Each Fake News is Fake in its Own Way: An Attribution Multi-Granularity Benchmark for Multimodal Fake News Detection

Hao Guo[*1], **Zihan Ma**[*2,3,4],
Zhi Zeng[2,3,4], **Minnan Luo**[2,3,4], **Weixin Zeng**[1], **Jiuyang Tang**[1], **Xiang Zhao**[†1],
[1]Laboratory for Big Data and Decision, Nation University of Defense Technology,
[2]School of Computer Science and Technology, Xi'an Jiaotong University,
[3]Ministry of Education Key Laboratory of Intelligent Networks and Network Security, Xi'an Jiaotong University,
[4]Shaanxi Province Key Laboratory of Big Data Knowledge Engineering, Xi'an Jiaotong University.
guo_hao@nudt.edu.cn, mazihan880@stu.xjtu.edu.cn, zhizeng@stu.xjtu.edu.cn, minnluo@xjtu.edu.cn,
dexterdervish@foxmail.com, jiuyang_tang@nudt.edu.cn, xiangzhao@nudt.edu.cn

Abstract

Social platforms, while facilitating access to information, have also become saturated with a plethora of fake news, resulting in negative consequences. Automatic multimodal fake news detection is a worthwhile pursuit. Existing multimodal fake news datasets only provide binary labels of real or fake. However, real news is alike, while each fake news is fake in its own way. These datasets fail to reflect the mixed nature of various types of multimodal fake news. To bridge the gap, we construct an attributing multi-granularity multimodal fake news detection dataset AMG, revealing the inherent fake pattern. Furthermore, we propose a multi-granularity clue alignment model MGCA to achieve multimodal fake news detection and attribution. Experimental results demonstrate that AMG is a challenging dataset, and its attribution setting opens up new avenues for future research.

Code and Datasets —
https://github.com/mazihan880/AMG-An-Attributing-Multi-modal-Fake-News-Dataset.

Extended version — https://arxiv.org/pdf/2412.14686

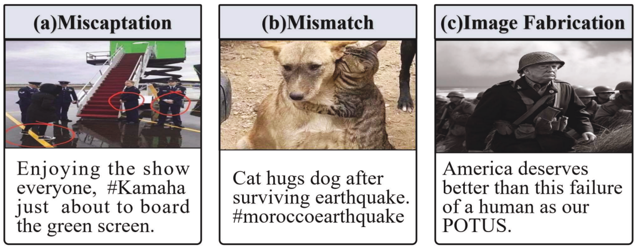

(a)Miscaptation
Enjoying the show everyone, #Kamaha just about to board the green screen.
(b)Mismatch
Cat hugs dog after surviving earthquake. #moroccoearthquake
(c)Image Fabrication
America deserves better than this failure of a human as our POTUS.

Figure 1: Various types of multimodal fake news in Twitter. "Miscaption" means that the caption of the image does not match the text. "Mismatch" indicates the image is related to the text, but from previous similar event. "Image Fabrication" indicates that the image comes from deepfake technology but is not stated.

Introduction

Fake news is false or misleading information presented as news (Rubin et al. 2016; Molina et al. 2021). Social media platforms are inundated with fake news, exerting a significant impact on public health, governance, and societal equilibrium (Zannettou et al. 2019; Allcott and Gentzkow 2017; Apuke and Omar 2021). In recent years, the media-rich nature of these platforms has led to a gradual shift in the type of information shared by the public, encompassing not only textual content but also a plethora of visual elements such as images and videos (Zeng et al. 2024). Because of the "Multimedia Effect" (Mayer 2002), multimedia content such as images and videos exerts a heightened allure on individuals (Jamet, Gavota, and Quaireau 2008; Mayer 2014). Furthermore, visual content is commonly utilized as substantiating evidence within storytelling, thus augmenting the credibility of news narratives. Regrettably, fake news publishers have adeptly utilized this opportunity to captivate attention and enhance credibility, leading to an evolution towards multimodal formats (Cao et al. 2020). The task of multimodal fake news detection has grown progressively intricate, which is the focal point of our research.

In contrast to news that relies solely on textual content, multimodal fake news encompasses visual, textual, and cross-modal correlation, allowing fabricators to craft deceptive narratives from multiple perspectives. We have observed that *real news is alike, each fake news is fake in its own way*. In popular social platforms Twitter, multimodal fake news manifests in various distinct types[1], as depicted in Figure 1. However, existing multimodal fake news detection methods typically focus on only one type. **First**, some methods incorporate visual-textual consistency features into the basis for detection (Zhou, Wu, and Zafarani 2020; Qi et al. 2021a), which aim to capture the correlation between the textual and visual content. The methods focus on detecting types like Figure 1(a), where the key person "Kamala" does not appear in attached image, while such methods overlook the tem-

[*] These authors contributed equally.

[†] Corresponding Author.

[1] Disclaimer. All examples of fake news in this paper are for illustrative purposes only and do not depict real incidents or accurate information. Any resemblance to actual persons or events is purely coincidental.

poral information. **Second**, a plethora of fake news utilizes images from other times and places for the latest trending events, as depicted in Figure 1(b): a picture of Turkey earthquake in Feb. 2023 is used to describe the Morocco earthquake in Sep. 2023, creating a strong association between the image and the accompanying text. **Third**, manipulated images directly impact the authenticity of news (Jin et al. 2016). Current methods exploit the frequency domain (Wu et al. 2021) and pixel domain (Qi et al. 2019) features of images to detect multimodal fake news, which tend to fall short due to the proliferation of Artificial Intelligence Generated Content (AIGC) (Huang et al. 2023; Rombach et al. 2022) that poses a significant challenge in combating deepfake images (Xu, Fan, and Kankanhalli 2023; Shao, Wu, and Liu 2023). As shown in Figure 1(c), the image of "Trump while serving in the military" is deepfake.

