Multimodal Misinformation Detection by Learning from Synthetic Data with Multimodal LLMs

Fengzhu Zeng[1], **Wenqian Li**[2], **Wei Gao**[1], **Yan Pang**[2]
[1]School of Computing and Information Systems, Singapore Management University
[2]Institute of Operations Research and Analytics, National University of Singapore
fzzeng.2020@phdcs.smu.edu.sg, wenqian@u.nus.edu
weigao@smu.edu.sg, jamespang@nus.edu.sg

Abstract

Detecting multimodal misinformation, especially in the form of image-text pairs, is crucial. Obtaining large-scale, high-quality real-world fact-checking datasets for training detectors is costly, leading researchers to use synthetic datasets generated by AI technologies. However, the generalizability of detectors trained on synthetic data to real-world scenarios remains unclear due to the distribution gap. To address this, we propose learning from synthetic data for detecting real-world multimodal misinformation through two model-agnostic data selection methods that match synthetic and real-world data distributions. Experiments show that our method enhances the performance of a small MLLM (13B) on real-world fact-checking datasets, enabling it to even surpass GPT-4V (OpenAI, 2023b).

1 Introduction

Multimodal misinformation, which appears more credible and spreads faster than text-only misinformation, has become a significant concern. About one-third of verification claims include multimodal data, with the primary modality being image-text pairs (Akhtar et al., 2023). This underscores the importance of Multimodal Misinformation Detection (MMD), which involves determining the veracity of such image-text pairs. These pairs can be created by pairing a textual claim with an out-of-context image, or by manipulating the content of the image, the text, or both. An illustrative example is depicted in Figure 1, where an image shows Elon Musk holding a flag bearing the statement "Trump Won, Democrats Cheated," a false claim debunked by fact-checkers[1].

The success of training a MMD model highly depends on the availability of large-scale, high-quality datasets, especially in the era of Multimodal Large Language Models (MLLMs) that require instruction tuning for downstream tasks. However, acquiring such real-world fact-checking datasets presents significant challenges as labor-intensive annotation has to be done by fact-checkers.

Figure 1: Elon Musk holding a flag that says "Trump Won, Democrats Cheated."

Given this data scarcity, a cost-effective solution is to utilize the advancements in generative AI technology to generate synthetic multimodal misinformation based on a vast repository of readily accessible real news (Luo et al., 2021; Shao et al., 2023; Jia et al., 2023). For instance, the synthetic dataset NewsCLIPings (Luo et al., 2021) contains more than 1 million instances, whereas the popular real-world dataset MediaEval (Boididou et al., 2016), annotated by fact-checkers, only has around 10,000 instances with just 514 unique images. Studies have proposed different multimodal misinformation detectors trained on synthetic datasets and have achieved reasonable performance (Abdelnabi et al., 2022; Wang et al., 2024a; Qi et al., 2024).

However, it is unclear how well detectors trained on large synthetic datasets can generalize to real-world fact-checking data, as there remain salient gaps between them (Visualization example is shown in Appendix C). Since synthetic and real-world data are out-of-distribution (OOD) relative to each other, the distributional gap between the two types of datasets can lead to significant discrep-

[1]https://www.snopes.com/fact-check/elon-musk-trump-won-flag/

Findings of the Association for Computational Linguistics: EACL 2024, pages 10467–10484
November 12-16, 2024

ancies in detection performance. Detectors directly trained on synthetic datasets may yield limited results or even fail in real-world applications.

To address this gap, we propose to select valuable synthetic data instances based on similarity metrics that are oblivious to the downstream detection tasks, and utilize it to improve the generalization capability of detectors to real-world fact-checking data. We first compile a large-scale synthetic dataset by amalgamating three distinct synthetic datasets, ensuring coverage of diverse multimodal misinformation categories. Given this synthetic dataset as training set and *a very small number of unlabeled* real-world instances as the validation set, our goal is to effectively select *a small subset* of relevant and valuable synthetic data to enhance the models' capacity for detecting real-world multimodal misinformation.

We approach this using two model-agnostic data selection methods: 1) semantic similarity-based selection, which prioritizes synthetic instances with the highest similarity scores to the validation set; and 2) distributional similarity-based selection using gradient information derived from the Optimal Transport (OT) problem (Villani et al., 2009) to increase the density of data around the desirable region by choosing data points in the synthetic dataset that are close to the target real-world distribution. These methods evaluate the relevance of synthetic data points without requiring model retraining, ensuring efficiency and scalability. Moreover, the small selected subset of synthetic training instances makes the fine-tuning of MLLMs feasible in computing resource-constrained scenarios.

Our main contributions are four-fold [2]:

- We propose a new task to tackle the scarcity of real-world fact-checking data for MMD by learning from synthetic data.
- We frame a new setting for MMD in the era of MLLMs: how can we effectively select relevant and valuable instances from a large-scale synthetic dataset to improve MLLM's detection of real-world multimodal misinformation?
- We employ two model-agnostic data selection methods to handle different distributions by selecting a small number of synthetic instances for fine-tuning open-source MLLMs on the MMD task.
- Evaluation on real-world fact-checking datasets demonstrates the effectiveness of data selection methods across MLLM scales and families. This approach enables a small MLLM (13B) to even surpass GPT-4V (OpenAI, 2023b).

[2]Code and dataset are released at https://github.com/znhy1024/MLLMs-for-MMD.

2 Problem Formulation

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}|}$ be a multimodal misinformation dataset with $|\mathcal{D}|$ instances, where each instance consists of the input x_i and the ground-truth label $y_i \in \{0, 1\}$. We denote $x_i = (\mathbf{m}_i, \mathbf{t}_i)$ which is a pair of information containing an image $\mathbf{m}_i$ and a corresponding claim text $\mathbf{t}_i$, and y_i indicates the veracity of x_i is true if $y_i = 0$ or false otherwise. The task of MMD aims to determine the veracity label of the given piece of information x_i.

We consider a large dataset $\mathcal{D}_s$ consisting of synthetic instances that cover common categories of multimodal misinformation, and a target dataset $\mathcal{D}_t$ which contains high-quality, annotated, and diverse real-world fact-checking data, where $|\mathcal{D}_s| \gg |\mathcal{D}_t|$. Our goal is to select a subset $\mathcal{D}_v = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}_v|}$ from $\mathcal{D}_s$ to fine-tune a pre-trained multimodal model $\mathcal{M}$, such that it generalizes well to the real-world test dataset $\mathcal{D}_t$. To reflect real-world constraints related to practical computational resource limitations, we also restrict the size of the selected subset being small, i.e. $|\mathcal{D}_v| \ll |\mathcal{D}_s|$.

A *small unlabeled* validation set $\mathcal{D}_u = \{(x_i)\}_{i=1}^{|\mathcal{D}_u|}$ with a few randomly sampled instances from $\mathcal{D}_t$ is given to facilitate data selection. We assume the unlabeled validation set is sampled i.i.d from the target distribution as they originate from the same source. It is expected that MLLMs trained on the selected synthetic instances based on $\mathcal{D}_u$ could generalize well on unseen test instances.

3 Multimodal Misinformation Data

3.1 Common Categories

There is no standard categorizations for multimodal misinformation. We categorize it into two types based on whether the image content are falsely altered or counterfeited, as briefly described below. Appendix A.1 contains a more formulated and detailed description.

Out-of-Context (OOC) Misuse occurs when the textual claim misrepresents the original context or intent of a genuine image, conveying misleading information. OOC image-text pair can be obtained by pairing a textual claim with an image taken out-of-context (Jaiswal et al., 2017; Luo et al., 2021),

or manipulating the textual claim such as replacing named entities and altering sentiments, aiming at distorting its meaning to conform to a false narrative (Sabir et al., 2018; Müller-Budack et al., 2020). Such OOC pairs not only deceive human but also pose significant challenges for detection models (Luo et al., 2021).

Content Manipulation involves intentionally altering image and text modality with the aim of deceiving or misleading. As it can be difficult to consistently obtain real images to support non-factual claim, manipulating image and text content to generate misinformation becomes an alternative approach (Abdali, 2022). Image manipulation refers to altering an image to obtain its fake version by modifying the elements within it using techniques, such as removing content and face swap (Shao et al., 2023; Jia et al., 2023). Image and text manipulation techniques can be combined to ensure that both visual and textual elements complement each other, thereby strengthening the deception.

