CMIE: Combining MLLM Insights with External Evidence for Explainable Out-of-Context Misinformation Detection

Fanxiao Li[1,3], Jiaying Wu[2], Canyuan He[3], Wei Zhou[3] [*]

[1]School of Information Science and Engineering, Yunnan University
[2] National University of Singapore
[3]Engineering Research Center of Cyberspace, Yunnan University
lifanxiao15@outlook.com, jiayingwu@u.nus.edu, zwei@ynu.edu.cn

Abstract

Multimodal large language models (MLLMs) have demonstrated impressive capabilities in visual reasoning and text generation. While previous studies have explored the application of MLLM for detecting out-of-context (OOC) misinformation, our empirical analysis reveals two persisting challenges of this paradigm. Evaluating the representative GPT-4o model on direct reasoning and evidence augmented reasoning, results indicate that MLLM struggle to capture the deeper relationships—specifically, cases in which the image and text are not directly connected but are associated through underlying semantic links. Moreover, noise in the evidence further impairs detection accuracy. To address these challenges, we propose CMIE, a novel OOC misinformation detection framework that incorporates a Coexistence Relationship Generation (CRG) strategy and an Association Scoring (AS) mechanism. CMIE identifies the underlying coexistence relationships between images and text, and selectively utilizes relevant evidence to enhance misinformation detection. Experimental results demonstrate that our approach outperforms existing methods. Data and code are available at: https://github.com/fanxiao15/CMIE.

1 Introduction

The proliferation of multimodal misinformation on social media poses significant risks, including impairing public judgment, inciting panic, and causing economic losses (Lazer et al., 2023; Islam et al., 2021). Multimodal misinformation manifests in various forms, such as image-text mismatches (Wang et al., 2023; Yuan et al., 2023; Abdelnabi et al., 2022; Papadopoulos et al., 2024) and manipulated media (Qi et al., 2023; Yang et al., 2024). Among image-text mismatches, out-of-context (OOC) misinformation is particularly prevalent (Luo et al., 2021), where an unaltered

---
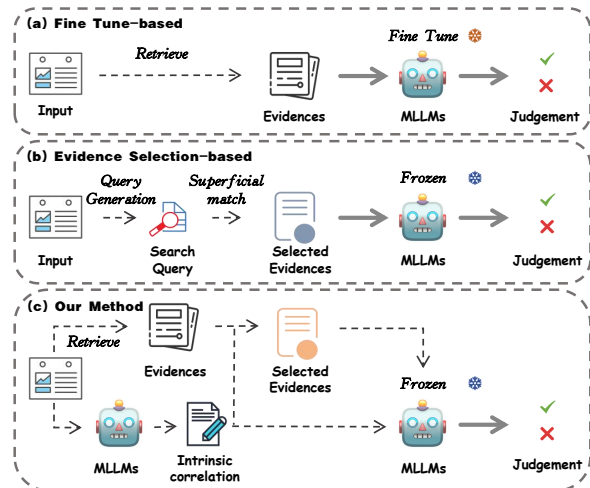[*] The corresponding author.

Figure 1: Comparison of different paradigms for MLLM-based OOC misinformation detection.

image is paired with a new but misleading textual context. Detecting such misinformation is challenging, as repurposing images with different textual descriptions can create seemingly strong yet deceptive correlations.

Previous studies (Luo et al., 2021; Papadopoulos et al., 2023; Aneja et al., 2021) typically detect OOC misinformation by training a small language model (SLM) to assess the consistency between image and text representations. Recently, some works (Abdelnabi et al., 2022; Müller-Budack et al., 2020; Yuan et al., 2023) attempt to verify information authenticity by retrieving external evidence from the Internet. However, these methods often lack explainability, providing only a final classification label without generating clear, human-readable justifications.

The advent of multimodal large language models (MLLMs) (Hurst et al., 2024; Anil et al., 2023) has driven significant advancements across various vision tasks (Li et al., 2024; Jin et al., 2024). Their advanced visual reasoning and text generation capabilities enable high-quality, human-readable expla-

nations for OOC misinformation detection. Additionally, their extensive world knowledge supports zero-shot transfer across diverse datasets, enhancing adaptability to novel misinformation instances.

Recent studies (Qi et al., 2024; Xuan et al., 2024; Zhou et al., 2024) integrate Retrieval-Augmented Generation (RAG) techniques to further improve misinformation detection by incorporating external evidence. SNIFFER (Qi et al., 2024) fine-tunes InstructBLIP (Dai et al., 2023) and uses ChatGPT (OpenAI, 2023) for explanation label generation, but as shown in Figure 1 (a), it is computationally expensive and requires extensive labeled data. In contrast, LEMMA (Xuan et al., 2024) and MUSE (Zhou et al., 2024) avoid fine-tuning and consider evidence relevance during retrieval (Figure 1 (b)). However, they primarily rely on superficial lexical matching and fail to explain how retrieved evidence influences MLLM's reasoning in detecting OOC misinformation.

To thoroughly investigate the capabilities of MLLM in OOC misinformation detection, this paper explores the following research questions: **(1) How well can MLLMs perform in misinformation detection without fine-tuning or external evidence? (2) How does external evidence influence MLLM-based misinformation detection?**

Through our empirical analysis using the representative GPT-4o model (detailed in Section 2), we find that MLLMs can achieve competitive performance in misinformation detection relying solely on their intrinsic knowledge, without requiring fine-tuning or external evidence. However, their judgments tend to be conservative, particularly when provided with image-text pairs that share similar contextual backgrounds but lack obvious inconsistencies. When external evidence is introduced, MLLMs often rely on superficial correlations rather than deeper semantic relationships, limiting their ability to make well-informed decisions. Additionally, noisy or irrelevant evidence further constrains performance improvements.