Despite of the various types of multimodal fake news, existing detection solutions have not fully considered the scenario where multiple types of fake news coexist and ignored the time consistency cross the image and text. Besides, most models can only output authenticity scores, which are compared with the authenticity labels in the datasets. However, the labels in the datasets are derived directly from fact-checking agencies, with the majority consisting of binary labels indicating only real or fake (Nan et al. 2021; Boididou et al. 2015), without providing fine-grained attribution labels that reveal the error patterns in multimodal fake news. Inspired by the idea of attributing unanswerable questions in the question answering domain (Rajpurkar, Jia, and Liang 2018; Liao et al. 2022), if we can attribute the types of multimodal fake news while detecting its authenticity, the credibility of the detection model will be further enhanced. Although a very recent study explores deception patterns in multimodal fake news (Dong et al. 2024), there still lack benchmarks and effective solutions for attributing multimodal fake news.

To surmount the constraints, we develop the first dataset for arrributing multimodal fake news with multi-granularity, namely AMG. To build the dataset, we collect fake news from multiple platforms. Then attribution rules are designed, and expert annotation is performed based on the rules and ruling articles from fact-checking websites. Finally, a three-fold cross-validation is conducted to achieve fine-grained attribution of fake news.

Furthermore, we propose a multimodal fake news detection and attribution model based on multi-granular clues alignment, namely MGCA. It extracts multi-view features from both visual and textual contents and incorporates consistency modeling of multi-granular clues to aid in authenticity detection and attribution. Extensive experimental results and analyses provide evidence for the increased challenge posed by our proposed dataset. Overall, our contributions are three-fold:

(1) To the best of our knowledge, we are among the first to elicit the notion and motivate the challenges of multi-granularity multimodal fake news attribution.

(2) Our proposed AMG is a first fine-grained attribution of multimodal fake news based on the causes of fake, attributing them to image fabrication, non-evidential image, entity inconsistency, event inconsistency and time inconsistency.

(3) We propose MGCA, a strong baseline to handle multimodal fake news detection and attribution, whose performance is demonstrated by inclusive experiments on AMG.

Dataset Construction

AMG, as the pioneering dataset for multimodal fake news detection and attribution, encompasses posts originating from diverse social platforms. In this section, the data collection, data processing and annotation, and the collation and analysis of AMG will be described in detail.

Data Collection

Fake News Collection. For gathering fake news, we intend to utilize existing fact-checking websites as initial sources of news. The ruling articles found on these websites can assist in fine-grained type annotation. Among them, Snopes[2] and CHECKYOURFACT[3] are widely recognized websites that verify and expose fake news. Professionals, including journalists, gather pertinent evidence and engage in evidence-based reasoning to formulate ruling articles, providing judgments on the authenticity of news. Instead of crawling short claims from the titles of fact-checking websites (Yao et al. 2023), we crawl the original posts associated with claims from various platforms, primarily including Instagram, Facebook, Twitter, TikTok, and YouTube, which aligns more closely with the reality of fake news on social platforms. Among them, we focus on Instagram, Facebook, and Twitter as the main sources of these posts.

Real News Collection. Initially, we crawl the verified true news from the same fact-checking websites. However, the quantity obtained is quite limited (only 126). Besides, to mitigate inherent biases between real and fake news (Zhu et al. 2022), it is essential to establish a relatedness between real news and the corresponding fake news. Therefore, we compensate for the shortage of real news by the following steps. Firstly, we employ the pre-trained Large Language Model Vicuna (Zheng et al. 2023) as our entity extraction tool. Then, based on the distribution proportion of fake news on social platforms, we crawl real news associated with these entities from authoritative and neutral media accounts[4], such as Reuters and NewsNation.

Due to an insufficient number of retrieved related real news, we have randomly selected a certain quantity of news articles from the aforementioned official account's archive to supplement the dataset. These news spans from 2016 to 2024, aligning with the temporal scope of the fake news. To maintain a similar ratio to the previous dataset, we set the number of real news at 1.5 times that of fake news.

Data Processing and Annotation

Filtering. In order to construct a multimodal fake news dataset, our first step involves filtering news articles based on

[2] https://www.snopes.com

[3] https://checkyourfact.com

[4] https://www.allsides.com/unbiased-balanced-news

Datasets	Time period	Class	#Post	#Image	Source	Attribution	Domain	Temporal Info
Weibo21 (Nan et al. 2021)	2014-2021	2	9,128	-	Weibo	✗	variety	✗
Weibo (Jin et al. 2017)	2012-2016	2	9,528	9,528	Weibo	✗	variety	✓
PolitiFact (Shu et al. 2020)	-2020	2	359	359	Twitter	✗	politics	✓
GossipCop (Shu et al. 2020)	-2020	2	10,010	10,010	Twitter	✗	gossip	✓
Twitter (Boididou et al. 2015)	-2014	2	13.924	514	Twitter	✗	11 events	✓
ReCOVery (Zhou et al. 2020)	-2020	2	2,017	2,017	Twitter	✗	covid-19	✓
Pheme (Zubiaga, Liakata, and Procter 2017)	2014-2015	2	5,802	3,670	Twitter	✗	5 events	✓
Fakeddit (Nakamura, Levy, and Wang 2020)	2008-2019	2/3/6	682,996	682,996	Reddit	✗	variety	✓
MR^2 (Hu et al. 2023b)	-2022	3	7,724 6,976	7,724 6,976	Twitter Weibo	✗	variety	✗
AMG	2016-2024	2/6	5,022	5,022	Ins/Twitter Facebook	✓	variety	✓

Table 1: Compilation of multimodal fake news datasets. #Post represents the number of multimodal news piece. #Image represents the number of unique image. MR^2 has both Twitter and Weibo datasets.

