# Enhancing Multimodal Misinformation Detection by Replaying the Whole Story from Image Modality Perspective

**Bing Wang[1,2], Ximing Li[1,2,3*], Yanjun Wang[2], Changchun Li[1,2], Lin Yuanbo Wu[4], Buyu Wang[5], Shengsheng Wang[1,2*]**

[1] College of Computer Science and Technology, Jilin University, China
[2] Key Laboratory of Symbolic Computation and Knowledge Engineering of the MoE, Jilin University, China
[3] RIKEN Center for Advanced Intelligence Project, Japan
[4] School of Engineering, University of Warwick, UK
[5] College of Computer and Information Engineering, Inner Mongolia Agricultural University, China
{wangbing1416,liximing86,changchunli93}@gmail.com, wss@jlu.edu.cn

## Abstract

Multimodal Misinformation Detection (MMD) refers to the task of detecting social media posts involving misinformation, where the post often contains text and image modalities. However, by observing the MMD posts, we hold that the text modality may be much more informative than the image modality because the text generally describes the whole event/story of the current post but the image often presents partial scenes only. Our preliminary empirical results indicate that the image modality exactly contributes less to MMD. Upon this idea, we propose a new MMD method named RETSIMD. Specifically, we suppose that each text can be divided into several segments, and each text segment describes a partial scene that can be presented by an image. Accordingly, we split the text into a sequence of segments, and feed these segments into a pre-trained text-to-image generator to augment a sequence of images. We further incorporate two auxiliary objectives concerning text-image and image-label mutual information, and further post-train the generator over an auxiliary text-to-image generation benchmark dataset. Additionally, we propose a graph structure by defining three heuristic relationships between images, and use a graph neural network to generate the fused features. Extensive empirical results validate the effectiveness of RETSIMD.

**Code** — https://github.com/wangbing1416/RETSIMD

## Introduction

Over the past decade, social media platforms, *e.g.,* Twitter and Weibo, have become the priority medium for delivering a plethora of messages and information to people worldwide. Unfortunately, they also promote the spread of misinformation about social topics, especially for hot events and stories, resulting in various negative effects (Vosoughi, Roy, and Aral 2018; Scheufele and Krause 2019). To alleviate such impacts, detecting misinformation timely becomes a primary demand in many real-world platforms, giving birth to the prevalent research topic, namely **M**isinformation **D**etection (**MD**) (Zhang et al. 2021b; Zhu et al. 2022; Wang et al. 2024a), in the data mining community.

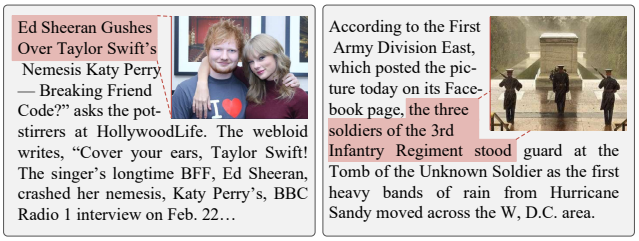


Figure 1: Two real-world MMD post examples. Each text describes the whole story of the current post but the image may only present a partial scene corresponding to the text segment marked in red.

To our knowledge, the previous studies regard MD as a binary classification problem, and they mainly train neural-based models, which can predict any future post content whether it describes a real or fake event/story. Nowadays, many posts from social media are accompanied by images, so one can formulate such posts as multimodal samples with text and image modalities. Inspired by the cognition that multiple modalities can provide various views to express the post content, the community has paid much attention to developing **M**ultimodal **M**isinformation **D**etection (**MMD**) methods (Jin et al. 2017; Chen et al. 2022; Wang et al. 2024c). Generally, the prevalent ideas of MMD methods include fusing text and image modalities to achieve more discriminative features (Ying et al. 2023; Wu, Liu, and Zhang 2023), and learning the semantic inconsistency between text and image modalities as auxiliary cues for verifying misinformation (Fung et al. 2021; Chen et al. 2022).

A potential assumption is that, in the MMD task, the text and image modalities are equally important (Hu et al. 2023; Wang et al. 2024d; Ma et al. 2024). However, by observing the MMD posts, we hold that the text modality may be much more informative than the image modality because the text generally describes the whole event/story of the current post but the image often presents partial scenes only (see examples in Fig. 1). Accordingly, we argue that the text modality must be more important than the image modality in MMD. To verify this viewpoint, we conduct extensive preliminary

*Ximing Li and Shengsheng Wang are corresponding authors.

experiments to evaluate the contributions of text and image modalities by predicting ablative variants of SOTA MMD methods. The results validate that, for each MMD method, the performance gap between the full variant and the variants with text modalities is significantly and consistently lower than the one between the full variant and the variants with image modalities, empirically proving the argument.

Based on these observations, a natural way of promoting MMD is, for each MMD sample, to augment a sequence of images that can also replay the whole story of the post, instead of employing the image that often presents partial scenes only. Upon this idea, we propose a new MMD method named **RETSIMD**. Specifically, we suppose that each text can be divided into several segments, and each text segment describes a partial scene that can be presented by an image. Accordingly, we split the text into a sequence of segments, and feed these segments into a pre-trained text-to-image generator to augment a sequence of images. To guarantee higher-quality augmented images, we incorporate two auxiliary objectives concerning text-image and image-label mutual information, and further post-train the generator over an auxiliary text-to-image generation benchmark dataset. With the augmented images, we propose a graph-based encoder to integrate their features. Specifically, we construct a graph structure by defining three heuristic relationships between images, and use a graph neural network to encode the graph to obtain the fused features.

We conduct extensive experiments across three benchmark MMD datasets and compare five SOTA MMD methods. The experimental results indicate that RETSIMD can consistently improve the performance of baseline models, which proves the effectiveness of RETSIMD. Meanwhile, we also report quantitative results demonstrating that RETSIMD effectively improves the contributions of the image modality to MMD.

Our contributions are the following three-fold:

- We empirically validate that the text modality contributes more to MMD models than the image modality.
- We propose a new MMD method RETSIMD by generating a sequence of augmented images that can also replay the whole story of the text.
- Extensive experiments are conducted to validate the effectiveness of RETSIMD.

## Preliminary Empirical Evaluations on Modality Contributions

In this evaluation, we employ five SOTA MMD methods including ResNet + BERT **(R&B)** (He et al. 2016; Devlin et al. 2019), MAE + DeBERTaV3 **(M&D)** (He et al. 2022; He, Gao, and Chen 2023), **CLIP** (Radford et al. 2021), **CAFE** (Chen et al. 2022), and **BMR** (Ying et al. 2023). To evaluate the modality contributions in MMD, for each MMD method, we compare the classification accuracy scores between the following ablative variants:

- **Full variant** trains using the original text-image pairs.
- **Text-only variant** removes the images by setting them to a completely white image, leaving only the text.
- **Image-only variant** removes the texts by setting them to the same-length `[PAD]` tokens, leaving only the images.
- **Text-replaced variant** replaces the texts in the original image-text pairs with the ones in other samples.
- **Image-replaced variant** replaces the images in the original image-text pairs with the ones in other samples.