Motivated by these observations, we propose CMIE, an OOC misinformation detection framework designed to uncover deeper connections between image-text pairs and refine evidence selection accordingly, as illustrated in Figure 1(c). CMIE introduces a **coexistence relationship generation** (CRG) strategy that leverages MLLMs to identify and extract the underlying relationships between images and text. Since images are expected to support the credibility of the associated text, they should share common expressive elements. These extracted relationships serve as a core validation thread, enabling a more structured approach to misinformation detection. To incorporate high-quality external evidence, CMIE further incorporates an **association scoring** (AS) mechanism that evaluates external evidence based on its deeper semantic alignment with the identified image-text relationships. Rather than relying on superficial lexical matches, this mechanism ensures that only the most relevant evidence is considered, mitigating the impact of noisy or misleading information.

The key contributions of this paper can be summarized as follows:

- **In-Depth Exploration**: We examined the performance of MLLM in detecting OOC misinformation without fine-tuning and analyzed the impact of integrating external evidence directly into MLLM. Our findings revealed the importance of incorporating external evidence into the task reasonably.

- **Novel Solution**: We introduced CMIE, a novel framework consist of coexistence relationship generation strategy and the association scoring mechanism. This innovation effectively models the relationship between evidence and the task, significantly enhancing MLLM's performance without fine-tuning.

- **Effectiveness Verification**: Extensive experiments demonstrate that our method not only generates human-readable explanations but also achieves superior performance compared to state-of-the-art approaches across multiple evaluation metrics.

2 Exploratory Study

2.1 Task Definition

This paper seeks to utilize MLLM for detecting OOC misinformation, producing both final predictions and human-readable explanations. Given an image-text pair (I, T), knowledge G, we expect MLLM to produce a label $Y_p \subseteq \{Real, Fake\}$ and a human-readable explanation Y_e. Formally, this is expressed as:

$$Y_p, Y_e = MLLM((I, T), G), \quad (1)$$

which G may include only internal knowledge solely or a combination of internal knowledge and retrieved external evidence.

2.2 Evaluation Setting

we evaluate the performance in detecting OOC misinformation of representative MLLM, specifically GPT-4o (Hurst et al., 2024).

Dataset. We conduct our experiments on NewsCLIPpings (Luo et al., 2021), the largest benchmark for OOC misinformation detection, where both images and texts originate from real-world sources. Following prior research, we select the test set from the Merged/Balance subset, which consists of 71,072 samples in the training set, 7,024 in the validation set, and 7,264 in the test set.

Metrics. We measure performance using precision, recall, F1 score, and test accuracy for both real and fake samples.

2.3 Evaluation Types

We primarily explored the MLLM's ability without fine-tuning for OOC detection by leveraging its internal knowledge and combining it with external evidence.

- **Direct reasoning with internal knowledge (DR)**: In this method, MLLM relies solely on its internal knowledge to detect OOC misinformation. The image-text pair is input directly into the MLLM, which then generates a label (real or fake) and a corresponding explanation.

- **Augmented reasoning with external evidence (AR)**: This approach enhances MLLM's reasoning by incorporating external evidence retrieved from the internet. We provide the titles of posts containing the image-text pair, allowing the MLLM to make more informed judgments.

2.4 Findings

In Table. 1, we report the exploratory results for the two evaluation types mentioned above. Our findings are as follows:

Table 1: OOC misinformation detection performance of direct reasoning (DR) using internal knowledge and augmented reasoning (AR) with external knowledge.

-	Type	Precision	Recall	F1	Accuracy
real	DR	0.79	0.89	0.84	0.83
	AR	0.86	0.83	0.84	0.85
fake	DR	0.87	0.76	0.81	0.83
	AR	0.83	0.86	0.85	0.85

(1) Both DR and AR achieve competitive performance, highlighting **the feasibility of directly using MLLMs for OOC misinformation detection without fine-tuning.**

(2) When relying solely on internal knowledge for detection, **MLLM tend to produce conservative responses.** For instance, as illustrated in Figure 2, as long as there is a slight connection of the image-text pair and the MLLM cannot identify obvious errors via its internal knowledge, the pair is often considered as real. This observation is supported by the results in Table. 1, which shown a high recall rate for real samples. In OOC misinformation, images are used to enhance the credibility of the texts, leading to a certain degree of correlation of the pair. This correlation complicates the accurate identification of misinformation.
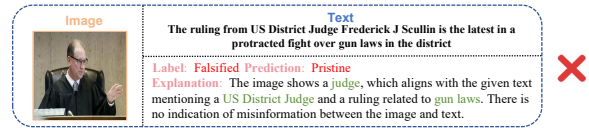

Figure 2: Example of direct reasoning in OOC misinformation detection using MLLM's internal knowledge.