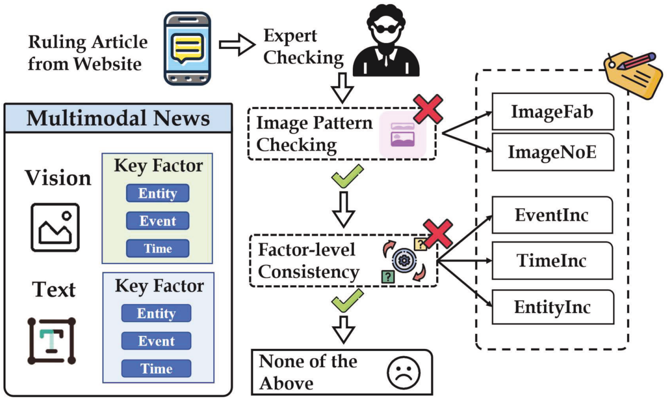


Figure 2: The process of multimodal fake news attribution.

the presence of relevant multimodal news. Both images and videos are included within the scope of our dataset. Moving forward, we utilize visual content similarity to eliminate news articles with high resemblance to one another, thus preserving the diversity among each piece and safeguarding against potential data leakage.

Expert Annotation. Diverging from our previous approach of directly crawling websites with authenticity labels, we have embarked on a more detailed annotation process for news articles, based on these labels and verified articles. Our annotation work is carried out by a team of experts who possess relevant domain knowledge. A comprehensive annotation guideline has been developed, along with specialized training for the annotators. The team consists of a total of 17 individuals (Details in Extended version).

Annotation Process. Binary labels indicating the truthfulness of news can be easily obtained from the verification websites. For fake news, we have meticulously designed each step of the attribution process, as shown in Figure 2. Ruling articles serve as our basis for judgment. Firstly, we perform image pattern checking on the image itself to identify any signs of fabrication or non-evidential content. Secondly, we examine the consistency between the image and the accompanying text across various key factors, attributing entity, event and time inconsistency. And it is also possible that some instances do not belong to any of the above categories.

Attribution Foundation. The specific explanations and theoretical foundations for each attribution type are as follows (See examples in Figure 3):

Image Fabrication (ImageFab): The authenticity of an image is questionable. This can encompass the application of cutting-edge deepfake techniques as well as simpler forms of manipulation such as image splicing or PS. Furthermore, it also includes the simulation of images imitating official websites or tweets, representing a unique circumstance within the realm of image forgery. Previous research (Wu et al. 2021; Xue et al. 2021) has already highlighted the use of the authenticity of the image for detection, while (Shao, Wu, and Liu 2023) established a dataset for detecting AIGC-based fake images. So image fabrication is one typical fake attribution of multimodal news.

Non-Evidential Image (ImageNoE) refers to cases where the image consists of textual information that cannot provide evidence or proof for news content. A notable characteristic of real news is that its images provide support for the accompanying text, such as on-site photos of breaking events. On the other hand, images that solely consist of text are a common image pattern found in multimodal fake news.

Entity Inconsistency (EntityInc) refers to a phenomenon where there is a discrepancy between the key entities depicted in the textual and visual modalities. In other words, there is a lack of alignment or coherence between the entities described in the text and those visually represented, which has been validated as an effective clue in previous study (Qi

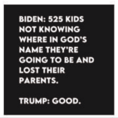
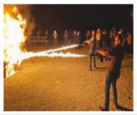
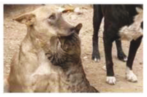
Image	Text	Attribution
	FIRST IMAGES!!! from "Adolf" the Netflix series starring Will Smith.	Image Fabrication
 BIDEN: 525 KIDS NOT KNOWING WHERE IN GOD'S NAME THEY'RE GOING TO BE AND LOST THEIR PARENTS. TRUMP: GOOD.	There is no excuse for crimes against humanity. Period."#vote #joebiden	Non-Evidential Image
	Kamala just about to board the green screen.	Entity Inconsistency
	Missouri Republicans at a literal book burning.	Event Inconsistency
	Cat hugs dog after surviving #moroccoearthquake	Time Inconsistency

Figure 3: Examples of various attributions.

et al. 2021b; Li et al. 2021).

Event Inconsistency (EventInc): Despite the presence of associated entities in both text and image, there is a event-level discrepancy. News always describes events, the alignment of textual and visual events serves as a vital criterion for assessing the authenticity of news (Wei et al. 2022; Wang et al. 2018). Within this category, the images themselves are not forged, the inconsistency often arises from excessive inference and misrepresentation in the written text for attached image.

Time Inconsistency (TimeInc) maintains consistency at the entity or event level, a disparity arises at the temporal information level. It refers to the practice of using unaltered images or videos that depict past events, like natural disasters or gatherings, but falsely presenting them as recent events. Most of out-of-context misinformation (Luo, Darrell, and Rohrbach 2021; Abdelnabi, Hasan, and Fritz 2022; Qi et al. 2024) or image-repurposing (Jaiswal et al. 2019; Sabir et al. 2018) can be attributed to TimeInc.

During the labeling process, we acknowledge that there may be special cases that do not fit into our predefined categories. To account for such situations, we include the label "**None of the Above**" to accommodate those instances. The specific examples that fall outside our attribution categories, as well as the analysis of this particular category, can be found in Extended version.

