# Invariant Meets Specific: A Scalable Harmful Memes Detection Framework


Chuanpeng Yang
Institute of Information Engineering, Chinese Academy of Sciences & School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
yangchuanpeng@iie.ac.cn

Fuqing Zhu*
Institute of Information Engineering, Chinese Academy of Sciences & School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
zhufuqing@iie.ac.cn

Jizhong Han
Institute of Information Engineering, Chinese Academy of Sciences
Beijing, China
hanjizhong@iie.ac.cn

Songlin Hu
Institute of Information Engineering, Chinese Academy of Sciences & School of Cyber Security, University of Chinese Academy of Sciences
Beijing, China
husonglin@iie.ac.cn



## ABSTRACT

Harmful memes detection is a challenging task in the field of multimodal information processing due to the semantic gap between different modalities. Current research on this task mainly focuses on multimodal dual-stream models. However, the existing works ignore the misalignment of the memes caused by the modality gap. Moreover, the cross-modal interaction in the dual-stream models is insufficient to identify harmful memes. To this end, this paper proposes a scalable invariant and specific modality (ISM) representations framework via graph neural networks. The proposed ISM framework provides a comprehensive and disentangled view for memes and promotes inter-modal interaction. Specifically, ISM projects each modality to two distinct spaces. The first space is modality-invariant, learning the corresponding commonalities and reducing the modality gap. The second space is modality-specific, holding the distinctive characteristics of each modality and complementing the common latent features captured in invariant spaces. Then, we construct fully connected visual and textual graphs for each space. The unimodal graphs are fused to dynamically balance inter-modal and intra-modal relationships, which are complementary to the dual-stream models. Finally, an adaptive module is designed to weigh the proportion of each fusion graph for memes. Moreover, the mainstream multimodal dual-stream models could be employed as the backbone flexibly. Extensive experiments on five publicly available datasets show that the proposed ISM provides a stable improvement over baselines and produces a competitive performance compared with the existing harmful memes detection methods.



*Corresponding author.





*MM '23, October 29-November 3, 2023, Ottawa, ON, Canada*



ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3611761



## CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Computer vision representations**;

## KEYWORDS

harmful memes detection, modality invariant, modality specific, cross-modal interaction

**ACM Reference Format:**
Chuanpeng Yang, Fuqing Zhu, Jizhong Han, and Songlin Hu. 2023. Invariant Meets Specific: A Scalable Harmful Memes Detection Framework. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3611761


**Disclaimer:***This paper contains discriminatory content that may be disturbing to some readers.*

## 1 INTRODUCTION

The growing popularity of social media has enabled users to share and spread ideas at a prodigious rate. The information exchanges on social media platforms can guarantee freedom of expression and enhance an individual sense of connection with real and virtual communities. But platforms are increasingly used as tools to express hate, sarcasm, offense and other harmful content that directly or indirectly attacks people based on characteristics, including religion, gender, race, etc [16]. Due to their viral nature, memes have become a prevalent way of propagating harmful content in current social media. Typically, a meme is an image embedded with a short piece of text that is humorous in nature. However, in the context of contemporary political and socio-cultural divisions, a seemingly harmless meme can easily become a source of multimodal harm by using an adroit combination of images and texts. Therefore, classifying harmful memes turns out to be a very challenging task.

As shown in Figure 1, the memes of H1 and H2 correspond to the same text. Combined with various images, the sentiment tendencies of expression are opposite. Similarly, the memes of H3 and H4 correspond to the same image, and the sentiment tendencies of expression are also completely opposite due to different additional


**Figure 1: The examples of memes with harmful speech. "H" for Hate memes, "S" for Sarcasm memes and "O" for Offense memes.**

texts. Thus, images or texts alone are insufficient to identify the sentiment tendencies in the aforementioned memes. The diversity and interactivity of modality information make the conventional single modality detection ineffective. Recent harmful memes detection methods mainly focus on visual and textual feature fusion [16, 17], as well as directly fine-tuning multimodal pre-trained models [7, 16, 24]. Despite their advances, these detection techniques are often challenged by the persistent gap between heterogeneous modalities, which result in the misalignment of images and texts in memes (*e.g.*, S1, S2). Moreover, for any meme, each modality holds distinctive characteristics and represents different sentiment tendencies. For example, both memes O1 and O2 express harmful speech about *black race*. But the overall sentiment tendency of O1 memes is ultimately determined by the text, while the overall sentiment tendency of O2 memes is ultimately determined by the image. Therefore, the cross-modal interaction modeling in terms of modality-invariant and modality-specific could be significant for the harmful memes detection task.

According to the model architecture, existing multimodal models can be generally divided into two categories, which are single-stream and dual-stream [57]. In single-stream models, both texts and images are fed into an encoder module, such as VisualBERT [20], VL-BERT [38], and UNITER [5]. The semantic interaction of inter-modal and intra-modal could be simultaneously captured by the self-attention mechanism. However, it is unreasonable to exchange cross-modal information by treating the two kinds of information equally due to the large variations between inter-modal and intra-modal distributions. In dual-stream models, both texts and images are respectively fed into the corresponding encoder modules, such as ViLBERT [29], X-LXMERT [6], and ERNIE-VIL [50]. The semantic interaction of inter-modal is captured by the cross-attention mechanism, and the semantic interaction of intra-modal is mainly accomplished through the linear transformation layer. In this situation, the discrepancy of inter-modal information and the consistency of intra-modal information could be guaranteed. However, the ability of modeling inter-modal interaction is insufficient compared to the single-stream models. With the emergence of Vision Transformer (ViT) [10], more and more research works tend to design the dual-stream models, such as CLIP [35], ALBEF [19] and BLIP [18]. Therefore, this paper focuses on the dual-stream architecture to exploit cross-modal interaction and discover potential semantic relationships in the harmful memes detection task.

When the dual-stream models are employed, each modality is separately treated and classified in multimodal detection tasks. However, the cross-modal interaction in dual-stream models is insufficient for the harmful memes detection task. Recently, graph neural networks (GNNs) are widely applied for learning node correlation representations, promoting the information interaction between various modalities in multimodal-related tasks (*e.g.*, social networks [14, 40], recommendation systems [41, 54] and rumor detection [1, 23]). Rather than merely focus on the similarities in content, GNNs can discover the potential semantic relationships by establishing connections between different modalities and integrating the information of neighbor nodes. Inspired by these advantages, we attempt to leverage GNNs to facilitate the inter-modal interaction modeling of the dual-stream models.