We employ three benchmark MMD datasets including *GossipCop* (Shu et al. 2020), *Weibo* (Jin et al. 2017), and *Twitter* (Boididou et al. 2018). The descriptions of benchmark datasets, compared methods, and implementation details are presented in the Supplementary Material.

### Results and Discussion

Because the full variants perform the best in all the cases, we show the accuracy gap between the full variants and the other four ablative variants in Fig. 2. First, we can observe that the gaps of text-only/image-replaced variants are lower than those of image-only/text-replaced variants by a large margin and, in some cases, the image-only/text-replaced variants are even ineffective to some extent. These results directly validate the argument that the text modality contribution is much more than the image modality to MMD. Further, we can see that the accuracy gaps of text-only and image-replaced variants are consistently insignificant in most cases. That is because the text can provide sufficient information to describe the content, further supporting the argument.

### Quantifying Modality Contributions

In the above section, we conclude that the text modality contributes more to MMD than the image modality. Meanwhile, in Fig. 1, we analyze that the different contributions of the two modalities are attributed to the fact that text always describes the whole story but images only depict partial scenes. This inspires us to investigate the modality contributions from the information theory perspective. Accordingly, we are motivated by the information gain (Quinlan 1986; Donoho 1995), and design a quantitative metric named **contribution degree** for our experiments. Formally, given an image-text pair $(\mathbf{x}^v, \mathbf{x}^t)$, we calculate its information entropy *w.r.t* the veracity label $y$ as follows:

$$\mathcal{H}(y|\mathbf{x}^v, \mathbf{x}^t) = -\sum p(y|\mathbf{x}^v, \mathbf{x}^t) \log p(y|\mathbf{x}^v, \mathbf{x}^t). \quad (1)$$

Upon it, we can calculate the information gains carried by image and text modalities as follows:

$$\begin{aligned} \mathcal{G}(y, \mathbf{x}^v) &= \mathcal{H}(y|\varnothing, \mathbf{x}^t) - \mathcal{H}(y|\mathbf{x}^v, \mathbf{x}^t), \\ \mathcal{G}(y, \mathbf{x}^t) &= \mathcal{H}(y|\mathbf{x}^v, \varnothing) - \mathcal{H}(y|\mathbf{x}^v, \mathbf{x}^t). \end{aligned} \quad (2)$$

In addition to the information gains brought by such two variants, we also investigate the information gains of text-replaced and image-replaced variants $\mathcal{G}(y, \hat{\mathbf{x}}^t)$ and $\mathcal{G}(y, \hat{\mathbf{x}}^v)$, the experimental results are reported in Fig. 3.

Generally, the information gains brought by the image-only/text-replaced variants are larger than those of text-only/image-replaced variants. The observation indicates that text contributes the most substantial information gain for veracity predictions, while images contribute the least, or

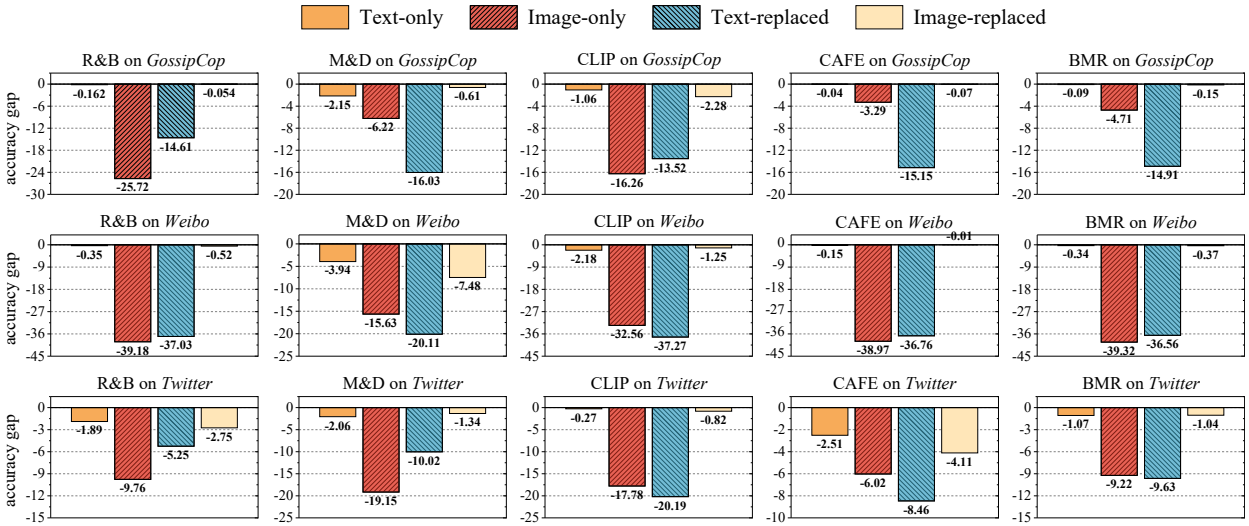


Figure 2: Preliminary empirical evaluations on accuracy gaps between four ablative variants and the full variant.

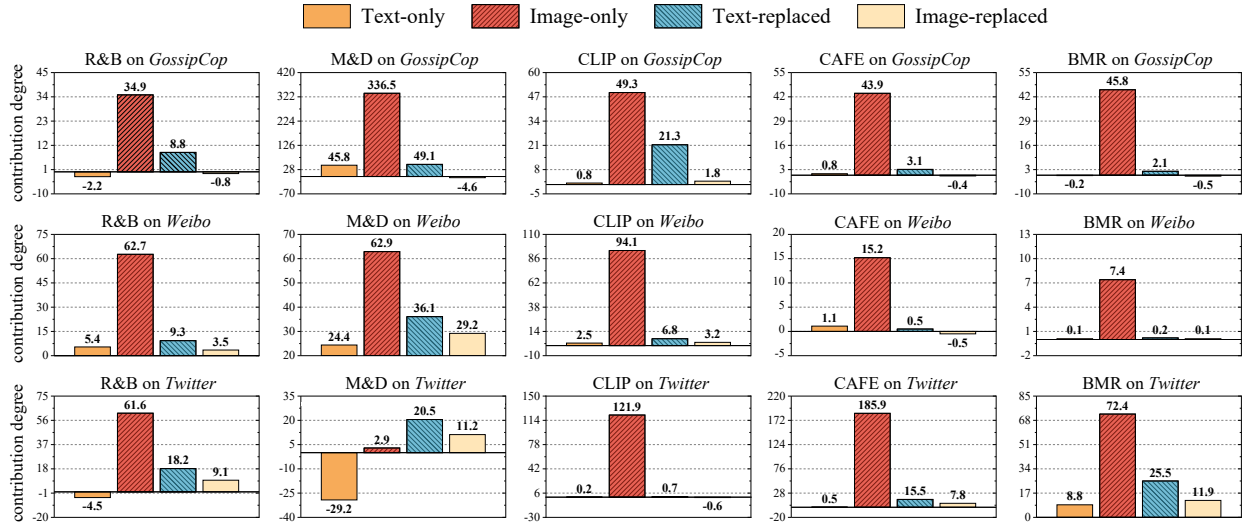


Figure 3: Preliminary empirical evaluations on contribution degrees of four ablative variants.

even negative gain. Therefore, these findings quantitatively demonstrate the greater importance of the text modality over the image modality, and provide support for our hypothesis that the contribution gap exists between the two modalities. Inspired by the evidence, we aim to generate more images that depict the whole stories presented in the text modality to bridge the information gap between the two modalities.

