HiEAG: Evidence-Augmented Generation for Out-of-Context Misinformation Detection

Junjie Wu[1] Yumeng Fu[3] Nan Yu[1] Guohong Fu[1,2*]
[1]School of Computer Science and Technology, Soochow University
[2]Institute of Artificial Intelligence, Soochow University
[3]School of Computer Science and Technology, Harbin Institute of Technology
{20224027010, nyu}@stu.suda.edu.cn, 24b303004@stu.hit.edu.cn, ghfu@suda.edu.cn

Abstract

Recent advancements in multimodal out-of-context (OOC) misinformation detection have made remarkable progress in checking the consistencies between different modalities for supporting or refuting image-text pairs. However, existing OOC misinformation detection methods tend to emphasize the role of internal consistency, ignoring the significant of external consistency between image-text pairs and external evidence. In this paper, we propose HiEAG, a novel Hierarchical Evidence-Augmented Generation framework to refine external consistency checking through leveraging the extensive knowledge of multimodal large language models (MLLMs). Our approach decomposes external consistency checking into a comprehensive engine pipeline, which integrates reranking and rewriting, apart from retrieval. Evidence reranking module utilizes Automatic Evidence Selection Prompting (AESP) that acquires the relevant evidence item from the products of evidence retrieval. Subsequently, evidence rewriting module leverages Automatic Evidence Generation Prompting (AEGP) to improve task adaptation on MLLM-based OOC misinformation detectors. Furthermore, our approach enables explanation for judgment, and achieves impressive performance with instruction tuning. Experimental results on different benchmark datasets demonstrate that our proposed HiEAG surpasses previous state-of-the-art (SOTA) methods in the accuracy over all samples.

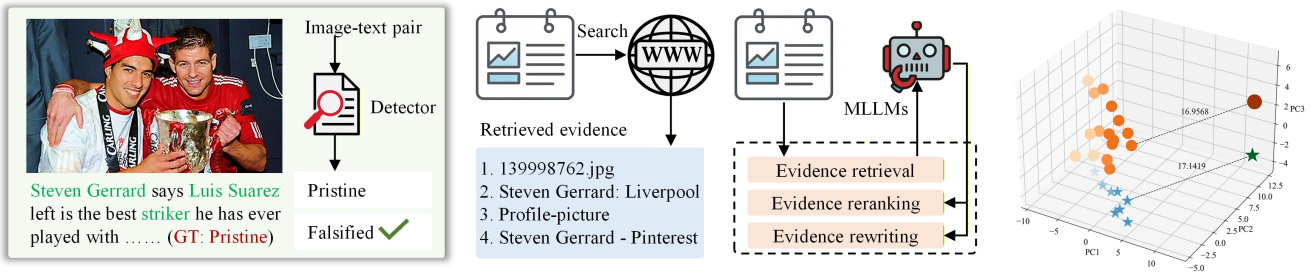
1. Introduction

Out-of-context (OOC) misinformation on social media is particularly prevalent, and often induces significant risks, such as conspiracy theories and societal safety [28, 31]. This prominent type of misinformation refers to a genuine image used in a misleading or incorrect textual context, posing a unique challenge in the digital era [3, 16]. For example, malicious actors pair election images with unrelated textual context during the U.S. presidential election, thereby deceiving or misleading voters [47]. To address these potential risks, it is essential to develop effective approaches for detecting OOC misinformation.

In the existing literature, the goal of OOC misinformation detection is to identify the misuse of a genuine image within a certain textual context. Early OOC misinformation detection methods primarily underscore internal consistency [2, 26], *i.e.*, the alignment between images and their textual context at the semantic level. Despite their impressive performance, these methods struggle to capture nuanced discrepancies, such as cross-referencing temporal and entity details. Fig. 1(a) provides an example of multimodal OOC misinformation detection, in which a textual context includes "Steven Gerrard" and "Luis Suarez", and a predicted label denotes the authenticity of an image-text pair. To further enhance fact-checking, some researchers attempt to retrieve external evidence into OOC misinformation detectors with different architectures, including attached classifiers [14], pre-trained small models [37], and Multimodal Large Language Models (MLLMs) [6]. As shown in Fig. 1(b), external evidence of image-text pairs is collected through an inverse image search, and is then used to improve detection performance. Though they can work together with web-retrieved evidence, their capacity is limited due to the lack of external consistency, *i.e.*, the alignment between image-text pairs and external evidence. Hence, refining external consistency has become a pressing necessity in the task of OOC misinformation detection.

To reach the target above, some research leverages cosine similarity to calculate the relationships between image-text pairs and external evidence [30, 39]. On the other hand, recent advancements in MLLMs, Qi et al. [36] introduced a two-pronged analysis for checking both internal and exter-

*Corresponding author.



(a) OOC misinformation detection (b) Evidence retrieval (c) Complete workflow (d) Evidence distribution

Figure 1. **Multimodal OOC misinformation detection consists of critical components.** Subfigure (a) presents internal consistency checking for identifying the authenticity of an image-text pair. Subfigure (b) presents evidence retrieval through tool usage. Subfigure (c) presents core parts within the complete workflow. Subfigure (d) visualizes the Euclidean distance between image-text pairs (dark points) and their retrieved evidence (shallow points).

nal consistencies via news-domain entity learning. Furthermore, inspired by the successful application of Retrieval-Augmented Generation (RAG) in large models [27, 51], recent studies combine this technique with lexical matching [44, 50], or association scoring [17, 45] to acquire the targeted evidence for judgment. However, they encounter the inherent limitations of external consistency checking as follows:

(a) **Incomplete workflow**: A standard workflow should include external evidence retrieval, reranking, and rewriting, as shown in Fig. 1(c). The former two steps have been considered in existing studies, but the last step is unexplored. This is attributed to preference alignment from model evolution [29]. MLLMs have strong understanding capacities towards evidence sentences, rather than evidence pieces.

(b) **Evidence entanglement**: External evidence with different degrees of relevance represents the relationships between image-text pairs and classification labels. Compared to weakly relevant, and even irrelevant evidence, relevant evidence is more useful for more accurate judgments. In Fig. 1(d), a qualitative analysis is provided to present an urgent requirement: How to design an effective evidence-augmented generation (EAG) strategy to enhance performance in detection?