In this paper, we propose a scalable invariant and specific modality (ISM) representations framework via graph neural networks for harmful memes detection. Specifically, two distinct feature representations are learned for each modality. The first representation is modality-invariant, which aims to reduce the modality gap. The second representation is modality-specific, which is private to each modality. Then, we construct fully connected visual and textual graphs for each space utilizing the obtained modality-invariant and modality-specific representations. The unimodal graphs are fused to balance inter-modal and intra-modal relationships dynamically. Each graph node contains sufficient interactive information by aggregation, which is complementary to the dual-stream models. Finally, we design an adaptive module to weigh the proportion of each fusion graph for memes.

The contributions of this paper are summarized as follows:

- A scalable harmful memes detection framework ISM is proposed to learn modality-invariant and modality-specific representations via graph neural networks, providing a comprehensive and disentangled view of memes by reducing the modality gap and aligning image-text pairs.

- We construct fully connected graphs of visual and textual to balance the inter-modal and intra-modal relationships dynamically, promoting inter-modal information interaction and complementing the dual-stream models.
- Extensive experiments on five publicly available datasets demonstrate that the proposed ISM provides a stable improvement over baselines and produces a competitive performance compared with existing harmful memes detection methods.

## 2 RELATED WORK

### 2.1 Harmful Memes Detection

Harmful memes detection is an emerging multimodal classification task that aims to identify negative information, including hate, sarcasm and offense speech. Existing studies have explored classic dual-stream models that combine the visual and textual features learned from image and text encoders using attention-based mechanisms and other fusion methods to perform harmful memes detection.

For hate memes detection, some early feature fusion methods [16, 39] are adopted to concatenate textual and visual features for classification. Besides, the large-scale pre-trained multimodal models [7, 16, 24] are directly fine-tuned for feature learning. Recent studies have also attempted to utilize data augmentation [4, 17, 48, 55, 56] and ensemble methods [37, 45] to improve the hate memes classification performance. However, the above methods require extracting additional features from the image, such as entities and demographic information. For sarcasm memes detection, Cai *et al.* [3] construct a new dataset from image-text tweets and propose a hierarchical fusion model. Relying on the dataset, some models [31, 47] are designed to explore implicit associations between images and texts in sarcasm. Liang *et al.* [21, 22] deploy a heterogeneous graph structure to learn the sarcastic features from both intra- and inter-modality perspectives. For offense memes detection, an offensive dataset containing abusive messages against a person or minority group is constructed [39]. The visual and textual representations are disentangled for offense memes understanding [17]. For harm memes detection, a harmful dataset about COVID-19 is constructed and further expanded by adding US politics-related [33], and MOMENTA is proposed based on intra-modal attention [34]. However, the above works neglect the image-text misalignment caused by the persistent modality gap in harmful memes. Therefore, we propose a scalable harmful memes detection framework ISM to learn modality-invariant and modality-specific representations, providing a comprehensive and disentangled view of memes by reducing the gap between modalities and aligning image-text pairs.

### 2.2 Multimodal Models

Multimodal models aim to improve the performance of downstream vision and language tasks by pretraining the model on large-scale image-text pairs and have received tremendous success on various multimodal tasks. Existing multimodal models could be divided into two types (*i.e.*, single-stream and dual-stream) from the architectural perspective. The single-stream models utilize a single transformer [43] encoder to model both image and text representations in a unified semantic space. The semantic interaction of inter-modal and intra-modal could be simultaneously captured by the self-attention mechanism. The representative models include VisualBERT [20], VL-BERT [38] and UNITER [5]. However, it is unfavorable to exchange cross-modal heterogeneous information by treating two types of information equally. The dual-stream models encode images and texts with the individual image encoder and text encoder. The semantic interaction of inter-modal is captured by the cross-attention mechanism and the semantic interaction of intra-modal is mainly accomplished through the linear transformation layer. The representative models include ViLBERT [29], X-LXMERT [6] and ERNIE-VIL [50]. However, the inter-modal interaction is insufficient compared to the single-stream models. This paper mainly focuses on the dual-stream models for harmful memes detection and relies on knowledge from off-the-shelf pre-trained dual-stream models (*e.g.*, CLIP [35], ALBEF [19] and BLIP [18]) to discriminate memes. To solve the problem of insufficient interaction between modalities, we utilize the powerful ability of graph neural networks to capture node information interaction to complement the dual-stream models.

### 2.3 Graph Neural Networks

Graph neural networks (GNNs) could learn node correlation representations and promote the semantic interaction between various modalities. Rather than merely focus on the similarities in content, GNNs can discover the potential semantic relationships among different modalities and integrate the information of neighbor nodes. Recently, GNNs have been widely applied in multimodal-related fields, such as social networks [14, 40], recommendation systems [41, 54] and rumor detection [1, 23]. Furthermore, Wen *et al.* [46] design a dual semantic relation module through the graph attention network to enhance the regional and global relations for more accurate multimodal representations. Yin *et al.* [49] and Zhang *et al.* [52] propose a multimodal graph fusion encoder for neural machine translation and named entity recognition respectively to enhance the representation of text features, combining the strengths of structural information with semantic information. Inspired by the above GNNs-based semantic interaction methods, we apply GNNs to the field of harmful memes detection to construct fully connected visual and textual graphs. The graphs are fused to dynamically balance the inter-modal and intra-modal relationships, which are also complementary to the dual-stream models.

## 3 METHODOLOGY

### 3.1 Problem Statement

For the harmful memes detection task, each meme is associated with an image $I$ and a text $T$ composed of a word sequence. Both the visual and textual modalities correspond to the class label $y$. Our goal is to design a classification model that can predict the label of the harmful meme (harmful or non-harmful) by integrating information from both the visual and textual modalities. Specifically, $y_0$ represents the prediction probability of meme harmlessness while $y_1$ represents the probability of the meme being harmful. If $y_1 > y_0$, the meme is predicted as harmful, otherwise, non-harmful.