## Our Proposed Method

In this section, we introduce the task formulation of MMD and our proposed MMD framework RETSIMD, which aims to improve the contribution of the image modality to MMD.

**Task formulation.** Typically, an MMD dataset is formulated as $\mathcal{D} = \{\mathcal{X}_i, y_i\}_{i=1}^{N}$, where $\mathcal{X}_i = (\mathbf{x}_i^t, \mathbf{x}_i^v)$ denotes the text-image pair, and $y_i \in \{0, 1\}$ is its ground-truth veracity label, *e.g.,* real and fake. Accordingly, the general goal of MMD is to train a detector $\mathcal{F}_{\boldsymbol{\theta}}(\cdot)$, which can predict the veracity labels of unseen text-image pairs.

### Model Overview

In Sec. , we empirically observe that the text modality contribution is much more than the image modality to MMD. And the experimental results support the argument that the text may completely describe the whole story, but the image can only partially present one scene in the story. Accordingly, we attempt to improve the contribution of the image modality by generating a sequence of augmented images *w.r.t* the text modality. Based on this idea, we propose a new MMD framework RETSIMD, which alternately optimizes the misinformation detector and a text-to-image generator, which generates images to enhance the image modality, and then we design a fusion model to integrate these images by mining their potential relationships. Specifically, RETSIMD

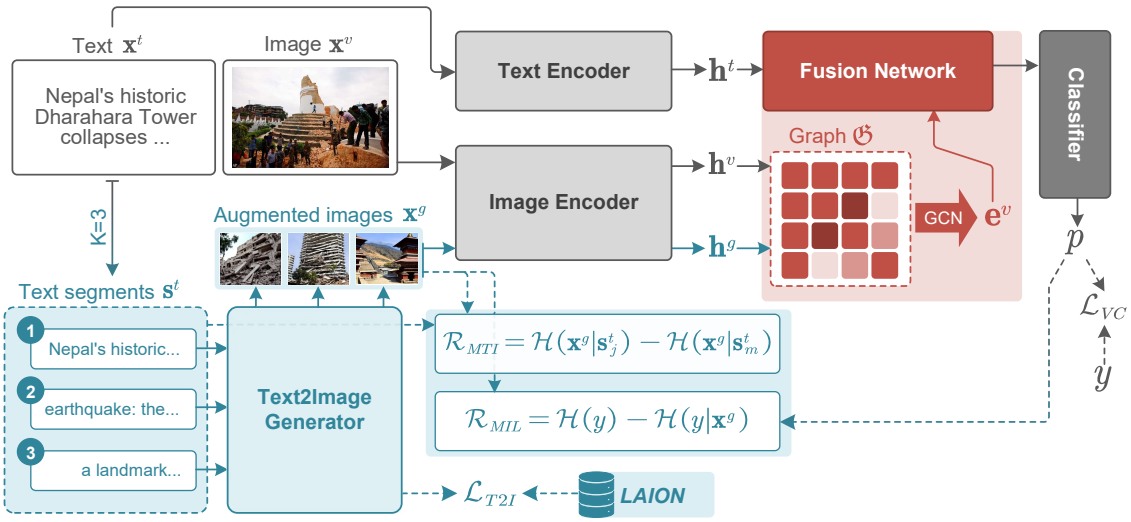


Figure 4: Overall framework of RETSIMD. We split $\mathbf{x}^t$ into $K$ segments, and use a text-to-image generator to generate corresponding images $\mathbf{x}^g$. Then, we design a multimodal fusion network to integrate these features and predict the veracity label.

generally includes four modules: **feature encoders**, **text-to-image generator**, **multimodal fusion network**, and **veracity classifier**. For clarity, the overall framework of RETSIMD is depicted in Fig. 4.

**Feature encoders.** Given text $\mathbf{x}_i^t$ and an image $\mathbf{x}_i^v$, the feature encoders utilize the text encoder and image encoder to extract their shared semantic features $\mathbf{h}_i^t$ and $\mathbf{h}_i^v$. Specifically, we first leverage a pre-trained language model BERT (Devlin et al. 2019) and a ResNet34 model (He et al. 2016) pre-trained by ImageNet (Krizhevsky, Sutskever, and Hinton 2012) to obtain the hidden text and image features $\mathbf{z}_i^t = \mathcal{F}_{\boldsymbol{\theta}^t}(\mathbf{x}_i^t)$ and $\mathbf{z}_i^v = \mathcal{F}_{\boldsymbol{\theta}^v}(\mathbf{x}_i^v)$, respectively. And then, to align these features into a shared space, we use linear layers to obtain two shared features $\mathbf{h}_i^t = \mathbf{z}_i^t \mathbf{W}_a^t$ and $\mathbf{h}_i^v = \mathbf{z}_i^v \mathbf{W}_a^v$.

**Text-to-image generator.** To improve the contribution of the image modality to MMD, we design a text-to-image generator $\mathcal{G}_{\boldsymbol{\phi}}(\cdot)$ to generate a sequence of augmented images corresponding to the text. Specifically, given a piece of text $\mathbf{x}_i^t$, we aim for the augmented images to completely express its described story. Therefore, we divide $\mathbf{x}_i^t$ into $K$ segments with a fixed-number sliding window[1] as $\{\mathbf{s}_{ij}^t\}_{j=1}^K$. Then, we feed these text segments into the generator to generate images $\mathbf{x}_{ij}^g = \mathcal{G}_{\boldsymbol{\phi}}(\mathbf{s}_{ij}^t)$.

Meanwhile, to ensure that the augmented images $\mathbf{x}_{ij}^g$ contain sufficient information and can fully express the information in the text $\mathbf{s}_{ij}^t$, we design two information theory-based regularizations $\mathcal{R}_{MTI}$ and $\mathcal{R}_{MIL}$ to tune the generator, which learn the mutual information between the text and its generated images, and between the generated images and the veracity labels, respectively. We also utilize prevalent text-to-image datasets, *e.g., LAION-2B* (Schuhmann et al. 2022), to post-train the generator by an objective $\mathcal{L}_{T2I}$. In summary, the overall training objective of the text-to-image generator is as follows:

$$\min_{\boldsymbol{\phi}} \mathcal{L}_{GEN}(\boldsymbol{\phi}) = \mathcal{L}_{T2I} + \alpha_1 \mathcal{R}_{MTI} + \alpha_2 \mathcal{R}_{MIL}, \quad (3)$$

where $\alpha_1$ and $\alpha_2$ represent two trade-off parameters to balance multiple objectives.