3.2 Training and Evaluation Datasets

Given that synthetic datasets often lack the comprehensive diversity of misinformation categories found in real-world data and typically focus on specific types, such as out-of-context misuse, we curate a large-scale synthetic training dataset by including three representative synthetic datasets. This ensures coverage of common multimodal misinformation categories and includes diverse techniques for creating more deceptive multimodal misinformation.

For a comprehensive assessment of model's performance across various sources and distributions of misinformation, we utilize two real-world fact-checking datasets collected from social media and fact-checking websites, both of which provide sufficient data for reliable evaluation. Appendix A.2 contains more datasets details.

Training Datasets: 1) NewsCLIPpings (Luo et al., 2021) is an automatically generated OOC multi-modal dataset that contains both pristine and falsified instances. A falsified instance is an unmanipulated but mismatched image-caption pair, constructed by pairing an image with a caption from an inconsistent context. The falsified instances could portray subjects in an image as different entities, depict specific individuals in a misleading context, and mislabel the event described within a particular scene. By utilizing different embeddings models during the image-text mismatch procedure, around 988k unique instances are automatically generated. 2) DGM[4] (Shao et al., 2023) is a multimodal misinformation dataset, where synthetic instances are automatically generated by applying image and text manipulation techniques to pristine instances sourced from news data. By employing diverse manipulation techniques (e.g., face swap, emotion manipulation, name entity replacement), eight manipulated instances comprising both OOC and manipulation instances are generated for each pristine instance, resulting in a total of 230k synthetic instances. 3) Autosplice (Jia et al., 2023) is a manipulated image dataset that utilizes a language-image model DALLE-2 (Ramesh et al., 2022), guided by textual prompts to edit images. Unlike DGM[4], where image and text manipulation are initially employed separately and then combined, Autosplice alters the image based on modified text, enabling a more direct and integrated manipulation process.

Evaluation Datasets: 1) MediaEval (Boididou et al., 2016) collects a set of tweets associated images/videos with manually verified veracity (i.e., fake or real), which were spread around 17 widely attention-grabbing events. The number of instances for each event is small, and the classes are imbalanced. We utilize the image-text instances for evaluation, consisting of 702 instances, in which 292 are real and 410 are fake. 2) Snopes, a dataset derived from Fauxtography (Zlatkova et al., 2019) and MOCHEG (Yao et al., 2023), which contains image-related claims from the popular fact-checking website Snopes[3]. These image-claim pairs are labeled as either True or False by fact-checkers, consisting of 756 instances, where 376 are true and 380 are false.

4 Methodology

Data selection aims to find a set of training instances that improve model performance (Albalak et al., 2024). Some approaches use data valuation and attribution methods such as influence function (Hampel, 1974; Koh and Liang, 2017; Xia et al., 2024), Leave-One-Out (Hastie et al., 2009), and Shapley value (Ghorbani and Zou, 2019). These approaches perturb one data point from the training set to get a new subset, with or without permutations, to trace the change of the model validation performance, and quantify the effect of that point. However, such process requires re-training the model on each subset that

[3] Snopes.com

excludes each perturbed data point to be evaluated, which will cause expensive computational costs (Park et al., 2023). Additionally, it requires labeled validation sets to trace performance changes, which differs from our setting.

Recently, distributional distance metrics such as Wasserstein distance and KL divergence, have emerged as effective proxies for conducting data selections by providing an upper bound on performance change (Courty et al., 2017; Nguyen et al., 2021; Just et al., 2023; Li et al., 2023, 2024). As model-agnostic approaches, they are more desirable in the context of LLMs. There have been some successful applications for either pre-training or pre-fine-tuning (Xie et al., 2023; Everaert and Potts, 2023; Kang et al., 2024). For example, DSIR (Xie et al., 2023) proposes importance resampling on the hashed n-gram features of text data, and measure the distributional similarity based on KL-reduction. GIO (Everaert and Potts, 2023) uses iterative gradient methods to prune training samples by minimizing KL-divergence. GOT-D (Kang et al., 2024) enjoys the geometric property of Optimal Transport to select data that nudges the pre-training distribution closer to the target distribution. However, these methods are not directly applicable to multimodal data due to the difficulties in quantifying the probability density of multimodal misinformation data, typically necessitating tens of thousands of data points. Additionally, applying these textual data selection methods in the context of MMD task is incorrect. The factuality of a piece of information is determined by the combined information from text and image in MMD. Relying solely on a single modality fails to capture cases like out-of-context misuse, where the standalone textual claim might be accurate while the image-text pair together conveys false information.

Several studies in multimodal data selection primarily focus on sampling large corpora, ranging from millions to billions of samples, from vast volumes of noisy, web-curated datasets (Rao et al., 2020; Gadre et al., 2023; Yu et al., 2023; Maini et al., 2023; Wang et al., 2024b). These methods involve non-trivial data-dependent filtering strategies to select high-quality data for pre-training, such as CLIPScore (Hessel et al., 2021) that ranks each data point based on the cosine similarity scores between its CLIP image and text embeddings. Different from these works, our goal is to select a small number of useful synthetic instances for fine-tuning, with the aim of aligning distributions between synthetic and real-world datasets for MMD. Another line of work focuses on selecting features or modalities to boost model performance (Kamyab and Eftekhari, 2016; Zhang et al., 2019; Lu et al., 2021; Zhang et al., 2023; He et al., 2024), which is beyond the scope of this paper.

We employ two model-agnostic data selection approaches, i.e., semantic similarity-based selection and distributional similarity based selection, which provide straightforward, efficient, and scalable estimates of synthetic data relevance, facilitating the tuning of MLLMs for detecting real-world multimodal misinformation.

Feature Extraction. We adopt a mixed-modal encoder, consisting of a text encoder and an image encoder, to extract multimodal features for data selection. It is a simple extension of the off-the-shelf CLIP model (Radford et al., 2021a), which has demonstrated good performance in multimodal retrieval tasks (Yasunaga et al., 2023). Specifically, given an image-claim pair x_i, we obtain the visual features and textual features via the image encoder and text encoder. This pair's multimodal features $\mathbf{e}_i$ are then computed by averaging these two modalities, with the L_2 norm scaled to 1.

Semantic Similarity (SemSim). Similarity between the synthetic training set $\mathcal{D}_s$ and the validation set $\mathcal{D}_u$ can be used for data selection. Specifically, the semantic features of instances in each set are obtained using the mixed-modal encoder. Then, we calculate the similarity between the multimodal features $\mathbf{e}_i$ of a training instance and the averaged multimodal features $\mathcal{E}_u = \frac{1}{|\mathcal{D}_u|}\sum_{j=1}\mathbf{e}_j$ of the validation set $\mathcal{D}_u$. Subsequently, we select the data points with the highest similarity score to construct $\mathcal{D}_v$. We use the cosine similarity as the measure, defined as follows:

$$\mathcal{G}_{\text{ss}}(\mathbf{e}_i, \mathcal{E}_u) = \frac{\mathbf{e}_i \cdot \mathcal{E}_u}{||\mathbf{e}_i|| \cdot ||\mathcal{E}_u||}. \quad (1)$$