In this paper, we propose a novel Hierarchical Evidence-Augmented Generation (HiEAG) framework, designed to refine consistency checking for the task of multimodal OOC misinformation detection. Specifically, at the process of evidence reranking, we introduce Automatic Evidence Selection Prompting (AESP) to autonomously acquire the most relevant item once evidence retrieval for image-text pairs. Subsequently, we design Automatic Evidence Generation Prompting (AEGP). This utilizes large models to understand and generate an alignment sentence based on multimodal content and the selected evidence item. In addition, to enable the model for both explanation and judgment, we employ instruction-tuning on a constructed OOC misinformation dataset. Experiments on both synthetic and real-world datasets demonstrate the effectiveness and robustness of our proposed HiEAG in the realm of multimodal OOC misinformation detection.

In summary, our major contributions are three-fold:

- We propose HiEAG, a novel multimodal OOC misinformation detection framework that effectively refines consistency checking between multimodal content and external evidence, significantly improving detection performance.
- We develop the AESP to leverage the knowledge of MLLM's parameters, selecting the relevant item from retrieved evidence. Additionally, the AEGP is designed to achieve the alignment sentence for task adaptation on MLLM-based OOC misinformation detection models.
- We conduct extensive experiments on the NewsCLIPpings and VERITE benchmark datasets, demonstrating that HiEAG outperforms the SOTA methods in the accuracy over all samples.

The rest of this paper is organized as follows. Section 2 summarizes the related works of OOC misinformation detection. Section 3 presents the details of our proposed approach HiEAG. Section 4 provides the experimental settings and results. Finally, we provide a conclusion and future research directions in Section 5.

2. Related Work

2.1. OOC misinformation detection

Out-of-context misinformation (OOC) detection is the task of identifying the misuse of a genuine image within a certain textual context. Due to the rapid expansion of forgery content on social media, this detection task is urgent to alleviate the potential risk. Recently, a large amount of OOC misinformation detection methods have been developed. The methods are categorized as: internal checking

[2, 26, 32] and external checking [1, 11, 12, 30, 38, 52, 58]. For the internal checking that focuses on modality features for capturing the semantic differences between multiple modalities, Papadopoulos et al. [32] used CLIP [37] and auxiliary Transformer layers to intensify multimodality feature extraction and interaction. Aneja et al. [2] utilized visual grounding with textual descriptions for interpreting semantic conflicts. For the external checking that introduces web-retrieved evidence into detectors for cross-modal consistency reasoning, Abdelnabi et al. [1] established a CNN-based model to perform the cycle consistency between image-text pairs and web-retrieved evidence. Yuan et al. [52] followed this work to further learn the stances of external evidence through neural networks.

Despite remarkable progress made in previous OOC misinformation detection methods, the accuracy of detection is still limited by the scale of pre-trained multimodal models.

2.2. MLLM-based OOC misinformation detection

Against the backdrop of the rapid development of Large Language Models (LLMs), MLLMs have emerged as a promising alternative for visual and multimodal tasks [13, 23, 24]. MLLMs likewise LLaVA [21], MiniGPT-v2 [4], and LLaVA-1.5 [22], have demonstrated remarkable multimodal understanding capabilities, and presented superior zero-shot performance in various tasks, such as Visual Question Answer (VQA) [9, 53] and Anomaly Segmentation (AS) [19, 35]. For instance, Xu et al. [46] designed prompts to extract visual information for answering visual questions. Xu et al. [48] introduced a tuning-free pipeline to leverage frozen MLLMs for video moment retrieval. Some studies [8, 17, 44, 56] leverage the closed-source MLLMs and expensive API calls to detect OOC misinformation in the zero-shot or few-shot setting. Such manner is impractical and inefficient in low-resource scenarios. A Parameter-Efficient Fine-Tuning (PEFT) technique, namely instruction tuning, is introduced for MLLMs on task-specific adaptation. Qi et al. [36] introduced InstructBLIP [6] with Vicuna [5] to conduct both internal and external consistency checking. Inspired by this, Xuan et al. [49] adopted multi-query generation to gather comprehensive evidence for enhancing the detection capabilities of MLLMs. Chat-OOC [40] based on MiniGPT-4 [57] with 13B parameters employs instruction tuning to improve the controllability of MLLMs for both judgments and explanations. In this work, our proposal considers the extensive knowledge of MLLMs to refine consistency checking between image-text pairs and external evidence, thus effectively debunking multimodal OOC misinformation.

3. Methodology

3.1. Problem setting & overview

Given an out-of-context misinformation dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_i^N$, where $\mathbf{x}_i = \{x_i^I, x_i^T\}$ denotes a news image and its attached textual context. Each pair of image-text is assigned with the ground-truth label $y_i \in \{0, 1\}$, being either Pristine (Not out-of-context) or Falsified (Out-of-context). N denotes the size of $\mathcal{D}$. The goal of OOC misinformation detection is to train a detector $f(\mathbf{x}_i) \rightarrow \hat{y}_i$, which provides a judgment $\hat{y}_i$ regarding the authenticity of an image-text pair $\mathbf{x}_i$.

As depicted in Fig. 2, we present a novel hierarchical framework HiEAG for OOC misinformation detection. HiEAG begins with evidence retrieval that stems from the Internet. External evidence is then processed through rerank-and-rewrite prompting. In the Automatic Evidence Selection Prompting (AESP), the most relevant evidence item is capturing by query reranking, rather than calculating the Euclidean distance between multimodality vectors. In the Automatic Evidence Generation Prompting (AEGP), the generated sentence aligns the original meaning of the selected evidence, thereby assisting in checking the consistencies between image-text pairs and external evidence. The two prompting mechanisms jointly contribute to the final judgment. The former enhances evidence retrieval while the latter highlights evidence optimization for task adaptation on large models. This designed hierarchical decomposition refines consistency checking between multimodal content and external information while optimizing the utilization of external evidence for debunking OOC misinformation.

