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

Peng Qi Zehong Yan Wynne Hsu Mong Li Lee
National University of Singapore
peng.qi@nus.edu.sg, e0787894@u.nus.edu, {whsu, leeml}@comp.nus.edu.sg

https://pengqi.site/Sniffer

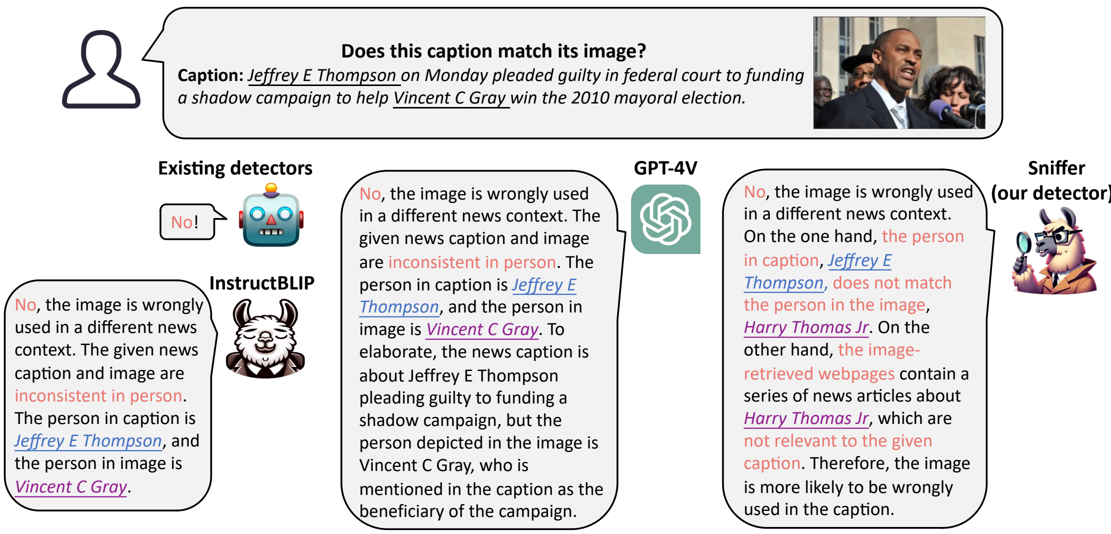


Figure 1. Comparison between the proposed SNIFFER and other detectors[1]. In this out-of-context misinformation, the individual in the image is *Harry Thomas Jr*, which contradicts the caption. Existing detectors often give a judgment without explanation. While InstructBLIP and GPT-4V correctly identify the inconsistent news element (*i.e*. person) in the given image-text pair, they mistakenly associate the person in the image with a different individual mentioned in the caption. In contrast, SNIFFER analyzes both the consistency of the image-text content and the claim-evidence relevance, and accurately identifies the person in the image as *Harry Thomas Jr*, thereby providing a precise and persuasive explanation.

Abstract

Misinformation is a prevalent societal issue due to its potential high risks. Out-Of-Context (OOC) misinformation, where authentic images are repurposed with false text, is one of the easiest and most effective ways to mislead audiences. Current methods focus on assessing image-text consistency but lack convincing explanations for their judgments, which are essential for debunking misinformation. While Multimodal Large Language Models (MLLMs) have rich knowledge and innate capability for visual reasoning and explanation generation, they still lack sophistication in understanding and discovering the subtle cross-modal differences. In this paper, we introduce SNIFFER, *a novel multimodal large language model specifically engineered for OOC misinformation detection and explanation.* SNIFFER *employs two-stage instruction tuning on InstructBLIP. The first stage refines the model's concept alignment of generic objects with news-domain entities and the second stage leverages OOC-specific instruction data generated by language-only GPT-4 to fine-tune the model's discriminatory powers. Enhanced by external tools and retrieval,* SNIFFER *not only detects inconsistencies between text and image but also utilizes external knowledge for contextual verification. Our experiments show that* SNIFFER *surpasses the original MLLM by over 40% and outperforms state-of-the-art methods in detection accuracy.* SNIFFER *also provides accurate and persuasive explanations as validated by quantitative and human evaluations.*

1. Introduction

In recent years, Deepfake and other media manipulation technologies [43, 46, 47, 53] have garnered considerable at-

[1]Due to space constraints, we have made minor edits to the models' responses for brevity without altering their original meaning.

tention from both the computer vision community and the general public, due to their lifelike qualities and their significant influence in accelerating the spread of fake news. Nevertheless, one of the easiest and most prevalent ways to mislead audiences is to use unaltered images in a new but false or misleading context, known as out-of-context (OOC) misinformation [13]. For example, during the recent Israel-Hamas war, numerous instances of OOC misinformation were observed on social media. These often involved repurposing old images from unrelated armed conflicts or even military footage from video games [2]. Detecting such OOC misinformation poses a unique challenge since the visual content remains authentic, and the deception stems solely from the context created by combining these images with misleading or incorrect text.

Faced with this challenge, current studies [3, 16, 17, 29, 31, 36, 42] focus on learning a unified latent representation space to assess the consistency of image-text pairs or compare them with external references. While these approaches have made some progress, the development of convincing explanations for these judgments, critical for establishing public trust and more effectively debunking misinformation [8, 26], still remains unexplored. Although some works [3, 29] visualize salient objects and words based on the model's attention weights, it is still unclear why these highlighted regions signify misinformation. Therefore, it is important to develop technologies capable of not only detecting but also explaining out-of-context misinformation.

