SNIFFER: Multimodal Large Language Model for Explainable Out-of-Context Misinformation Detection

Supplementary Material

A. Behaviors of Open-Sourced General-Purpose MLLMs

We evaluated three representative open-sourced general-purpose MLLMs, including InstructBLIP [1], LLaVA-Adapter-V2 [2] and Woodpecker [6], for their performance in detecting out-of-context misinformation. As demonstrated in Table 4, despite explicit instructions regarding the desired output format, these MLLMs exhibit a *failure to follow instructions accurately*. InstructBLIP, in particular, does not provide a clear judgment, instead merely paraphrasing the given caption, which indicates a *misunderstanding of the user's intent*.

Furthermore, both LLaVA-Adapter-V2 and Woodpecker display *hallucination* phenomena: LLaVA-Adapter-V2 erroneously attributes the content of the caption to the image and assigns a new, erroneous meaning (*i.e.* "the Apollo 11 mission, which took place in 1969") to the caption; Woodpecker attributes an incorrect new context to the image (*i.e.* "a rocket launching in the 1980s"). This hallucination effect may be attributed to the training data, where text and image are generally aligned to depict the same event.

We also analyzed InstructBLIP's lexical preferences based on the descriptions it generated for images within the NewsCLIPpings dataset. Our statistics reveal that in samples containing person nouns, only 27% of the responses utilize fine-grained proper nouns (*i.e.* specific names of individuals), while the remaining 73% employ coarse-grained common nouns (such as "person", "woman" and "man"). This suggests that InstructBLIP favors general nouns over specific proper nouns in its responses.

B. Instruction Data Construction

Instructions for brief image description. Table 5 shows the ChatGPT-generated questions to construct the diverse instruction data for news domain alignment. They present the same meaning with natural language variance.

Prompt to generate the OOC instruction. Figure 10 illustrates the prompt utilized for asking GPT-4 to identify inconsistencies between cap_1 and cap_2. From the responses, we extract $element$, ent_t, and ent_v to formulate the OOC instruction data, as depicted in Figure 4 in the main paper.

C. Evaluation on Real-World OOC Misinformation

Since SNIFFER is trained on the synthetic OOC dataset, we further performed additional experiments on the VERITE [4] dataset to evaluate its performance on detecting real-world OOC misinformation. VERITE [4] is a newly released dataset collected from fact-checking websites (*i.e.* Snopes and Reuters). It also excludes asymmetric multimodal misinformation and employs modality balancing to avoid unimodal bias. Results show that SNIFFER achieves 74.0% accuracy in distinguishing the True and OOC samples, surpassing both the baseline (72.7%) proposed in [4] and the SOTA method (*i.e.* RED-DOT [3]) which uses additional visual evidence (73.9%).

D. Case Studies

We show three representative cases in Table 6 to reveal the behavioral differences between GPT-4V and SNIFFER.

In the first example, SNIFFER identifies the claim as fake based on both the image-text inconsistency and the claim-evidence irrelevance. In contrast, GPT-4V fails to detect any inconsistencies. This highlights SNIFFER's superiority in recognizing news entities and utilizing external knowledge.

In the second example, the image depicts a scene that aligns with the caption's description of "people cheer and take pictures" and no relevant webpages were retrieved for this image, leading SNIFFER to incorrectly classify this claim as real. However, GPT-4V, drawing on the Pope's attire and the security conditions at the event, deduces that the image does not depict a scene of the Pope participating in an event, and thus correctly classifies the claim as fake. This example demonstrates GPT-4V's superior world knowledge and reasoning capabilities, far surpassing those of smaller models.

The third example presents a real news story about passengers stranded outside the Guangzhou railway station, accompanied by an image of people queuing. SNIFFER, evaluating both the image-text consistency and the claim-evidence relevance, correctly identifies this news as real. In contrast, GPT-4V erroneously classifies this news as fake, reasoning that the orderly crowd in the image does not resemble the expected chaos of stranded passengers and noting the absence of any railway station signage. This instance illustrates GPT-4V's overly cautious and sensitive judgment, predicated on the expectation that key elements

Table 4. Behaviors of existing open-sourced general-purpose MLLMs in detecting OOC misinformation.