3.2. Evidence retrieval

Evidence retrieval (ER) refers to knowledge around an image-text pair $\mathbf{x}$ by querying external knowledge bases, such as the Google Vision API. The retrieved evidence is regarded as potential clues for consistency checking. In practice, we follow the previous collection pipeline [1], using the API to perform evidence retrieval for image captions. The API returns the containing page's URLs associated with the image. A web crawler is designed to visit these pages, and then saves the captions if found. The captions may contain the image's content and further contain the image's context. However, in rare cases where the captions is not available, we employ the MLLM to directly describe the content of the image for subsequent consistency checking.

3.3. Automatic evidence selection prompting

After evidence retrieval for image captions, as shown in Fig. 2, we design Automatic Evidence Selection Prompting (AESP) to choose the most relevant evidence item and further leverage the MLLM's inherent knowledge. In prac-

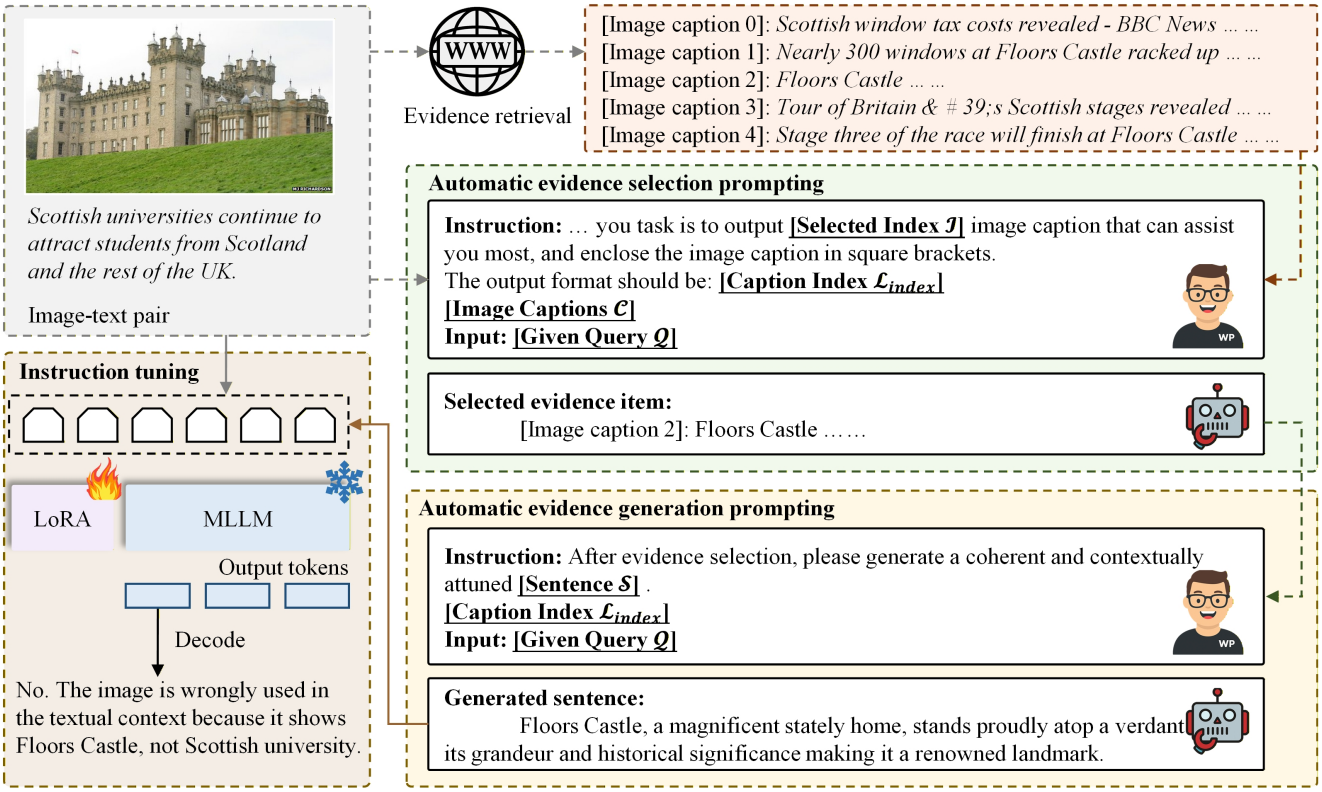


Figure 2. **The overview of our proposed framework HiEAG.** (a) Automatic evidence selection prompting acquires the most relevant evidence item regarding the image-text pair. (b) Automatic evidence generation prompting achieves a novel sentence that aligns the selected item. (c) Instruction tuning enables the MLLM-based OOC misinformation detector for both judgment and explanation.

tice, the prompt is as follows:

> **Instruction:** ...your task is to output [Selected Index $\mathcal{I}$] image caption that can assist you most, and enclose the image caption in square brackets.
> The output format should be: [Caption Index $\mathcal{L}_{index}$] [Image Captions $\mathcal{C}$]
>
> **Input:** [Given Query $\mathcal{Q}$]

In practice, the AESP directly inputs the [Given Query $\mathcal{Q}$] and the set of [Image Captions $\mathcal{C}$] into the MLLM for analyzing the relationships between image-text pairs and external information. After that, the [Selected Index $\mathcal{I}$] of image captions is integrated into the final output [Caption Index $\mathcal{L}_{index}$]. Formally, the evidence selection process is presented as follows:

$$\mathcal{L}_{index} = \underset{\mathcal{I} \in \mathcal{C}_{range}}{\operatorname{argmax}} \; p(\mathcal{I} \mid \mathcal{Q}, \mathcal{C}), \tag{1}$$

where $\mathcal{C}_{range}$ represents the number of image captions, and $\mathcal{L}_{index}$ represents the selected item to facilitate the final decision-making.