Distributional Similarity (DisSim). Wasserstein distance (Kolouri et al., 2017) is a distributional measure and has been shown to provide an upper bound on the difference in a model's performance when it is trained on one distribution and evaluated on another. With this in mind, the selection strategy for the featurized synthetic set $\mathbf{e}^s$ is to prioritize data points that could increase the density around the featurized target set $\mathbf{e}^u$. Formally, the p-Wasserstein distance between two probability

measures $\mathbf{e}^s$ and $\mathbf{e}^u$ is defined as follows:

$$\begin{aligned}&\mathcal{W}_p(\mathbf{e}^s, \mathbf{e}^u)\\ &\overset{(a)}{=}\Big(\inf_{\pi\in\Pi(\mathbf{e}^s,\mathbf{e}^u))}\int c^p(\mathbf{e}^s_i,\mathbf{e}^u_j)d\pi(\mathbf{e}^s_i,\mathbf{e}^u_j)\Big)^{\frac{1}{p}},\\ &\overset{(b)}{=}\max_{(f,g)\in C^0(\mathcal{X})^2}\langle\, f, \mathbf{e}^s\,\rangle + \langle\, g, \mathbf{e}^u\,\rangle \qquad (2)\end{aligned}$$

where the equality (a) comes from the definition of the Wasserstein distance, which is the primal problem. Specifically, $c^p(\mathbf{e}^s_i, \mathbf{e}^u_j) = ||\mathbf{e}^s_i - \mathbf{e}^u_j||^p_p$ and $p \geq 1$ for $i \in \{1, \cdots, |\mathcal{D}_s|\}$ and $j \in \{1, \cdots, |\mathcal{D}_u|\}$, represents the pairwise distance metric. Without loss of generality, we set $p = 2$. $\pi \in \Pi(\mathbf{e}^s, \mathbf{e}^u)$ is the joint distribution of $\mathbf{e}^s$ and $\mathbf{e}^u$, and any $\pi^\star$ attains the minimum value of equality (a) is considered as an OT plan (Villani et al., 2009), as it is the most cost-effective strategy to make synthetic set be transformed into the target set. Kantorovich formulation (Kantorovich, 1942) defines the OT problem as a Linear Program. Then, based on the duality theorem, the equality (b) holds if $\pi^\star$ and $(f^\star, g^\star)$, are optimal variables of the corresponding primal and dual problem, respectively. Specifically, $f \in \mathbb{R}^{|\mathcal{D}_s|}$ and $g \in \mathbb{R}^{|\mathcal{D}_u|}$, and $C^0(\mathcal{X})$ is the set of all continuous functions over the feature space $\mathcal{X}$.

We select the highest-scored data points with the largest negative calibrated *wasserstein gradient*, defined as follows:

$$\mathcal{J}_{\text{wg}}(\mathbf{e}^s_i, \mathbf{e}^u) = f^\star_i - \sum_{j\in\{1,\cdots,|\mathcal{D}_s|\}\setminus i}\frac{f^\star_j}{|\mathbf{e}^s| - 1}, \quad (3)$$

which measures the sensitivity of the i-th data point of the synthetic training set for $\mathcal{W}_p(\mathbf{e}^s, \mathbf{e}^u)$. This gradient value determines the shifting direction based on whether it is positive or negative. If the value is positive (negative), shifting more probability mass to that datum will result in an increase (decrease) of the distance between the synthetic set and the validation set.

5 Experimental Evaluation

5.1 Experimental Settings

Datasets. As introduced in §3, we uses NewsCLIPpings (Luo et al., 2021), DGM[4] (Shao et al., 2023) and Autosplice (Jia et al., 2023) to construct a large synthetic dataset as training set. We evaluate models on the **MediaEval** (Boididou et al., 2016) and the **Snopes** datasets. Since the Snopes dataset only includes instances labeled as True/False but excludes those labeled as Miscaptioned[4] (Zlatkova et al., 2019) that fall into OOC category, we observe a disproportionately high volume of manipulation instances. For a balanced evaluation, we additionally create a variant dataset **Snopes (O+)** with increased OOC ratio based on the original one by mismatching the image and text of around 50% instances of Snopes.

Base MLLMs. We employ LLaVA-NeXT-13B[5] (Liu et al., 2024a) as the base MLLM for fine-tuning, given its excellent performance on various multimodal tasks compared to other MLLMs. Additionally, we perform an ablation study on different base models, including LLaVA-7B and mPLUG-Owl2 (Ye et al., 2023). More details are in Appendix B.1.1.

Baselines. We compare with full-dataset fine-tuned LLaVA (Liu et al., 2024a) models: $\text{LLaVA}_{\text{M}=\text{S}}$, which are the models 1) trained on full set of MediaEval and evaluated on Snopes and Snopes (O+); and 2) trained on Snopes and evaluated on MediaEval. We also compare with the random selection, where we randomly sample instances from the synthetic dataset for fine-tuning. We also include baselines using strong MLLMs as misinformation detectors, including LLaVA (Liu et al., 2024a) and GPT-4V (OpenAI, 2023b) , where we directly prompt them to generate predictions.

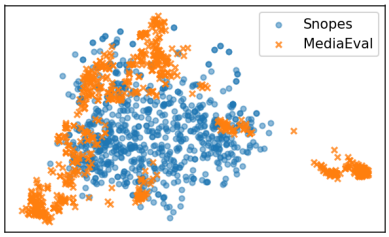
Default Settings. We utilize the off-the-shelf CLIP model (ViT-L/14) (Radford et al., 2021b) for feature extraction. We use the same prompt template for all models to ensure a fair comparison. Each data selection method empirically selects 750 synthetic instances from the synthetic dataset $\mathcal{D}_s$ to construct $\mathcal{D}_v$ for fine-tuning, and we ensure the class distribution of selected set is balanced. The unlabeled validation set $\mathcal{D}_u$ contains 5% instances randomly sampled from each test set. We conduct all experiments with the selection methods three times using different random seeds, and report the mean macro-F1 and standard deviation of each metric across these three runs. More details are provided in Appendix B.1.2.

[4] https://www.snopes.com/fact-check/rating/miscaptioned/

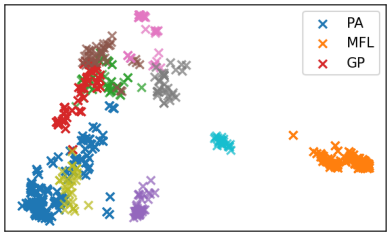
[5] In the rest of the paper, we refer to LLaVA-NeXT-13B as LLaVA for brevity.

Setting		Direct Prompting		Real-world	Synthetic		
Method	Majority	LLaVA	GPT-4V	$\text{LLaVA}_{\text{M}\rightleftharpoons\text{S}}$	LLaVA_{R}	LLaVA_{S}	LLaVA_{D}
MediaEval	0.368	0.480	0.595	0.431	$0.568_{(0.020)}$	$\mathbf{0.687}_{(0.015)}$	$\underline{0.611}_{(0.009)}$
Snopes	0.335	0.407	**0.614**	0.511	$\underline{0.540}_{(0.020)}$	$0.521_{(0.022)}$	$0.496_{(0.027)}$
Snopes (O+)	0.335	0.399	0.394	0.548	$\underline{0.758}_{(0.024)}$	$0.716_{(0.037)}$	$\mathbf{0.813}_{(0.017)}$

Table 1: Results of Majority, direct prompting LLaVA and GPT-4V, full-dataset fine-tuning on the other real-world dataset ($\text{LLaVA}_{\text{M}\rightleftharpoons\text{S}}$), and fine-tuning on a small set of selected synthetic data (LLaVA_{R}: Random, LLaVA_{S}: SemSim, LLaVA_{D}: DisSim). We report the macro-F1 averaged over 3 trials with different random seeds. The best results for each dataset are in **bold** while the second-best results are underlined. The standard deviation is in (.).



(a)



(b)

Model	PA	MFL	GP
SemSim	0.781 (0.020)	0.356 (0.115)	0.323 (0.078)
DisSim	0.822 (0.027)	0.333 (0.077)	0.375 (0.044)

(c)

Figure 2: The 2D projection of the multimodal features using PCA. **(a)** MediaEval and Snopes datasets; **(b)** Top-10 events with the most instances in MediaEval. Each colored group represents an event, and three highlighted groups are top-3 events; **(c)** The F1 score of semantic and distributional selection methods on top-3 events with the most instances from MediaEval dataset. (.) encloses standard deviation. PA: Paris Attack; MFL: Mt Fuji Lenticular; GP: German Protest.

5.2 Experimental Results

5.2.1 Main results and analysis

We present main results in Table 1, and provide an in-depth analysis.

Directly prompt base model. In Table 1, we observe that directly prompting the LLaVA model does not yield satisfactory performance, indicating that merely relying on the inherent knowledge of base MLLM is insufficient for MMD, especially when the model size is not large enough. We conjecture that task-specific information might be necessary to induce this capability in the base model.