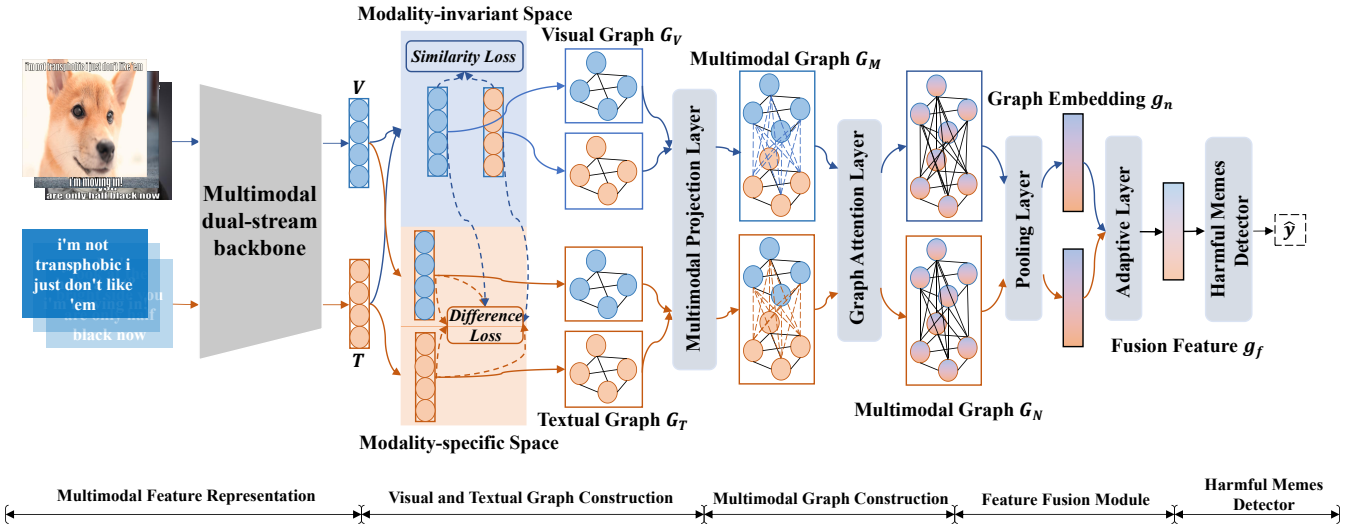


**Figure 2: The architecture of the proposed ISM.**

## 3.2 Model Overview

In this section, we describe our scalable invariant and specific modality (ISM) representations framework for harmful memes detection in detail. As demonstrated in Figure 2, the architecture of the proposed ISM contains five main components: 1) *Multimodal feature representation*, which employs the multimodal dual-stream models to capture the image and text features; 2) *Visual and textual graph construction*, which projects the obtained image-text representations into two distinct spaces to learn the modality-invariant and modality-specific features of memes; 3) *Multimodal graph construction*, which establishes the connections between visual nodes and text nodes, and introduces the graph attention layer to weigh the inter-modal and intra-modal relationships dynamically; 4) *Feature fusion module*, which designs an adaptive layer to balance the graph-level representation of different spaces in each meme; 5) *Harmful memes detector*, which feeds the fused features into the detector to identify whether memes are harmful. In the proposed ISM, the multimodal dual-stream backbone is replaceable, where CLIP, ALBEF and BLIP could be employed for feature representation.

## 3.3 Multimodal Feature Representation

Given a meme consisting of an image and text, we first resize the image to a fixed size and segment it into patches. Then, the patches are fed into the transformer model to encode them into the image hidden state vectors $\{v_0, v_1, ..., v_s\}$. For the text, we first tokenize and embed it into the word vectors following [8]. Then, we apply the transformer model to the word vectors to encode them as a list of hidden state vectors $\{t_0, t_1, ..., t_l\}$.

## 3.4 Visual and Textual Graph Construction

Harmful memes are metaphorical, which put forward a higher requirement for complex semantics understanding. Therefore, we argue for learning modality-invariant and modality-specific representations to provide a holistic view of memes. Inspired by [12], we project each meme into two distinct spaces. The first is the modality-invariant component to learn a shared representation in a common subspace with distribution similarity constraint. The constraint is to minimize the gap between modalities. Then we introduce a similarity loss. The discrepancy between the shared representations of each modality can be further reduced by minimizing the loss of similarity. The common cross-modal features in the meme are aligned in the shared subspace. Domain adaptation [26] has shown a superior ability for aligning the feature distribution, which mainly includes two aspects. The first is the statistic moment matching-based method, *e.g.*, MMD, KL-divergence [27, 28, 58]. Another is the adversarial learning-based method, *e.g.*, domain adversarial adaptation [11, 13]. We utilize the central moment difference (CMD) [51] metric as the similarity loss to align the feature distribution. The discrepancy between the distribution of two representations is measured by matching the corresponding order-wise moment differences. CMD distance decreases as the two distributions become similar. Compared to the MMD or KL-divergence method, CMD performs the explicit match of higher-order moments without expensive distance and kernel matrix computations. Compared to the adversarial training method, the CMD formula expression is more straightforward, since there is no discriminator with extra parameters.

Let $X$ and $Y$ be bounded random samples with respective probability distributions $p$ and $q$ on the interval $[a, b]^N$. The central moment discrepancy regularizer $\mathrm{CMD}_K$ is defined as an empirical estimate of the CMD metric, by

$$\begin{aligned}\mathrm{CMD}_K(X, Y) = &\frac{1}{|b-a|}\|\mathbf{E}(X) - \mathbf{E}(Y)\|_2 \\ &+ \sum_{k=2}^{K} \frac{1}{|b-a|^k} \|C_k(X) - C_k(Y)\|_2,\end{aligned} \tag{1}$$

where $\mathbf{E}(X) = \frac{1}{|X|}\sum_{x\in X} x$ is the empirical expectation vector computed on the sample $X$ and $C_k(X) = \mathbf{E}((x - \mathbf{E}(X))^k)$ is the vector

of all the $k^{th}$ order sample central moments of the coordinates of $X$. In this paper, CMD loss is calculated between the invariant representations of each image-text pair as follows:

$$\mathcal{L}_{\text{sim}} = \text{CMD}_K(V_{mi}, T_{mi}), \tag{2}$$

where $V_{mi}$ and $T_{mi}$ are the vectors projected into the modality-invariant space by the image representation $V$ and the text representation $T$, respectively.

The second is the modality-specific component to capture the unique characteristics of the corresponding modality. Motivated by recent works [2, 12, 25, 36] on shared-private latent space analysis, we design a difference loss that enforces a soft orthogonality constraint between the two representations. It penalizes redundant latent representations and ensures the modality-invariant and -specific representations capture different aspects of the input. Taking the image as an example, let $H_{V_{mi}}$ and $H_{V_{ms}}$ be the matrices whose rows denote the hidden vectors $V_{mi}$ and $V_{ms}$, respectively. The orthogonality constraint is calculated as $\|H_{V_{mi}}^{\top} H_{V_{ms}}\|_F^2$, where $\|\cdot\|_F^2$ is the squared Frobenius norm. In addition to the constraints between the invariant and specific representations, we also add orthogonality constraints between the modality-specific representations. The overall difference loss is computed as:

$$\mathcal{L}_{\text{diff}} = \|H_{V_{mi}}^{\top} H_{V_{ms}}\|_F^2 + \|H_{T_{mi}}^{\top} H_{T_{ms}}\|_F^2 + \|H_{V_{ms}}^{\top} H_{T_{ms}}\|_F^2, \tag{3}$$

where $V_{ms}$ and $T_{ms}$ are the vectors projected into the modality-specific space by the image representation $V$ and the text representation $T$, respectively.