3.4. Automatic evidence generation prompting

Once achieving the most relevant evidence item through the AESP, as shown in Fig. 2, we further introduce Automatic Evidence Generation Prompting (AEGP). In practice, our carefully devised prompt for guiding the MLLM to automatically generate a novel sentence while retaining the original meaning of the selected evidence item as follows:

> **Instruction:** After evidence selection, please generate a coherent and contextually attuned [Sentence $\mathcal{S}$] .
> [Caption Index $\mathcal{L}_{index}$]
>
> **Input:** [Given Query $\mathcal{Q}$]

In practice, the process depends on the [Given Query $\mathcal{Q}$] and the [Caption Index $\mathcal{L}_{index}$] to directly guide the MLLM to generate a coherent and contextually attuned [Sentence $\mathcal{S}$]. This enhances the performance of detectors by aligning the image caption to the evidence sentence generated from the MLLM. Formally, the evidence generation process is presented as

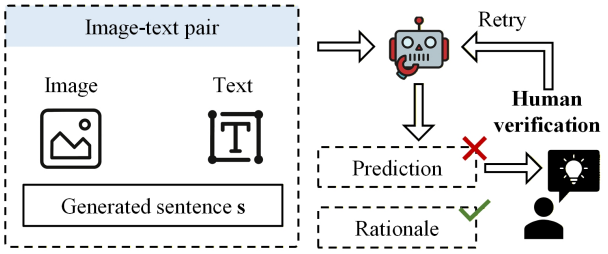


Figure 3. The construction process of evidence-augmented instruction dataset.

follows:

$$\mathcal{S} = \text{argmax } p(s \mid \mathcal{Q}, \mathcal{L}_{index}), \tag{2}$$

where $\mathcal{S}$ represents the alignment sentence for the selected item of the AESP. The proposed hierarchical framework consists of both AESP and AEGP, which contributes to evidence-augmented generation for refining consistency checking. Ablation study demonstrates that utilizing this proposal can effectively improve detection performance.

3.5. Evidence-augmented instruction dataset

To enable the explanation capabilities of detectors for more accurate judgments, we provide a detailed process of constructing the instruction dataset for multimodal OOC misinformation detection, as shown in Fig. 3. The instruction dataset is built upon the NewsCLIPpings dataset [26]. In practice, we present two categories of targeted outputs y'. For each pristine sample, we directly provide the targeted language without any rationales (*i.e.*, "Yes. The image is rightly used in the textual context."). For each falsified sample, we achieve the predicted language with a rationale (*i.e.*, "No. The image is wrongly used in the textual context. *rationale*"). Formally, the predicted language can be calculated as follows:

$$y' = \text{MLLMs}(\mathbf{x}, \mathbf{s}, q), \tag{3}$$

where $\mathbf{x}$ represents the image-text pair, $\mathbf{s}$ represents the alignment sentence, and q represents the given query. After achieving the predicted output, we further perform a verification process between the predicted output y' and the correct label y. This is to ensure that the predicted output matches the correct label. If there are inconsistencies between them, we retry Eq. (3) until they are equal. Finally, the constructed instruction dataset $\mathcal{D}' = \{\mathbf{x}_i, y_i'\}_i^N$ consists of 35,536 pristine samples and 35,536 falsified samples with rationales.

3.6. Evidence-augmented instruction tuning

After constructing the instruction dataset, we further adopt instruction tuning with the LoRA adapter [10] to the task of multimodal OOC misinformation detection. This preserves the original parameters θ of the model, thereby reducing the computational resource usage:

$$\theta' = \theta + \Delta\theta, \;\; \Delta\theta = \text{LoRA}(A, B), \tag{4}$$

where $\Delta\theta$ represents the updated model parameters, $A \in \mathbb{R}^{r\times d}$ and $B \in \mathbb{R}^{d\times r}$ represent low rank matrices. The signs d and r denote the dimension and the rank within a single matrix, respectively.

In accordance with the optimization objective of MLLMs, we apply the next token prediction loss to assess the output error of the task-specific MLLM, which can be formulated as follows:

$$\mathcal{L} = -\frac{1}{B}\frac{1}{T}\sum_{i=1}^{B}\sum_{t=1}^{T}\log P(\epsilon_t \mid \epsilon_{<t}, x'), \tag{5}$$

where B and T represent the batch size and the number of tokens, respectively. The sign x' represents the input sequence that integrates both textual and visual representations. This serves as the context for next token predictions.

4. Experiment

In this section, we conduct extensive experiments to evaluate the effectiveness of our proposed HiEAG in the task of OOC misinformation detection. First, we briefly introduce benchmark datasets and evaluation metrics used in the experiments, followed by a discussion on the compared methods and training details.

4.1. Experimental setup

Datasets. We conduct experiments on two widely used OOC misinformation datasets, including NewsCLIPpings [26] and VERITE [33].

- **NewsCLIPpings** [26] is the largest benchmark dataset for OOC misinformation detection. It is derived from the dataset VisualNews [20], which stems from four news agencies, *i.e.,* The Guardian, BBC, The Washington Post and USA Today. The out-of-context samples in this dataset are generated by replacing original images with semantically related images but present distinct news events. The NewsCLIPpings Merged-Balanced subset is divided into 71,072 samples for training, 7,024 samples for validation, and the remaining 7,264 samples for testing, respectively. Each sample contains a pair of image-text, and the dataset is evenly balanced with respect to class labels.
- **VERITE** [33] is a real-world dataset, which consists of 1,000 annotated samples in the form of image-text pairs. These samples are collected from fact-checking websites, such as Snopes and Reuters. In these websites, human experts verify the authenticity of content in cycle. Each

image is paired with different textual context. One is regarded as pristine, and other is verified as falsified. Moreover, to avoid uni-modal bias, this dataset excludes asymmetric multimodal content, and adopts modality balancing strategy. To align with previous works in experimental settings, we also train the proposed HiEAG on the training set of NewsCLIPpings, and then report its results on the NewsCLIPpings test set and VERITE, respectively.

Evaluation metrics. To evaluate the performance of our proposed method HiEAG in the task of OOC misinformation detection, we utilize accuracy across all samples (All), separately for the OOC (Falsified) samples and the not OOC (Pristine) samples as performance metrics. These metrics are regarded as a comprehensive assessment of detectors efficiency [26]. The accuracy metric denotes the proportion of correctly classified instances. In judgment, a professional OOC misinformation detector should maintain a good balance between Falsified samples and Pristine samples.