**Multimodal fusion network.** Given augmented images $\{\mathbf{x}_{ij}^g\}_{j=1}^K$ generated by the text-to-image generator, the multimodal fusion network integrates their features with the original image $\mathbf{x}_i^v$ and the text $\mathbf{x}_i^t$. Specifically, to grasp the potential relationships among different images, we construct a graph structure $\mathfrak{G}$, where the nodes represent all the images, and the edges indicate our designed relationships between the image nodes. Upon this graph, we leverage a prevalent graph neural network (Kipf and Welling 2017) to capture a fused image feature $\mathbf{e}_i^v = \mathcal{F}_{\boldsymbol{\theta}^g}(\mathbf{h}_i^v, \mathbf{h}_{ij}^g, \mathfrak{G})$, where $\mathbf{h}_{ij}^g = \mathcal{F}_{\boldsymbol{\theta}^v}(\mathbf{x}_{ij}^g)\mathbf{W}_a^v$ denotes the shared feature of the augmented images. Given the text feature $\mathbf{h}_i^t$ and the fused image feature $\mathbf{e}_i^v$, we use a cross-attention network to obtain the overall fused feature $\mathbf{e}_i = \mathcal{F}_{\boldsymbol{\theta}^c}(\mathbf{h}_i^t, \mathbf{e}_i^v)$.

**Veracity classifier.** Utilizing the fused feature $\mathbf{e}_i$, we feed it into a veracity classifier consisting of two linear layers and an activation function, to predict its veracity label $p_i = \mathbf{e}_i \mathbf{W}_v$. Accordingly, the training objective of the veracity classification is as follows:

$$\mathcal{L}_{VC} = \frac{1}{N} \sum\nolimits_{i=1}^{N} \ell_{CE}(p_i, y_i), \quad (4)$$

where $\ell_{CE}(\cdot\ ,\cdot)$ represents a cross-entropy loss function. Upon the aforementioned modules, the overall training objective of the misinformation detector is as follows:

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{DET}(\boldsymbol{\theta}) = \mathcal{L}_{VC} + \beta \mathcal{R}_{CA}, \quad (5)$$

where $\beta$ is a hyper-parameter. In summary, the training of RETSIMD is to alternately optimize Eq. (3) and Eq. (5) *w.r.t* $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, respectively. In the following sections, we describe the details of the text-to-image generator and the multimodal fusion network, respectively.

[1]We evaluate multiple text segmentation strategies in the Appendix, and the fixed-number sliding has the best performance.

## Text-to-image Generator

In general, we tune the text-to-image generator $\mathcal{G}_{\boldsymbol{\phi}}(\cdot)$ , which generates images $\{\mathbf{x}_{ij}^g\}_{j=1}^K$ with the text $\mathbf{x}_i^t$ to make the image modality contributing. Specifically, we use the stable diffusion model (Rombach et al. 2022) pre-trained across text-image pairs as the basic generator. Given $\mathbf{x}_i^t$, we split it into $K$ equal-length segments $\{\mathbf{s}_{ij}^t\}_{j=1}^K$, and feed them into the generator to generate augmented images $\{\mathbf{x}_{ij}^g\}_{j=1}^K$. Upon the generated ones, we propose the following two mutual information objectives to further tune the generator.

**Text-image mutual information.** Typically, the augmented images should be semantically consistent to their conditioned text descriptions. Therefore, we design an objective to constrain the mutual information between the text $\mathbf{s}_{ij}^t$ and its corresponding augmented image $\mathbf{x}_{ij}^g$. Previous mutual information based method (Zhang et al. 2021a) achieve this via contrastive losses among different texts and images. In fact, in our method, different segments in the same text naturally form positive and negative samples. Formally, our text-image mutual information objective is as follows:

$$\mathcal{R}_{MTI} = \frac{1}{NK}\sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{K}\mathfrak{R}\big(\mathbf{s}_{ij}^t, \mathbf{x}_{ij}^g\big), \tag{6}$$

where,

$$\mathfrak{R}\big(\mathbf{s}_{ij}^t, \mathbf{x}_{ij}^g\big) = \frac{1}{K-1}\sum\nolimits_{m\neq j}\mathcal{H}\big(\mathbf{x}_{ij}^g|\mathbf{s}_{ij}^t\big) - \boldsymbol{\xi}_{jm}\mathcal{H}\big(\mathbf{x}_{ij}^g|\mathbf{s}_{im}^t\big), \tag{7}$$

where $\mathcal{H}(\mathbf{x}_{ij}^g|\cdot)$ indicates the entropy of $\mathbf{x}_{ij}^g$ conditioned on different text segments. With this objective, we aim for $\mathbf{s}_{ij}^t$ to bring more information about $\mathbf{x}_{ij}^g$ than $\mathbf{s}_{im}^t$. Additionally, $\boldsymbol{\xi}_{jm}$ denotes an adaptive weight. Since there is a temporal relationship between different text segments, the semantics of segments with close spatial distances are more relevant. Therefore, when $m$ is closer to $j$, the weight $\boldsymbol{\xi}_{jm}$ is smaller.

**Image-label mutual information.** The goal of our augmented images is to make the image modality more informative regarding the veracity labels. Therefore, we need to propose a objective to constrain the generator to achieve this goal. Accordingly, our image-label mutual information objective is formulated as follows:

$$\begin{aligned}\mathcal{R}_{MIL} &= \frac{1}{N}\sum\nolimits_{i=1}^{N}\mathscr{G}\left(y_i|\{\mathbf{x}_{ij}^g\}_{j=1}^K\right)\\ &= \frac{1}{N}\sum\nolimits_{i=1}^{N}\mathcal{H}(y_i) - \mathcal{H}\left(y_i|\{\mathbf{x}_{ij}^g\}_{j=1}^K\right).\end{aligned} \tag{8}$$

This objective utilizes the information gain brought by the generated images to the veracity labels, and uses it to train the generator in an end-to-end scheme.

The success of the aforementioned objectives always depends on the zero-shot capability of the pre-trained generator (Clark and Jaini 2023). To maintain high-quality images generated by the generator, we introduce a text-to-image dataset $\overline{\mathcal{D}} = \{\overline{\mathbf{x}}_i, \overline{y}_i\}_{i=1}^{|\overline{\mathcal{D}}|}$, *e.g., LAION*, to continuously post-train the generator, where $(\overline{\mathbf{x}}_i, \overline{y}_i)$ is a gold text-image pair. The post-training objective is as follows:

$$\mathcal{L}_{T2I} = \frac{1}{|\overline{\mathcal{D}}|}\sum\nolimits_{i=1}^{|\overline{\mathcal{D}}|}\ell_{MSE}\big(\mathcal{G}_{\boldsymbol{\phi}}(\overline{\mathbf{x}}_i), \overline{y}_i\big), \tag{9}$$

where $\ell_{MSE}(\cdot,\cdot)$ means the mean squared error function. Upon the objectives, we optimize the generator with Eq. (3).