(3) When combining external evidence for augmentation, **overall performance improves, but the improvement is limited.** The upper part of Figure 3 provides an example where external evidence effectively counteracts incorrect matches. As shown in Table. 1, adding evidence significantly improves the recall rate for fake samples. However, the improvement is constrained, which we attribute to two main factors. First, MLLM often struggled to grasp the core thread of verification after integrate evidence, as illustrated in the middle part of Figure 3. When there is no clear correlation between the evidence and the content to be verified, MLLM fails to find implicit relationships of the image-text pair as seen in direct reasoning. Instead, it only determines whether there is a correlation between the image-text pair and the evidence. Second, evidence can be noisy. As shown in the lower part of Figure 3, part of the evidences may be related to the image, but not to the core thread being verified. For effective verification, it is crucial to prioritize highly relevant evidence and treat less relevant evidence as noise. According these analyses, directly using evidence for reasoning leads to limited performance improvement. Results in Table. 1 further illustrated that while incorporating evidence

improves average accuracy, other metrics for both real and fake samples show unstable improvements, and may decline due to these aforementioned issues.

Figure 3: Examples of augmented reasoning by integrating external evidence.

3 Methodology

Based on the conclusions drawn from the the above exploratory analysis, we introduce CMIE: **C**ombining **M**LLM **I**nsights with **E**xternal Evidence for Explainable Out-of-Context Misinformation Detection. As illustrated in Figure 4, the framework takes multimodal news as input and produces final predictive label and explanation. In our approach, when relevant evidence is unavailable, we rely on the internal knowledge of the MLLM for judgement. When evidence can be retrieved, we focus on rationally integrating the evidence into MLLM's judgement process, which involves two key components: coexistence relationship generation strategy and association scoring mechanism. The first component identifies the deeper connection of image-text pair, while the second evaluates the significance of the evidence in relation to the connection.

3.1 Evidence Collection

When processing a image-text pair, we first need to retrieve relevant evidence. In CMIE, we select the titles associated with the image and the entities extracted from the image as evidence. Our evidence collection pipeline follows the method described in CCN (Abdelnabi et al., 2022). For image titles, we use a web crawler to retrieve the pages where the image appears and save the titles. For entities, we use the Google Vision API [1] to perform extraction. The API returns a list of entities associated with the image, which may describe the content of the image and further describe the context of where these images appear. However, in rare cases where such evidence is not available, we rely on the MLLM's internal knowledge to directly assess the consistency of the image-text pair and produce a human-readable explanation.

3.2 Coexistence Relationship Generation

Our analysis reveals that, after incorporating external evidence, MLLM struggle to identify deeper connection of image-text pair. Instead, it focus on whether the image-text pair align with the evidence. In OOC misinformation, the image's role is to enhance the credibility of the text and should have potential relationship with the text. The proper process for OOC misinformation detection should prioritize the image-text relationship, using evidence to support or refute this relationship. Inspired by this, we propose a coexistence relationship generation strategy that identifies the reasons why image and text might coexist (including direct and potential correlation). Specifically, given an image-text pair (I, T), we use MLLM to first identify and extract potential coexistence relationship R_{co} and coexistence score S_{co}. The formulas are as follows:

$$(R_{co}, S_{co}) = \text{MLLM}(I, T), \qquad (2)$$

here, R_{co} represents "why the image and text can appear together", and S_{co} indicates the extent to which image and text can appear together. We follow the paradigm of *(M)LLM-as-a-judge*(Gu et al., 2024; Chen et al., 2024) to prompt the model to generate corresponding scores without providing explicit definitions.

3.3 Association Scoring Mechanism

For retrieved evidences, they can be strongly related, weakly related, or even unrelated. Considering human thought processes, we should prioritize extracting useful information from strongly related evidence to assist the judgment process. Weaker

---

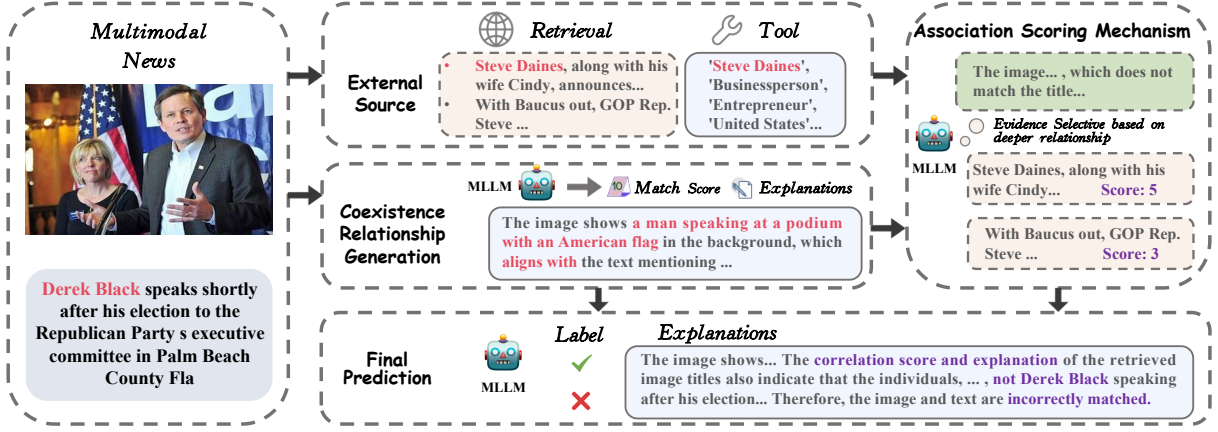[1] https://cloud.google.com/vision/docs/detecting-web

Figure 4: Overall architecture of the proposed CMIE, which effectively models the relationship between external evidence, generating predictive labels and interpretable explanations.

related evidence can be used as supplementary information, helping to further validate the alignment between the image and text.