Utilize OOD real-world data. However, incorporating task-specific knowledge by training on one real-world fact-checking dataset does not consistently enhance performance on another; it may even impede it. As shown in Table 1, training LLaVA on the MediaEval dataset enhances its detection performance on Snopes and Snopes (O+) with an absolute increase of 0.104 and 0.149 respectively, but the reverse adversely affects its performance, where we observe a decline of 0.09. We examine this inconsistency by checking the distributions of Snopes and MediaEval datasets.

Specifically, we employ Principal Component Analysis (PCA) (Pearson, 1901) to project the multimodal features of each instance from both two datasets. As depicted in Figure 2a, the MediaEval dataset exhibits a feature distribution characterized by multiple clusters, displaying OOD traits when compared across different clusters. This is attributed to MediaEval encompassing various events, each comprising posts with high internal density. This observation is further supported by the visualization of the top-10 events with the most instances in MediaEval in Figure 2b.

In contrast, the distribution of the Snopes dataset lacks such characteristics, as the instances comprise independent claims from fact-checking websites. Therefore, training on such diverse yet insufficient instances might induce catastrophic forgetting behaviors, leading to dramatic performance degradation. Conversely, when there are more similar instances within each cluster of MediaEval, the model can learn the general semantic patterns from each group more effectively, thereby enhancing its capability in misinformation detection. These findings suggest that direct training on real-world data cannot guarantee robust generalization, particularly

when these real-world data are often OOD relative to each other.

Utilize synthetic data. Compared to using OOD real-world data, a small amount of selected synthetic data consistently enhances base model's performance on various real-world datasets, with absolute F1 score improvements ranging from 0.133 to 0.414, as observed in Table 1. This highlights the promising potential of leveraging synthetic data for real-world MMD. Next, we provide a detailed analysis of specific results from the standpoint of data distribution, offering insights for selecting synthetic data for real-world MMD.

Firstly, both similarity-based data selection methods, SemSim and DisSim, enable the base MLLM to outperform GPT-4V on MediaEval and Snopes (O+) datasets respectively. This confirms the effectiveness of these methods and demonstrate their applicability in real-world scenarios. In Table 1, SemSim achieves the best F1 score on the MediaEval dataset (0.687) surpassing DisSim (0.611), while DisSim exhibits the best on the Snopes (O+) dataset (0.813) outperforming SemSim (0.716). We hypothesize this phenomenon is influenced by the choice of similarity metrics.

Semantic similarity and Wasserstein distance, provide distinct criteria for evaluating the similarities of data points. Semantic similarity primarily assesses the resemblance of individual points in feature space, selecting data samples that closely match the representations of the target set. In contrast, Wasserstein distance excels at discerning similarities across different distributions, prioritizing the construction of the mean distribution among all probability measures to minimize the total transport cost, rather than matching individual clusters. To delve deeper, we perform data selection and evaluation for each of the top-3 events (as other events lack sufficient instances) and report the macro-F1 in Table 2c. We find that when selecting synthetic data for a specific event, DisSim outperforms SemSim by 5% on average, aligning with our earlier hypothesis. These insights suggest the importance of employing appropriate criteria for data selection to enhance results on real-world data.

Secondly, all the data selection methods improve the base model on Snopes dataset, narrowing the performance gap between the base MLLM and GPT-4V. However, the magnitude of the improvements become smaller compared to the MediaEval and Snopes (O+) datasets. For example, as shown in Table 1, DisSim improves the base model performance from 0.399 to 0.813 on Snopes (O+), whereas its improvement is from 0.407 to 0.496 on Snopes. The reason is likely two-fold: 1) These two datasets have different evaluation focuses–Snopes requires models to have strong manipulation detection capabilities while Snopes (O+) emphasizes OOC detection. Additionally, OOC instances are easier to obtain than manipulation instances requiring data-specific adjustments, which may result in more OOC instances in the synthetic data. Consequently, when the synthetic data distribution has more OOC instances as support, DisSim can better identify valuable instances based on gradient direction and magnitude, aligning the training distribution of the base MLLM closer to the target distribution. 2) Detecting image manipulation usually requires the base MLLMs have strong visual grounding capabilities, while most open-source MLLMs exhibit some systematic visual shortcomings because their pre-trained CLIP vision encoders might overlook visual details in images (Tong et al., 2024). These observations suggest the importance of increasing diversity of misinformation data in synthetic datasets and enhancing the capability of vision encoder for better real-world MMD.

5.2.2 Impact of base MLLMs

In Table 2, we report the results of SemSim and DisSim across different base models. Fine-tuning on a small number of synthetic data selected by both methods consistently improves the performance of the base models, including MLLMs with smaller model size (i.e., LLaVA-7B) and from different families (i.e., mPLUG-Owl2 (Ye et al., 2023)). The maximum F1 score improvement on all datasets for LLaVA-7B is 0.26 on average and for mPLUG-Owl2 is 0.19, which indicates that the selected synthetic data is generalizably useful, enhancing the MMD performance of various MLLMs. Notably, as SemSim and DisSim are model-agnostic data selection approaches, the selected synthetic data can be reused without further selection costs, endowing efficiency to their deployment on real-world data.

5.2.3 Impact of the number of selected data

We investigate whether increasing the number of selected synthetic instances improves the performance of the base MLLM. The overall results in Figure 3 indicate that increasing the number of selected synthetic instances does not necessarily improve the performance of the base MLLM. A

	LLaVA-7B				mPLUG-Owl2			
	Prompt	SemSim	DisSim	Δ	Prompt	SemSim	DisSim	Δ
MediaEval	0.410	**0.667** (0.019)	0.621 (0.057)	0.257↑	0.294	**0.685** (0.069)	0.412 (0.037)	0.391↑
Snopes	0.337	**0.489** (0.033)	0.462 (0.034)	0.152↑	0.332	0.368 (0.036)	**0.405** (0.058)	0.073↑
Snopes (O+)	0.373	0.689 (0.009)	**0.752** (0.042)	0.379↑	0.285	0.386 (0.014)	**0.402** (0.037)	0.117↑

Table 2: Results of MLLMs with different model scales and families. (.) encloses standard deviation.

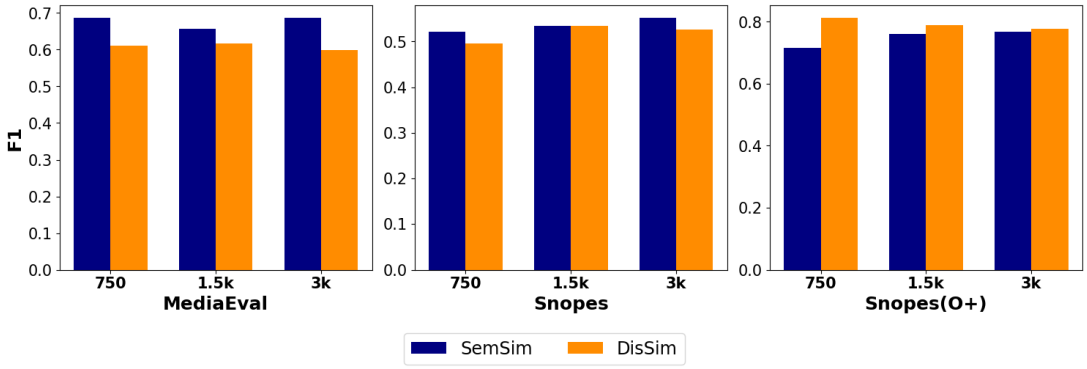


Figure 3: The F1 score with increasing the number of selected synthetic instances for training.

small number of selected synthetic data already contain data points with most of the relevant information for MMD on real-world data. We observe that increasing selected training instances slightly hinders its performance on MediaEval and Snopes (O+). This is because, on MediaEval, as the size increases, the selected data instances tend to converge towards a "mean" distribution of all events as discussed before, thus deviating further from the distribution of individual cluster. And we hypothesize that the small number of selected data by DisSim on Snopes (O+) already encapsulates the traits of OOC misinformation, and more data may introduce additional irrelevant or detrimental data points. Determining an optimal number is a problem beyond the scope of our current work.