## Multimodal Fusion Network

Generally, the multimodal fusion network aims to integrate the features of the augmented images $\{\mathbf{x}_{ij}^g\}_{j=1}^K$ and the original text-image pair $(\mathbf{x}_i^t, \mathbf{x}_i^v)$. Specifically, by using the feature encoders, we can first obtain the hidden features $\{\mathbf{h}_{ij}^g\}_{j=1}^K$ and $\mathbf{h}_i^v$ of images. To utilize these features more effectively, we capture the potential relationships among images, and construct a graph structure $\mathfrak{G}_i = (\boldsymbol{\Lambda}_i, \mathbf{H}_i)$, where $\boldsymbol{\Lambda}_i$ represents the adjacency matrix, the nodes are $K+1$ images, the edges represent their relationships, and $\mathbf{H}_i = \{\mathbf{h}_i^v, \mathbf{h}_{i1}^g, \cdots, \mathbf{h}_{iK}^g\}$ denotes the node features. Specifically, we design the following three heuristic relationships to construct the graph.

**Central relationship.** We suggest that although the original images may only express the partial scene in an event, they always present the most central scene compared to images generated using different text segments. Therefore, based on this assumption, we directly connect all the augmented images with the original image to construct the initial graph.

**Temporal relationship.** Our augmented images originate from the original text segments, and we expect that they depict the scenes of each text segment. In fact, a natural temporal relationship exists between the text segments, meaning that adjacent segments have a strong semantic correlation. Therefore, to learn this temporal relationship, we connect the augmented images according to the order of their corresponding text segments in the original text.

**Dependency relationship.** The augmented images correspond to text segments that not only exhibit inherent temporal relationships, but also logical dependency relationships, *e.g.,* coordination relationships. To integrate these dependencies into the graph $\mathfrak{G}$, we first utilize the prevalent spaCy toolkit to construct a dependency tree structure from the original text. These dependency tree structures typically represent token-level relationships, therefore, to consolidate these into text segment-level relationships, we merge all token nodes within a segment into a single node while retaining the connected edges. Based on this operation, we can obtain a new segment-level graph structure and incorporate it into $\mathfrak{G}$.

Upon these relationships, we obtain a graph structure $\mathfrak{G}$, and we use a graph neural network to encode a fused image feature as follows:

$$\mathbf{e}_i^{v(l)} = \boldsymbol{\sigma}\left(\sum\nolimits_{j=1}^{K+1}\boldsymbol{\Lambda}_{ij}\boldsymbol{\theta}^{g(l)}\mathbf{e}_i^{v(l-1)} + b^{(l)}\right), \tag{10}$$

where $\boldsymbol{\sigma}(\cdot)$ is a activation function, and $\boldsymbol{\theta}^{g(l)}$ and $b^{(l)}$ are a linear weight matrix and a bias term, respectively. And we initialize $\mathbf{e}_i^{v(0)}$ as $\mathbf{H}_i$, and denote the feature $\mathbf{e}_i^{v(l)}$ at the final GCN layer as $\mathbf{e}_i^v$. In our model, we empirically fix the number of GCN layers to 2. Given the fused feature $\mathbf{e}_i^v$ and text and image features $\mathbf{h}_i^t$ and $\mathbf{h}_i^v$, we employ a cross-attention