Baseline methods. To execute a comprehensive evaluation of our proposed method HiEAG, we compare it with several representative OOC misinformation detection methods. The compared methods can be categorized as follows:

Auxiliary learning methods. This category contains **SAFE** [55] and **EANN** [43]. The former method designs an auxiliary loss to calculate sentence similarity, and translates images into descriptive sentences. The latter method adopts an adversarial learning framework to combine the event discrimination loss with the detection loss. These methods leverage auxiliary tasks to effectively train detectors for identifying multimodal OOC misinformation.

Pre-trained methods. This category includes **VisualBERT** [18], **CLIP** [37], **Neu-Sym Detector** [54], **DT-Transformer** [32], **CCN** [1], and **SEN** [52]. VisualBERT leverages a unified transformer to optimize the alignment of multimodal content. CLIP adopts contrastive learning to acquire similar representations of image-text pairs. Neu-Sym Detector performs neural-symbolic reasoning through text decomposition and aggregation. DT-Transformer introduces more transformer layers to refine the interaction of multimodal content. CCN based on CLIP conducts multimodal cycle-consistency check. SEN models the stance extraction and calculates support-refutation scores for judgments.

MLLM-based methods. This group encompasses **Chat-OOC** [40] and **SNIFFER** [36]. Chat-OOC leverages a cross entropy loss to fine-tune a MLLM in the realm of OOC misinformation detection. SNIFFER adopts a two-stage instruction tuning, and directly integrates retrieved evidence for the final judgment. The methods rely on MLLM's promising capabilities for detecting OOC misinformation.

Implementation details. We employ PandaGPT (7B) [41] and Qwen2-VL (7B) [42] models as the base MLLMs. We implement HiEAG on PyTorch [34] version 2.3.1 with CUDA 12.2, and conduct all experiments with 4 NVIDIA GeForce RTX 3090 (24G) GPUs. The model is optimized by AdamW [25] with a liner warm-up learning rate strategy. We set the batch size to 8, the learning rate to 2e-4, and train the model for 5 epochs. In addition, we leverage FlashAttention-2 [7] to replace the original attention layers of large models for efficient training.

Table 1. **Comparison of OOC misinformation detection methods on the test set of NewsCLIPpings [26].** For each metric, we report the average of three runs without any hyperparameter searching. The best results are indicated in **bold**.

Method	**All** ↑	**Falsified** ↑	**Pristine** ↑
SAFE [55]	52.8	54.8	52.0
EANN [43]	58.1	61.8	56.2
VisualBERT [18]	58.6	38.9	78.4
CLIP [37]	66.0	64.3	67.7
Neu-Sym detector [54]	68.2	-	-
DT-Transformer [32]	77.1	78.6	75.6
CCN [1]	84.7	84.8	84.5
SEN [52]	87.1	85.5	88.6
Chat-OOC (13B) [40]	80.0	-	-
InstructBLIP (13B) [6]	82.5	75.3	89.7
SNIFFER [36]	88.4	86.9	**91.8**
PandaGPT (7B) [41]	72.7	72.1	73.2
w/ HiEAG	84.2	84.0	84.4
Qwen2-VL (7B) [42]	78.9	73.8	84.2
w/ HiEAG	**89.9**	**90.3**	89.4

4.2. Results on NewsCLIPpings dataset

Table 1 provides performance comparison between HiEAG and other methods on the NewsCLIPpings [26] test set. From the results, we can obtain the following observations as: (1) HiEAG achieves superior performance, which outperforms other methods in terms of all samples. (2) HiEAG with different base MLLMs consistently exhibits significant gains in all evaluation metrics. (3) HiEAG employs 7B parameters of large models, which is less than that of other MLLM-based methods, such as Chat-OOC and SNIFFER. (4) HiEAG presents a good balance between falsified samples and pristine samples. Overall, our proposed method HiEAG has strong detection accuracy in OOC misinformation, achieving performance significantly better than the base MLLMs. This demonstrates the efficacy of HiEAG.

4.3. Results on VERITE dataset

Furthermore, we compare our HiEAG method with other methods such as VERITE [33] and SNIFFER [36] on a real-world benchmark dataset VERITE [33]. As shown in Table 2, HiEAG outperforms all methods. We attribute this

Table 2. **Comparison of OOC misinformation detection methods on the test set of VERITE [33].** T/O denotes true versus out-of-context. For each metric, we report the average of three runs without any hyperparameter searching. The best results are indicated in **bold**.

Method	**T/O** ↑
VERITE [33]	72.7
SNIFFER (InstructBLIP 13B) [36]	74.0
HiEAG (Qwen2-VL 7B)	**74.4**

Table 3. **Ablation experiments for the HiEAG with the base MLLM PandaGPT (7B) on the NewsCLIPpings [26] test set.** ER and RE denote evidence retrieval and reasoning explanation, respectively.

ER	AESP	AEGP	RE	Tuning	**All** ↑	**Falsified** ↑	**Pristine** ↑
					49.4	56.1	42.8
✓					51.1	57.2	45.0
				✓	72.7	72.1	73.2
✓				✓	79.9	78.9	80.9
✓	✓			✓	82.1	79.3	85.2
✓	✓	✓		✓	83.4	80.8	86.0
✓	✓	✓	✓	✓	84.2	84.0	84.4

Table 4. **Ablation experiments for the HiEAG with the base MLLM Qwen2-VL (7B) on the NewsCLIPpings [26] test set.** ER and RE denote evidence retrieval and reasoning explanation, respectively.