6 Related Work

Multimodal Misinformation Datasets. Real-world datasets are typically collected from social media posts like Twitter (Gupta et al., 2013; Boididou et al., 2016), or fact-checking websites like Snopes (Zlatkova et al., 2019; Yao et al., 2023), but such collection processes are labor-intensive and costly. Another line of datasets consist of images and captions collected from Reddit[6] posts, with each instance labeled based on subreddit themes (Heller et al., 2018; Huh et al., 2018; Nakamura et al., 2020). For example, all posts from the subreddit "usnews" are labeled as real since they are authentic news, while the posts from the subreddit "photoshopbattles" are labeled as fake since this subreddit is for users to battle using image manipulation software. However, these datasets do not reflect the complexities of real-world circulating misinformation.

Recently, synthetic datasets have emerged as a cost-effective data solution (Aneja et al., 2023; Luo et al., 2021; Shao et al., 2023; Jia et al., 2023), which can be generated in a large scale by using generative AI technology. For instance, the synthetic datasets NewsCLIPings (Luo et al., 2021) swap the captions of different news images, and MEIR manipulates the news captions (Sabir et al., 2018) to create OOC misinformation. On the other hand, some datasets manipulate the image of news and combine it with text manipulation to generate fake versions (Shao et al., 2023; Jia et al., 2023). However, such datasets might not reflect the data distribution of real-world misinformation. In this paper, we aim at exploring how well the models trained on synthetic data can generalize to real-

[6] https://www.reddit.com/

world fact-checking datasets.

Multimodal Misinformation Detection. Most existing methods for MMD are typically trained in a fully supervised manner using the entire training set and evaluated on the test set from the same small-scale real-world fact-checking dataset (Jin et al., 2017; Wang et al., 2018; Khattar et al., 2019; Wu et al., 2021; Qian et al., 2021; Ghorbanpour et al., 2023; Zhou et al., 2023; Liu et al., 2023c; Chen et al., 2023), or from the same large-scale synthetic dataset (Aneja et al., 2023; Abdelnabi et al., 2022; Shao et al., 2023; Jia et al., 2023; Mu et al., 2023; Wang et al., 2024a; Qi et al., 2024). Meanwhile, by integrating external world knowledge, Xuan et al. (2024) improve the performance of GPT-4V (OpenAI, 2023b) on real-world fact-checking datasets. Liu et al. (2024b) improve cross-domain performance of PandaGPT (Su et al., 2023) on a synthetic dataset, which allows the model trained on instances generated using news from one agency (e.g., BBC) to generalize to those generated using news from other sources (e.g., USA TODAY). Given the discrepancy in size between synthetic and real-world datasets, our work aims to enhance the generalizability of models trained on synthetic data to test on real-world misinformation datasets using data selection methods, which provides an cost-effective solution for MMD.

7 Conclusion and Future Work

We propose leveraging synthetic data to address the MMD task in the scenarios of real-world data scarcity, in which relevant and valuable instances can be selected from a large-scale synthetic dataset for fine-tuning MLLMs. We advocate for utilizing two model-agnostic similarity-based data selection approaches for this task. Results demonstrate that even using a very small number of selected synthetic training instances can significantly boost MLLMs' detection performance on real-world fact-checking data, enabling a small MLLM, LLaVA-13B, to outperform GPT-4V.

In future research, we plan to determine the optimal number of selected instances for different data selection methods for maximizing the effectiveness of useful training data. Additionally, we plan to explore more design solutions to obtain better multimodal features for further improving misinformation detection performance.

8 Limitations

Although a small number of selected synthetic data can significantly boost the base MLLM's detection performance, there might be an optimal number of synthetic instances that yield best results. Therefore, determining such optimal numbers is crucial to optimize the effectiveness of training instances. Additionally, our experiments show that both similarity-based selection methods, SemSim and DisSim, enhance performance. However, SemSim performs better on MediaEval, while DisSim excels on Snopes. This indicates the need for further exploration to develop a unified solution capable of handling different data distributions. Although experimental results show that using a simple extension on the frozen CLIP model help select useful synthetic data for improving MMD, relying on the CLIP model without fine-tuning for measuring semantic similarity introduces issues like potential bias. This suggests the need for further exploration on how to obtain better multimodal features for improving detection performance. Lastly, although all data selection methods consistently improve the performance of the base MLLM on the Snopes dataset, they still underperform compared to GPT-4V. This suggests further study on strategy of constructing synthetic datasets for different types of multimodal misinformation.

9 Ethics Statement

Our research focuses on detecting real-world multimodal misinformation using synthetic datasets generated by AI technologies, offering a solution to address the scarcity of real-world fact-checking data and enhances the effectiveness of misinformation detection. Our method is intended for research purposes. To ensure responsible use and prevent potential misuse, we emphasize the necessity of human oversight during utilization to avoid unintended consequences.

All datasets used in our experiments, both synthetic and real-world, are obtained from publicly available sources commonly used in multimodal misinformation detection research. The licenses for public datasets are listed in Appendix D. There are some image-text pairs in the used datasets include misleading content that may be disturbing to certain celebrity identities.

References

Sara Abdali. 2022. Multi-modal misinformation detection: Approaches, challenges and opportunities. *CoRR*, abs/2203.13883v5.

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

Mubashara Akhtar, Michael Schlichtkrull, Zhijiang Guo, Oana Cocarascu, Elena Simperl, and Andreas Vlachos. 2023. Multimodal automated fact-checking: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5430–5448, Singapore. Association for Computational Linguistics.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. 2024. A survey on data selection for language models. *CoRR*, abs/2402.16827.

Shivangi Aneja, Chris Bregler, and Matthias Nießner. 2023. 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.

Christina Boididou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, Stuart E. Middleton, Andreas Petlund, and Yiannis Kompatsiaris. 2016. Verifying multimedia use at mediaeval 2016. In *Working Notes Proceedings of the MediaEval 2016 Workshop, Hilversum, The Netherlands, October 20-21, 2016*, volume 1739 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. 2020. Simswap: An efficient framework for high fidelity face swapping. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 2003–2011. ACM.

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 627–638, Toronto, Canada. Association for Computational Linguistics.

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. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. 2017. Joint distribution optimal transportation for domain adaptation. *Advances in neural information processing systems*, 30.

Dante Everaert and Christopher Potts. 2023. Gio: Gradient information optimization for training dataset selection. *arXiv preprint arXiv:2306.11670*.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah M. Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander J. Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. 2023. Datacomp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. 2021. Information bottleneck disentanglement for identity swapping. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3404–3413. Computer Vision Foundation / IEEE.

Amirata Ghorbani and James Zou. 2019. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pages 2242–2251. PMLR.

Faeze Ghorbanpour, Maryam Ramezani, Mohammad A. Fazli, and Hamid R. Rabiee. 2023. FNR: a similarity and transformer-based approach to detect multimodal fake news in social media. *Soc. Netw. Anal. Min.*, 13(1):56.

Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13 Companion, page 729–736, New York, NY, USA. Association for Computing Machinery.

Frank R Hampel. 1974. The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Yifei He, Runxiang Cheng, Gargi Balasubramaniam, Yao-Hung Hubert Tsai, and Han Zhao. 2024. Efficient modality selection in multimodal learning. *Journal of Machine Learning Research*, 25(47):1–39.

Silvan Heller, Luca Rossetto, and Heiko Schuldt. 2018. The ps-battles dataset - an image collection for image manipulation detection. *CoRR*, abs/1804.04866.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Minyoung Huh, Andrew Liu, Andrew Owens, and Alexei A. Efros. 2018. Fighting fake news: Image splice detection via learned self-consistency. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI*, volume 11215 of *Lecture Notes in Computer Science*, pages 106–124. Springer.

Ayush Jaiswal, Ekraam Sabir, Wael AbdAlmageed, and Premkumar Natarajan. 2017. Multimedia semantic integrity assessment using joint embedding of images and text. In *Proceedings of the 25th ACM International Conference on Multimedia*, MM '17, page 1465–1471, New York, NY, USA. Association for Computing Machinery.