| Method | Accuracy | Macro F1 | $P_{real}$ | $R_{real}$ | $F1_{real}$ | $P_{fake}$ | $R_{fake}$ | $F1_{fake}$ | $\Delta$ | $\mathscr{G}(y, \mathbf{x}^v)$ |
|---|---|---|---|---|---|---|---|---|---|---|
| **Dataset: *GossipCop*** (Shu et al. 2020) | | | | | | | | | | |
| ResNet + BERT | 87.17±0.4 | 78.30±0.7 | 90.84±0.6 | 93.55±1.1 | 92.17±0.3 | 69.25±2.6 | **60.41**±3.5 | 64.43±1.4 | - | .0349 |
| + **RETSIMD** | **88.13**±0.2* | **79.47**±0.4* | **90.89**±0.2 | **94.79**±0.3* | **92.80**±0.1* | **73.38**±0.9* | 60.18±1.2 | **66.13**±0.7* | **1.21** | **.0301** |
| R&B + SAFE | 87.14±0.6 | 78.74±0.3 | 91.35±0.6 | 92.88±1.6 | 92.10±0.4 | 68.22±3.6 | **63.08**±3.8 | 65.38±0.5 | - | .0325 |
| + **RETSIMD** | **88.30**±1.1* | **79.79**±0.7* | **91.02**±0.7 | **94.88**±1.3* | **92.91**±0.8* | **73.88**±3.3* | 61.73±3.1 | **66.67**±0.7* | **1.29** | **.0287** |
| R&B + MCAN | 87.29±0.7 | 78.73±0.2 | 91.18±0.8 | 93.31±2.1 | 92.21±0.5 | 69.50±3.1 | **62.05**±3.0 | 65.26±0.6 | - | .0200 |
| + **RETSIMD** | **88.22**±0.1* | **79.72**±0.3* | **91.04**±0.2 | **94.73**±0.3* | **92.85**±0.0 | **73.44**±0.9* | 61.91±1.4 | **66.57**±0.6* | **1.12** | **.0129** |
| R&B + CAFE | 87.16±0.8 | 78.89±0.6 | 90.80±0.7 | 92.73±1.9 | 92.10±0.6 | 68.15±3.4 | **62.81**±3.3 | 65.48±0.9 | - | .0439 |
| + **RETSIMD** | **88.38**±0.3* | **79.37**±0.3 | **91.02**±0.6 | **94.65**±1.3* | **92.78**±0.2* | **71.87**±2.8* | 62.75±2.7 | **65.99**±0.8 | **1.09** | **.0402** |
| R&B + BMR | 87.32±0.3 | 78.87±0.4 | 91.26±0.4 | 93.24±0.8 | 92.23±0.2 | 68.92±1.9 | 62.53±2.5 | 65.51±0.8 | - | .0458 |
| + **RETSIMD** | **88.42**±0.4* | **79.71**±0.5* | **91.42**±0.8 | **94.25**±1.5* | **92.90**±0.3* | **71.81**±2.9* | **62.73**±1.5 | **66.82**±1.1* | **1.02** | **.0409** |
| R&B + GAMED | 87.03±0.5 | 78.81±0.3 | **91.55**±0.5 | 92.47±1.4 | 92.01±0.5 | 67.05±0.4 | 64.22±0.3 | 65.60±0.4 | - | .0350 |
| + **RETSIMD** | **88.30**±0.7* | **79.79**±0.7* | 91.02±0.7 | **94.88**±0.2* | **92.91**±0.8* | **73.88**±0.5* | **64.73**±0.4 | **67.67**±0.7* | **1.81** | **.0257** |
| **Dataset: *Weibo*** (Jin et al. 2017) | | | | | | | | | | |
| ResNet + BERT | 90.38±0.7 | 90.37±0.7 | 88.89±1.0 | 91.68±1.2 | 90.22±0.8 | 91.95±1.0 | 89.95±1.1 | 90.53±0.7 | - | .0760 |
| + **RETSIMD** | **91.48**±0.7* | **91.48**±0.7* | **89.98**±1.4* | **92.80**±1.9* | **91.34**±0.6* | **93.09**±1.6* | **90.24**±1.4 | **91.61**±0.8* | **1.01** | **.0632** |
| R&B + SAFE | 90.52±0.6 | 90.51±0.6 | 89.86±1.2 | 90.68±2.3 | 90.25±0.7 | 91.24±1.9 | 90.36±1.5 | 90.78±0.5 | - | .0629 |
| + **RETSIMD** | **91.60**±0.3* | **91.59**±0.3* | **91.50**±1.2* | **91.11**±1.0 | **91.31**±0.4* | **91.70**±1.5 | **92.06**±1.4* | **91.88**±0.2* | **1.07** | **.0583** |
| R&B + MCAN | 90.58±0.5 | 90.58±0.5 | 88.43±2.1 | 92.67±1.7 | 90.50±0.4 | 92.80±1.3 | 88.62±2.4 | 90.66±0.6 | - | .0330 |
| + **RETSIMD** | **91.76**±0.3* | **91.76**±0.3* | **90.48**±1.0* | **92.77**±1.8 | **91.59**±0.4* | **93.10**±1.5 | **90.81**±1.3* | **91.92**±0.2* | **1.17** | **.0280** |
| R&B + CAFE | 90.72±0.5 | 90.71±0.4 | 89.19±1.5 | 91.96±1.1 | 90.56±0.4 | 92.23±0.9 | 89.55±1.8 | 90.87±0.5 | - | .0952 |
| + **RETSIMD** | **91.76**±0.2* | **91.76**±0.2* | **89.53**±1.4 | **94.02**±1.4* | **91.70**±0.2* | **94.14**±1.2* | **89.65**±1.6 | **91.82**±0.3* | **1.07** | **.0773** |
| R&B + BMR | 90.86±0.7 | 90.86±0.7 | 89.11±2.2 | 92.52±1.8 | 90.75±0.6 | 92.77±1.4 | 89.31±2.7 | 90.97±0.9 | - | .0119 |
| + **RETSIMD** | **92.01**±1.6* | **92.01**±1.6* | **90.44**±0.8* | **93.37**±2.7* | **91.88**±0.9* | **93.59**±2.7* | **90.74**±2.2* | **92.14**±1.3* | **1.13** | **.0101** |
| R&B + GAMED | 90.24±0.5 | 90.19±0.5 | 89.27±1.4 | 91.04±1.6 | 89.51±0.7 | 91.79±1.8 | 89.18±1.8 | 90.87±0.3 | - | .0798 |
| + **RETSIMD** | **92.08**±0.2* | **92.08**±0.2* | **90.34**±1.4* | **93.65**±1.4* | **91.97**±0.2* | **93.84**±1.2* | **90.61**±1.6* | **92.19**±0.3* | **1.84** | **.0616** |
| **Dataset: *Twitter*** (Boididou et al. 2018) | | | | | | | | | | |
| ResNet + BERT | 66.02±1.9 | 65.84±1.8 | 58.40±2.4 | **68.09**±3.9 | **63.63**±1.6 | 74.14±1.7 | 63.02±3.9 | 68.04±2.7 | - | .1428 |
| + **RETSIMD** | **68.15**±2.4* | **67.46**±3.9* | **61.46**±3.5* | 66.45±4.4 | 63.49±3.7 | **74.40**±2.3 | **69.40**±2.8* | **71.46**±2.3* | **1.89** | **.1002** |
| R&B + SAFE | 66.41±3.5 | 66.13±3.5 | 58.69±3.6 | 66.61±3.7 | 62.43±3.7 | 72.62±3.2 | 63.01±3.8 | 68.24±3.6 | - | .0927 |
| + **RETSIMD** | **68.28**±2.6* | **68.03**±2.5* | **60.87**±2.6* | **72.32**±4.1* | **65.86**±2.7* | **76.43**±3.5* | **65.31**±3.3* | **70.19**±3.1* | **2.89** | **.0707** |
| R&B + MCAN | 66.66±2.4 | 66.50±2.3 | 59.02±2.9 | 68.88±1.6 | 64.37±1.6 | 74.72±1.3 | 63.56±2.7 | 68.64±3.1 | - | .0459 |
| + **RETSIMD** | **68.60**±2.1* | **68.43**±1.8* | **61.02**±3.0* | **75.26**±2.9* | **67.12**±1.5* | **77.79**±1.4* | **63.68**±3.5 | **69.75**±2.4* | **2.41** | **.0382** |
| R&B + CAFE | 66.73±1.2 | 66.35±1.1 | 59.74±1.7 | 66.75±4.1 | 62.97±1.5 | 73.20±1.4 | 66.71±4.1 | 69.73±1.1 | - | .0872 |
| + **RETSIMD** | **69.16**±1.7* | **68.73**±1.7* | **62.66**±2.4* | **68.52**±3.2* | **65.31**±1.8* | **75.11**±1.4* | **69.64**±2.9* | **72.15**±1.7* | **2.39** | **.0663** |
| R&B + BMR | 66.71±1.4 | 66.33±1.5 | 59.62±1.4 | 66.79±3.8 | 62.94±2.5 | 73.24±2.2 | 66.65±2.6 | 69.73±1.2 | - | .0145 |
| + **RETSIMD** | **69.57**±2.7* | **69.00**±2.3* | **64.29**±3.5* | **68.39**±3.6* | **65.71**±2.2* | **75.24**±2.1* | **70.45**±4.3* | **72.30**±3.1* | **2.87** | **.0137** |
| R&B + GAMED | 66.53±2.5 | 66.21±2.5 | 58.93±2.0 | 66.61±2.5 | 62.57±2.4 | 73.25±2.2 | 63.52±2.3 | 68.84±2.6 | - | .1122 |
| + **RETSIMD** | **69.77**±2.5* | **69.77**±2.3* | **63.96**±1.9* | **69.11**±1.2* | **65.57**±2.4* | **75.42**±2.5* | **69.41**±2.0* | **71.97**±1.8* | **3.56** | **.0838** |

Table 1: Experimental results of RETSIMD across three prevalent MD datasets *GossipCop*, *Weibo*, and *Twitter*. The results marked by * are statistically significant compared to its baseline models, satisfying p-value $< 0.05$.

network to obtain the multimodal feature as follows:

$$\begin{aligned} \mathbf{e}_i &= \boldsymbol{\mu}\left(\mathbf{o}_i \mathbf{W}^Q \left(\mathbf{h}_i^t \mathbf{W}^K\right)^\top / \sqrt{d_n}\right) \mathbf{o}_i \mathbf{W}^V, \\ \mathbf{o}_i &= \boldsymbol{\mu}\left(\mathbf{e}_i^v \mathbf{W}^Q \left(\mathbf{h}_i^v \mathbf{W}^K\right)^\top / \sqrt{d_n}\right) \mathbf{e}_i^v \mathbf{W}^V, \end{aligned} \tag{11}$$

where $\boldsymbol{\theta}^c = \{\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V\}$ represents the parameter of the attention network.