Multimodal Large Language Models (MLLMs) have made great advancements in a variety of multimodal tasks [19, 50]. With their extensive world knowledge and robust visual reasoning and generation capabilities, MLLMs have the potential to detect factual inconsistency in image-text pairs and to generate coherent, natural language-based explanations. However, *applying existing MLLMs to the task of OOC misinformation detection is non-trivial*. *On the one hand*, initial experiments on some open-sourced MLLMs (such as InstructBLIP [10]) revealed common shortcomings, including failure to follow instructions, misunderstanding of user's intent, and hallucination (detailed in Supplementary). These shortcomings may stem from the fundamental differences between the training corpus for MLLMs and the requirements of the specific task. In OOC detection, the model must discern when text and images describe disparate news events, which are typically not present in MLLMs' training data. *On the other hand*, the news event from which the image originates may not be discernible from the image itself. For example, news images accompanying reports of public figures speaking at various events are generally similar close-up shots of the individuals, making it almost impossible to distinguish from the image content alone which news scene or event the image is from. MLLMs have inherent limitations as they lack the ability to access real-time information or utilize external tools for identifying and contextualizing events.

To address these challenges, we introduce SNIFFER, an MLLM specifically engineered for detecting OOC misinformation. We utilize language-only GPT-4 [34] to generate the instruction data that includes both judgments and explanations, and employ two-stage instruction tuning on InstructBLIP. We first refine the model's concept alignment of generic objects with news-domain entities by image captioning data, and then leverage the OOC-specific data to fine-tune the model's discriminatory powers. Through task-specific tuning, the inherent knowledge within the model can be activated and reorganized to align with the specific logic required by our task. This is the first attempt to extend multimodal instruction tuning to the news domain for end-to-end training of an OOC misinformation detector.

Moreover, SNIFFER is augmented with external knowledge through retrieval and tool usage. It conducts *internal checking* to spot inconsistencies between text and image, and *external checking* to reason between the given text and retrieved image context. The judgments and explanations derived from these two perspectives are then integrated into the LLM module for a more comprehensive and unified output. Empowered by multimodal instruction tuning and enhanced with external knowledge, SNIFFER demonstrates a significant performance improvement, exceeding the original MLLM by over 40% and surpassing current state-of-the-art methods. Beyond accurate judgments, SNIFFER also excels in providing precise and persuasive explanations, as evidenced by quantitative analysis and human evaluations.

Our contribution can be summarized as follows:

- We design a novel data reformation pipeline assisted by language-only GPT-4 to convert the given OOC image-text pairs into the appropriate instruction-following format with judgments and explanations simultaneously.
- We propose a practical approach to adapt existing general-purpose MLLMs for out-of-context misinformation detection through two-stage instruction tuning. Our task-specific MLLM, SNIFFER, enhanced with external tools and retrieval, effectively models both the internal image-text clues and the external claim-evidence clues for simultaneous OOC detection and explanation.
- Extensive experiments show that SNIFFER significantly surpasses the original MLLM and SOTA methods in detection performance, achieves comparable results with just 10% of the training data, and provides precise and persuasive explanations validated by both quantitative and human assessments.

2. Related Work

Out-of-Context Misinformation Detection. Existing multimodal misinformation detection methods extract and fuse features from different perspectives, such as linguistic patterns [37], image tampered patterns [7, 38, 43], multi-modal inconsistency [39, 54], user response [44], propagation structure [30], and publisher profile [40, 45] to classify the given news as real or fake. Out-of-context misinformation, also known as image repurposing and cheapfake, is a specific form of misinformation that is easy to create and highly misleading. It reuses authentic images within a similar yet incorrect context, rendering many traditional detection methods ineffective [52].

To detect OOC misinformation, some methods utilize knowledge-rich pre-trained models to conduct *internal checking* for the given image-text pair. For example, [29, 36] use the multimodal pre-trained models CLIP [41] and VisualBERT [23] to classify. Other methods employ external resources to do *external checking*. For example, [16, 17, 42] use a reference dataset that contains unmanipulated related claims to mimic world knowledge, and then detect the OOC use by comparing the given claim with the retrieved one. [3] uses the text and image to retrieve related Web evidence separately, and compute the claim-evidence consistency under both textual and visual modalities. Similarly, [31] proposes an unsupervised cross-modal entity consistency verification method, which retrieves images using entities extracted from the given text, and then calculates the similarity between the entities in the given image and retrieved images. Different from judging the veracity of the given image-caption claim, [5] re-formulates this task as: If two captions refer to the same objects in the same image but are semantically different, then it indicates an out-of-context use of this image. They introduce a self-supervised training strategy to train the model's visual grounding ability and evaluate the model on OOC samples.

Although existing works have made some progress, they often fail to explain their judgment. This lack of transparency hinders the effectiveness of these methods in debunking misinformation. [52] proposes an interpretable decontextualization detector, which uses MLLM to verify the sub-questions decomposed from the given text and selects supported question-answer pairs as the explanation. However, the effectiveness of this method is heavily limited by the ability of existing general-domain MLLM. In contrast, we use a two-stage instruction tuning to adapt the general-domain MLLM for the OOC detection task, enabling it to concurrently generate both judgment and explanation.

Instruction Tuning for MLLMs. Multimodal large language models typically use a lightweight visual prompt generator that produces soft prompts for the input images to connect to an existing LLM. Early models [4, 22, 55] focus on large-scale pre-training, while recent works [10, 18, 27] employ instruction tuning for pre-trained MLLMs. Instruction tuning is a crucial technique to enhance the capabilities and controllability of large language models, which involves further training pre-trained models on a collection of instruction-formatted datasets to enhance models' generalization to unseen tasks. One challenge of instruction tuning is how to construct high-quality instructions that properly cover the desired target behaviors. In addition to adapting existing benchmarks [10], some works [18, 27] collect samples through self-instruction [49], which bootstraps LLMs to generate textual instruction-following data using a few hand-annotated samples. The success of LLMs in the general domain has inspired interest in applications such as biomedical [20], law [15] and education [51]. In this work, we extend the ability of general-domain MLLM for OOC misinformation detection via instruction tuning.