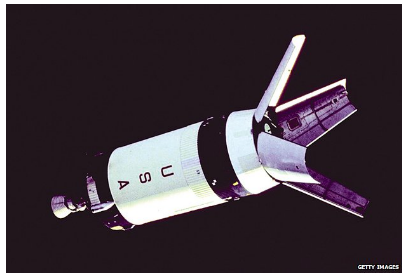
User	Some rumormongers use images from other events as illustrations of the current news event to make up multimodal misinformation. Given a news caption and a news image, judge whether the given image is rightly used in the given news context. Let's analyze their inconsistency from perspectives of main news elements, including time, place, person, event, artwork, etc. You should answer in the following forms: "Yes, the image is rightly used." or "No, the image is wrongly used in a different news context. The given news caption and image are inconsistent in <element>. The <element>in the caption is <ent_t>, and the <element>in the image is <ent_v>." News caption: Saturn V booster was used in Nasa space missions between 1967 and 1972. The answer is
	
InstructBLIP	1967-1972, which is the time period when the Saturn V booster was used in Nasa space missions.
LLaVA-Adapter-V2	No, the image is wrongly used in a different news context. The image is of a Saturn V rocket, which was used in NASA space missions between 1967 and 1972. However, the given news caption refers to a different event, the Apollo 11 mission, which took place in 1969. The image and caption are inconsistent in time, as the Saturn V rocket was not used in the Apollo 11 mission.
Woodpecker	No, the image is wrongly used in a different news context. The given news caption and image are inconsistent in time. The caption states that the Saturn V booster was used in Nasa space missions between 1967 and 1972, while the image depicts a rocket launching in the 1980s. The image is not related to the Saturn V booster, which was used in the 1960s and 1970s.

Table 5. The list of instructions for brief news image description.

- "Analyze the news image and provide a brief summary of the event it depicts."
- "Give a short description of the scene shown in the news photograph."
- "Offer a concise report based on the news-related image provided."
- "Summarize the news story as represented by the image."
- "Interpret the journalistic image and detail the key elements in a succinct manner."
- "Provide a brief journalistic overview of the news picture."
- "Construct a short narrative to convey the news event shown in the image."
- "Deliver a compact exposition of the incident captured in the news photo."
- "Express the main news points illustrated by the given photograph."
- "Condense the news context of the image into a clear, brief description."
- "What news event does this image describe?"

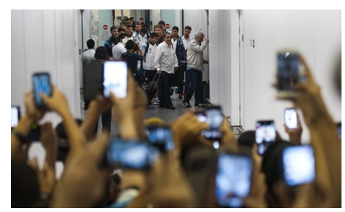
of a real news story should be visibly represented in the accompanying image. In fact, existing research [5] has found that more than half of social media news stories do not have their content words represented in the images. This tendency also accounts for GPT-4V's lower recall rate for real news samples, as outlined in Table 3 in the main paper.

In summary, SNIFFER integrates clues from both text-image inconsistency and claim-evidence relevance, leading to more comprehensive judgments. Benefiting from task-specific tuning, it demonstrates a deeper understanding of the complex logic needed for the OOC detection task. On the other hand, GPT-4V, with its vast repository of world knowledge, is adept at detecting subtle anomalies, yet this same attribute can lead to the misclassification of real news.

```
# system message
You are an expert in fact-checking. Some news captions and accompanying images are
inconsistent in terms of key news elements (5W1H) because rumormongers have taken images from
 other news and used them as illustrations for current news to make up multimodal
misinformation. Given the original news caption (i.e. caption_ori) for the image (i.e.
image_ori), the new news caption (i.e. caption_new), and a basic description of image_ori's
content, I need you to analyze the inconsistencies between caption_ori and caption_new in key
 news elements, and select from them the one most likely inconsistency between image_ori and
caption_new based on the description of image_ori. You should answer in the following forms:
 "They are inconsistent in <element>. The <element> in caption_new is <ent_t>, and the <
element> in image_ori is <ent_v>.\n Element: <element> \n Entity_caption: <ent_t> \n
Entity_image: <ent_v>\". Please answer only one inconsistent element.

# in-context examples
Caption_ori:  John Constable's Brightwell Church and Village was part of the 2013 exhibition
Caption_new:  From J Charles Eichhorn's American Skat or The Game of Skat Defined
Basic description of image_ori: This image describes a rural landscape with a farmhouse, a
barn, and a field. The farmhouse is situated in the middle of the field, surrounded by the
barn and the open land. The painting captures the essence of a peaceful, pastoral setting,
with the farmhouse serving as the central focus of the scene.
The answer is: They are inconsistent in artwork. The artwork in caption_new is American Skat
or The Game of Skat Defined, and the artwork in image_ori is Brightwell Church and Village. \
n Element: artwork \n Entity_caption: American Skat or The Game of Skat Defined \n
Entity_image: Brightwell Church and Village

Caption_ori:  Chris Huhne is among the ministers expected to address delegates at next week's
 Lib Dem conference
Caption_new:  Urs Rohner CEO of Credit Suisse participates in a panel session in Bern
Switzerland on Tuesday
Basic description of image_ori: The image depicts a man wearing a suit and tie, standing at a
 podium with a microphone in front of him. He appears to be giving a speech or addressing an
audience. In the background, there is a black screen or backdrop.
The answer is: They are inconsistent in person. The person in caption_new is Urs Rohner, and
the person in image_ori is Chris Huhne. \n Element: person \n Entity_caption: Urs Rohner \n
Entity_image: Chris Huhne

# query
Caption_ori:  <CAPTION_ORI>
Caption_new:  <CAPTION_NEW>
Basic description of image_ori:  <BASIC_DESCRIPTION_OF_IMAGE_ORI>
The answer is:
```