Cross Validation and Discussion. Each fake news is assigned to three annotators, and the final attribution is determined through a majority vote following (Feng et al. 2022). Furthermore, controversial cases undergo discussion and then secondary round of annotation.

Data Collation and Analysis

After integrating the collected news, we filter out fake news that does not fall under our attribution types. And the quantities for each attribution type are as follows: 434, 295, 133, 667, 475. The number of multimodal fake news from Instagram, Twitter, and Facebook are 142, 558, and 1,304, respectively. In addition, the final number of real news is set to approximately 1.5 times the number of fake news. The counts for real news and fake news are 3,018 and 2,004. More statistic are listed in Extended version.

Train/Val/Test Split. We split the whole dataset into training (Train), validation (Val), and test (Test) sets with the number of 3,532, 517 and 973, respectively. The percentage is nearly 7:1:2. Furthermore, we maintain consistent proportions within each subcategory during the dataset's split.

Rationality of our attribution rules. Upon analyzing the final statistics, we make an exciting observation: the samples that fall outside our attribution categories account for only around 3% of the total dataset, comprising approximately 60 instances. This observation suggests that our classification rules effectively cover almost all cases of fake news, thereby confirming the soundness of our attribution guidelines.

Legal and Ethical. Firstly, we adhere to the data scraping rules of each platform. Additionally, all annotators underwent rigorous training and were well-versed in data privacy and security regulations. During the annotation process, the annotators conducted a screening, selecting only posts related to public figures or public events, without involving ordinary users. Furthermore, any associated personal user information was anonymized, including id and name. We also took measures during data processing and training to prevent any leakage of user privacy((Details in Extended version). All collected data is stored on secure servers, with access restricted to our research team members only.

The strength of AMG. (1) Up-to-date and Temporal-inclusive. The fake news in AMG originate from the period between 2020 and 2024, with a small portion encompassing February 2024. AMG includes the publication timestamps of news posts, whereas MR^2 (Hu et al. 2023a) does not. (2) Multiple platforms. AMG is platform-agnostic which incorporating content from the three major mainstream social platforms. (3) Multiple domains. Upon a simple aggregation, we find that it encompasses multiple fields such as healthcare, elections, military, entertainment, and more. (4) Multi-granularity attribution labels. Different from Fakeddit (Nakamura, Levy, and Wang 2020), the fine-grained labels of AMG reveals the attribution for fake pattern.

Methodology

The section primarily discusses our proposed detection and attribution model. (Preliminary in Extended version)

Model Outline. As depicted in Figure 4, MGCA first gathers multi-perspective clues from both images and text. Next, it performs multimodal feature learning and aligns the

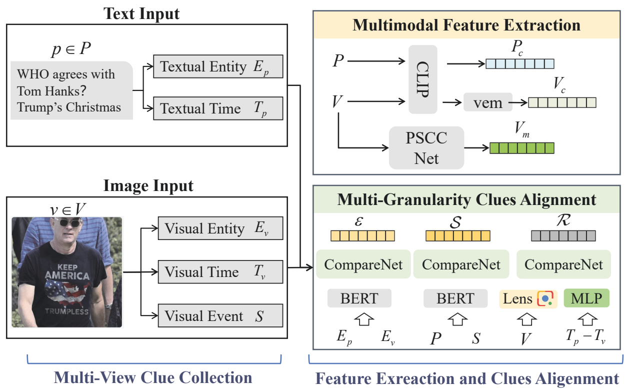


Figure 4: Model outline of MGCA.

collected clues. Finally, it integrates the extracted features to conduct inference of detection and attribution.

Multi-View Clue Collection

We extract the multi-view clues from both the textual input and visual input, which includes time, entity and event.

Textual Entity. Due to the narrative style typically present in news articles, which includes crucial named entities such as characters and locations, the association between these key entities can be instrumental in detecting fake news (Qi et al. 2021b). To enhance this process, we employ the pre-trained Large Language Model Vicuna (Zheng et al. 2023). By designing prompt templates and utilizing the capabilities of the large-scale model's In-context Learning (Mann et al. 2020; Xie et al. 2021), we incorporate examples of entity extraction within these templates, guiding the process. We denote the entity in the text as E_p.

Visual Entity. Corresponding to the textual content, certain news articles also contain valuable visual entities within their visual content. For the extraction of visual entities, we utilize Baiduan APIs[5] that specializes in extracting three types of entities: individuals, landmarks, and organizations. We denote this extraction of visual entities as E_v.

Textual Time. Temporal mismatch is a significant type of multimodal fake news. In this article, we consider the temporal information of news as a crucial factor in determining its authenticity. Firstly, we extract the time label of the news, denoted as t_1. As news articles often describe past events, we also extract the mentioned time, t_2, from the textual content. We then select the earlier time as the temporal reference for the text, which we refer to as $T_p = min\{t_1, t_2\}$.

Visual Time. Retrieving the original publication time of an image, along with its relevant content, can be helpful in identifying temporal inconsistencies in multimodal fake news. We employ GoogleLens[6] for performing reverse image searches. By conducting such searches, we obtain the earliest corresponding time T_v and title R of the related image.

Image Event. In addition to visual entities, we believe that the event present in images is also a valuable auxiliary clue. We utilize multimodal large language model LLaVA (Liu et al. 2023) for extracting image events denoted as S (Conducting details in Extended version).

Multimodal Feature Learning

To enhance the consistency representation, we employ CLIP (Radford et al. 2021) to extract features P_c and V_c from the total of news text P and news image V. To obtain the rich semantic clue representations, we exploit utilize BERT(Kenton and Toutanova 2019) to acquire C_s, C_r, C_p and C_v from event clues S. Also, we use Bert to encode entity clues E_p , E_v and the retrieval clues R.