3. Method

Our goal is to develop an explainable multimodal out-of-context misinformation detection model that can jointly output the prediction and explanation. Figure 2 illustrates the overall architecture of the proposed SNIFFER. Specifically, for a given image-text pair, we feed it into the multimodal large language model for checking the internal cross-modal inconsistencies. The image-retrieved textual evidence will be fed into the embedded LLM module with the input text to judge the external claim-evidence relevance. Similar to ensemble learning, the LLM module will output the final judgment and explanation drawing on the results from both internal and external checks. We first introduce the base MLLM and the two-stage instruction tuning procedures, before describing the reasoning process which includes internal/external checking, and composed reasoning.

3.1. Base MLLM

We employ InstructBLIP [10], a general-purpose vision-language model as the base MLLM for further tuning. InstructBLIP consists of an image encoder, an LLM, and a Query Transformer (Q-Former). As shown in Figure 2, Q-Former extracts instruction-aware visual features based on the output of the frozen image encoder, and feeds the visual features as soft prompt input to the frozen LLM. The learnable query vectors interact with the instruction text through self-attention layers, and interact with the frozen image embeddings through cross-attention layers. Similar to BLIP-2 [22], Q-Former is pre-trained for vision-language representation learning and vision-to-language generative learning and then is tuned for multi-task vision-language instruction learning in [10].

3.2. Instruction Tuning

We use a two-stage training procedure to adapt the general-domain InstructBLIP to the news domain and OOC detection task sequentially (see Figure 3).

Stage 1: News Domain Alignment. We observe that InstructBLIP tends to respond with coarse-grained nouns (*e.g.*

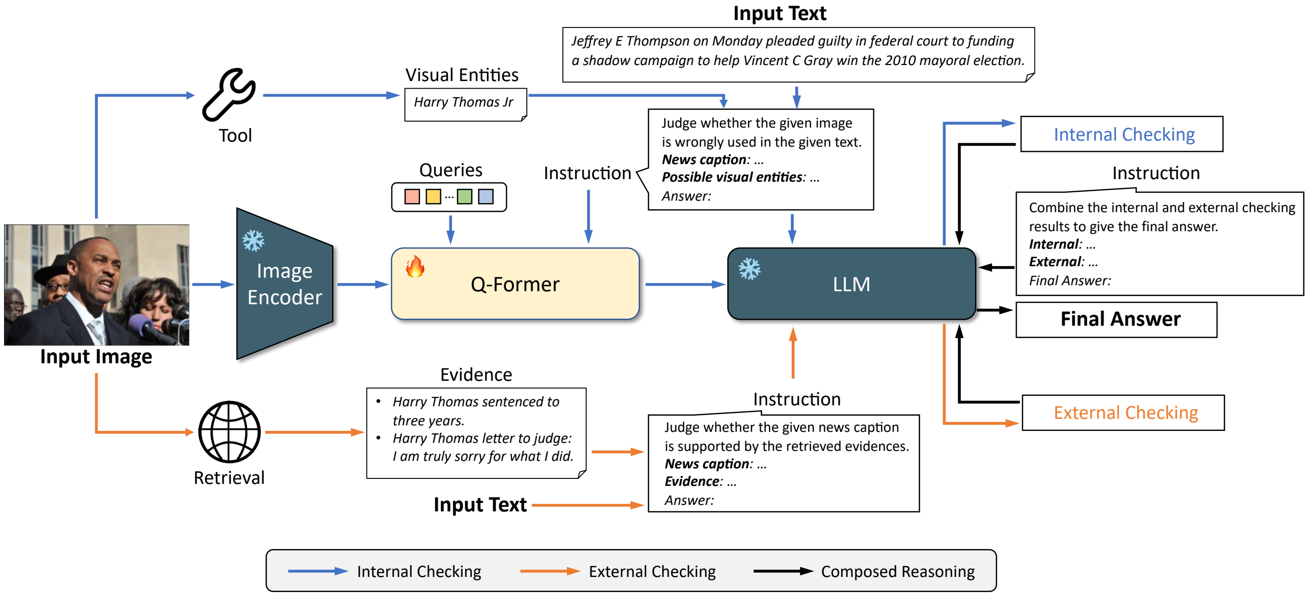


Figure 2. Architecture of the proposed framework SNIFFER. For a given image-text pair, SNIFFER conducts a two-pronged analysis: (1) it checks the consistency of the image and text content (*internal checking*), and (2) it examines the relevance between the image-retrieved evidence and the provided text (*external checking*). The outcomes of both these verification processes are then considered by SNIFFER to arrive at a final judgment and explanation (*composed reasoning*).

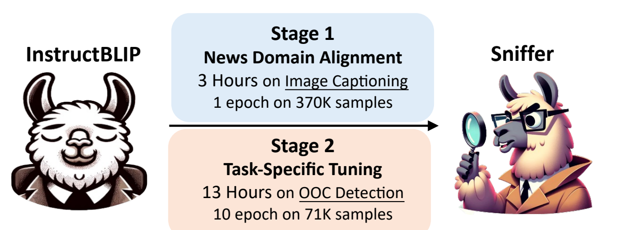


Figure 3. SNIFFER was initialized with the general-domain InstructBLIP and then continuously trained to adapt it to the news domain and OOC misinformation detection task sequentially.

"person", "woman", and "man") rather than fine-grained, specific names such as "Donald Trump". OOC samples are typically created by replacing original entities with ones that are similar but not identical, and these subtle distinctions appear to be difficult for InstructBLIP to capture.