## Experiments

In this section, we evaluate the performance of our proposed RETSIMD by answering the following questions:

- **EQ1.** How does RETSIMD perform on MMD?
- **EQ2.** Are all modules in RETSIMD effective?
- **EQ3.** Can RETSIMD enhance the contribution of the image modality to MMD?

For clarity, the experimental settings about datasets, baselines and experimental details can be seen in the Appendix.

### Main Results (EQ1, EQ3)

We select five baseline models to compare the performance of RETSIMD across three prevalent MMD datasets. Meanwhile, we employ nine metrics to evaluate the prediction capability of the model, where eight metrics, *e.g.,* accuracy (Acc.) and F1 score, assess the accuracy between the model's predictions and the gold veracity labels, and one metric $\mathscr{G}(y, \mathbf{x}^v)$ represents the information gain brought by $\mathbf{x}^v$ to quantitatively measure our model's ability to improve the image modality contribution. The experimental results are reported in Table 1. Generally, RETSIMD surpasses its baseline models over most datasets and metrics. For example, on the *Twitter* dataset, RETSIMD outperforms its base-

| Model | Acc. | F1 | $F1_{real}$ | $F1_{fake}$ | Avg.$\Delta$ |
|---|---|---|---|---|---|
| **Dataset**: *GossipCop* | | | | | |
| **RETSIMD** | 88.42 | 79.71 | 92.90 | 66.82 | / |
| *w/o* GF | 87.67 | 79.06 | 92.34 | 66.27 | **0.63** |
| *w/o* MI | 87.60 | 78.93 | 92.35 | 65.82 | **0.79** |
| *w/o* AI | 87.32 | 78.87 | 92.23 | 65.51 | **0.98** |
| **Dataset**: *Weibo* | | | | | |
| **RETSIMD** | 92.01 | 92.01 | 91.88 | 92.14 | / |
| *w/o* GF | 91.33 | 91.33 | 91.17 | 91.48 | **0.68** |
| *w/o* MI | 91.06 | 91.05 | 90.92 | 91.38 | **0.91** |
| *w/o* AI | 90.86 | 90.86 | 90.75 | 90.97 | **1.15** |
| **Dataset**: *Twitter* | | | | | |
| **RETSIMD** | 69.57 | 69.00 | 65.71 | 72.30 | / |
| *w/o* GF | 67.86 | 67.41 | 63.55 | 71.26 | **1.63** |
| *w/o* MI | 67.18 | 66.94 | 63.33 | 70.54 | **2.15** |
| *w/o* AI | 66.71 | 66.33 | 62.94 | 69.73 | **2.72** |

Table 2: Ablation study on three ablative versions.

line model BMR, by approximately 2.86 on the average of all metrics. This sufficiently demonstrates the effectiveness of the method proposed in this paper for enhancing the detection performance of MMD models.

Turning to the quantitative metric $\mathscr{G}(y, \mathbf{x}^v)$, our method outperforms the baseline models on most datasets. This result demonstrates that our method can improve the image modality contribution to MMD and enable the model to give more confident predictions by introducing additional augmented images. Exceptionally, $\mathscr{G}(y, \mathbf{x}^v)$ did not achieve the optimal result on the BMR model across *Twitter*. Upon observing the results of the experiments, we find that this model, when the text modality is removed, tends to predict all samples as the real class with high confidence, resulting in a low prediction entropy. However, our method mitigates this tendency, further validating its effectiveness.

## Ablative Study (EQ2)

To answer EQ2, we conduct the ablative study on three datasets, and the results are shown in Table 2. Specifically, we investigate the following four ablative versions:

- **w/o graph fusion (GF)** $\mathfrak{G}$ directly concatenate three features $\mathbf{h}^t$, $\mathbf{h}^v$, and $\mathbf{h}^g$, instead of fusing them with a graph $\mathfrak{G}$ and the graph neural network;
- **w/o mutual information (MI)** $\mathcal{R}_{MTI}$ **and** $\mathcal{R}_{MIL}$ does not utilize mutual information regularizations $\mathcal{R}_{MTI}$ and $\mathcal{R}_{MIL}$ to tune the text-to-image generator;
- **w/o augmented image (AI)** $\mathbf{x}^g$ removes the augmented images to enhance the MMD model, which is equivalent to the baseline model compared in Table 1.

In Table 2, we report the decrease in accuracy for four variants compared to the BMR baseline model. In general, the model's performance consistently declines when any module is removed, demonstrating the importance of these modules. Meanwhile, in more detail, the accuracy ranking of the four variants is consistently w/o graph fusion $>$w/o mutual information $>$w/o augmented image. Specifically, removing the mutual information regularizations, although still outperforming the baseline model, shows a significant decrease in performance compared to the full version of RETSIMD. This suggests that our designed mutual information method can indeed significantly increase the information *w.r.t* the veracity label in the generated images, thereby improving detection performance. The graph fusion utilizes the three types of relationships to construct and encode the graph, and the ablative results also demonstrate the effectiveness of this fusion approach.

## Related Works

Generally, recent MMD methods achieve this by learning the potential relationships between the multimodal content and their veracity labels (Wu, Liu, and Zhang 2023; Hu et al. 2023; Ying et al. 2023; Wang et al. 2024c). Specifically, these methods present various external features (Wang et al. 2024c; Dong et al. 2024), inconsistency measures (Chen et al. 2022; Ma et al. 2024), knowledge-augmented methods (Fung et al. 2021; Xuan et al. 2024), and multimodal fusion strategies (Ying et al. 2023; Wang et al. 2024b,d) to enhance the models. For example, Wang et al. (2024c) capture image manipulation features and their intention features to identify harmfully manipulated images in MMD datasets; Ma et al. (2024) learn the event-level inconsistency by constructing a dependency graph structure and utilize its feature with a multi-view learning framework. Additionally, with the development of large vision-language models, recent MMD works use these models to enhance detection models by synthesizing data (Zeng et al. 2024), integrating evidence (Tahmasebi, Müller-Budack, and Ewerth 2024) and supplementing knowledge (Liu et al. 2024; Xuan et al. 2024).