Therefore, we propose a association scoring mechanism to effectively evaluate the relevance of evidence for the final judgment. Given the coexistence relationship R_{co} and the set of evidence $E_t = \{E_{t1}, E_{t2}, \ldots, E_{tn}\}$, we calculate the relevance score S_i and scoring explanation exp_i for each piece of evidence according to the coexistence relationship:

$$(s_i, exp_i) = \text{MLLM}(R_{co}, E_{ti}), \quad (3)$$

By considering both the relevance score and scoring explanation, our approach effectively models evidence and image-text pair. Strongly related evidence significantly influence the final judgment, while weaker related evidence play a supplementary role. This approach ensures that various types of evidence are utilized effectively in evaluating the consistency between the image and text, thereby improving the accuracy and reliability of the judgment.

3.4 Final Judgement Prediction

To determine the final prediction for an image-text pair, we integrate the coexistence relationship R_{co}, relevance score S_i, evidence set $E_t = \{E_{t1}, E_{t1}, \ldots, E_{tn}\}$, and image-related entities $E_e = \{E_{e1}, E_{e1}, \ldots, E_{em}\}$ into the MLLM. The final prediction label Y_p and the corresponding explanation Y_e are generated as follows:

$$(Y_p, Y_e) = \text{MLLM}(R_{co}, \{S, exp, E_t\}, \{E_e\}, (I, T)). \quad (4)$$

Here, Y_p is the final prediction label, indicating the correctness of the image-text pair match, and Y_e provides the explanation for this prediction. Incorporating entities E_e extracted from image help to contextualize the evidence and refine the analysis, enhancing the accuracy and transparency of the final judgment.

4 Experiments

4.1 Experimental Setup

4.1.1 Dataset

We evaluate CMIE and all baselines on the largest benchmark NewsCLIPpings (Luo et al., 2021), as introduced in exploratory Study.

4.1.2 Baselines

We compared the proposed method with the following state-of-the-art methods. **CLIP** (Radford et al., 2021) was fine-tuned on image-caption classification task by adding a linear classifier on the top. **DT-transformer** (Papadopoulos et al., 2023) use CLIP to separate extract image and text feature, and adds a transformer layer to enhance the features interaction. The above methods rely on the internal knowledge of the model to complete the judgement. **CCN** (Abdelnabi et al., 2022) uses external evidence as supporting information, performs circular consistency check on the image-text pair as well as on the external evidence. **SEN** (Yuan et al., 2023) propose a stance extraction network, modeling the stances of evidence from different modalities and fully leverages the roles of various types of stance. **SNIFFER** (Qi et al., 2024) utilizes explanations generated by GPT-4 to fine-tune a InstructBLIP (Dai et al., 2023), and direct inte-

grates evidence to generated predictions. **LEMMA** (Xuan et al., 2024) enhances the detection capabilities of MLLM by retrieving more relevant evidence through a query-based generation method. The judgement of such methods combines internal knowledge with external evidence.

4.1.3 Implementation Details

For CLIP and DT-Transformer, feature extraction component is ViT/B-32. For SNIFFER, we directly use the results reported in their paper. Our method is build upon the OpenAI API interface [2] and temperature is set to 0.1.

4.2 Main Results

Table. 2 reports the performance comparison between our proposed method and baselines, along with results from our exploratory experiments. Based on these results, we observe the following: **1)** Compared to previous methods that not generate human-readable explanations, MLLM relying solely on internal knowledge to assess inconsistencies between image and text outperform CLIP and DT-transformer but lag behind CCN and SEN, which incorporate external evidence. This indicates that, despite MLLM has the potential to detect OOC misinformation via world knowledge, external evidence remains crucial. **2)** After incorporating evidence, the performance of MLLM is comparable to CCN and SEN, which also use external evidence, but falls short of SNIFFER. This suggests that while MLLM augmented by evidence but without fine-tuning has ability to detect OOC misinformation, simply combining MLLM with evidence does not achieve optimal performance. This because the MLLM can not fully grasp the core thread or effectively model the relationship between evidence and the misinformation to be verified. **3)** By using CMIE to identify core thread and thoroughly consider the importance of evidence, performance improves significantly, surpassing almost all baselines. This demonstrates the importance of effectively utilizing evidence and optimizing the reasoning process. CMIE effectively filters out noise and highlights important evidence related to core thread, thereby enhancing the MLLM's judgment accuracy.

4.3 Ablation Study

Table. 3 presents the precision performance of our method with various component combinations

Table 2: Performance comparison of CMIE and representative baselines on the NewsCLIPpings benchmark.

Method	Accuracy	Precision	
		Real	Fake
CLIP	0.65	0.65	0.64
DT-Transformer	0.78	0.77	0.79
CCN	0.84	0.84	0.85
SEN	0.87	0.86	0.88
SNIFFER	0.88	**0.91**	0.87
LEMMA	0.81	0.74	0.92
DR	0.83	0.79	0.87
AR	0.85	0.86	0.83
CMIE	**0.91**	0.88	**0.93**

(Image Title, coexistence Relationship Generation(CRG), Association Scoring (AS), and Entities).

Table 3: Ablation study of CMIE components in terms of accuracy (%).

