - Make a judgement according to the inconsistency between the caption and the
textual evidence.
Step 4 - According to the previous steps, you will first think out loud about your
eventual conclusion, enumerating reasons why the image does or does not match the
given caption. After thinking out loud, you should output either 'Real' or 'Fake'
depending on whether you think the image is faithful to the caption.

# query
<image>
Caption: <caption>
Visual Evidence: <visual evidence>
Textual Evidence: <textual evidence>
Your judgement:
```

Figure 11: Prompt used to ask GPT-4o to detect out-of-context misinformation.

of just 41.3%. We found that LLMs focus on grammar, sentence structure, and specific contextual details such as events and people, as well as vocabulary usage. Such traditional linguistic scopes are not enough because advanced large generative technologies, like GPT-4, are exceptionally proficient at mimicking human-like text, underscoring the need for more sophisticated approaches.

I Prompt to Detect the OOC Misinformation

Figure 11 illustrates the prompt utilized for asking GPT-4o to identify inconsistencies between the claim image and its caption. The preliminary step is to retrieve multimodal evidence. For each claim, we retrieve textual and visual evidence (converted to text via image captioning) separately and then pass them to GPT-4o to process.

References

- [Abdelnabi and Fritz, 2023] Sahar Abdelnabi and Mario Fritz. Fact-Saboteurs: A taxonomy of evidence manipulation attacks against fact-verification systems. In Joseph A. Calandrino and Carmela Troncoso, editors, *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, pages 6719–6736, 2023.
- [Abdelnabi *et al.*, 2022] Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14940–14949, 2022.
- [Aneja *et al.*, 2021] Shivangi Aneja, Christoph Bregler, and Matthias Nießner. Catching out-of-context misinformation.

- tion with self-supervised learning. *CoRR*, abs/2101.06278, 2021.
- [Atanasova *et al.*, 2020] Pepa Atanasova, Dustin Wright, and Isabelle Augenstein. Generating label cohesive and well-formed adversarial claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3168–3177, November 2020.
- [Babbar and Schölkopf, 2019] Rohit Babbar and Bernhard Schölkopf. Data scarcity, robustness and extreme multi-label classification. *Machine Learning*, 108(8):1329–1351, 2019.
- [Cao *et al.*, 2022] Meng Cao, Yue Dong, and Jackie Cheung. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, May 2022.
- [Chakraborty *et al.*, 2023] Megha Chakraborty, Khushbu Pahwa, Anku Rani, Shreyas Chatterjee, Dwip Dalal, Harshit Dave, Ritvik G, Preethi Gurumurthy, Adarsh Mahor, Samahriti Mukherjee, Aditya Pakala, Ishan Paul, Janvita Reddy, Arghya Sarkar, Kinjal Sensharma, Aman Chadha, Amit Sheth, and Amitava Das. FACTIFY3M: A benchmark for multimodal fact verification with explainability through 5W question-answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15282–15322, December 2023.
- [Chen and Shu, 2024] Canyu Chen and Kai Shu. Can llm-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.
- [Dai *et al.*, 2023] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instruct-BLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023*, 2023.
- [Du *et al.*, 2022] Yibing Du, Antoine Bosselut, and Christopher D Manning. Synthetic disinformation attacks on automated fact verification systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10581–10589, 2022.
- [Guo *et al.*, 2022] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022.
- [Jaiswal *et al.*, 2017] Ayush Jaiswal, Ekraam Sabir, Wael AbdAlmageed, and Premkumar Natarajan. Multimedia semantic integrity assessment using joint embedding of images and text. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1465–1471, 2017.
- [Ji *et al.*, 2023] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12):248:1–248:38, 2023.
- [Kim *et al.*, 2023] Sungdong Kim, Sanghwan Bae, Jamin Shin, Soyoung Kang, Donghyun Kwak, Kang Yoo, and Minjoon Seo. Aligning large language models through synthetic feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13677–13700, December 2023.
- [Li *et al.*, 2019] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557, 2019.
- [Luo *et al.*, 2021] Grace Luo, Trevor Darrell, and Anna Rohrbach. 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, November 2021.