After obtaining a comprehensive and disentangled view of memes, we utilize them to construct fully connected visual and textual graphs for each space. For the visual graph $G_V$, any two nodes are connected to construct a homogeneous graph, and the relationship between nodes is assumed according to [9]. The visual graph is fully connected with unweighted and bidirectional edges. For the textual graph $G_T$, the construction pattern is similar to the visual graph. The textual graph is also fully connected with unweighted and bidirectional edges. Visual graph nodes and text graph nodes represent similar semantics. To alleviate feature redundancy and fully capture the semantic correlation between modalities, we employ a weight-sharing fully connected network to learn multimodal representations in a unified feature space [32]. A single fully-connected feed-forward layer followed by a non-linearity exponential linear unit (ELU) [42] is utilized, and the node features of visual and textual graphs are projected to the common feature space.

$$m_s^v = \text{ELU}(W_m v_s + b_m), \tag{4}$$

$$m_l^t = \text{ELU}(W_m t_l + b_m), \tag{5}$$

where $W_m$ and $b_m$ are learnable parameters of the multimodal projection layer. $m_s^v$ and $m_l^t$ are visual and textual feature representations in the unified space.

## 3.5 Multimodal Graph Construction

In order to alleviate the inadequacy of cross-modal interaction in multimodal dual-stream models, we simultaneously model the inter-modal and intra-modal semantic relationships. The unweighted and bidirected edges are introduced to connect nodes in the visual and textual graphs. Different from the dual-stream models that formulate cross-modal relationships from a global perspective, we construct the inter-modal and intra-modal edges to learn from dependencies generated both within and across modalities concurrently at a fine-grained level. For the multimodal graph $G_M$, the nodes consist of nodes in the visual and textual graphs, which are represented as $\{m_0^v, m_1^v, ..., m_s^v, m_0^t, m_1^t, ..., m_l^t\}$. The edge is a matrix in which the diagonal elements are zeros and the rest are ones.

To learn the discriminative representation of each node in the multimodal graph, we employ the graph attention networks (GAT) [44]. Unlike the existing dual-stream models in which separate blocks are used for inter-modal and intra-modal fusion, the GAT layer allows nodes to adaptively select between inter-modal and intra-modal connections simultaneously. Specifically, we first apply projection transformation to transform nodes $\{m\}$ into nodes $\{n'\}$ in the multimodal graph $G_M$ according to the Eq.(6). Then, attention scores are calculated, *i.e.*, an attention score $attn_{ij}$ between $i^{th}$ node $n_i'$ and $j^{th}$ node $n_j'$ $(i \neq j)$ is calculated through Eq.(7) and Eq.(8). We derive $e_{ij}$, by first concatenating two nodes $n_i'$ and $n_j'$, and then projecting to a scalar value through a dot product with a learnable parameter $\gamma$, followed by a Leaky ReLU non-linearity [30]. Based on the calculated attention scores, the GAT performs self-attention on node $n_i'$ as in Eq.(9). The GAT is applied to each node $m$ transforming $n'$ to $n$, and then the output $n$ is fed into a graph pooling layer. The GAT process could be described as follows:

$$n' = mW_n, \tag{6}$$

$$e_{ij} = \text{LeakyReLU}(\gamma[n_i' \parallel n'_j]), \tag{7}$$

$$attn_{ij} = \text{Softmax}(e_{ij}), \tag{8}$$

$$n_i = \sum_{j=0}^{N} attn_{ij} n_j', \tag{9}$$

where $\parallel$ represents the concatenation operation. $N$ represents the number of nodes in the multimodal graph $G_N$.

## 3.6 Feature Fusion Module

We utilize a graph pooling layer to obtain the graph-level feature $g_n$. Similar to pooling layers in convolutional neural networks, graph pooling can reduce the size of features to enable high-level feature encoding and receptive field enlargement:

$$g_n = \text{Pool}(\{n_0, n_1, ..., n_i\}). \tag{10}$$

The inconsistent degree of image and text in each meme is different. In order to further promote the alignment of image-text pairs while maintaining the characteristics of each modality, we design the adaptive module to measure the proportion of modality-invariant and modality-specific representations. For the image representation $V$ and text representation $T$ in each meme, we utilize the cosine distance to calculate semantic similarity and weight for each graph-level representation to learn more robust fusion feature:

$$S = cosine(V, T) = \frac{VT}{\|V\|_2 \|T\|_2}, \tag{11}$$

$$g_f = [S g_{n_{mi}} \parallel (1 - S) g_{n_{ms}}], \tag{12}$$

where $g_{n_{mi}}$ and $g_{n_{ms}}$ are graph-level representations on modality-invariant and modality-specific spaces, respectively. $g_f$ is the graph fusion feature representation.

### 3.7 Harmful Memes Detector

The fusion feature $g_f$ is fed into the harmful memes detector for discrimination. The detector consists of a two-layer fully connected feed-forward network with intermediate ReLU non-linearity, and a softmax layer to estimate the harmful probability.

$$h_1 = \text{ReLU}(W_1 g_f + b_1), \tag{13}$$

$$\hat{y} = \text{Softmax}(W_2 h_1 + b_2), \tag{14}$$

where $W_1$, $W_2$, $b_1$ and $b_2$ are learnable parameters. $\hat{y}$ is the estimated probability. The cross-entropy loss $\mathcal{L}_{\text{task}}$ is utilized for harmful memes detection task:

$$\mathcal{L}_{\text{task}} = -\frac{1}{L}\sum_{c=1}^{L} y_c \log(\hat{y}_c), \tag{15}$$

where $L$ is the total number of memes in the training set. $y_c$ is the ground-truth one-hot label. The total objective function of ISM framework is:

$$\mathcal{L}_{\text{Loss}} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{sim}} + \mathcal{L}_{\text{diff}}. \tag{16}$$

## 4 EXPERIMENTS

### 4.1 Datasets

The experiment is conducted on five publicly available datasets. The details are briefly described as follows:

**Hate memes** are constructed as part of the Hateful Memes Challenge 2020 for multimodal hate speech detection and published in [16], containing 10$K$ memes with binary labels (*i.e.*, hateful or non-hateful).

**Sarcasm memes** consist of image-text tweets collected in [3] for multimodal sarcasm detection, containing nearly 25$K$ memes with binary labels (*i.e.*, sarcasm or non-sarcasm).

**Offense memes** are related to the 2016 United States presidential election and published in [39] for multimodal offensive detection, containing nearly 1$K$ memes with binary labels (*i.e.*, offensive or non-offensive).

**Harm-C memes** are related to COVID-19 and published in [33] for multimodal harmful detection, containing nearly 3.5$K$ memes with binary labels (*i.e.*, harmful or non-harmful).

**Harm-P memes** are related to United States politics and published in [34] for multimodal harmful detection, containing nearly 3.5$K$ memes with binary labels (*i.e.*, harmful or non-harmful).