ImageTitle	CRG	AS	Entity	All	Real	Fake
✗	✗	✗	✗	0.83	0.79	0.87
✓	✗	✗	✗	0.85	0.86	0.83
✓	✓	✗	✗	0.87	0.83	0.91
✓	✓	✓	✗	0.90	0.87	0.94
✓	✓	✓	✓	0.91	0.88	0.93

When using direct reasoning, the accuracy of real samples is lower compared to fake samples. This occurs because the model tends to produce conservative outputs, classifying fake samples without clear errors as real. After adding the title evidence, the accuracy of real samples decreases because the verification core thread becomes unclear, leading to more real samples being misclassified as fake. Incorporating CRG significantly improves the accuracy of fake samples but does not enhance the accuracy of real samples due to evidence noise. With the addition of the AS, which evaluates the importance of evidences based on coexistence relationships, overall accuracy improves. This demonstrates the effectiveness of combining CRG and AS. Further inclusion of entities extracted from the image results in a slight overall performance improvement, indicating that accurate entity information also contributes to OOC misinformation detection.

In conclusion, the progressive addition of each

---

[2]https://api.openai.com/v1/chat/completions

component enhances performance. The combination of CRG, AS, and entity information significantly improves misinformation detection accuracy, emphasizing the importance of leveraging multiple information sources in a rational manner.

4.4 Adaptability to Different MLLMs

In addition to GPT-4o, we also evaluated the performance improvements brought by CMIE compared to evidence-augmented reasoning on Gemini-pro. As shown in Figure 5, CMIE enhances performance for both models. However, due to the inherent limitations in understanding and reasoning capabilities, Gemini-pro's overall performance in the OOC misinformation detection task is lower than GPT-4o.
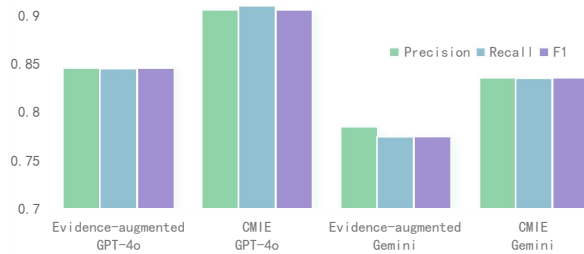

Figure 5: Performance comparison of GPT-4o and Gemini-Pro under evidence-augmented reasoning and CMIE.

4.5 Model Stability Analysis

Our framework adopts a multi-stage pipeline, raising the potential risk of error propagation, where inaccuracies in earlier stages may compromise downstream performance. To assess the stability of our approach under such risks, we conducted an analysis focused on errors introduced during Stage 1: coexistence relationship generation. We observed two common types of errors at this stage: (i) misinformation instances misclassified as having strong coexistence relationships when they should be weak, and (ii) true information instances misclassified as weak when they should be strong. According to our statistics, 1,278 misinformation samples were affected by such errors, compared to only 36 true information samples, with misinformation cases accounting for 97% of all error-propagated instances.

Table 4 presents the performance of our method on the affected misinformation samples. While some degradation in recall and F1 score is observed, the overall performance remains within an acceptable range. These results suggest that, despite up-stream errors, the association scoring mechanism in Stage 2 can still effectively leverage relevant evidence, thus contributing to the overall robustness of the framework.

Table 4: Detection performance on misinformation instances impacted by coexistence relationship misclassification in Stage 1. The results indicate that CMIE effectively mitigates upstream errors through the association scoring mechanism in Stage 2.

Precision	Recall	F1
0.98	0.70	0.82

4.6 Robustness to Prompt Variations

To evaluate the impact of different prompts on model performance, we applied two modifications to the original prompt: task description rewriting and label reversal (see Appendix A.1 for details). This design aims to assess the model's understanding of the task and evaluate the stability of CMIE. Under these two settings, the model achieved 88% and 89% accuracy, respectively, indicating that although there are slight performance differences across prompts, the overall performance remains largely stable.

4.7 Explanation Quality Assessment

We conducted human assessments to evaluate the quality of the explanations generated by MLLM. We randomly selected 50 image-text pairs and invited ten evaluators to participate. The evaluators rated the explanations generated by MLLM on a scale from 1 to 5, with prior knowledge of the labels (see Appendix A.2 for details). Higher scores were given if the evaluators considered the generated explanations to be more accurate and comprehensive; otherwise, lower scores were assigned.
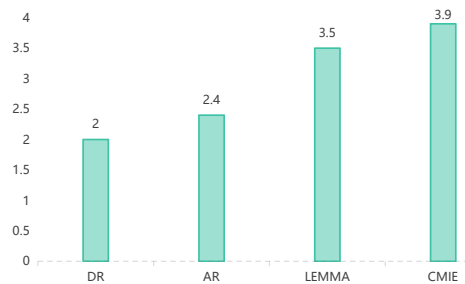

Figure 6: Human evaluation scores for OOC detection explanation quality (scale: 1–5).

Multimodal Input	Prediction

Some schools like Cleve House in Bristol say it is important to keep playground games alive

Evidences:
- Glasgow 2014: One year until city hosts Commonwealth Games - BBC News (Source：BBC)
- Unveiling of the 2014 Commonwealth Games Queen's Baton Relay ... (Source：pinterest.es)
- Glasgow 2014: The Gambia confirms Games withdrawal - Bantaba in Cyberspace (Source：gambia.dk)

DR
Label: Fake   Prediction: Real
Explanation: The text describes the importance of keeping playground games alive at Cleve House in Bristol. Without additional context or conflicting information from the image, there is no apparent misinformation detected between the text and the image.