Given the cross-domain differences in the lexical preference, we construct the news-domain instruction dataset to adapt the general-domain InstructBLIP to the news domain. This instruction dataset is curated from the NewsCLIPpings dataset[2] [29] which consists of 368,013 unique news image-caption pairs covering a diverse and representative set of news-domain concepts. Inspired by LLaVA [27], we leverage ChatGPT4 [33] to construct the instruction-following data from these image-caption pairs. Specifically, to keep the diversity of instruction data, we prompt ChatGPT4 to generate 11 questions (see Supplementary) with the intent to instruct the model to describe the image content. Given that these captions have less than 30 words, we stipulate that the generated questions should explicitly specify a "brief" description of the images. For an image $\mathbf{I}$ and its associated caption $\mathbf{T}_c$, we randomly sample one question $\mathbf{T}_q$ to form the corresponding instruction as

$$\textbf{Human}\text{: } \mathbf{I}\mathbf{T}_{\mathrm{q}}\langle\mathrm{STOP}\rangle;\ \ \textbf{Model}\text{: } \mathbf{T}_{\mathrm{c}}\langle\mathrm{STOP}\rangle. \tag{1}$$

During training, we keep both the image encoder and LLM frozen, and only update the Q-former. In this way, the image features of vast news-domain visual concepts can be aligned to the textual embeddings of their corresponding fine-grained entity names in the pre-trained LLM.

Stage 2: Task-Specific Tuning. In out-of-context misinformation, images are repurposed with false text, which is different from the typical visual-language tasks that general-purpose MLLMs focus on. These tasks, such as image captioning and visual question answering, typically involve text and images that correspond to the same event. To address this, we construct the task-specific instruction data to progressively adapt the news-domain MLLM (developed in Stage 1) to the task of OOC misinformation detection. Similar to Stage 1, we only update the parameters in Q-former.

One of the primary challenges in creating an explainable OOC misinformation detector is the lack of supervised data that includes both judgments and explanations. The NewsCLIPpings dataset generates fake pairs by replacing the img_1 in a real pair (cap_1, img_1) with another img_2 from

[2] We exclude the validation and testing set to prevent data leakage.

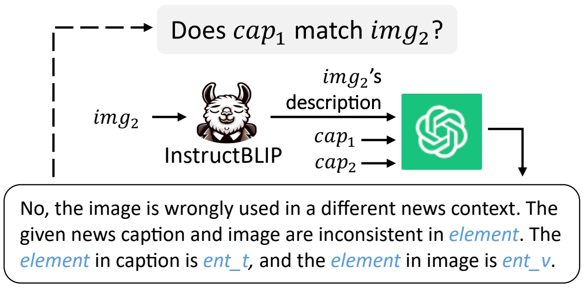


Figure 4. Process of OOC instruction generation.

a similar but different pair (cap_2, img_2), resulting in a fabricated pair (cap_1, img_2). While labels indicating their authenticity are provided for these pairs, the specific inconsistencies between cap_1 and img_2—the crux of the misinformation—are not explicitly identified. As such, we innovatively extract the inconsistencies between cap_1 and cap_2 by prompting language-only GPT-4 as an alternative. As shown in Figure 4, given cap_1, cap_2, and the InstructBLIP-generated img_2's basic description, we prompt GPT-4 to generate the inconsistencies between cap_1 and img_2 as if it could see the image (even though it only has access to the text). We curate few-shot examples in the prompt and also restrict the output format, requiring it to spot the news elements that are inconsistent between cap_1 and img_2, as well as the specific entities. The complete prompt is shown in Supplementary. Although there may be multiple inconsistent elements between cap_1 and img_2, we require GPT-4 to generate only the most likely one for clarity. In total, we obtain 35,536 GPT-assisted instructions for out-of-context samples and supplement the real samples with an equal number of instructions (*i.e.*, "No, the image is rightly used in the given news context. ")

3.3. Reasoning Process

To effectively tackle the challenge of OOC misinformation detection, the reasoning process in SNIFFER employs a comprehensive strategy that integrates both internal and external verification methods.

Internal Checking. The two-stage instruction tuning process equips SNIFFER with the ability to identify image-text inconsistency for internal checking. However, the model does not have access to up-to-date information as it is limited by the training corpus. Therefore, we employ Google Entity Detection API [1] to recognize visual entities in images and then incorporate them into the instruction as supplementary information.

External Checking. In addition to image-text internal checking, leveraging retrieved web evidence for external verification is a crucial step, as highlighted in [3, 56]. The context of an image, particularly the news text in which the image has previously appeared, serves as a vital supplement to the content of the image itself. Reverse image searches can often reveal the original event where the image was first reported, to help verify if the given text matches the true context of the image. Given the strong analytical reasoning ability of LLM, we input both the news caption and the text from webpages retrieved via image search [1] into the LLM module in SNIFFER. The LLM is then tasked with determining whether the provided news caption is supported by the retrieved evidence.

Composed Reasoning. The internal and external verification steps analyze the input image-text pairs from different perspectives and may yield different conclusions. Therefore, we employ the LLM once again for an interpretable model ensemble, tasking it to deliver its final decision based on both outcomes and the initial caption, and to clearly explain its decision-making process.

4. Performance Study

In this section, we conducted experiments to evaluate the effectiveness of SNIFFER. Specifically, we aim to answer the following evaluation questions:

Q1: Can SNIFFER improve the performance of out-of-context misinformation detection?
Q2: How effective are the different modules of SNIFFER in detection?
Q3: Can SNIFFER generate accurate and convincing explanations for their judgment?
Q4: How does SNIFFER perform in early detection, when the number of training samples is limited?
Q5: How does SNIFFER perform on other datasets?
Q6: How does SNIFFER perform compared to GPT-4V?

4.1. Experimental Setup

Dataset. We use the largest out-of-context misinformation detection benchmark NewsCLIPpings [29]. This dataset is built based on VisualNews [25], a large-scale corpus that contains image-caption pairs from four news agencies (The Guardian, BBC, USA Today, and The Washington Post). The out-of-context samples are automatically generated by replacing the images in the original image-caption pairs with retrieved images that are semantically related but belong to different news events. [3] extends the NewsCLIPpings dataset by supplementing the retrieved textual and visual evidence. Here, we use part of the textual evidence and detected visual entities provided in [3].