### 4.2 Experimental Settings

**Implementation details.** The architecture of the proposed ISM framework is shown in Figure 2, where the multimodal dual-stream backbone can be selected from CLIP [35], ALBEF [19] and BLIP [18]. During the multimodal feature representation stage, the output images and texts are both 768-dimensional feature vectors. During the visual and textual graph construction stage, the input of the multimodal projection layer is 768 dimensions and the output is 384 dimensions. During the multimodal graph construction stage, the GAT attention layer adopts ELU non-linearity, 0.4 feature dropout rate and one attention head. The output dimension is 256. During the feature fusion module stage, the output of the pooling layer is 128 dimensions. During the harmful memes detector stage, the intermediate feature dimension of the detector is 64 and the dropout rate is 0.4. For the above backbone models, the initial learning rate is set to 1e-5, 3e-5 and 2e-5, respectively. The size of the minibatch is set to 16. Each dataset is trained for 20 epochs.

**Table 1: The performance (%) comparison on Hate memes.**

| Models | Acc. ↑ | AUROC ↑ |
|---|---|---|
| Late Fusion [15] | 63.20 | 69.30 |
| Concat BERT [15] | 61.53 | 67.77 |
| MMBT-Region [15] | 67.66 | 73.82 |
| ViLBERT [29] | 65.27 | 73.32 |
| Visual BERT [20] | 66.67 | 74.42 |
| DisMultiHate [17] | 71.26 | 79.89 |
| PromptHate [4] | 72.98 | 81.45 |
| CLIP | 59.00 | 68.30 |
| ISM(CLIP) w/o adapt | 64.20 | 73.60 |
| ISM(CLIP) | 66.20 | 75.45 |
| ALBEF | 68.30 | 80.79 |
| ISM(ALBEF) w/o adapt | 72.60 | 81.35 |
| ISM(ALBEF) | **74.70** | **83.71** |
| BLIP | 68.80 | 74.93 |
| ISM(BLIP) w/o adapt | 71.80 | 77.62 |
| ISM(BLIP) | 73.50 | 80.44 |

**Evaluation metrics.** For Hate memes, we follow the evaluation method adopted by [16], utilizing Area Under the Receiver Operating Characteristic curve (AUROC) and accuracy (Acc.) as evaluation metrics. The AUROC is the primary metric. For Sarcasm memes, we follow the evaluation method adopted by [3], using F1, precision (Pre.), recall (Rec.) and Acc. as evaluation metrics. For Offense memes, we follow the evaluation method adopted by [39], using F1, Pre. and Rec. as evaluation metrics. For Harm-C and Harm-P memes, we follow the evaluation method adopted by [34], using Acc., F1, and MMAE as evaluation metrics.

### 4.3 Experimental Results

**Comparison with the baselines.** To evaluate the effectiveness of the proposed ISM framework, three strong multimodal dual-stream models (*i.e.*, CLIP, ALBEF and BLIP) are employed as the backbone of ISM, which are also the baselines in this paper. It can be observed from Table 1-4 that ISM outperforms the corresponding baselines. Specifically, for hate memes, AUROC is increased by +7.15%, +2.92% and +5.51% on each backbone. The best performance is achieved when ALBEF is employed the backbone. For sarcasm memes, F1 is increased by +2.44%, +2.58% and +3.23%, respectively. For offense meme, F1 is increased by +4.28%, +5.42% and +5.88%, respectively. For harm-C memes, MMAE is improved by 0.0663, 0.0222 and 0.0189, respectively. For harm-P memes, MMAE is improved by 0.0216, 0.0046 and 0.0212, respectively. The best performance is achieved when BLIP is employed as the backbone in the above four datasets. The stable improvement demonstrates the effectiveness of

**Table 2: The performance (%) comparison on Sarcasm memes.**

| Models | F1 ↑ | Pre. ↑ | Rec. ↑ | Acc. ↑ |
|---|---|---|---|---|
| HFM [3] | 80.18 | 76.57 | 84.15 | 83.44 |
| D&R Net [47] | 80.60 | 77.97 | 83.42 | 84.02 |
| Res-Bert [31] | 80.85 | 77.80 | 84.15 | 84.80 |
| MIII-MMSD [31] | 80.90 | 78.63 | 83.31 | 86.05 |
| InCrossMGs [21] | 82.84 | 81.38 | 84.36 | 86.10 |
| CMGCN [22] | 84.16 | 83.63 | **84.69** | 87.55 |
| CLIP | 78.32 | 76.83 | 79.87 | 82.40 |
| ISM(CLIP) w/o adapt | 79.16 | 77.92 | 80.44 | 84.91 |
| ISM(CLIP) | 80.76 | 79.85 | 81.70 | 85.07 |
| ALBEF | 80.60 | 80.59 | 80.62 | 84.20 |
| ISM(ALBEF) w/o adapt | 82.13 | 80.71 | 83.60 | 85.66 |
| ISM(ALBEF) | 83.18 | 81.91 | 84.50 | 86.33 |
| BLIP | 81.24 | 81.48 | 81.01 | 84.90 |
| ISM(BLIP) w/o adapt | 82.85 | 84.20 | 81.54 | 86.95 |
| ISM(BLIP) | **84.47** | **85.21** | 83.74 | **88.20** |

learning modality-invariant and modality-specific representations via graph neural networks. Meanwhile, the experimental results on multiple backbones also show the flexible scalability of ISM.

**Comparison with the harmful memes detection methods.** We are the first to evaluate five datasets in the harmful memes detection community simultaneously. Thus, the most advanced comparison methods are different for each dataset.

- In hate memes, PromptHate [4] is a prompt-based model that leverages the implicit knowledge in the pre-trained language models to perform hateful memes classification.
- In sarcasm memes, CMGCN [22] is a graph model based on auxiliary object detection for modeling critical textual and visual information.
- In offense memes, DisMultiHate [17] disentangles target information from the meme to improve the offense content classification.
- In harm memes, TOT [53] deciphers the implicit harm in memes scenario in the way of topology-aware optimal transport.

Compared to the above methods, the proposed ISM could produce higher performance without extracting additional features (such as entities and demographic information) and build graphs in a simpler way.

To verify that different memes require various proportions of two representations, we remove the adaptive layer (w/o adapt). Overall, the degradation of performance on the five datasets indicates that the inconsistent degree of image and text in each meme is different (*i.e.*, the ratio of modality-invariant and modality-specific representations required by each meme is different).