Shan Jia, Mingzhen Huang, Zhou Zhou, Yan Ju, Jialing Cai, and Siwei Lyu. 2023. Autosplice: A text-prompt manipulated image dataset for media forensics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pages 893–903. IEEE.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 795–816. ACM.

Hoang Anh Just, Feiyang Kang, Tianhao Wang, Yi Zeng, Myeongseob Ko, Ming Jin, and Ruoxi Jia. 2023. LAVA: data valuation without pre-specified learning algorithms. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Shima Kamyab and Mahdi Eftekhari. 2016. Feature selection using multimodal optimization techniques. *Neurocomputing*, 171:586–597.

Feiyang Kang, Hoang Anh Just, Yifan Sun, Himanshu Jahagirdar, Yuanzhi Zhang, Rongxing Du, Anit Kumar Sahu, and Ruoxi Jia. 2024. Get more for less: Principled data selection for warming up fine-tuning in llms. *arXiv preprint arXiv:2405.02774*.

Leonid V Kantorovich. 1942. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, WWW '19, page 2915–2921, New York, NY, USA. Association for Computing Machinery.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR.

Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde. 2017. Optimal mass transport: Signal processing and machine-learning applications. *IEEE signal processing magazine*, 34(4):43–59.

Wenqian Li, Shuran Fu, Fengrui Zhang, and Yan Pang. 2023. Data valuation and detections in federated learning. *arXiv preprint arXiv:2311.05304*.

Wenqian Li, Haozhi Wang, Zhe Huang, and Yan Pang. 2024. Private wasserstein distance with random noises. *arXiv preprint arXiv:2404.06787*.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2021. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6761–6771, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Hui Liu, Wenya Wang, and Haoliang Li. 2023c. Interpretable multimodal misinformation detection with logic reasoning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9781–9796, Toronto, Canada. Association for Computational Linguistics.

Xuannan Liu, Peipei Li, Huaibo Huang, Zekun Li, Xing Cui, Jiahao Liang, Lixiong Qin, Weihong Deng, and Zhaofeng He. 2024b. Fakenewsgpt4: Advancing multimodal fake news detection through knowledge-augmented lvlms. *CoRR*, abs/2403.01988.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv*, abs/1907.11692v1.

Xiaopeng Lu, Zhen Fan, Yansen Wang, Jean Oh, and Carolyn P. Rosé. 2021. Localize, group, and select: Boosting text-vqa by scene text modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2631–2639.

Grace Luo, Trevor Darrell, and Anna Rohrbach. 2021. NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6801–6817, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pratyush Maini, Sachin Goyal, Zachary C. Lipton, J. Zico Kolter, and Aditi Raghunathan. 2023. T-MARS: improving visual representations by circumventing text feature learning. *CoRR*, abs/2307.03132.

Michael Mu, Sreyasee Das Bhattacharjee, and Junsong Yuan. 2023. Self-supervised distilled learning for multi-modal misinformation identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2819–2828.

Eric Müller-Budack, Jonas Theiner, Sebastian Diering, Maximilian Idahl, and Ralph Ewerth. 2020. Multimodal analytics for real-world news using measures of cross-modal entity consistency. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, ICMR '20, page 16–25, New York, NY, USA. Association for Computing Machinery.

Kai Nakamura, Sharon Levy, and William Yang Wang. 2020. Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6149–6157, Marseille, France. European Language Resources Association.

A Tuan Nguyen, Toan Tran, Yarin Gal, Philip HS Torr, and Atılım Güneş Baydin. 2021. Kl guided domain adaptation. *arXiv preprint arXiv:2106.07780*.

OpenAI. 2023a. GPT-4 technical report. *CoRR*, abs/2303.08774v4.

OpenAI. 2023b. Gpt-4v(ision) system card.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. 2023. TRAK: attributing model behavior at scale. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 27074–27113. PMLR.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. Styleclip: Text-driven manipulation of stylegan imagery. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 2065–2074. IEEE.

Karl Pearson. 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Peng Qi, Zehong Yan, Wynne Hsu, and Mong-Li Lee. 2024. SNIFFER: multimodal large language model for explainable out-of-context misinformation detection. *CoRR*, abs/2403.03170.

Shengsheng Qian, Jinguang Wang, Jun Hu, Quan Fang, and Changsheng Xu. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 153–162, New York, NY, USA. Association for Computing Machinery.

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. 2021a. 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*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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. 2021b. 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*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125.

Roshan Rao, Sudha Rao, Elnaz Nouri, Debadeepta Dey, Asli Celikyilmaz, and Bill Dolan. 2020. Quality and relevance metrics for selection of multimodal pretraining data. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14-19, 2020*, pages 4109–4116. Computer Vision Foundation / IEEE.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Ekraam Sabir, Wael AbdAlmageed, Yue Wu, and Prem Natarajan. 2018. Deep multimodal image-repurposing detection. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, page 1337–1345, New York, NY, USA. Association for Computing Machinery.

Rui Shao, Tianxing Wu, and Ziwei Liu. 2023. 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.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *CoRR*, abs/2305.16355.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *CoRR*, abs/2401.06209.

Cédric Villani et al. 2009. *Optimal transport: old and new*, volume 338. Springer.

Bin Wang and C.-C. Jay Kuo. 2020. SBERT-WK: A sentence embedding method by dissecting bert-based word models. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:2146–2157.

Jiazhen Wang, Bin Liu, Changtao Miao, Zhiwei Zhao, Wanyi Zhuang, Qi Chu, and Nenghai Yu. 2024a. Exploiting modality-specific features for multi-modal manipulation detection and grounding. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4935–4939.

Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. 2022. High-fidelity GAN inversion for image attribute editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11369–11378. IEEE.

Weizhi Wang, Khalil Mrini, Linjie Yang, Sateesh Kumar, Yu Tian, Xifeng Yan, and Heng Wang. 2024b. Finetuned multimodal language models are high-quality image-text data filters. *CoRR*, abs/2403.02677.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 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, KDD 2018, London, UK, August 19-23, 2018*, pages 849–857. ACM.

Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2560–2569, Online. Association for Computational Linguistics.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. 2023. Data selection for language models via importance resampling. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

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

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. 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*, page 2733–2743, New York, NY, USA. Association for Computing Machinery.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-Tau Yih. 2023. Retrieval-augmented multimodal language modeling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 39755–39769. PMLR.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *CoRR*, abs/2311.04257v2.

Haichao Yu, Yu Tian, Sateesh Kumar, Linjie Yang, and Heng Wang. 2023. The devil is in the details: A deep dive into the rabbit hole of data filtering. *CoRR*, abs/2309.15954.

Dong Zhang, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Effective sentiment-relevant word selection for multi-modal sentiment analysis in spoken language. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 148–156, New York, NY, USA. Association for Computing Machinery.

Xinyu Zhang, Zhiwei Li, Zhenhong Zou, Xin Gao, Yijin Xiong, Dafeng Jin, Jun Li, and Huaping Liu. 2023. Informative data selection with uncertainty for multimodal object detection. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–13.

Bolei Zhou, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1452–1464.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. 2022. Detecting twenty-thousand classes using image-level supervision. *CoRR*, abs/2201.02605.

Yangming Zhou, Yuzhou Yang, Qichao Ying, Zhenxing Qian, and Xinpeng Zhang. 2023. Multi-modal fake news detection on social media via multi-grained information fusion. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, ICMR '23, page 343–352, New York, NY, USA. Association for Computing Machinery.

Dimitrina Zlatkova, Preslav Nakov, and Ivan Koychev. 2019. 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)*, pages 2099–2108, Hong Kong, China. Association for Computational Linguistics.

A Multimodal Misinformation Data

A.1 Common Categories

There is no standard categorizations for multimodal information. We categorize them into two types based on whether the image content are falsely altered or counterfeited. Next, we provide a more formulated and detailed introduction. Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{|\mathcal{D}|}$ be a multimodal misinformation dataset with $|\mathcal{D}|$ instances, where each instance consists of the input x_i and the ground-truth label $y_i \in \{0, 1\}$. We denote $x_i = (\mathbf{m}_i, \mathbf{t}_i)$ which is a pair of information containing an image $\mathbf{m}_i$ and a corresponding claim text $\mathbf{t}_i$, and y_i indicates the veracity of x_i is true if $y_i = 0$ or false otherwise.