Unlike these MMD methods, we observe a problem that the image modality contributes less in existing MMD models . In the community, a few MMD efforts focus on similar phenomena. For example, Chen et al. (2023) found that only using images to detect fake news is unreliable, hence they proposed a casual approach. Nevertheless, a comprehensive empirical study on this issue remains a blank.

## Conclusion

In this work, we empirically observe that the text modality contributes more to MMD than the image modality since the text describes the whole story of the MMD post but the image always presents partial scenes only. Therefore, to boost the contribution of the image modality, we propose a new MMD framework RETSIMD, which aims to generate a sequence of augmented images that replay the whole story in the text. To achieve this goal, we split the text into a sequence of segments, which present partial scenes, and feed them into a pre-trained text-to-image generator to generate a sequence of augmented images. To ensure the quality of the augmented images, we tune the generator with two text-image and image-label mutual information. Meanwhile, to effectively integrate the image features, we construct a graph with their potential relationships and employ a graph neural network to fuse them. The experiments can demonstrate that RETSIMD improves the performance of the baseline models, and alleviates the information gap between text and images.

## Acknowledgements

We acknowledge support for this project from the National Science and Technology Major Project (No. 2021ZD0112500), the National Natural Science Foundation of China (No. 62276113), and the China Postdoctoral Science Foundation (No. 2022M721321).

## References

Boididou, C.; Papadopoulos, S.; Zampoglou, M.; Apostolidis, L.; Papadopoulou, O.; and Kompatsiaris, Y. 2018. Detection and visualization of misleading content on Twitter. *International Journal of Multimedia Information Retrieval*, 7(1): 71–86.

Chen, Y.; Li, D.; Zhang, P.; Sui, J.; Lv, Q.; Lu, T.; and Shang, L. 2022. Cross-modal Ambiguity Learning for Multimodal Fake News Detection. In *The ACM Web Conference*, 2897–2905.

Chen, Z.; Hu, L.; Li, W.; Shao, Y.; and Nie, L. 2023. Causal Intervention and Counterfactual Reasoning for Multi-modal Fake News Detection. In *Annual Meeting of the Association for Computational Linguistics*, 627–638.

Clark, K.; and Jaini, P. 2023. Text-to-Image Diffusion Models are Zero Shot Classifiers. In *Advances in Neural Information Processing Systems*.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186.

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. In *AAAI Conference on Artificial Intelligence*, 8354–8362.

Donoho, D. L. 1995. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3): 613–627.

Fung, Y. R.; Thomas, C.; Reddy, R. G.; Polisetty, S.; Ji, H.; Chang, S.; McKeown, K. R.; Bansal, M.; and Sil, A. 2021. InfoSurgeon: Cross-Media Fine-grained Information Consistency Checking for Fake News Detection. In *Annual Meeting of the Association for Computational Linguistics*, 1683–1698.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. B. 2022. Masked Autoencoders Are Scalable Vision Learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15979–15988.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

He, P.; Gao, J.; and Chen, W. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In *International Conference on Learning Representations*.

Hu, L.; Chen, Z.; Zhao, Z.; Yin, J.; and Nie, L. 2023. Causal Inference for Leveraging Image-Text Matching Bias in Multi-Modal Fake News Detection. *IEEE Transactions on Knowledge and Data Engineering*, 35(11): 11141–11152.

Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; and Luo, J. 2017. Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs. In *ACM on Multimedia Conference*, 795–816.

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, 1106–1114.

Liu, X.; Li, P.; Huang, H.; Li, Z.; Cui, X.; Liang, J.; Qin, L.; Deng, W.; and He, Z. 2024. FKA-Owl: Advancing Multimodal Fake News Detection through Knowledge-Augmented LVLMs. In *ACM International Conference on Multimedia*, 10154–10163.

Ma, Z.; Luo, M.; Guo, H.; Zeng, Z.; Hao, Y.; and Zhao, X. 2024. Event-Radar: Event-driven Multi-View Learning for Multimodal Fake News Detection. In *Annual Meeting of the Association for Computational Linguistics*, 5809–5821.

Quinlan, J. R. 1986. Induction of Decision Trees. *Machine Learning*, 1(1): 81–106.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*, volume 139, 8748–8763.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10674–10685.

Scheufele, D. A.; and Krause, N. M. 2019. Science audiences, misinformation, and fake news. *Proceedings of the National Academy of Sciences*, 116(16): 7662–7669.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; et al. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*.

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.

Tahmasebi, S.; Müller-Budack, E.; and Ewerth, R. 2024. Multimodal Misinformation Detection using Large Vision-Language Models. In *ACM International Conference on Information and Knowledge Management*, 2189–2199.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *science*, 359(6380): 1146–1151.

Wang, B.; Li, X.; Li, C.; Fu, B.; Pei, S.; and Wang, S. 2024a. Why Misinformation is Created? Detecting them by Integrating Intent Features. In *ACM International Conference on Information and Knowledge Management*, 2304–2314.

Wang, B.; Li, X.; Li, C.; Wang, S.; and Gao, W. 2024b. Escaping the Neutralization Effect of Modality Features Fusion in Multimodal Fake News Detection. *Information Fusion*, 111: 102500.

Wang, B.; Wang, S.; Li, C.; Guan, R.; and Li, X. 2024c. Harmfully Manipulated Images Matter in Multimodal Misinformation Detection. In *ACM International Conference on Multimedia*, 2262–2271.

Wang, J.; Zhang, H.; Liu, C.; and Yang, X. 2024d. Fake News Detection via Multi-scale Semantic Alignment and Cross-modal Attention. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2406–2410.

Wu, L.; Liu, P.; and Zhang, Y. 2023. See How You Read? Multi-Reading Habits Fusion Reasoning for Multi-Modal Fake News Detection. In *AAAI Conference on Artificial Intelligence*, 13736–13744.

Xuan, K.; Yi, L.; Yang, F.; Wu, R.; Fung, Y. R.; and Ji, H. 2024. LEMMA: Towards LVLM-Enhanced Multimodal Misinformation Detection with External Knowledge Augmentation. *CoRR*, abs/2402.11943.

Ying, Q.; Hu, X.; Zhou, Y.; Qian, Z.; Zeng, D.; and Ge, S. 2023. Bootstrapping Multi-View Representations for Fake News Detection. In *AAAI Conference on Artificial Intelligence*, 5384–5392.

Zeng, F.; Li, W.; Gao, W.; and Pang, Y. 2024. Multimodal Misinformation Detection by Learning from Synthetic Data with Multimodal LLMs. In *Findings of the Association for Computational Linguistics: EMNLP*, 10467–10484.

Zhang, H.; Koh, J. Y.; Baldridge, J.; Lee, H.; and Yang, Y. 2021a. Cross-Modal Contrastive Learning for Text-to-Image Generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 833–842.

Zhang, X.; Cao, J.; Li, X.; Sheng, Q.; Zhong, L.; and Shu, K. 2021b. Mining Dual Emotion for Fake News Detection. In *The Web Conference*, 3465–3476.

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 *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2120–2125.