Figure 10. Prompt used to ask GPT-4 to generate the OOC instruction data.

Table 6. Comparison between GPT-4V and SNIFFER. For each case, we annotate the ground truth (GT) and the correctness of the model's judgment, indicated by either ✓(correct) or ✗ (wrong).

Caption: Nick Clegg (left) in Oldham with Liberal Democrat candidate Elwyn Watkins (centre) and Lib Dem deputy leader Simon Hughes. [GT: FAKE]		
GPT-4V: Yes, the image is rightly used. ✗	**SNIFFER**: No, the image is wrongly used in a different news context. On the one hand, the person in the caption, Nick Clegg, Simon Hughes and Elwyn Watkins, do not match the person in the image, Tim Henman. On the other hand, the caption refers to Nick Clegg, Simon Hughes, and Elwyn Watkins attending an event, while the image-retrieved webpages are about the funeral of a tennis player named Elena Baltacha. The two entities are not connected, and there are no common elements or context that link them. Therefore, the image is more likely to be wrongly used in the caption. ✓	
Caption: People cheer and take pictures as the pope arrives to meet with representatives of the World of Work organization in Ciudad Juarez. [GT: FAKE]		
GPT-4V: No, the image is wrongly used in a different news context. The given news caption and image are inconsistent in person. The person in the caption is the pope, and the person in the image is a man who does not appear to be the pope. The man seen in the image is not dressed in papal attire, which is distinctive and would typically include white robes and sometimes a zucchetto (skullcap) for the pope. Additionally, the context of the picture does not seem to match the typical scenarios where the pope is present, which often feature heightened security and more ceremonial settings, depending on the occasion. ✓	**SNIFFER**: Yes, the image is rightly used. ✗	
Caption: Passengers are stranded outside the Guangzhou railway station in China's Guangdong province on Tuesday. [GT: REAL]		
GPT-4V: No, the image is wrongly used in a different news context. The event in the caption is passengers stranded outside the Guangzhou railway station, and the event in the image is a large gathering of people at an event that is not related to a transportation disruption. This can be inferred from the fact that the people appear to be standing very close to one another in an organized manner, rather than having the disorganized, possibly frustrated behavior one might expect from stranded passengers. Additionally, there are no clear signs of a railway station, such as platforms, tracks, or trains, visible in the image. The setting does not match the context described in the news caption. ✗	**SNIFFER**: Yes, the image is rightly used. On the one hand, the image depicts a group of people, which is consistent with the caption. On the other hand, the image-retrieved webpages mention the Guangzhou railway station in China, the Chinese New Year and a situation where passengers were stranded outside the railway station due to snow, which are relevant to the caption. Therefore, the image is likely to be correctly used in the caption. ✓	

References

[1] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500, 2023. 1

[2] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. LLaMA-Adapter V2: parameter-efficient visual instruction model. *CoRR*, abs/2304.15010, 2023. 1

[3] Stefanos-Iordanis Papadopoulos, Christos Koutlis, Symeon Papadopoulos, and Panagiotis C. Petrantonakis. RED-DOT: multimodal fact-checking via relevant evidence detection. *CoRR*, abs/2311.09939, 2023. 1

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

[5] Alakananda Vempala and Daniel Preotiuc-Pietro. Categorizing and inferring the relationship between the text and image of Twitter posts. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2830–2840. Association for Computational Linguistics, 2019. 2

[6] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *CoRR*, abs/2310.16045, 2023. 1