Following previous works [3, 36, 52], we report the results on the Merged/Balance subset, which has a balanced proportion of different retrieval strategies and positive/negative samples. The number of samples in the training, validation, and testing sets are 71072, 7024, and 7264, respectively. As in [29], we use the accuracy over all samples (All) and separately for the Fake (Out-of-Context) and Real (Not Out-of-Context) samples as evaluation metrics.

Implementation Details. We select InstructBLIP [10] as the base MLLM, of which the image encoder is ViT-G/14 [12] and the LLM is Vicuna-13B [9]. During training, we initialize the model from the pre-trained InstructBLIP and only finetune the parameters of Q-Former while keeping both the image encoder and the LLM frozen. To reduce the memory cost, we use FlashAttention-2 [11] to replace the standard Attention layer in LLM. Our implementation uses the LAVIS library [21]. The batch size is set as 8 and 4 in the stages of news domain alignment and task-specific tuning, respectively. The max input sequence length is 550 and the output length is 256. We use the AdamW [28] optimizer and apply a linear warmup of the learning rate, increasing from 10^{-8} to 10^{-5}, followed by a cosine decay. The models are trained utilizing 4 Nvidia A100 (40G) GPUs.

Baselines. We compared SNIFFER with two representative multimodal misinformation detectors trained from scratch: 1) **SAFE** [54] which translates the input image into a sentence, and computes the multimodal relevance based on the sentence similarity as the auxiliary loss; 2) **EANN** [48] which uses adversarial training to guide the model to learn event-invariant multimodal features for detection.

We also compared SNIFFER with pre-trained multimodal baselines: 1) **VisualBERT** [23], one of the earliest works on multimodal pre-training, concatenates the bounding box features and textual tokens. The combined features are fed into a unified encoder, consisting of a series of transformer layers, to align them into one embedding space. The pre-training of VisualBERT involves masked language modeling and image-text matching objectives; 2) **CLIP** [41] passes image and text through separate encoders and uses contrastive loss to guide the multimodal encoders to generate similar representations for related concepts; 3) **DT-Transformer** [36] uses CLIP as the multimodal encoder and adds auxiliary Transformer layers to enhance the multimodal features interaction. It combines different types of generated data as the training corpus; 4) **CCN** [3] proposes a consistency-checking network assisted by CLIP, which considers the consistency of the claim-evidence (image-image and text-text), in addition to the image-text pairing. This is the only baseline that utilizes retrieved external evidence; 5) **Neu-Sym detector** [52] proposes a neural-symbolic model which symbolically disassembles the text into a set of fact queries, and forwards the query-image pairs into a pre-trained MLLM. The output answers are further selected and combined to obtain the final judgment.

4.2. Performance Comparison (Q1)

Table 1 shows the performance of SNIFFER and various baselines. We observe that: 1) SNIFFER outperforms all baselines, which validates that it can effectively capture the subtle inconsistency in the OOC samples; 2) Even though SNIFFER only considers part of the textual evidence, it still surpasses CCN by over 3.7%; 3) The inferior performance of the Neu-Sym detector suggests that general-purpose MLLM may not be well-suited for the OOC detection task; 4) Baselines trained from scratch (*i.e.* EANN and SAFE) are worse than pre-trained multimodal baselines, verifying the importance of world knowledge in the pre-trained models in distinguishing the OOC samples.

Table 1. Performance (%) comparison between SNIFFER and baselines. The best results are in **boldface**.

Method	All	Fake	Real
SAFE	52.8	54.8	52.0
EANN	58.1	61.8	56.2
VisualBERT	58.6	38.9	78.4
CLIP	66.0	64.3	67.7
DT-Transformer	77.1	78.6	75.6
CCN	84.7	84.8	84.5
Neu-Sym detector	68.2	-	-
SNIFFER (*Ours*)	**88.4**	**86.9**	**91.8**

Table 2. Ablation studies on each component in SNIFFER.

InstructBLIP	PT	OOC Tuning	VisEnt	Evidence	All	Fake	Real
✓	✗	✗	✗	✗	47.4	4.6	90.3
✓	✓	✗	✗	✗	49.3	9.4	89.2
✓	✗	✓	✗	✗	82.5	75.3	89.7
✓	✗	✓	✓	✗	87.6	83.9	91.3
✓	✓	✓	✗	✗	83.1	76.5	89.6
✓	✓	✓	✓	✗	88.2	84.9	94.0
✓	✗	✗	✗	✓	84.5	**92.9**	76.0
✓	✓	✓	✓	✓	**88.4**	86.9	**91.8**

4.3. Ablation Studies (Q2)

We conducted ablation experiments to analyze the importance of each component of SNIFFER in detecting OOC misinformation. Specifically, we first tested the original **InstructBLIP**, and then incrementally integrated various components: pre-training (**PT**) with news-domain data in Stage 1, task-specific tuning based on OOC data in Stage 2 (**OOC tuning**), visual entities (**VisEnt**), and retrieved external evidence (**Evidence**). From Table 2, we can see that:

- The original InstructBLIP exhibits a classification accuracy of only 47.4%, which is even lower than random guessing. Furthermore, the recall for fake samples was a mere 4.6%. These results indicate that the general-purpose InstructBLIP tends to misclassify OOC misinformation as real, likely due to the similarity in the text and image composition of such challenging misinformation.
- All components in SNIFFER are important for achieving its best performance, especially the OOC tuning. Specifically, it improves the performance of InstructBLIP by over 35 percentage points, suggesting that task-specific tuning can aid the model in learning the logic necessary to accurately judge OOC samples.