## 4.4 Ablation Study

To analyze the effectiveness of each component in ISM framework, we carry out a series of ablation studies on Table 5, it can

**Table 3: The performance (%) comparison on Offense memes.**

| Models | F1 ↑ | Pre. ↑ | Rec. ↑ |
|---|---|---|---|
| StackedLSTM+VGG16 [39] | 46.30 | 37.30 | 61.10 |
| BiLSTM+VGG16 [39] | 48.00 | 48.60 | 58.40 |
| CNNText+VGG16 [39] | 46.30 | 37.30 | 61.10 |
| ERNIE-VIL [50] | 53.10 | 54.30 | 63.70 |
| DisMultiHate [17] | 64.60 | 64.50 | 65.10 |
| CLIP | 57.35 | 57.22 | 57.49 |
| ISM(CLIP) w/o adapt | 59.81 | 58.90 | 60.74 |
| ISM(CLIP) | 61.63 | 60.85 | 62.44 |
| ALBEF | 59.87 | 59.13 | 60.62 |
| ISM(ALBEF) w/o adapt | 62.86 | 62.40 | 63.32 |
| ISM(ALBEF) | 65.29 | 64.40 | **66.20** |
| BLIP | 60.30 | 60.50 | 60.10 |
| ISM(BLIP) w/o adapt | 63.97 | 65.30 | 62.70 |
| ISM(BLIP) | **66.18** | **67.30** | 65.10 |

**Table 4: The performance (%) comparison on Harm-C and Harm-P memes.**

| Models | Harm-C | | | Harm-P | | |
|---|---|---|---|---|---|---|
| | Acc. ↑ | F1 ↑ | MMAE ↓ | Acc. ↑ | F1 ↑ | MMAE ↓ |
| ViLBERT [29] | 78.53 | 78.06 | 0.1881 | 87.25 | 86.03 | 0.1276 |
| Visual BERT [20] | 81.36 | 80.13 | 0.1857 | 86.80 | 86.07 | 0.1318 |
| MOMENTA [34] | 83.82 | 82.80 | 0.1743 | 89.84 | 88.26 | 0.1314 |
| TOT [53] | 87.01 | 85.93 | 0.1634 | 91.55 | 91.29 | 0.1245 |
| CLIP | 73.45 | 72.61 | 0.2508 | 83.02 | 82.83 | 0.1604 |
| ISM(CLIP) w/o adapt | 77.22 | 76.48 | 0.2042 | 86.74 | 85.87 | 0.1432 |
| ISM(CLIP) | 78.62 | 78.16 | 0.1845 | 88.01 | 87.22 | 0.1388 |
| ALBEF | 78.75 | 77.67 | 0.1944 | 87.86 | 87.04 | 0.1330 |
| ISM(ALBEF) w/o adapt | 81.76 | 81.11 | 0.1805 | 88.64 | 87.90 | 0.1302 |
| ISM(ALBEF) | 83.46 | 81.92 | 0.1722 | 89.88 | 88.46 | 0.1284 |
| BLIP | 82.77 | 80.93 | 0.1774 | 89.45 | 88.19 | 0.1297 |
| ISM(BLIP) w/o adapt | 85.38 | 84.13 | 0.1677 | 91.15 | 90.19 | 0.1114 |
| ISM(BLIP) | **87.08** | **86.23** | **0.1585** | **92.29** | **91.61** | **0.1085** |

be observed: 1) The proposed ISM consisting of all modules produces the best performance on the above five datasets; 2) Removing the modality-invariant representation (w/o sim), performance decreases the most, verifying that reducing the gap between modalities to align image and text is particularly significant for identifying harmful memes; 3) Removing the modality-specific representation (w/o diff), performance drops a little, demonstrating that holding distinctive characteristics of each modality can complement the common latent features captured in the invariant space and provide a comprehensive multimodal representation of memes; 4) Removing the multimodal projection layer (w/o proj), performance degrades to some extent, illustrating that the projection before graph fusion can reduce the feature redundancy and fully capture the semantic correlation between modalities; 5) Removing the graph attention networks (w/o gat), performance degrades greatly, indicating that modeling the inter-modal and intra-modal relationships simultaneously is effective, and further proving the complementary effect for the dual-stream models. Regardless of which module is retained,

**Table 5: Ablation study evaluated on Hate, Sarcasm, Offense, Harm-C and Harm-P memes.**

| Models | Hate | | Sarcasm | | | | Offense | | | Harm-C | | | Harm-P | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc. ↑ | AUROC ↑ | F1 ↑ | Pre. ↑ | Rec. ↑ | Acc. ↑ | F1 ↑ | Pre. ↑ | Rec. ↑ | Acc. ↑ | F1 ↑ | MMAE ↓ | Acc. ↑ | F1 ↑ | MMAE ↓ |
| ISM(CLIP) | **66.20** | **75.45** | **80.76** | **79.85** | **81.70** | **85.07** | **61.63** | **60.85** | **62.44** | **78.62** | **78.16** | **0.1845** | **88.01** | **87.22** | **0.1388** |
| ISM(CLIP) w/o sim | 62.10 | 71.25 | 78.53 | 77.05 | 80.07 | 84.21 | 60.03 | 59.62 | 60.45 | 76.94 | 75.22 | 0.2109 | 85.87 | 85.58 | 0.1527 |
| ISM(CLIP) w/o diff | 64.50 | 74.23 | 80.16 | 79.02 | 81.33 | 84.92 | 59.96 | 60.62 | 59.31 | 78.10 | 77.59 | 0.2028 | 86.46 | 85.69 | 0.1456 |
| ISM(CLIP) w/o proj | 65.70 | 74.81 | 79.92 | 79.08 | 80.77 | 85.01 | 60.59 | 59.87 | 61.33 | 78.24 | 77.87 | 0.1943 | 87.03 | 86.71 | 0.1401 |
| ISM(CLIP) w/o gat | 63.40 | 73.47 | 79.03 | 78.18 | 79.89 | 84.82 | 59.21 | 60.01 | 58.44 | 77.53 | 76.69 | 0.2036 | 86.12 | 85.65 | 0.1505 |
| CLIP | 59.00 | 68.30 | 78.32 | 76.83 | 79.87 | 82.40 | 57.35 | 57.22 | 57.49 | 73.45 | 72.61 | 0.2508 | 83.02 | 82.83 | 0.1604 |
| ISM(ALBEF) | **74.70** | **83.71** | **83.18** | **81.91** | **84.50** | **86.33** | **65.29** | **64.40** | **66.20** | **83.46** | **81.92** | **0.1722** | **89.88** | **88.46** | **0.1284** |
| ISM(ALBEF) w/o sim | 70.10 | 80.85 | 81.61 | 80.62 | 82.63 | 85.61 | 63.88 | 62.70 | 65.10 | 80.49 | 78.85 | 0.1905 | 88.07 | 87.53 | 0.1307 |
| ISM(ALBEF) w/o diff | 73.90 | 83.42 | 82.82 | 81.60 | 84.07 | 85.83 | 64.43 | 63.57 | 65.32 | 81.70 | 81.03 | 0.1824 | 89.16 | 88.04 | 0.1295 |
| ISM(ALBEF) w/o proj | 73.10 | 82.54 | 82.90 | 81.42 | 84.44 | 86.08 | 64.96 | 64.10 | 65.85 | 82.74 | 81.69 | 0.1783 | 89.41 | 88.13 | 0.1290 |
| ISM(ALBEF) w/o gat | 72.70 | 81.44 | 82.39 | 81.05 | 83.77 | 85.72 | 64.06 | 63.12 | 65.03 | 81.23 | 80.67 | 0.1876 | 88.74 | 87.95 | 0.1301 |
| ALBEF | 68.30 | 80.79 | 80.60 | 80.59 | 80.62 | 84.20 | 59.87 | 59.13 | 60.62 | 78.75 | 77.67 | 0.1944 | 87.86 | 87.04 | 0.1330 |
| ISM(BLIP) | **73.50** | **80.44** | **84.47** | **85.21** | **83.74** | **88.20** | **66.18** | **67.30** | **65.10** | **87.08** | **86.23** | **0.1585** | **92.29** | **91.61** | **0.1085** |
| ISM(BLIP) w/o sim | 70.40 | 76.71 | 82.12 | 82.33 | 81.92 | 85.23 | 63.16 | 64.22 | 62.13 | 85.11 | 83.74 | 0.1692 | 90.42 | 89.84 | 0.1187 |
| ISM(BLIP) w/o diff | 73.20 | 79.95 | 83.26 | 83.11 | 83.41 | 87.01 | 64.78 | 65.84 | 63.76 | 86.77 | 85.80 | 0.1599 | 91.75 | 90.98 | 0.1091 |
| ISM(BLIP) w/o proj | 73.00 | 79.12 | 83.64 | 84.33 | 82.96 | 87.44 | 65.11 | 66.20 | 64.05 | 86.43 | 85.75 | 0.1602 | 91.30 | 90.20 | 0.1105 |
| ISM(BLIP) w/o gat | 72.40 | 78.04 | 82.75 | 83.14 | 82.37 | 86.86 | 64.08 | 66.00 | 62.26 | 85.78 | 84.66 | 0.1648 | 90.88 | 90.16 | 0.1125 |
| BLIP | 68.80 | 74.93 | 81.24 | 81.48 | 81.01 | 84.90 | 60.30 | 60.50 | 60.10 | 82.77 | 80.93 | 0.1774 | 89.45 | 88.19 | 0.1297 |