AR
Label: Fake   Prediction: Fake
Explanation: The provided text is about the importance of playground games at Cleve House in Bristol, which is unrelated to the titles related to the Commonwealth Games and Queen's Baton Relay. Thus, there is misinformation between the given image and text.

CMIE
Label: Fake   Prediction: Fake
Explanation: The image shows children holding flags and a baton, which is related to the Commonwealth Games relay baton. The text description mentions schools keeping playground games alive, which does not directly align with the image of children participating in a Commonwealth Games-related event. The coexistence relationship analysis score is 8, indicating some alignment, but the correlation scores and explanations of the retrieved image titles suggest a stronger connection to the Commonwealth Games rather than playground games. Therefore, there is clear evidence to refute the match between the image and text.

LEMMA
Label: Fake   Prediction: Real
Explanation: The image shows children with flags, possibly at a school event. The text mentions Cleve House School in Bristol and its focus on playground games. The external resources confirm Cleve House School's existence in Bristol and mention activities like gardening, netball, and chess, but don't specifically mention traditional playground games. While the image doesn't depict playground games, it doesn't contradict the text's claim that the school values them. The image could be from a separate school event. Therefore, there is no direct misinformation between the image and the text.

Figure 7: Examples of predictions and explanations generated by the proposed CMIE.

The experimental results in Figure 6 show that the explanations generated by LEMMA and CMIE outperform those of DR and AR, with CMIE performing better than LEMMA, indicating that CMIE can produce more accurate and comprehensive explanations.

4.8 Case Study

Figure 7 presents examples of explanations generated by different methods. As shown in this figure, CMIE and LEMMA can generate more comprehensive explanations, but LEMMA is affected by evidence and fails to identify deep connections, whereas CMIE is able to uncover deeper information, resulting in more accurate outcomes.

5 Related Work

5.1 Multimodal Misinformation Detection

In the field of OOC misinformation detection, early studies (Radford et al., 2021; Papadopoulos et al., 2023) utilise pre-trained models to perform out-of-context image-text pair detection. i.e. CLIP. These methods detect misinformation using the internal knowledge of pre-trained models. Other works (Abdelnabi et al., 2022; Müller-Budack et al., 2020) aid the detection of misinformation by introducing external evidence, but such approaches do not take into account the relevance of the evidence to the task. SEN(Yuan et al., 2023) indicate that evidence stance represents a bias against results and propose a stance extraction network, modeling the stances of evidence from different modalities and fully

leverages the roles of various types of stance. However, unlike our approach, none of these methods can generate human-readable explanations, which somewhat diminishes the confidence of humans to rely on such methods for decision-making.

5.2 Retrieval Augmented Generation for OOC Misinformation Detection

Despite the significant potential demonstrated by LLMs and MLLM across various tasks (Li et al., 2024; Jin et al., 2024), they still struggle to accomplish complex tasks (Cao et al., 2023; Han et al., 2024) due to their inability to timely update knowledge and their limited understanding of domain-specific knowledge. Research has shown that leveraging Retrieval-Augmented Generation (RAG) can enhance the performance of LLMs (Ding et al., 2024; Wang and Shu, 2023; Chen and Shu, 2023). By combining retrieved external information such as wikipedia and internet, RAG is able to generate more accurate outputs. For OOC misinformation detection, existing studies (Qi et al., 2024; Xuan et al., 2024; Zhou et al., 2024) also employ Retrieval-Augmented Generation (RAG) techniques to assist in detecting misinformation by retrieving relevant external evidence. Among these studies, SNIFFER (Qi et al., 2024) requires fine-tuning based on InstructBLIP (Dai et al., 2023) to generate explanations and relies on ChatGPT (OpenAI, 2023) to generate explanation labels. This method demands significant computational resources and a large amount of labeled data,

making it both time- and resource-intensive. In contrast, LEMMA (Xuan et al., 2024) and MUSE (Zhou et al., 2024) do not require model fine-tuning. These methods also consider the relevance of external evidence to the task. However, they primarily focus on superficial lexical matching between the retrieved results and the query, lacking deeper alignment between the evidence and the image-text pair. When faced with complex OOC misinformation, they may introduce irrelevant or misleading information.

6 Conclusion

This paper explores and analyzes the performance of multimodal large language models in out-of-context misinformation detection, both using the internal knowledge and in combination with external evidence. We found that while MLLM exhibit significant potential for OOC misinformation detection, it tend to overlook key validation aspects and susceptible to noise in external evidence. Based on these findings, we proposed the coexistence relationship generation strategy and the association scoring mechanism, which are designed to help MLLM identify deep connections of image-text pair and effectively utilize relevant evidence. Experimental results demonstrate that our method outperforms baselines in most cases. Our research provides a foundation for using MLLM combined with external evidence for OOC misinformation detection, aiding in the development of more effective detection methods.

Limitations

The known limitations and possible solutions (plans) of this work are as follows:

- MLLM suffer from the hallucination problem, which may affect the detection results. To address this issue, our future work plans to further analyze the generated explanations and apply methods to reduce hallucinations.

- Like other MLLM-based methods, some of the explanations generated by CMIE differ from human evaluation standards, even though it mimics the human recognition process. We will introduce intelligent agents in future work to further improve this consistency.

Ethical Considerations

This work aims to explore the difficulty of detecting multimodal misinformation using MLLM. Therefore, our work can have a positive impact by mitigating the harm caused by the spread of misinformation in society.

The proposed CMIE relies on the capabilities of MLLM. Therefore, CMIE may produce explanations that deviate from human judgment in certain cases. This is an inherent drawback of MLLM, rather than a limitation of our method.