ER	AESP	AEGP	RE	Tuning	**All** ↑	**Falsified** ↑	**Pristine** ↑
					69.1	54.4	83.9
✓					76.7	68.0	85.3
				✓	78.9	73.8	84.2
✓				✓	83.0	77.1	88.9
✓	✓			✓	87.7	86.5	88.7
✓	✓	✓		✓	88.5	87.7	89.1
✓	✓	✓	✓	✓	89.9	90.3	89.4

to superior evidence-augmented generation approach in the realm of OOC misinformation detection. Typically, compared to the method VERITE, SNIFFER and HiEAG underscore the significance of evidence retrieval related to multimodal content. In evidence entanglement, our method acquires the relevant evidence, rather than all evidence, thereby refining consistency checking between multimodal content and external information. This superior performance indicates the effectiveness and generalizability capabilities of our proposed method HiEAG, even in a real-world OOC misinformation dataset.

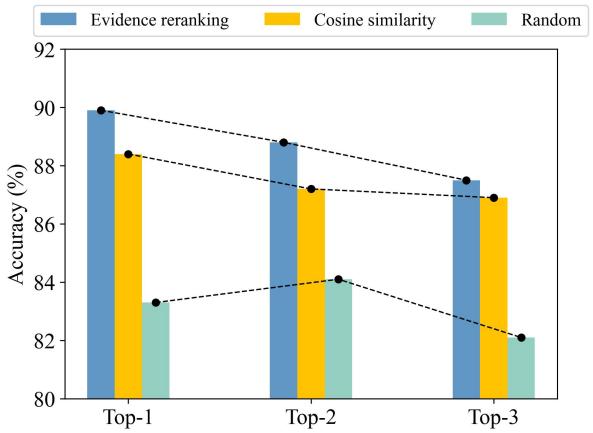


Figure 4. **Ablation experiments for HiEAG with distinct evidence reranking strategies across the NewsCLIPpings [26] dataset.** The axis denoting the number (Top-k) of evidence increases gradually, whereas the axis for the accuracy (%) across all samples declines.

4.4. Ablation study

In this study, we provide an ablation experiment to evaluate the effectiveness of critical components in our proposed method HiEAG. The experiments are conducted on the test set of NewsCLIPpings dataset, and the results are reported in Table 3 and Table 4.

First, the experiment that presents the comparison between w ER and w/o ER in a zero-shot setting (see 1-th row and 2-th row in both Tables). Introducing evidence retrieval into the base MLLMs achieves significant gains in performance. This indicates that external evidence is potential to assist advanced detectors for detecting OOC misinformation, even in zero-shot scenarios. Second, in instruction tuning, the experiment that removes the proposed HiEAG approach demonstrates a decline in all evaluation metrics (see 3-th row to 6-th row in both Tables). This approach plays a critical role in extracting and modeling the relationships between multimodal content. Finally, the utilization of reasoning explanation also results in a gain in detection performance (see 7-th row in both Tables). This enhances the detector's capacity to provide possible explanations for judgments.

The results above demonstrate that each component in the proposed HiEAG framework is complementary. Together, the HiEAG achieves outstanding performance in multimodal OOC misinformation detection. This confirms the effectiveness of each component in learning the relevant evidence item for supporting or refuting multimodal content.

Table 5. A multimodal OOC misinformation detectors with distinct LVLMs.

Setting	Tuning	Parameters	**All** ↑
Random		-	50.3
Qwen2-VL-2B		1.5B	65.7
Qwen2-VL-7B		7.6B	79.1
Qwen2-VL-72B		72B	79.8
Qwen2-VL-2B	✓	1.5B	81.4
Qwen2-VL-7B	✓	7.6B	89.9

4.5. Further analysis

In this section, we provide an in-depth analysis of the proposed HiEAG from the perspectives of evidence reranking strategy, model size and training data size, as shown in Fig. 4, Fig. 5, and Table 5, respectively. This is to discuss the efficiency of HiEAG.

Analysis of evidence reranking. We further investigate the role of evidence reranking in our proposed method HiEAG, as shown in Fig. 4. In this investigation, we vary the number of external evidence from 1 to 3. This avoids the degradation of performance due to the increasement of context input. Specifically, the proposed evidence reranking strategy surpasses others consisting of random and cosine similarity across different settings. This achieves peak performance with top-1 evidence item in the term of accuracy metric. More external items exhibit limited capacity to conduct consistency checking between multimodal content and supplementary information, resulting in suboptimal performance. Compared with random and cosine similarity, our approach leverages the extensive knowledge of MLLMs parameters to acquire the relevant evidence item, thereby effectively reducing the disturbance of weakly relevant items and even irrelevant items. This suggests that the MLLM reranks external evidence based on the consistencies between multimodal content and evidence items, facilitating multimodal OOC misinformation detection.

Analysis of model size. To assess the influence of different model sizes in the HiEAG, we start the investigation of HiEAG with Qwen2-VL family [42] by varying model parameters from 2B to 72B in various experimental settings. As presented in Table 5, we achieve the following findings: In the setting of zero-shot (see 2-th row and 4-th row in Table 5), as model size increases, the detection performance across all samples improves. The results align with prior studies in scaling laws [15]. However, the base MLLM Qwen2-VL-72B just achieves a performance improvement of 0.7%, and introduces around 10 times parameters, compared to Qwen2-VL-7B. This ignores the compromise between computation burden and detection performance. In the setting of instruction tuning, the general-purpose base MLLMs are extended to the task of OOC misinformation detection, resulting in significant improvements in performance. This encourages us to select Qwen2-VL-7B as the base MLLM.

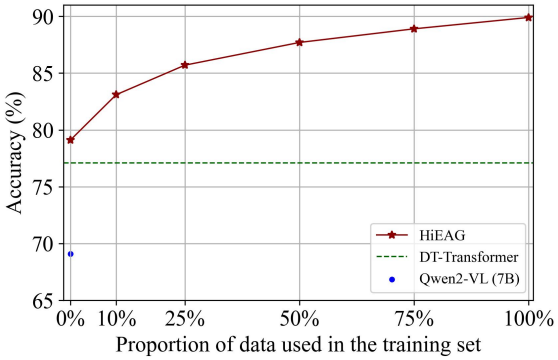


Figure 5. Performance of HiEAG on the NewsCLIPpings [26] using different training data proportions.