Out-of-Context (OOC) Misuse occurs when the textual claim misrepresents the original context or intent of the image, conveying misleading information. Obtaining OOC image-text pairs by pairing a textual claim with an image taken out-of-context is a straightforward and effective method (Luo et al., 2021; Akhtar et al., 2023). Specifically, a mismatched pair $(\mathbf{m}_i, \mathbf{t}_j)$ is created by using two pristine instances x_i and x_j $(j \neq i)$. An alternative method involves manipulating the textual claim (Sabir et al., 2018; Müller-Budack et al., 2020), by editing the original textual claim $\mathbf{t}_i$ to produce $\tilde{\mathbf{t}}_i$ that is inconsistent with $\mathbf{m}_i$, which may include replacing named entities, altering sentiments or stances, selectively quoting, and other techniques aimed at distorting its meaning to conform to a false narrative. In this way, a genuine image is accompanied with inconsistent text, conveying misleading information as the text may falsely describe the image's origin, context, and meaning. Such OOC pairs not only deceive human but also pose significant challenges for automatic misnformation detection models (Luo et al., 2021).

Content Manipulation involves intentionally altering image and text modality with the aim of deceiving or misleading. As it can be difficult to consistently obtain real images to support non-factual claim, manipulating image and text content to generate misinformation becomes an alternative approach (Abdali, 2022). Image manipulation refers to altering an image $\mathbf{m}_i$ to obtain a fake image $\tilde{\mathbf{m}}_i$ by modifying the elements within it using techniques such as copying and pasting specific regions, splicing images, removing content, and face swap (Shao et al., 2023; Jia et al., 2023). Image and text manipulation techniques can be combined to obtain $(\tilde{\mathbf{m}}_i, \tilde{\mathbf{t}}_i)$, ensuring that both visual and textual elements complement each other, thereby strengthening the deceptive information.

A.2 Training and Evaluation Datasets

A.2.1 Training Datasets

NewsCLIPpings (Luo et al., 2021) is an automatically generated OOC multi-modal dataset that contains both pristine and falsified instances. It is derived from the VisualNews corpus (Liu et al., 2021), which contains image-caption pairs from popular news agencies. A falsified instance is an unmanipulated but mismatched image-caption pair, constructing by pairing an image with a caption from an inconsistent context. Given a pristine instance $(\mathbf{m}_i, \mathbf{t}_i)$ as a query, another instance $(\mathbf{m}_j, \mathbf{t}_j)$ is retrieved from the news corpus to form the falsified instance $(\mathbf{m}_j, \mathbf{t}_i)$ $(j \neq i)$. The mismatching process is based on the similarity between $(\mathbf{m}_i, \mathbf{t}_i)$ and $(\mathbf{m}_j, \mathbf{t}_j)$, utilizing the CLIP semantic embeddings (Radford et al., 2021a), SBERT-WK text embedding (Wang and Kuo, 2020), and scene embeddings (Zhou et al., 2018). Such image-text mismatch procedure automatically generates around 988k unique synthetic instances. The falsified instances could portray subjects in an image as differ-

ent entities, depict specific individuals in a misleading context, and mislabel the event depicted within a particular scene.

DGM[4] (Shao et al., 2023) is a multimodal misinformation dataset, where synthetic instances are automatically generated by applying image and text manipulation techniques to pristine instances sourced from the VisualNews corpus. Specifically, for a pristine instance $(\mathbf{m}_i, \mathbf{t}_i)$, a manipulated image $\tilde{\mathbf{m}}_i$ is generated by employing (1) *face swap manipulation* via two representative face swap methods SimSwap (Chen et al., 2020) and InfoSwap (Gao et al., 2021), and (2) *face attribute manipulation*, which involves editing the emotion of the main character's face while preserving the identity using GAN-based methods, HFGI (Wang et al., 2022) and StyleCLIP (Patashnik et al., 2021). For text manipulation, $\tilde{\mathbf{t}}_i$ is generated using (1) Sentence-BERT (Reimers and Gurevych, 2019) to obtain the text embeddings of $\mathbf{t}_i$ and retrieve a different caption $\mathbf{t}_{j,j\neq i}$ that contains the same person entity as $\mathbf{t}_i$ but has low cosine similarity to the embeddings of $\mathbf{t}_i$, and (2) RoBERTa (Liu et al., 2019) to classify the sentiment of the $\mathbf{t}_i$ and then replace all sentiment words with the opposite sentiment text to get $\tilde{\mathbf{t}}_i$. After applying these manipulation methods, two manipulated images and two manipulated text captions are generated, along with the original image and caption, resulting in a total of 8 synthetic instances consisting of OOC and manipulation instances for each pristine instance, resulting in a total of 230k synthetic instances.

Autosplice (Jia et al., 2023) is a manipulated image dataset that utilizes a language-image model DALLE-2 (Ramesh et al., 2022), guided by textual prompts to edit images. Unlike DGM[4], where image and text manipulation are initially employed separately and then combined, Autosplice alters the image based on modified text, enabling a more direct and integrated manipulation process. Specifically, given a pristine image-caption pair $(\mathbf{m}_i, \mathbf{t}_i)$ from the VisualNews corpus, an object detection model, Detic (Zhou et al., 2022), is utilized to extract a list of object regions in $\mathbf{m}_i$, while a text parsing tool, spaCy[7], is employed to segment text terms in $\mathbf{t}_i$. Subsequently, human annotators select an object region and its corresponding text term, and then substitute this text term with a target generation term that is similar but inconsistent with $(\mathbf{m}_i, \mathbf{t}_i)$. The modified caption $\tilde{\mathbf{t}}_i$ and image $\mathbf{m}_i$ with the selected object region masked are constructed as inputs for the DALLE-2 model. The model performs local image editing on the masked region and generates the falsified image $\tilde{\mathbf{m}}_i$ based on the text prompt $\tilde{\mathbf{t}}_i$. To facilitate our task, we pair each modified image $\tilde{\mathbf{m}}_i$ with its altered caption $\tilde{\mathbf{t}}_i$, and label it with $y_i = 1$ to indicate false veracity. The pristine instance $(\mathbf{m}_i, \mathbf{t}_i)$ is labeled as $y_i = 0$ to denote truthfulness.

A.3 Evaluation Datasets

MediaEval (Boididou et al., 2016) collects a set of tweets associated images/videos with manually verified veracity (i.e., fake or real), which were spread around 17 widely attention-grabbing events such as the November 2015 Paris Attacks[8]. The number of instances for each event is small, and the classes are imbalanced. Some of these events were hoaxes, hence all instances related to them are fake. Also, there are several events that only contain real instances. For evaluation purposes, we utilize the image-text instances in the test set consisting of 702 instances, in which 292 of them are real and 410 are fake, and remove any replicated instances.

Snopes contains image-related claims from the popular fact-checking website Snopes[9]. It is initially based on the Fauxtography dataset (Zlatkova et al., 2019), which consists of image-claim pairs labeled as either True or False on Snopes and image-caption pairs from news agencies. However, using such news data in evaluation is inappropriate, as it may leak ground-truth labels given that synthetic data is generated using news data. Moreover, news data from trusted news agencies do not require fact-checking. Therefore, we filter the news instances and augment the dataset with more True-labeled instances by combining the fact-checked true image-related claims from the MOCHEG (Yao et al., 2023) dataset, designed to provide multimodal evidence for claims on Snopes, and additional image-related claims we collected directly from the Snopes website. These image-claim pairs are labeled as either True or False by fact-checkers on Snopes, consisting of 756 instances where which 376 of them are true and 380 are false.