References

Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. 2022. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *CVPR*, pages 14940–14949.

Shivangi Aneja, Chris Bregler, and Matthias Nießner. 2021. Cosmos: Catching out-of-context misinformation with self-supervised learning. *arXiv preprint arXiv:2101.06278*.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Han Cao, Lingwei Wei, Mengyang Chen, Wei Zhou, and Songlin Hu. 2023. Are large language models good fact checkers: A preliminary study. *CoRR*, abs/2311.17355.

Canyu Chen and Kai Shu. 2023. Combating misinformation in the age of llms: Opportunities and challenges. *CoRR*, abs/2311.05656.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NIPS*.

Yujuan Ding, Wenqi Fan, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on RAG meets llms: Towards retrieval-augmented large language models. *CoRR*, abs/2405.06211.

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

Tianyang Han, Qing Lian, Rui Pan, Renjie Pi, Jipeng Zhang, Shizhe Diao, Yong Lin, and Tong Zhang. 2024. The instinctive bias: Spurious images lead to hallucination in mllms. *CoRR*, abs/2402.03757.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, and et al. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

Md Saiful Islam, Abu-Hena Mostofa Kamal, Alamgir Kabir, Dorothy L Southern, Sazzad Hossain Khan, SM Murshid Hasan, Tonmoy Sarkar, Shayla Sharmin, Shiuli Das, Tuhin Roy, et al. 2021. Covid-19 vaccine rumors and conspiracy theories: The need for cognitive inoculation against misinformation to improve vaccine adherence. *PloS one*, 16(5):e0251605.

Yizhang Jin, Jian Li, Yexin Liu, Tianjun Gu, Kai Wu, Zhengkai Jiang, Muyang He, Bo Zhao, Xin Tan, Zhenye Gan, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. 2024. Efficient multimodal large language models: A survey. *CoRR*, abs/2405.10739.

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David M. Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2023. The science of fake news. *CoRR*, abs/2307.07903.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. 2024. Multimodal foundation models: From specialists to general-purpose assistants. *Found. Trends Comput. Graph. Vis.*, 16(1-2):1–214.

Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. Newsclippings: Automatic generation of out-of-context multimodal media. In *EMNLP*, pages 6801–6817.

Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, and Ralph Ewerth. 2020. Multimodal analytics for real-world news using measures of cross-modal entity consistency. In *ICMR*, pages 16–25.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis Petrantonakis. 2023. Synthetic misinformers: Generating and combating multimodal misinformation. In *ICMR*, pages 36–44.

Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C. Petrantonakis. 2024. VERITE: a robust benchmark for multimodal misinformation detection accounting for unimodal bias. *Int. J. Multim. Inf. Retr.*, 13(1):4.

Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. 2023. Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms. In *AAAI*, pages 14444–14452.

Peng Qi, Zehong Yan, Wynne Hsu, and Mong Li Lee. 2024. Sniffer: Multimodal large language model for explainable out-of-context misinformation detection. In *CVPR*, pages 13052–13062.

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

Haoran Wang and Kai Shu. 2023. Explainable claim verification via knowledge-grounded reasoning with large language models. In *EMNLP*, pages 6288–6304.

Longzheng Wang, Chuang Zhang, Hongbo Xu, Yongxiu Xu, Xiaohan Xu, and Siqi Wang. 2023. Cross-modal contrastive learning for multimodal fake news detection. In *MM*, pages 5696–5704.

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

Linyao Yang, Hongyang Chen, Xiao Wang, Jing Yang, Fei-Yue Wang, and Han Liu. 2024. Two heads are better than one: Integrating knowledge from knowledge graphs and large language models for entity alignment. *CoRR*, abs/2401.16960.

Xin Yuan, Jie Guo, Weidong Qiu, Zheng Huang, and Shujun Li. 2023. Support or refute: Analyzing the stance of evidence to detect out-of-context mis- and disinformation. In *EMNLP*, pages 4268–4280.

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

A Appendix

A.1 Prompt Sensitive Evaluation

To evaluate the impact of different prompts on model performance, we applied two types of modifications to the prompt used for the final prediction: task description rewriting and label reversal. For task description rewriting, we modified the task instruction by rephrasing it.

Below is the original task description:

> You will receive an image and a text description, and your task is to determine if the image matches the text description. In addition, you will receive additional auxiliary information, including titles retrieved from images and matching analysis lists with images, entities extracted from images, and a simple coexistence relationship analysis. Coexistence relationship analysis shows a simple reason why images and text are matched together. Please combine other information and a given confidence score to comprehensively determine whether the given image and text are incorrectly matched.

And below is the rewritten version:

> You are provided with an image and a corresponding text description. Your task is to determine whether the image and the text are appropriately matched. To assist in this judgment, you will also receive auxiliary information, including: image-retrieved titles, image-based matching analysis lists, entities extracted from the image, and a brief analysis of their coexistence relationship—which offers a basic rationale for the image-text pairing. Please evaluate all available information, along with a given confidence score, to make a comprehensive judgment on whether the image and text are mismatched.

For label reversal, we modified the output format. In the original prompt, the model was instructed to output "Yes" when misinformation was detected. With label reversal, however, the model is prompted to output "No" under the same condition. This modification helps evaluate the model's task

comprehension and stability.