To ensure mathematical distribution consistency, we also utilize BERT to obtain the semantic representation P_b of the news text. As for the timeline, we calculate the temporal gap T_g between the images and the text, denoted as $T_g = (T_P - T_V)$ to characterize the temporal inconsistency.

To detect the manipulated image, we employ the effective manipulation detection network PSCC-NET (Liu et al. 2022) for detecting image manipulation. Specifically, by freezing the feature extraction layer of PSCC-NET, we obtain the manipulation features V_m for the news images.

Multi-Granularity Clues Alignment

To detect the entity-level and event-level consistency between news image and text, we utilize a Compare-Net(Shen et al. 2018) to obtain consistency features $\mathcal{E}$ and $\mathcal{S}$,*i.e.*,

$$\begin{aligned} \mathcal{E} &= f_{cmp} = (C_p, C_v), \\ \mathcal{S} &= f_{cmp} = (C_s, P_b), \end{aligned} \tag{1}$$

where f_{cmp} denotes the Compare-Net(Shen et al. 2018). To measure the embedding closeness and relevance, we design the comparison function as:

$$f_{cmp}(C_1, C_2) = W_c[C_1, C_2, C_1 - C_2, C_1 * C_2], \tag{2}$$

where W_c is a transformation matrix and $*$ is Hadamard product. C_1 and C_2 are the features to be compared. Additionally, we compare the news text with the results obtained from reverse search to verify the presence of temporal alignments. In particular, we concurrently splice temporal features in the vectors of the Compare-Net.

$$\begin{aligned} \mathcal{T} &= W_t T_g, \\ \mathcal{R} &= f_{cmp}(C_r, P_b, \mathcal{T}) \\ &= W_r[C_r, P_b, C_r - P_b, C_r * P_b, \mathcal{T}], \end{aligned} \tag{3}$$

where $\mathcal{R}$ represents the temporal consistency features, W_t is is a 1-dimensional learnable matrix, W_r refer to learnable transformation matrix.

Training and Inference

To obtain a better fake news representation of various attributions, we incorporate a classification head before each category of features to perform a binary classification task

[5] https://ai.baidu.com/tech/imagerecognition/general
[6] https://lens.google.com/

for distinguishing between real and fake news. The label for this task is denoted as y_b. In particular, we also separately perform a binary classification task on the images feature V_c to better distinguish samples of visual effectiveness. We use binary cross-entropy loss to individually optimize these five feature categories:

$$\begin{aligned}&\hat{y}_n = MLP(n),\ n = \mathcal{E}, \mathcal{S}, \mathcal{R}, V_m, V_c,\\ &\mathcal{L}_n = -(y_b \cdot \log \hat{y}_n + (1 - y_b) \cdot \log(1 - \hat{y}_n)).\end{aligned} \tag{4}$$

Simultaneously, we concatenate the features and multiply the probability ϕ_n of a single judgment network indicating the news as fake with the corresponding network's feature. When the probability approaches 1, it signifies a higher likelihood of the news being false due to that particular feature. Meanwhile, we splice text clip semantic features to better obtain a global multimodal representation of the news. After passing through a Multilayer Perceptron (MLP), we obtain the final prediction result $\hat{y}_b$, *i.e.*,

$$\begin{aligned}\hat{y}_b = MLP([&P_c, \mathcal{E} * \phi_{\mathcal{E}}, \mathcal{S} * \phi_{\mathcal{S}},\\ &\mathcal{R} * \phi_{\mathcal{R}}, V_m * \phi_m, V_c * \phi_c]).\end{aligned} \tag{5}$$

Then, we consider the minimization of the standard binary cross-entropy loss value as the objective function,*i.e.*,

$$\begin{aligned}\mathcal{L}_b(y_b, \hat{y}_b) = -(y_b \log \hat{y}_b + (1 - y_b) \log(1 - \hat{y}_b))&\\ + \frac{1}{5} \sum_n \mathcal{L}_n,&\end{aligned} \tag{6}$$

where y_b denotes the actual label and $y_b \in \{0, 1\}$; $\hat{y}_b$ represent the predicted label. In attributing inference, we define the downstream task as a 6-classification task, and obtain the final attributing prediction $\hat{y}$ through the MLP,*i.e.*,

$$\begin{aligned}\hat{y} = MLP([&P_c, \mathcal{E} * \phi_{\mathcal{E}}, \mathcal{S} * \phi_{\mathcal{S}},\\ &\mathcal{R} * \phi_{\mathcal{R}}, V_m * \phi_m, V_c * \phi_c]),\end{aligned} \tag{7}$$

and optimize the classification results using cross-entropy loss,*i.e.*,

$$\mathcal{L} = -\sum_{i=1}^{6} y_i \log \hat{y}_i + \frac{1}{5} \sum_n \mathcal{L}_n, \tag{8}$$

where $\hat{y}_i$ is the probability of predicting the post as class i.

Experiment

Experimental Settings. Experimental settings can be found in Extended version, which includes compared methods, implementation details, and evaluation metrics. All experiments are conducted on a cluster of 8 RTX3090 GPUs. Additionally, we also analyze the **computational complexity** of the model. Details can be found in the Extended version.

Results on Multimodal Fake News Detection. According to Table 2, our proposed model exhibits the best performance across various metrics. MGCA achieves an approximately 2.5% improvement in overall accuracy (acc) and a 2.8% improvement in F1 score. Additionally, to demonstrate the **generalization** of MGCA, we conduct experiments on public datasets Twitter (Boididou et al. 2015), Weibo (Jin et al. 2017), and Weibo21 (Nan et al. 2021). MGCA outperforms the compared baselines, achieving F1-scores of 0.905, 0.899, and 0.901, respectively. Related table can be found in Extended version.