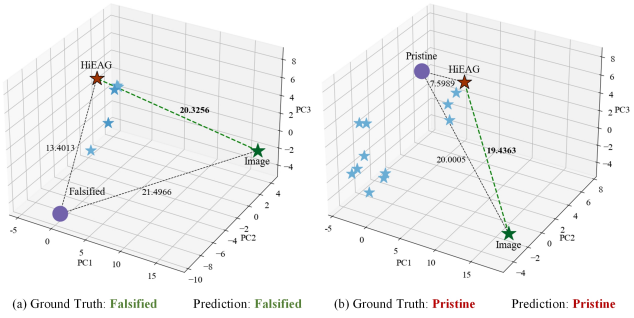


Figure 6. **Visualization of different data distributions.** Subfigure (a) represents a falsified image-text pair. Subfigure (b) represents a pristine image-text pair. Blue points denote the retrieved evidence regarding image-text pairs.

Analysis of training data size. To evaluate the feasibility of rapid deployment of HiEAG at the stage of early detection, we vary the size of training data from 10% to 75% while fixing other configurations during the training process. As shown in Fig. 5, our proposed method HiEAG achieves remarkable performance in any experimental settings. With the growth of training data, the discernment capabilities of HiEAG improves, resulting in a rising tendency in the term of accuracy across all samples. This highlights that HiEAG can be rapidly deployed while ensuring comparable detection accuracy, even at the early stage.

4.6. Case study and visualization

As presented in Fig. 6, we provide two instances from the validation set of NewsCLIPpings [26] to demonstrate the effectiveness of HiEAG through visualization. The instances provide valuable insights regarding the contributions of HiEAG in the OOC misinformation detection task. In the first instance, by analyzing an OOC image-text pair

(see subfigure (a) in Fig. 6), we note that the novel generated sentence is far further away from the image-text pair. This prevents the model from modeling the relationships between the image and its misleading textual context. In the second instance, by observing a pristine image-text pair (see subfigure (b) in Fig. 6), we find that HiEAG maintains the generated sentence close to the image-text pair, thereby facilitating an accurate judgment of the targeted content. The observation above demonstrates the effectiveness and superiority of our proposed HiEAG in refining consistency checking between multimodal content and external information, highlighting its capacity to the task of OOC misinformation detection.

5. Conclusion

In this paper, we presented HiEAG, a novel hierarchical evidence-augmented generation framework designed to the task of multimodal out-of-context misinformation detection. In the evidence reranking, we designed Automatic Evidence Selection Prompting to learn the relevant evidence item, enhancing evidence retrieval for multimodal content. In the evidence rewriting, we devised Automatic Evidence Generation Prompting to achieve the alignment sentence, which improves task adaptation on model architectures. Additionally, to enable the model for both explanation and judgment, we based on the instruction data, extended instruction tuning to the task for training a detector. Through extensive experiments on synthetic and real-world datasets, we demonstrated the effectiveness and robustness of the proposed HiEAG in addressing the challenges of multimodal out-of-context misinformation detection.

References

[1] 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 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14940–14949, 2022. 3, 6

[2] Shivangi Aneja, Chris Bregler, and Matthias Niessner. Cosmos: Catching out-of-context image misuse using self-supervised learning. In *Proceedings of the 39th AAAI Conference on Artificial Intelligence (AAAI)*, pages 9342–9354, 2025. 1, 3

[3] Kevin Aslett, Zeve Sanderson, William Godel, Nathaniel Persily, Jonathan Nagler, and Joshua A Tucker. Online searches to evaluate misinformation can increase its perceived veracity. *Nature*, 625(7995):548–556, 2024. 1

[4] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *2023 arXiv preprint arXiv: 2310.09478*. 3

[5] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 3

[6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2023. Curran Associates Inc. 1, 3, 6

[7] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024. 6

[8] Yimeng Gu, Zhao Tong, Ignacio Castro, Shu Wu, and Gareth Tyson. Multi-mllm knowledge distillation for out-of-context news detection. *2025, arXiv preprint arXiv: 2505.22517*. 3

[9] Seunghoon Han, Mingyu Choi, Hyewon Lee, SoYoung Park, Jong-Ryul Lee, Sungsu Lim, and Tae-Ho Kim. Diverse knowledge selection for enhanced zero-shot visual question answering. In *Proceedings of the ACM on Web Conference 2025 (WWW)*, page 2161–2169, 2025. 3

[10] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022. 5

[11] 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 (MM)*, pages 1465–1471, 2017. 3

[12] Ayush Jaiswal, Yue Wu, Wael AbdAlmageed, Iacopo Masi, and Premkumar Natarajan. Aird: Adversarial learning framework for image repurposing detection. In *Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11330–11339, 2019. 3

[13] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27036–27046, 2024. 3

[14] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia (MM)*, pages 795–816, 2017. 1

[15] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *2020, arXiv preprint arXiv: 2001.08361*. 8

[16] Kumud Lakara, Georgia Channing, Juil Sock, Christian Rupprecht, Philip Torr, John Collomosse, and Christian Schroeder de Witt. Llm-consensus: Multi-agent debate for visual misinformation detection. 2024, arXiv preprint arXiv: 2410.20140. 1

[17] Fanxiao Li, Jiaying Wu, Canyuan He, and Wei Zhou. CMIE: Combining MLLM insights with external evidence for explainable out-of-context misinformation detection. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025*, pages 9342–9354, 2025. 2, 3

[18] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. 2019, arXiv preprint arXiv: 1908.03557. 6

[19] Shengze Li, Jianjian Cao, Peng Ye, Yuhan Ding, Chongjun Tu, and Tao Chen. Clipsam: Clip and sam collaboration for zero-shot anomaly segmentation. *Neurocomputing*, 618: 129122, 2025. 3

[20] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6761–6771, 2021. 5

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS)*, pages 34892–34916, 2023. 3

[22] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306, 2024. 3

[23] Huan Liu, Zichang Tan, Qiang Chen, Yunchao Wei, Yao Zhao, and Jingdong Wang. Unified frequency-assisted transformer framework for detecting and grounding multi-modal manipulation. *International Journal of Computer Vision*, 133 (3):1392–1409, 2024. 3

[24] Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10770–10780, 2024. 3

[25] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *2019, arXiv preprint arXiv: 1711.05101*. 6