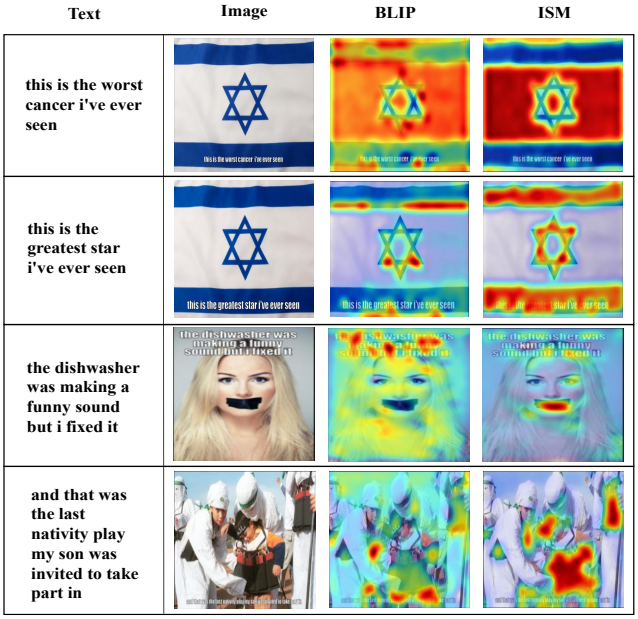


**Figure 3: Grad-CAM visualizations on the attention maps in the last layer of BLIP model.**

the result is above the baseline, demonstrating the reasonableness of the designed framework.

## 4.5 Case Study

The purpose of ISM is to provide a comprehensive and disentangled view of memes and promote semantic interaction between modalities for harmful memes detection. To understand ISM intuitively, we show the cases in Figure 3. Specifically, in the first sample, ISM associates the *cancer* with *white areas* in the image to indicate the severity of the *cancer*. Compared with BLIP, ISM has fine-grained semantic alignment between modalities. In the second sample, ISM focuses on the *star* and the areas that have the same color as the *star*. Compared with BLIP, ISM can understand the interaction between the text and image at a deeper level and narrow the gap between modalities. The above cases contain the same image with different texts and have completely opposite sentiment tendencies. ISM can pay differentiated attention to samples by learning modality-invariant and modality-specific representations.

In the third sample, the image contains harmful information about gender. BLIP is unable to identify due to interference from the embedded text in the image. However, ISM could accurately identify harmful information by associating text information with *mouth features* in the image. Similarly, in the last sample, ISM can focus more accurately on *gestures* and *dangerous objects* than BLIP. The above cases show that ISM could provide a comprehensive multimodal representation of memes and promote inter-modal interaction.

## 5 CONCLUSION

In this paper, a scalable framework for harmful memes detection (denoted as ISM) is proposed by learning modality-invariant and modality-specific representations via graph neural networks, which is complementary to the dual-stream models. ISM projects each modality into two different spaces, holding distinctive characteristics while learning commonalities to reduce the modality gap, providing a comprehensive and disentangled view for memes. Moreover, the mainstream multimodal dual-stream models (*e.g.*, CLIP, ALBEF and BLIP) could be employed as the backbone, verifying the scalability of ISM. Experimental results show that ISM provides a stable improvement over baselines and produces a competitive performance compared with the existing harmful memes detection methods. The ablation and case studies further demonstrate the effectiveness and rationality of each component.

# REFERENCES


[1] Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*. 549–556.

[2] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. 2016. Domain separation networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 343–351.

[3] Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 2506–2515.

[4] Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. Prompting for Multimodal Hateful Meme Classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 321–332.

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV : 16th European Conference, Glasgow, UK, August 23–28, 2020*. 104–120.

[6] Jaemin Cho, Jiasen Lu, Dustin Schwenk, Hannaneh Hajishirzi, and Aniruddha Kembhavi. 2020. X-LXMERT: Paint, Caption and Answer Questions with Multi-Modal Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 8785–8805.

[7] Abhishek Das, Japsimar Singh Wahi, and Siyao Li. 2020. Detecting hate speech in multi-modal memes. *arXiv preprint arXiv:2012.14891* (2020).

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[9] Mudit Dhawan, Shakshi Sharma, Aditya Kadam, Rajesh Sharma, and Ponnurangam Kumaraguru. 2022. Game-on: Graph attention network based multimodal fusion for fake news detection. *arXiv preprint arXiv:2202.12478* (2022).