Method	Fake News Detection		Fake News Attribution	
	Accuracy	F1 Score	Accuracy	F1 Score
CLIP	0.7812	0.7809	0.6469	0.5325
CAFE	0.7667	0.7628	0.6382	0.4665
MCAN	0.7740	0.7693	0.6115	0.4605
BMR	0.8079	0.8057	0.6687	0.5193
MGCA	**0.8323**	**0.8310**	**0.7385**	**0.5666**

Table 2: Results of multimodal fake news detection and attribution.

Method	Detection		Attribution	
	Acc	F1	Acc	F1
w/o PSCC-NET	0.8167	0.8166	0.6781	0.4660
w/o entity	0.8146	0.8138	0.6937	0.4283
w/o event	0.7917	0.7916	0.6813	0.4542
w/o temporal	0.8094	0.8085	0.7010	0.4294
w/o vem	0.8187	0.8177	0.6937	0.4407
MGCA	**0.8323**	**0.8310**	**0.7385**	**0.5666**

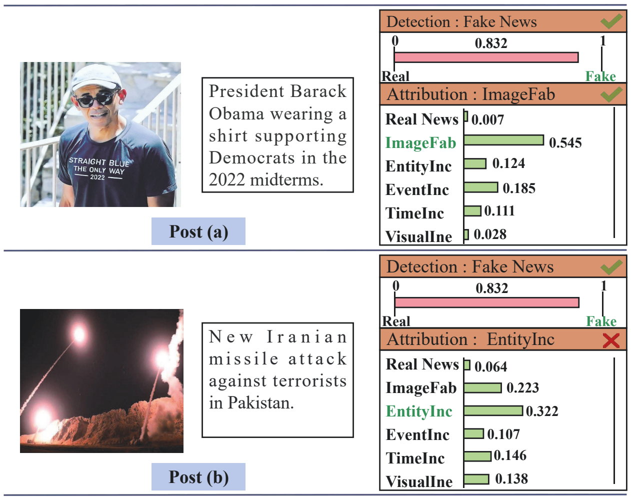
Table 3: F1 results of ablation study.

Discussion on Dataset Difficulty. Comparing the experimental results of the same model on previous datasets, we observe that AMG is a more challenging dataset. BMR achieves an accuracy (acc) of 90% on both the Weibo (Jin et al. 2017) and GossipCop (Shu et al. 2020), while the detection accuracy of AMG falls below 81%. Other models also exhibit varying degrees of performance decline. We have analyzed the reasons behind the increased challenge in AMG and arrived at a preliminary conclusion: the presence of **entity bias** (Zhu et al. 2022) in the collection process of real and fake news. However, our approach of collecting true news has successfully avoided this bias.

Results on Multimodal Fake News Attribution. We present the overall attribution accuracy and F1 scores in Table 2, while the detailed results on each attribution category are presented in Extended version. The experimental results show that our model outperforms the baseline model in terms of overall attribution accuracy and F1 scores. Compared to the suboptimal model BMR, our model achieves improvements of approximately 7% and 4.7% in accuracy and F1 score, respectively. Furthermore, MGCA demonstrates a significant enhancement of around 10% in accuracy compared to other methods.

Ablation Study. To demonstrate the effectiveness of the multi-granularity clue and various feature extraction modules we employed, we conducted ablation experiments. The results of these experiments are displayed in Table 3.

Removing individual modules leads to a certain degree of decline in both detection and attribution performance. Among them, the removal of event-level coherence features has the greatest impact on the detection of multimodal fake


President Barack Obama wearing a shirt supporting Democrats in the 2022 midterms.

New Iranian missile attack against terrorists in Pakistan.

Figure 5: The case study of MGCA in dataset AMG.

news, resulting in a decrease of approximately 4% in both accuracy and F1 score. Furthermore, the temporal coherence, which is the focus of MGCA, also has a significant impact on both detection and attribution, demonstrating the importance the temporal information between image and text.

Case Study. We select two representative samples to analyze the detection and attribution results of MGCA. As can be observed in Figure 5, both detection and attribution results of Post (a) are correct. Despite the high coherence between the image and text, MGCA is still able to draw the conclusion of image fabrication. However, there still exist some challenging samples. Post (b) claims that Iran used missiles to strike terrorists in Pakistan in 2018, but in reality, the image used in the article is from 2015, showcasing a typical case of time inconsistency. Although it is classified as fake news, it is categorized as entity inconsistency in the attribution process. In the image, the key entity of "terrorists" mentioned in the text is not detected, leading MGCA to make the judgment.

Discrimination Performance. We utilize heatmaps to visualize the discriminative power of MGCA on AMG. We randomly select 90 real news and 90 fake news. We then calculate the pairwise similarities between the 16-dimensional representations from the binary classification classifier and the attribution classifier. The darker colors indicate weaker correlation and lighter colors indicate stronger correlation.

From Figure 6, we can observe that our model demonstrates strong discriminative ability, with relatively clear intra-class similarity and inter-class differences. Additionally, it is evident that the binary classification representations of genuine news and fake news exhibit a higher level of distinctiveness, while the attribution learning shows a slightly reduced discriminative capacity. This observation indicates that capturing intra-class variations among the fake news instances represents the main challenge faced by AMG.