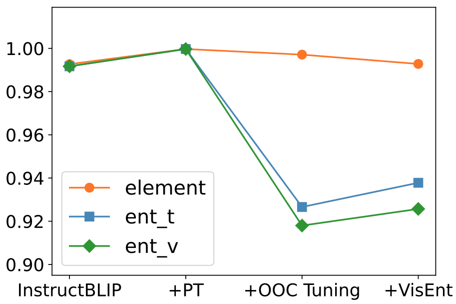


Figure 5. Response ratio.

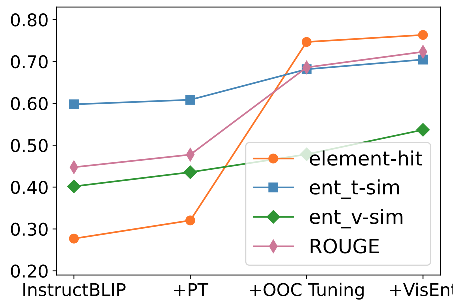


Figure 6. Explanation accuracy.

- The incorporation of visual entities contributed to a 5-point increase in accuracy. This indicates that the integration of external tools can significantly enhance the model's ability to identify visual elements.
- Relying solely on assessing the relevance between text and image-retrieved evidence can yield relatively good detection results. However, while this approach achieves high recall for fake samples, that of real samples is notably lower than other methods. This is largely attributable to the noise in the retrieved evidence. Even for real news, relevant evidence may not always be retrievable, leading to misclassification of real samples. Additionally, since only less than 60% of the samples have associated evidence, the improvement brought by this component in the model's accuracy is somewhat limited.

4.4. Explainability Analysis (Q3)

We evaluated the quality of the explanations generated by SNIFFER in terms of accuracy and persuasiveness.

Quantitative Analysis. We compared SNIFFER's generated explanation with the ground truth obtained in Section 3.2 on the test set. We focus on three critical information points within OOC samples, *i.e.* inconsistent news *element*, the entity of that element in the text (*ent_t*) and image (*ent_v*). We design two types of evaluation metrics: 1) Response ratio of these three points; 2) Accuracy of these three points. We use hard match and compute the hit ratio for *element*, and use similarity of the CLIP embedding for *ent_t* and *ent_v*. We also use ROUGE [24] to measure the accuracy of the generated explanations from an overall perspective. We compared the performance of various variants of SNIFFER on these metrics to demonstrate the role of different components in enhancing the model's explainability.

1) Response Ratio: Figure 5 shows the model's response ratio of *element*, *ent_t*, and *ent_v* for different training steps. We see that the response ratios are close to 1 initially but significantly decrease for *ent_t* and *ent_v* after OOC tuning. This indicates that the model has become more conservative in its outputs. Further, these response ratios improve slightly after incorporating visual entities, indicating that these entities have enhanced the model's confidence to analyze entities.

2) Explanation Accuracy: Although there is a decline in the response ratio of the model, the accuracy of its responses shows an increase. Figure 6 shows that all components contribute to the improvement of explanation accuracy in terms of the hit ratio for *element*, average similarity for *ent_t* and *ent_v*, and ROUGE value for the whole response. Specifically, the hit ratio of *element* increases by 4% and 44% by pre-training and OOC tuning with visual entities, respectively, demonstrating the model's increasing ability to capture inconsistencies between text and images. The accuracy of *ent_t* consistently outperforms *ent_v* by 17%. This disparity can be attributed to the relative simplicity of directly extracting *ent_t* from text, as opposed to recognizing *ent_v* from images.

Overall, after training, the model has become more conservative yet more accurate in spotting the key detection points. This indicates that the model has truly captured clues of inconsistencies between text and images, rather than merely fitting to real or fake labels.

Human Evaluation. To assess the effectiveness of SNIFFER in debunking misinformation through its generated explanations, we conducted a human evaluation. We randomly selected 20 OOC misinformation samples that are correctly identified by SNIFFER from the test set. Ten participants were asked to judge the veracity of each news item (*i.e.* fake or real) and their confidence level (*i.e.* no, somewhat, high) both before and after reading SNIFFER's explanations. Figure 7 shows that: (a) 69% of the items were correctly identified as fake by the users[3], a finding that aligns with the statistics in [29], indicating a certain level of user discernment in detecting OOC misinformation; (b) for the OOC samples initially misidentified as real by the users, 87% of their judgments changed to fake after reading SNIFFER's judgment and explanations, demonstrating the strong persuasive power of SNIFFER; (c) for the OOC samples that were initially identified as fake by the users, SNIFFER's explanations also increased the users' confidence in their judgments for 42% of these samples.

[3] We observed that users tended to label news items as fake more frequently when engaged in data annotation compared to their usual behavior when reading news on social media.

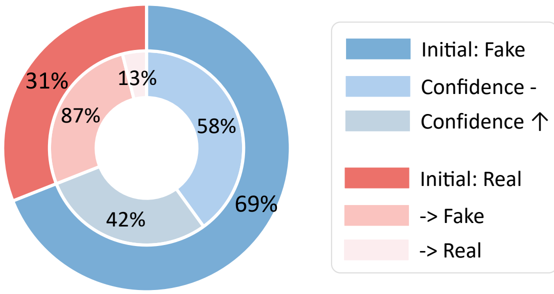


Figure 7. Results of human evaluation. The outer ring displays the initial distribution of test samples judged as real and fake, while the inner circle illustrates the changes in users' judgments and confidence levels after reading SNIFFER's explanations.