Below is the original output format:

> Generate a JSON object with two properties: 'label', 'explanation'. The return value of 'label' property should be selected from ["Yes", "No"]. Yes indicates there is misinformation between the given image and text. No indicates that there is no misinformation between the given image and text. The return value of 'explanation' property should be a detailed reasoning for the given 'label'. Please provide detailed steps for judging based on scores and other evidence.

And below is the label reversal version:

> Generate a JSON object with two properties: 'label', 'explanation'. The return value of 'label' property should be selected from ["Yes", "No"]. Yes indicates there is no misinformation between the given image and text. No indicates that there is misinformation between the given image and text. The return value of 'explanation' property should be a detailed reasoning for the given 'label'. Please provide detailed steps for judging based on scores and other evidence.

A.2 Human Evaluation Details

We conducted human evaluation in Section 4.7. Below are the instructions provided to the participants.

> You will be presented with a multimodal news item consisting of an image and its corresponding caption. Additionally, you will see the news label (true or false) along with the model-generated justification for this label. Please evaluate the justification based on its reasonableness and comprehensiveness, scoring it on a scale from 1 to 5, where a higher score indicates that you find the model's justification more thorough and convincing.

A.3 Prompts

The prompts used in the CMIE reported in the following tables respectively. The {·} in prompt represent the input content.

9352

Direct Reasoning

```
[Guidance]
You are given an image and a piece of text.  Your task is to predict whether there is
misinformation between the given image and text.

[Input]
The given text: {text}

[Output Format]
Generate a JSON object with two properties: 'label', 'explanation'. The return value of 'label'
property should be selected from ["Yes", "No"]. Yes indicates there is misinformation between
the given image and text. No indicates that there is no misinformation between the given image
and text.  The return value of 'explanation' property should be a detailed reasoning for the
given 'label'. Note that your response will be passed to the python interpreter, SO NO OTHER
WORDS! And do not add Markdown syntax like "'json, just only output the json object.

Your Response:
```

Figure 8: Prompt of direct reasoning using MLLM.

Evidence Augmented Reasoning

```
[Guidance]
You are given an image, a piece of text, and a list of titles related to the image content.
Your task is to predict whether there is misinformation between the given image and text.

[Input]
The given text: {text}

[Output Format]
Generate a JSON object with two properties: 'label', 'explanation'. The return value of 'label'
property should be selected from ["Yes", "No"]. Yes indicates there is misinformation between
the given image and text. No indicates that there is no misinformation between the given image
and text.  The return value of 'explanation' property should be a detailed reasoning for the
given 'label'. Note that your response will be passed to the python interpreter, SO NO OTHER
WORDS! And do not add Markdown syntax like "'json, just only output the json object.

Your Response:
```

Figure 9: Prompt of evidence enhanced reasoning using MLLM.

Coexistence Relationship Generation

```
[Guidance]
You are given an image, a piece of text. Your task is to provide the reason and score for the
coexistence of the picture and text.

[Input]
The given text: {text}

[Output Format]
Generate a JSON object with two properties: 'score','explanation'. The return value of 'score'
property should be the coexistence degree of image and text , the value of should be selected
from the range of [0,10].  The return value of 'explanation' property should be a detailed
reasoning. Note that your response will be passed to the python interpreter, SO NO OTHER WORDS!
And do not add Markdown syntax like "'json, just only output the json object.

Your Response:
```

Figure 10: Prompt of generating coexistence relationship.

9353

Association Scoring Mechanism

```
[Guidance]
You will receive an image, a title related to the image content, a text description as well
as a coexistence relationship about the image and text description. Your task is to score the
relevance level of each title based on images,text description and coexistence relationship.
Please first decompose the coexistence relationship into several topics to be validated, and
rate the title based on these topics combined with images. Please think step by step.

[Input]
The given text: {text}
The list of titles related to the image content: {text}
The coexistence relationship: {text}

[Output Format]
Generate a JSON object with three properties:'index', 'score', 'explanation','original title'.
The return value of 'index' property should be the index of title. The return value of 'score'
property should be the degree of correlation between the title and the image, the value of
should be selected from the range of [0,10]. The return value of 'explanation' property should
be a detailed reasoning for the given 'score'. The return value of 'original title' property
should be the original text of title.

Your Response:
```

Figure 11: Prompt of scoring the image related evidences.

Final Prediction

```
[Guidance]
You will receive an image and a text description, and your task is to determine if the image
matches the text description. In addition, you will receive additional auxiliary information,
including titles retrieved from images and matching analysis lists with images, entities
extracted from images, and a simple coexistence relationship analysis. Coexistence relationship
analysis shows a simple reason why images and text are matched together. Please combine other
information and a given confidence score to comprehensively determine whether the given image
and text are incorrectly matched.

[Input]
The given text: {text}
The list of titles related to the image content: {text}
The coexistence relationship analysis of image and text: {text}
The correlation score and explanation of retrieved image titles and images: {text}
The the entities extracted from image: {text}

[Output Format]
Generate a JSON object with two properties: 'label', 'explanation'. The return value of 'label'
property should be selected from ["Yes", "No"]. Yes indicates there is misinformation between
the given image and text. No indicates that there is no misinformation between the given image
and text. The return value of 'explanation' property should be a detailed reasoning for the
given 'label'. Please provide detailed steps for judging based on scores and other evidence.
Note that your response will be passed to the python interpreter, SO NO OTHER WORDS! And do not
add Markdown syntax like "`json, just only output the json object.

Your Response:
```

Figure 12: Prompt of generating final decision results.