- [Müller-Budack *et al.*, 2020] Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, and Ralph Ewerth. Multimodal analytics for real-world news using measures of cross-modal entity consistency. In *Proceedings of the 2020 international conference on multimedia retrieval*, pages 16–25, 2020.
- [OpenAI, 2023] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.
- [OpenAI, 2024] OpenAI. Hello GPT-4o, 2024. Accessed: 2024-06-07.
- [Pan *et al.*, 2023a] Liangming Pan, Wenhua Chen, Min-Yen Kan, and William Yang Wang. Attacking open-domain question answering by injecting misinformation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 525–539, November 2023.
- [Pan *et al.*, 2023b] Yikang Pan, Liangming Pan, Wenhua Chen, Preslav Nakov, Min-Yen Kan, and William Wang. On the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403, December 2023.
- [Papadopoulos *et al.*, 2023a] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis Petrantonakis. Synthetic misinformers: Generating and combating multimodal misinformation. In *Proceedings of the 2nd ACM International Workshop on Multimedia AI against Disinformation*, pages 36–44, 2023.
- [Papadopoulos *et al.*, 2023b] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C. Petrantonakis. RED-DOT: multimodal fact-checking via relevant evidence detection. *CoRR*, abs/2311.09939, 2023.
- [Papadopoulos *et al.*, 2024] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and

- Panagiotis C Petrantonakis. VERITE: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. *International Journal of Multimedia Information Retrieval*, 13(1):4, 2024.
- [Qi *et al.*, 2024] Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. 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, 2024.
- [Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [Ramesh *et al.*, 2022] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022.
- [Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [Russo *et al.*, 2023] Daniel Russo, Shane Kaszefski-Yaschuk, Jacopo Staiano, and Marco Guerini. Countering misinformation via emotional response generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11476–11492, December 2023.
- [Sabir *et al.*, 2018] Ekraam Sabir, Wael AbdAlmageed, Yue Wu, and Prem Natarajan. Deep multimodal image-repurposing detection. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1337–1345, 2018.
- [Simonyan and Zisserman, 2015] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [Thorne and Vlachos, 2021] James Thorne and Andreas Vlachos. Evidence-based factual error correction. 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 3298–3309, August 2021.
- [Villalobos *et al.*, 2024] Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbahn. Position: Will we run out of data? limits of LLM scaling based on human-generated data. In *Forty-first International Conference on Machine Learning, ICML 2024*, 2024.
- [Wu *et al.*, 2023] Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F. Wong, and Lidia S. Chao. A survey on llm-generated text detection: Necessity, methods, and future directions. *CoRR*, abs/2310.14724, 2023.
- [Wu *et al.*, 2024] Jiaying Wu, Jiafeng Guo, and Bryan Hooi. 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*, pages 3367–3378, 2024.
- [Xiang *et al.*, 2024] Chong Xiang, Tong Wu, Zexuan Zhong, David A. Wagner, Danqi Chen, and Prateek Mittal. Certifiably robust RAG against retrieval corruption. *CoRR*, abs/2405.15556, 2024.
- [Yao *et al.*, 2023] Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 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, SIGIR ’23*, page 2733–2743, New York, NY, USA, 2023.
- [Yerukola *et al.*, 2023] Akhila Yerukola, Xuhui Zhou, Elizabeth Clark, and Maarten Sap. Don’t take this out of context!: On the need for contextual models and evaluations for stylistic rewriting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11419–11444, December 2023.
- [Yuan *et al.*, 2023] Xin Yuan, Jie Guo, Weidong Qiu, Zheng Huang, and Shujun Li. 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, December 2023.
- [Zhang *et al.*, 2023a] Fanrui Zhang, Jiawei Liu, Qiang Zhang, Esther Sun, Jingyi Xie, and Zheng-Jun Zha. ECENet: Explainable and context-enhanced network for multi-modal fact verification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1231–1240, 2023.
- [Zhang *et al.*, 2023b] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. Instruction tuning for large language models: A survey. *CoRR*, abs/2308.10792, 2023.
- [Zhang *et al.*, 2023c] Yizhou Zhang, Loc Trinh, Defu Cao, Zijun Cui, and Yan Liu. Detecting out-of-context multimodal misinformation with interpretable neural-symbolic model. *CoRR*, abs/2304.07633, 2023.