[7] https://github.com/explosion/spaCy

[8] https://en.wikipedia.org/wiki/November_2015_Paris_attacks

[9] Snopes.com

B Experiments

B.1 Experimental Settings

B.1.1 Base MLLMs

LLaVA-NeXT-13B (Liu et al., 2024a), an improved version of LLaVA-1.5 (Liu et al., 2023a) and LLaVA (Liu et al., 2023b), is an end-to-end trained MLLM that connects the pre-trained CLIP vision encoder ViT-L/14 (Radford et al., 2021b) with an LLM Vicuna-13B (Chiang et al., 2023) using a two-layer MLP as the projection layer. It undergoes a two-stage instruction-tuning process to align two modalities for improving visual reasoning capabilities: (1) pre-training the projection layer with image-text pairs; (2) updating the weights of both the projection layer and LLM using multimodal instruction-tuning data generated by GPT-4 (OpenAI, 2023a). LLaVA-NeXT-7B is a smaller version based on an LLM Vicuna-7B.

mPLUG-Owl2 (Ye et al., 2023) is a versatile MLLM featuring a modularized network design and comprising a vision encoder, a visual abstractor, a text embedding layer, and a language decoder. In lieu of directly aligning the visual features with textual features, mPLUG-Owl2 integrates a modality-adaptive module within the language decoder. This module takes multimodal inputs, utilizing different parameters to project various modalities into a shared semantic space while preserving modality-specific features. The training process involves two stages: (1) pre-training the visual encoder, visual abstractor and newly added parameters of the modality-adaptive module within the language decoder using image-text pairs; (2) instruction-tuning the entire model on unimodal and multimodal instruction data.

B.1.2 Default Settings

We utilize the off-the-shelf CLIP model (ViT-L/14) (Radford et al., 2021b) for feature extraction. We use the original code and pre-trained checkpoint of LLaVA[10], with a vision encoder CLIP model (ViT-L/14) and an LLM Vicuna-13B (Chiang et al., 2023). We fine-tune the LLaVA model using the parameter-efficient fine-tuning method LoRA (Hu et al., 2022), with rank set to 128, α value to 256, and learned LoRA matrices for both the projector and the LLM. We set training epochs as 3, batch size as 16, and learning rate as 2×10^{-5} with cosine decay. Following the LLaVA model (Liu et al., 2023b), we convert each training instance into instruction-following data using its specified template. We use the API service of GPT-4V from OpenAI[11]. During inference, we prompt models using the same prompt template to ensure a fair comparison.

Given the large scale of the synthetic datasets and our computational constraints, we construct $\mathcal{D}_s$ by randomly sampling 6k instances from the complete training dataset of NewsCLIPings and DGM^4, and 3k instances from the Autosplice dataset since it is smaller, resulting in a total of 15k instances. For a robust evaluation, we repeat the sampling process using three different random seeds, resulting in three distinct $\mathcal{D}_s$. Each data selection method selects 750 instances from $\mathcal{D}_s$ to construct $\mathcal{D}_v$ for fine-tuning, and we ensure the class distribution of selected set is balanced. The unlabeled validation set $\mathcal{D}_u$ contains 5% instances of each test set, which amounts to 37 instances for MediaEval and 35 instances for Snopes and Snopes (O+).

We conduct experiments three times, each with a different training set, and report the mean macro-F1 and standard deviation of each metric across these three runs in all experiments. The seeds and training dataset are kept the same across different models. All the experiments use a server with 8 NVIDIA Tesla-V100 32GB GPUs.

C Visualization of Examples

We visualize a set of real-world instances from the 2015 Paris Attack event in Figure 4. In Figure 5, we present a set of synthetic instances with high cosine similarity to these real-world instances, while Figure 6 shows synthetic instances with low cosine similarity. We observe in Figure 4 that these instances contain relevant elements such as police, French and Paris, while the instances in 6 are irrelevant. These examples illustrate that synthetic datasets not only contain useful instances for multimodal misinformation detection on real-world datasets, but also have irrelevant instances, indicating the gap between synthetic and real-world data.

D Licenses of Datasets

- NewsCLIPpings: Unspecified
- DGM^4: Apache-2.0

[10] https://github.com/haotian-liu/LLaVA

[11] gpt-4-turbo: https://openai.com/index/gpt-4/

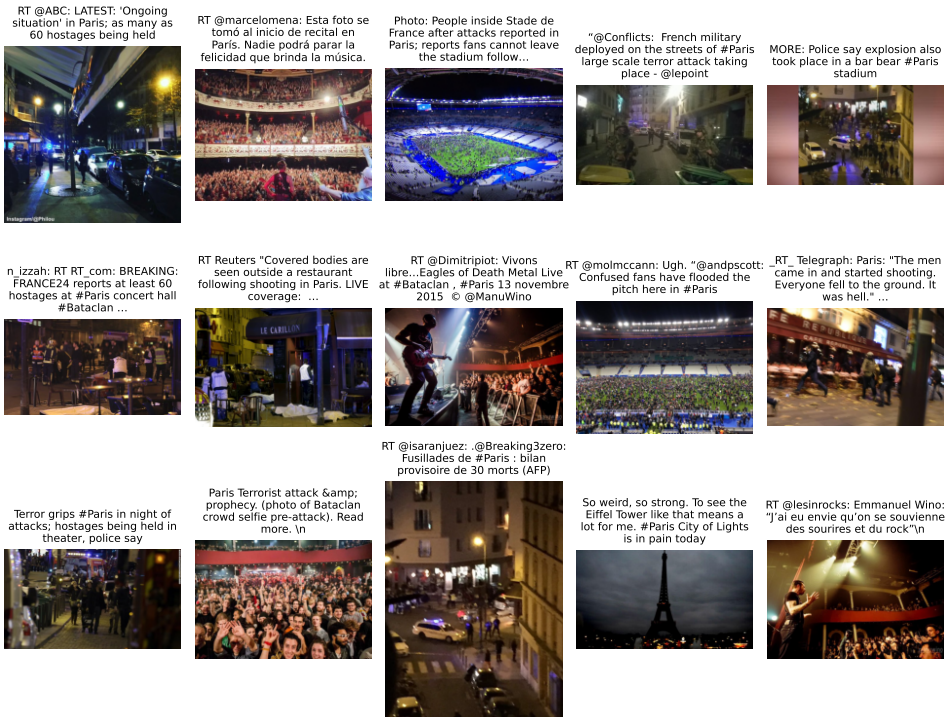


Figure 4: Visualization of real-world instances of 2015 Paris Attack event.

- AutoSplice: Only for academic research
- MediaEval: Apache-2.0
- Fauxtography: MIT License
- MOCHEG: CC BY 4.0

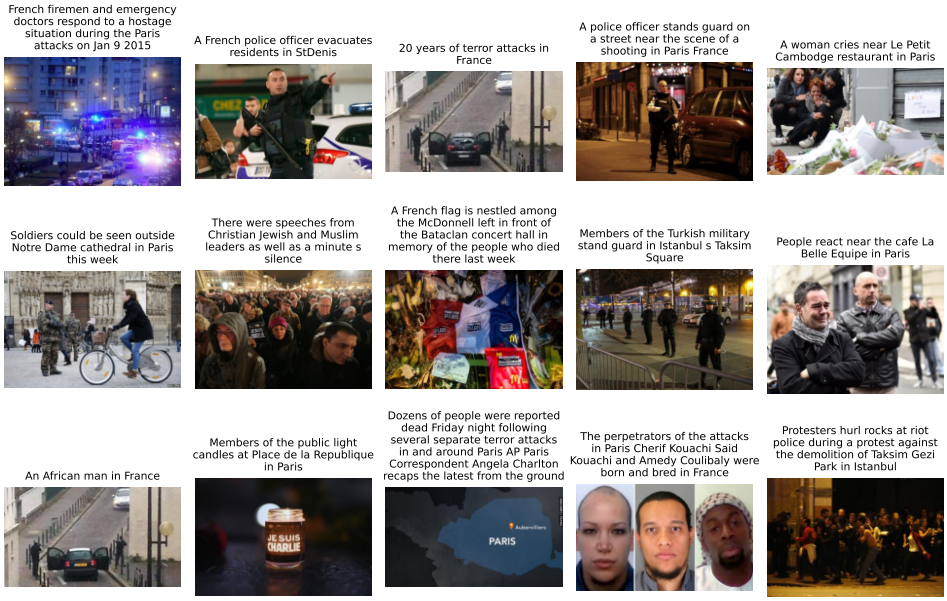


Figure 5: Visualization of synthetic instances with high similarity to real-world data.

Figure 6: Visualization of synthetic instances with low similarity to real-world data.