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

[11] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research* (2016), 2096–2030.

[12] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*. 1122–1131.

[13] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. 2018. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*. PMLR, 1989–1998.

[14] Chao Huang, Huance Xu, Yong Xu, Peng Dai, Lianghao Xia, Mengyin Lu, Liefeng Bo, Hao Xing, Xiaoping Lai, and Yanfang Ye. 2021. Knowledge-aware coupled graph neural network for social recommendation. In *Proceedings of the AAAI conference on artificial intelligence*. 4115–4122.

[15] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised Multimodal Bitransformers for Classifying Images and Text. *arXiv preprint arXiv:1909.02950* (2019).

[16] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: detecting hate speech in multimodal memes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. 2611–2624.

[17] Roy Ka-Wei Lee, Rui Cao, Ziqing Fan, Jing Jiang, and Wen-Haw Chong. 2021. Disentangling hate in online memes. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5138–5147.

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*. PMLR, 12888–12900.

[19] Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *Proceedings of Advances in Neural Information Processing Systems*. 9694–9705.

[20] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).

[21] Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. 2021. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In *Proceedings of the 29th ACM international conference on multimedia*. 4707–4715.

[22] Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1767–1777.

[23] Hongzhan Lin, Jing Ma, Mingfei Cheng, Zhiwei Yang, Liangliang Chen, and Guang Chen. 2021. Rumor Detection on Twitter with Claim-Guided Hierarchical Graph Attention Networks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 10035–10047.

[24] Phillip Lippe, Nithin Holla, Shantanu Chandra, Santhosh Rajamanickam, Georgios Antoniou, Ekaterina Shutova, and Helen Yannakoudakis. 2020. A multimodal framework for the detection of hateful memes. *arXiv preprint arXiv:2012.12871* (2020).

[25] Pengfei Liu, Xipeng Qiu, and Xuan-Jing Huang. 2017. Adversarial Multi-task Learning for Text Classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1–10.

[26] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. 2015. Learning transferable features with deep adaptation networks. In *International conference on machine learning*. 97–105.

[27] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. 2018. Conditional adversarial domain adaptation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 1647–1657.

[28] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. 2017. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*. PMLR, 2208–2217.

[29] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 13–23.

[30] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. 2013. Rectifier nonlinearities improve neural network acoustic models. In *International conference on machine learning*. PMLR, 3–13.

[31] Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 1383–1392.

[32] Yuxin Peng and Jinwei Qi. 2019. CM-GANs: Cross-modal generative adversarial networks for common representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 15, 1 (2019), 1–24.

[33] Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Detecting Harmful Memes and Their Targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2783–2796.

[34] Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 4439–4455.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*. 8748–8763.

[36] Sebastian Ruder and Barbara Plank. 2018. Strong Baselines for Neural Semi-Supervised Learning under Domain Shift. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1044–1054.

[37] Vlad Sandulescu. 2020. Detecting hateful memes using a multimodal deep ensemble. *arXiv preprint arXiv:2012.13235* (2020).

[38] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *International Conference on Learning Representations*.

[39] Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*. 32–41.

[40] Yu Tian, Xingliang Huang, Ruigang Niu, Hongfeng Yu, Peijin Wang, and Xian Sun. 2022. Hypertron: Explicit Social-Temporal Hypergraph Framework for Multi-Agent Forecasting. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. 1356–1362.

[41] Yijun Tian, Chuxu Zhang, Zhichun Guo, Chao Huang, Ronald Metoyer, and Nitesh V. Chawla. 2022. RecipeRec: A Heterogeneous Graph Learning Model for Recipe Recommendation. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. 3466–3472.

[42] Ludovic Trottier, Philippe Giguere, and Brahim Chaib-Draa. 2017. Parametric exponential linear unit for deep convolutional neural networks. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 207–214.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of Advances in Neural Information Processing systems*, Vol. 30.

[44] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *International Conference on Learning Representations*.

[45] Riza Velioglu and Jewgeni Rose. 2020. Detecting hate speech in memes using multimodal deep learning approaches: Prize-winning solution to hateful memes challenge. *arXiv preprint arXiv:2012.12975* (2020).

[46] Keyu Wen, Xiaodong Gu, and Qingrong Cheng. 2020. Learning dual semantic relations with graph attention for image-text matching. *IEEE transactions on circuits and systems for video technology* 31, 7 (2020), 2866–2879.

[47] Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 3777–3786.

[48] Chuanpeng Yang, Fuqing Zhu, Guihua Liu, Jizhong Han, and Songlin Hu. 2022. Multimodal hate speech detection via cross-domain knowledge transfer. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4505–4514.

[49] Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A Novel Graph-based Multi-modal Fusion Encoder for Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3025–3035.

[50] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 3208–3216.

[51] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. 2017. Central Moment Discrepancy (CMD) for Domain-Invariant Representation Learning. In *International Conference on Learning Representations*.

[52] Dong Zhang, Suzhong Wei, Shoushan Li, Hanqian Wu, Qiaoming Zhu, and Guodong Zhou. 2021. Multi-modal graph fusion for named entity recognition with targeted visual guidance. In *Proceedings of the AAAI conference on artificial intelligence*. 14347–14355.

[53] Linhao Zhang, Li Jin, Xian Sun, Guangluan Xu, Zequn Zhang, Xiaoyu Li, Nayu Liu, Shiyao Yan, and Qing Liu. 2023. TOT: Topology-Aware Optimal Transport For Multimodal Hate Detection. *arXiv preprint arXiv:2303.09314* (2023).

[54] Yixin Zhang, Yong Liu, Yonghui Xu, Hao Xiong, Chenyi Lei, Wei He, Lizhen Cui, and Chunyan Miao. 2022. Enhancing Sequential Recommendation with Graph Contrastive Learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. 2398–2405.

[55] Yi Zhou, Zhenhao Chen, and Huiyuan Yang. 2021. Multimodal learning for hateful memes detection. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. 1–6.

[56] Ron Zhu. 2020. Enhance multimodal transformer with external label and in-domain pretrain: Hateful meme challenge winning solution. *arXiv preprint arXiv:2012.08290* (2020).

[57] Yushan Zhu, Huaixiao Zhao, Wen Zhang, Ganqiang Ye, Hui Chen, Ningyu Zhang, and Huajun Chen. 2021. Knowledge perceived multi-modal pretraining in e-commerce. In *Proceedings of the 29th ACM International Conference on Multimedia*. 2744–2752.

[58] Yongchun Zhu, Fuzhen Zhuang, and Deqing Wang. 2019. Aligning domain-specific distribution and classifier for cross-domain classification from multiple sources. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 5989–5996.