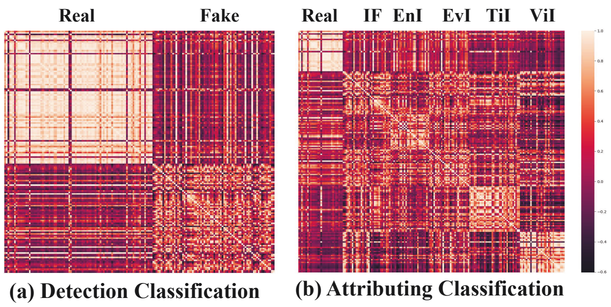
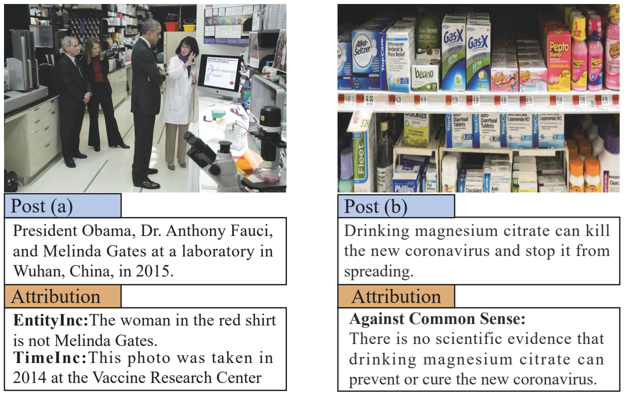


Figure 6: The discriminative power of MGCA.


Post (a)

President Obama, Dr. Anthony Fauci, and Melinda Gates at a laboratory in Wuhan, China, in 2015.

Attribution

EntityInc: The woman in the red shirt is not Melinda Gates.
TimeInc: This photo was taken in 2014 at the Vaccine Research Center

Post (b)

Drinking magnesium citrate can kill the new coronavirus and stop it from spreading.

Attribution

Against Common Sense:
There is no scientific evidence that drinking magnesium citrate can prevent or cure the new coronavirus.

Figure 7: Fake News out of our attributions.

Conclusion

In this study, we introduce a novel task, multimodal fake news attribution, which aims to enhance the credibility of model detection results. We believe it will provide promising and meaningful avenues for research. Furthermore, we develop AMG, the first multimodal fake news attribution dataset and make it open-sourced, which will facilitate future follow-up studies. We emphasize the significance of temporal information in the detection of multimodal fake information, highlighting it as a key factor for fake news detection. We also introduce a competitive method MGCA.

Limitation. AMG focuses solely on the content of multimodal fake news, excluding metadata like comments and social networks. We are collecting this data for future release. Additionally, besides dividing attributions into five categories, we include the label "Not fall into any of the above types", during the labeling process. Figure 7 shows several instances that fall outside the scope of our attributions. Post (a) delineates the occurrence of multiple overlapping attribution anomalies, encompassing both entity and temporal inconsistency. And Post (b) signifies instances that do not conform to any of our attribution categories.

Acknowledgments

This work was partially supported by National Key R&D Program of China (No. 2022YFB3102600), and the Natural Science Foundation of China (Nos. U23A20296, 62272469, 62302513, 62192781 and 62272374).

References

Abdelnabi, S.; Hasan, R.; and Fritz, M. 2022. Open-Domain, Content-based, Multi-modal Fact-checking of Out-of-Context Images via Online Resources. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Allcott, H.; and Gentzkow, M. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2): 211–236.

Apuke, O. D.; and Omar, B. 2021. Fake news and COVID-19: modelling the predictors of fake news sharing among social media users. *Telematics and Informatics*, 56: 101475.

Boididou, C.; Andreadou, K.; Papadopoulos, S.; Dang Nguyen, D. T.; Boato, G.; Riegler, M.; Kompatsiaris, Y.; et al. 2015. Verifying multimedia use at mediaeval 2015. In *MediaEval 2015*, volume 1436. CEUR-WS.

Cao, J.; Qi, P.; Sheng, Q.; Yang, T.; Guo, J.; and Li, J. 2020. Exploring the role of visual content in fake news detection. *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*, 141–161.

Dong, Y.; He, D.; Wang, X.; Jin, Y.; Ge, M.; Yang, C.; and Jin, D. 2024. Unveiling Implicit Deceptive Patterns in Multi-Modal Fake News via Neuro-Symbolic Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(8): 8354–8362.

Feng, S.; Tan, Z.; Wan, H.; Wang, N.; Chen, Z.; Zhang, B.; Zheng, Q.; Zhang, W.; Lei, Z.; Yang, S.; et al. 2022. Twibot-22: Towards graph-based twitter bot detection. *Advances in Neural Information Processing Systems*, 35: 35254–35269.

Hu, B.; Sheng, Q.; Cao, J.; Zhu, Y.; Wang, D.; Wang, Z.; and Jin, Z. 2023a. Learn over Past, Evolve for Future: Forecasting Temporal Trends for Fake News Detection. *arXiv preprint arXiv:2306.14728*.

Hu, X.; Guo, Z.; Chen, J.; Wen, L.; and Yu, P. S. 2023b. MR2: A Benchmark for Multimodal Retrieval-Augmented Rumor Detection in Social Media. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, 2901–2912. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394086.

Huang, B.; Wang, Z.; Yang, J.; Ai, J.; Zou, Q.; Wang, Q.; and Ye, D. 2023. Implicit Identity Driven Deepfake Face Swapping Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4490–4499.

Jaiswal, A.; Wu, Y.; AbdAlmageed, W.; Masi, I.; and Natarajan, P. 2019. Aird: Adversarial learning framework for image repurposing detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11330–11339.

Jamet, E.; Gavota, M.; and Quaireau, C. 2008. Attention guiding in multimedia learning. *Learning and instruction*, 18(2): 135–145.

Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; and Luo, J. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, 795–816.

Jin, Z.; Cao, J.; Zhang, Y.; Zhou, J.; and Tian, Q. 2016. Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia*, 19(3): 598–608.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, 2.

Li, P.; Sun, X.; Yu, H.; Tian, Y.; Yao, F.; and Xu, G. 2021. Entity-oriented multi-modal alignment and fusion network for fake news detection. *IEEE Transactions on Multimedia*, 24: 3455–3468.

Liao, J.; Zhao, X.; Zheng, J.; Li, X.; Cai, F.; and Tang, J. 2022. PTAU: Prompt Tuning for Attributing Unanswerable Questions. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1219–1229.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning.

Liu, X.; Liu, Y.; Chen, J.; and Liu, X. 2022. PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 7505–7517.

Luo, G.; Darrell, T.; and Rohrbach, A. 2021. emnlp21NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Mayer, R. E. 2002. Multimedia learning. In *Psychology of learning and motivation*, volume 41, 85–139. Elsevier.

Mayer, R. E. 2014. Incorporating motivation into multimedia learning. *Learning and instruction*, 29: 171–173.

Molina, M. D.; Sundar, S. S.; Le, T.; and Lee, D. 2021. "Fake news" is not simply false information: A concept explication and taxonomy of online content. *American behavioral scientist*, 65(2): 180–212.

Nakamura, K.; Levy, S.; and Wang, W. Y. 2020. Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6149–6157.

Nan, Q.; Cao, J.; Zhu, Y.; Wang, Y.; and Li, J. 2021. MDFEND: Multi-domain fake news detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 3343–3347.

Qi, P.; Cao, J.; Li, X.; Liu, H.; Sheng, Q.; Mi, X.; He, Q.; Lv, Y.; Guo, C.; and Yu, Y. 2021a. 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*, 1212–1220.

Qi, P.; Cao, J.; Li, X.; Liu, H.; Sheng, Q.; Mi, X.; He, Q.; Lv, Y.; Guo, C.; and Yu, Y. 2021b. 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*, 1212–1220.

Qi, P.; Cao, J.; Yang, T.; Guo, J.; and Li, J. 2019. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE international conference on data mining (ICDM)*, 518–527. IEEE.

Qi, P.; Yan, Z.; Hsu, W.; and Lee, M. L. 2024. SNIFFER: Multimodal Large Language Model for Explainable Out-of-Context Misinformation Detection. *arXiv preprint arXiv:2403.03170*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know what you don't know: Unanswerable questions for SQuAD. *arXiv preprint arXiv:1806.03822*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Rubin, V. L.; Conroy, N.; Chen, Y.; and Cornwell, S. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection*, 7–17.

Sabir, E.; AbdAlmageed, W.; Wu, Y.; and Natarajan, P. 2018. Deep Multimodal Image-Repurposing Detection. In *Proceedings of the 26th ACM international conference on Multimedia*.

Shao, R.; Wu, T.; and Liu, Z. 2023. Detecting and grounding multi-modal media manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6904–6913.

Shen, D.; Zhang, X.; Henao, R.; and Carin, L. 2018. Improved semantic-aware network embedding with fine-grained word alignment. *arXiv preprint arXiv:1808.09633*.

Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3): 171–188.

Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; and Gao, J. 2018. 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*, 849–857.

Wei, P.; Wu, F.; Sun, Y.; Zhou, H.; and Jing, X.-Y. 2022. Modality and event adversarial networks for multi-modal fake news detection. *IEEE Signal Processing Letters*, 29: 1382–1386.

Wu, Y.; Zhan, P.; Zhang, Y.; Wang, L.; and Xu, Z. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, 2560–2569.

Xie, S. M.; Raghunathan, A.; Liang, P.; and Ma, T. 2021. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Xu, D.; Fan, S.; and Kankanhalli, M. 2023. Combating misinformation in the era of generative AI models. In *Proceedings of the 31st ACM International Conference on Multimedia*, 9291–9298.

Xue, J.; Wang, Y.; Tian, Y.; Li, Y.; Shi, L.; and Wei, L. 2021. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management*, 58(5): 102610.

Yao, B. M.; Shah, A.; Sun, L.; Cho, J.-H.; and Huang, L. 2023. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2733–2743.

Zannettou, S.; Sirivianos, M.; Blackburn, J.; and Kourtellis, N. 2019. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *Journal of Data and Information Quality (JDIQ)*, 11(3): 1–37.

Zeng, Z.; Luo, M.; Kong, X.; Liu, H.; Guo, H.; Yang, H.; Ma, Z.; and Zhao, X. 2024. Mitigating World Biases: A Multimodal Multi-View Debiasing Framework for Fake News Video Detection. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 6492–6500.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.

Zhou, X.; Mulay, A.; Ferrara, E.; and Zafarani, R. 2020. Recovery: A multimodal repository for covid-19 news credibility research. In *Proceedings of the 29th ACM international conference on information & knowledge management*, 3205–3212.

Zhou, X.; Wu, J.; and Zafarani, R. 2020. : Similarity-Aware Multi-modal Fake News Detection. In *Pacific-Asia Conference on knowledge discovery and data mining*, 354–367. Springer.

Zhu, Y.; Sheng, Q.; Cao, J.; Li, S.; Wang, D.; and Zhuang, F. 2022. Generalizing to the future: Mitigating entity bias in fake news detection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2120–2125.

Zubiaga, A.; Liakata, M.; and Procter, R. 2017. Exploiting context for rumour detection in social media. In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part I 9*, 109–123. Springer.