[26] Grace Luo, Trevor Darrell, and Anna Rohrbach. NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In *Proceedings of 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6801–6817, 2021. 1, 3, 5, 6, 7, 8

[27] Yuanjie Lyu, Zhiyu Li, Simin Niu, Feiyu Xiong, Bo Tang, Wenjin Wang, Hao Wu, Huanyong Liu, Tong Xu, and Enhong Chen. Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models. *ACM Transactions on Information Systems*, 43(2):1–32, 2025. 2

[28] Zihan Ma, Minnan Luo, Hao Guo, Zhi Zeng, Yiran Hao, and Xiang Zhao. Event-radar: Event-driven multi-view learning for multimodal fake news detection. In *Proceedings of the 62th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5809–5821, 2024. 1

[29] Shrikant Malviya and Stamos Katsigiannis. Evidence retrieval for fact verification using multi-stage reranking. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7295–7308, 2024. 2

[30] 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 2020 International Conference on Multimedia Retrieval (ICMR)*, page 16–25, 2020. 1, 3

[31] Qiong Nan, Qiang Sheng, Juan Cao, Yongchun Zhu, Danding Wang, Guang Yang, and Jintao Li. Exploiting user comments for early detection of fake news prior to users' commenting. *Frontiers of Computer Science*, 19(10):1910354, 2025. 1

[32] 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 (MAD)*, page 36–44, 2023. 3, 6

[33] 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. 5, 6, 7

[34] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: an imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NIPS)*, pages 8026–8037, 2019. 6

[35] Yun Peng, Xiao Lin, Nachuan Ma, Jiayuan Du, Chuangwei Liu, Chengju Liu, and Qijun Chen. Sam-lad: Segment anything model meets zero-shot logic anomaly detection. *Knowledge-Based Systems*, 314:113176, 2025. 3

[36] 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 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13052–13062, 2024. 1, 3, 6, 7

[37] 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. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. 1, 3, 6

[38] Ekraam Sabir, Wael AbdAlmageed, Yue Wu, and Prem Natarajan. Deep multimodal image-repurposing detection. In *Proceedings of the 26th ACM International Conference on Multimedia (MM)*, page 1337–1345, 2018. 3

[39] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18 (11):613–620, 1975. 1

[40] Fatma Shalabi, Hichem Felouat, Huy H. Nguyen, and Isao Echizen. Leveraging chat-based large vision language models for multimodal out-of-context detection. In *Proceedings of the 38th International Conference on Advanced Information Networking and Applications (AINA)*, pages 86–98, 2024. 3, 6

[41] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. PandaGPT: One model to instruction-follow them all. In *Proceedings of the 1st Workshop on Taming Large Language Models (TLLM)*, pages 11–23, 2023. 6

[42] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *2024, arXiv preprint arXiv: 2409.12191*. 6, 8

[43] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, page 849–857, 2018. 6

[44] Yin Wu, Zhengxuan Zhang, Fuling Wang, Yuyu Luo, Hui Xiong, and Nan Tang. Exclaim: An explainable cross-modal agentic system for misinformation detection with hierarchical retrieval. 2025, arXiv preprint arXiv: 2504.06269. 2, 3

[45] Yuzhuo Xiao, Zeyu Han, Yuhan Wang, and Huaizu Jiang. Xfacta: Contemporary, real-world dataset and evaluation for multimodal misinformation detection with multimodal llms. *2025, arXiv preprint arXiv: 2508.09999*. 2

[46] Liyong Xu, Yifan Jiao, and Bing-Kun Bao. Bool prompt with decomposition and enhancement: Zero-shot vqa based on pvlms. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(9):1–21, 2025. 3

[47] Qingzheng Xu, Heming Du, Huiqiang Chen, Bo Liu, and Xin Yu. Mmooc: A multimodal misinformation dataset for out-of-context news analysis. In *Proceedings of the 29th Australasian Conference on Information Security and Privacy (ACISP)*, pages 444–459, 2024. 1

[48] Yifang Xu, Yunzhuo Sun, Benxiang Zhai, Ming Li, Wenxin Liang, Yang Li, and Sidan Du. Zero-shot video moment retrieval via off-the-shelf multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 8978–8986, 2025. 3

[49] Keyang Xuan, Li Yi, Fan Yang, Ruochen Wu, Yi R. Fung, and Heng Ji. Lemma: Towards lvlm-enhanced multimodal misinformation detection with external knowledge augmentation. *2024, arXiv preprint arXiv: 2402.11943*. 3

[50] Zehong Yan, Peng Qi, Wynne Hsu, and Mong Li Lee. Mitigating genai-powered evidence pollution for out-of-context multimodal misinformation detection. 2025, arXiv preprint arXiv: 2501.14728. 2

[51] Dayu Yang and Fuli Wang. Towards retrieval-augmented large language model-based conversational recommender system. In *Proceedings of the 29th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 317–330, 2025. 2

[52] 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 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4268–4280, 2023. 3, 6

[53] Yujian Yuan, Jiabei Zeng, and Shiguang Shan. Exp-vqa: fine-grained facial expression analysis via visual question answering. *Pattern Recognition*, 168:111783, 2025. 3

[54] Yizhou Zhang, Loc Trinh, Defu Cao, Zijun Cui, and Yan Liu. Interpretable detection of out-of-context misinformation with neural-symbolic-enhanced large multimodal model. 2024, arXiv preprint arXiv: 2304.07633. 6

[55] Xinyi Zhou, Jindi Wu, and Reza Zafarani. Safe: Similarity-aware multi-modal fake news detection. In *Proceedings of the 24th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 354–367, 2020. 6

[56] Xinyi Zhou, Ashish Sharma, Amy X. Zhang, and Tim Althoff. Correcting misinformation on social media with a large language model. 2024, arXiv preprint arXiv: 2403.11169. 3

[57] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. 2023, arXiv preprint arXiv: 2304.10592. 3

[58] Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. Fact-checking meets fauxtography: Verifying claims about images. In *Proceedings of 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2099–2108, 2019. 3