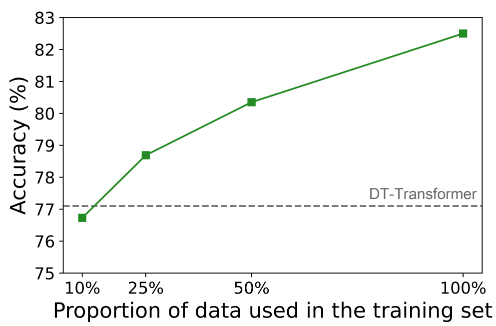
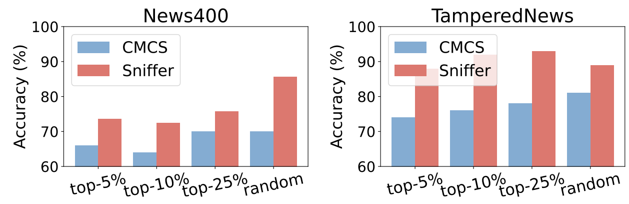


Figure 8. Performance in early detection.

4.5. Practical Setting

Early Detection (Q4). Detecting misinformation in its early stage is important for timely mitigating its negative influences [14]. We conducted experiments using different proportions of the training set to evaluate the model's performance. Specifically, we randomly selected 10%, 25%, 50% and 100% training samples and conducted experiments on SNIFFER-, a reduced version of SNIFFER that only employs OOC Tuning on InstructBLIP, for a fair comparison. Figure 8 shows that, with merely 10% of the training data, SNIFFER- achieves comparable detection performance as baselines that do not utilize external evidence. When trained with 25% of the data, SNIFFER- surpasses these baselines' performance, underscoring the superiority of MLLM in training efficiency and detection accuracy.

Generalization Analysis (Q5). To validate the generalizability of SNIFFER, we tested the model that was trained on the NewsCLIPpings dataset, on the other two datasets, *i.e.* News400 and TamperedNews [32]. In these datasets, img_1 is replaced by images of similar appearance (img_2) based on top-k similarity to img_1, creating different subsets according to the value of k. A smaller k indicates higher similarity between img_1 and img_2, and hence, a greater challenge in detection. We compared SNIFFER with Cross-Modal Context Similarity (CMCS), a competitive baseline proposed in [32]. CMCS quantifies the similarity between the textual context and visual scene labels within the textual feature space. Figure 9 shows that SNIFFER significantly outperforms CMCS at different difficulty levels in both datasets, confirming its cross-dataset generalizability.



(a) (b)

Figure 9. Cross-dataset detection performance of SNIFFER.

4.6. Comparison with GPT-4V (Q6)

As the most popular MLLM, GPT-4 with vision (GPT-4V) [35] has demonstrated unparalleled performance across a variety of multimodal tasks, which inspired us to test it on the OOC detection task. We randomly sampled 400 samples from the test set (200 fake and 200 real examples) and queried GPT-4V using the same prompts as those used for SNIFFER. Table 3 shows that SNIFFER outperforms GPT-4V by 11% in terms of classification accuracy. This demonstrates that task-specific (relatively) smaller models are fully capable of outperforming general-purpose larger models in specific tasks[4]. To further understand the behavioral differences between SNIFFER and GPT-4V, we compare their generated explanations in Supplementary.

Table 3. Comparison of classification accuracy (%) between SNIFFER and GPT-4V on randomly sampled test set.

Method	All	Fake	Real
GPT-4V	75.5	77.0	74.0
SNIFFER (*Ours*)	**86.8**	**79.0**	**94.5**

5. Conclusion

In this paper, we proposed SNIFFER, a multimodal large language model for out-of-context misinformation detection, providing both judgment and explanation. To develop SNIFFER, we constructed multi-perspective instruction data assisted by GPT-4, and employed instruction tuning to continuously adapt the general-purpose InstructBLIP for the news domain and OOC detection task. We further augmented the model by integrating external tools and retrieval methods. Experiments prove that SNIFFER not only achieves SOTA performance in detection, but also generates precise and persuasive explanations.

Acknowledgments. This work is supported by the Ministry of Education, Singapore, under its MOE AcRF TIER 3 Grant (MOE-MOET32022-0001).

[4]Currently, there is no literature explicitly stating the size of GPT-4V, but it is anticipated to exceed the 175 billion parameter scale of GPT-3 [6].

References

[1] Google Vision API. https://cloud.google.com/vision/docs/detecting-web. 5

[2] Reading about the Israel-Hamas war on X? Beware fake news. https://wired.me/technology/x-misinformation/, 2023. 2

[3] Sahar Abdelnabi, Rakibul Hasan, and Mario Fritz. Open-domain, content-based, multi-modal fact-checking of out-of-context images via online resources. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 14920–14929. IEEE, 2022. 2, 3, 5, 6

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 3

[5] Shivangi Aneja, Chris Bregler, and Matthias Nießner. COSMOS: catching out-of-context image misuse using self-supervised learning. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 14084–14092. AAAI Press, 2023. 3

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 8

[7] Juan Cao, Peng Qi, Qiang Sheng, Tianyun Yang, Junbo Guo, and Jintao Li. Exploring the role of visual content in fake news detection. *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*, pages 141–161, 2020. 3

[8] Lu Cheng, Kush R. Varshney, and Huan Liu. Socially responsible AI algorithms: Issues, purposes, and challenges. *J. Artif. Intell. Res.*, 71:1137–1181, 2021. 2

[9] 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. 6

[10] 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. 2, 3, 6

[11] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *CoRR*, abs/2307.08691, 2023. 6

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 6

[13] Lisa Fazio. Out-of-context photos are a powerful low-tech form of misinformation. https://theconversation.com/out-of-context-photos-are-a-powerful-low-tech-form-of-misinformation-129959, 2020. 2

[14] Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. The future of false information detection on social media: New perspectives and trends. *ACM Comput. Surv.*, 53 (4):68:1–68:36, 2021. 8

[15] Quzhe Huang, Mingxu Tao, Zhenwei An, Chen Zhang, Cong Jiang, Zhibin Chen, Zirui Wu, and Yansong Feng. Lawyer LLaMA technical report. *ArXiv*, abs/2305.15062, 2023. 3

[16] Ayush Jaiswal, Ekraam Sabir, Wael Abd-Almageed, 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 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 1465–1471. ACM, 2017. 2, 3

[17] Ayush Jaiswal, Yue Wu, Wael AbdAlmageed, Iacopo Masi, and Premkumar Natarajan. AIRD: adversarial learning framework for image repurposing detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 11330–11339. Computer Vision Foundation / IEEE, 2019. 2, 3

[18] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 3

[19] Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. Multimodal foundation models: From specialists to general-purpose assistants. *CoRR*, abs/2309.10020, 2023. 2

[20] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *CoRR*, abs/2306.00890, 2023. 3

[21] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C. H. Hoi. LAVIS: A one-stop library for language-vision intelligence. In *Proceedings of the*

61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2023, Toronto, Canada, July 10-12, 2023, pages 31–41. Association for Computational Linguistics, 2023. 6

[22] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, pages 19730–19742. PMLR, 2023. 3

[23] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. VisualBERT: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557, 2019. 3, 6

[24] C Lin. Recall-oriented understudy for gisting evaluation (rouge). *Retrieved August*, 20:2005, 2005. 7

[25] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6761–6771. Association for Computational Linguistics, 2021. 5

[26] Haoyang Liu, Maheep Chaudhary, and Haohan Wang. Towards trustworthy and aligned machine learning: A data-centric survey with causality perspectives. *CoRR*, abs/2307.16851, 2023. 2

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 3, 4

[28] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. 6

[29] Grace Luo, Trevor Darrell, and Anna Rohrbach. NewsCLIPpings: Automatic generation of out-of-context multimodal media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6801–6817. Association for Computational Linguistics, 2021. 2, 3, 4, 5, 7

[30] Jing Ma, Wei Gao, and Kam-Fai Wong. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 708–717. Association for Computational Linguistics, 2017. 3

[31] Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, and Ralph Ewerth. Multimodal analytics for real-world news using measures of cross-modal entity consistency. In *Proceedings of the 2020 on International Conference on Multimedia Retrieval, ICMR 2020, Dublin, Ireland, June 8-11, 2020*, pages 16–25. ACM, 2020. 2, 3

[32] Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, Sherzod Hakimov, and Ralph Ewerth. Multimodal news analytics using measures of cross-modal entity and context consistency. *Int. J. Multim. Inf. Retr.*, 10(2):111–125, 2021. 8

[33] OpenAI. ChatGPT. https://openai.com/blog/chatgpt/. 4

[34] OpenAI. GPT-4 technical report, 2023. 2

[35] OpenAI. GPT-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023. 8

[36] 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@ICMR 2023, Thessaloniki, Greece, June 12-15, 2023*, pages 36–44. ACM, 2023. 2, 3, 5, 6

[37] Piotr Przybyla. Capturing the style of fake news. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 490–497. AAAI Press, 2020. 3

[38] Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE International Conference on Data Mining, ICDM 2019, Beijing, China, November 8-11, 2019*, pages 518–527. IEEE, 2019. 3

[39] Peng Qi, Juan Cao, Xirong Li, Huan Liu, Qiang Sheng, Xiaoyue Mi, Qin He, Yongbiao Lv, Chenyang Guo, and Yingchao Yu. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. In *Proceedings of the 29th ACM International Conference on Multimedia , MM '21, Virtual Event, China, October 20 - 24, 2021*, pages 1212–1220. ACM, 2021. 3

[40] Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 14444–14452. AAAI Press, 2023. 3

[41] 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 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 3, 6

[42] 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 2018, Seoul, Republic of Korea, October 22-26, 2018*, pages 1337–1345. ACM, 2018. 2, 3

[43] Rui Shao, Tianxing Wu, and Ziwei Liu. Detecting and grounding multi-modal media manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 6904–6913. IEEE, 2023. 1, 3

[44] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. dEFEND: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 395–405. ACM, 2019. 3

[45] Kai Shu, Xinyi Zhou, Suhang Wang, Reza Zafarani, and Huan Liu. The role of user profiles for fake news detection. In *ASONAM '19: International Conference on Advances in Social Networks Analysis and Mining, Vancouver, British Columbia, Canada, 27-30 August, 2019*, pages 436–439. ACM, 2019. 3

[46] Ruben Tolosana, Rubén Vera-Rodríguez, Julian Fiérrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Inf. Fusion*, 64:131–148, 2020. 1

[47] Xueyu Wang, Jiajun Huang, Siqi Ma, Surya Nepal, and Chang Xu. DeepFake disrupter: The detector of DeepFake is my friend. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 14900–14909. IEEE, 2022. 1

[48] 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 & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 849–857. ACM, 2018. 6

[49] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-Instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13484–13508. Association for Computational Linguistics, 2023. 3

[50] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *CoRR*, abs/2306.13549, 2023. 2

[51] Jingsi Yu, Junhui Zhu, Yujie Wang, Yang Liu, Hongxiang Chang, Jinran Nie, Cunliang Kong, Ruining Cong, XinLiu, Jiyuan An, Luming Lu, Mingwei Fang, and Lin Zhu. Taoli LLaMA. `https://github.com/blcuicall/taoli`, 2023. 3

[52] Yizhou Zhang, Loc Trinh, Defu Cao, Zijun Cui, and Yan Liu. Detecting out-of-context multimodal misinformation with interpretable neural-symbolic model. *CoRR*, abs/2304.07633, 2023. 3, 5, 6

[53] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional Deepfake detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 2185–2194. Computer Vision Foundation / IEEE, 2021. 1

[54] Xinyi Zhou, Jindi Wu, and Reza Zafarani. SAFE: similarity-aware multi-modal fake news detection. In *Advances in Knowledge Discovery and Data Mining - 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11-14, 2020, Proceedings, Part II*, pages 354–367. Springer, 2020. 3, 6

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

[56] Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. Fact-checking meets fauxtography: Verifying claims about images. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2099–2108. Association for Computational Linguistics, 2019. 5