

# Ambiguity-aware Multi-level Incongruity Fusion Network for Multi-Modal Sarcasm Detection

Kuntao Li<sup>1\*</sup>, Yifan Chen<sup>1\*</sup>, Qiaofeng Wu<sup>1</sup>, Weixing Mai<sup>1</sup>  
Fenghuan Li<sup>2†</sup>, Yun Xue<sup>1†</sup>

<sup>1</sup>Guangdong Provincial Key Laboratory of Quantum Engineering and Quantum Materials,  
School of Electronic Science and Engineering (School of Microelectronics),  
South China Normal University

<sup>2</sup>School of Computer Science and Technology, Guangdong University of Technology  
{likuntao,chenyifan,maiwx,scnu\_wqf,xueyun}@m.scnu.edu.cn fhli20180910@gdut.edu.cn

## Abstract

Multi-modal sarcasm detection aims to identify whether a given image-text pair is sarcastic. The pivotal factor of the task lies in accurately capturing incongruities from different modalities. Although existing studies have achieved impressive success, they primarily committed to fusing the textual and visual information to establish cross-modal correlations, overlooking the significance of original unimodal incongruity information at the text-level and image-level. Furthermore, the utilized fusion strategies of cross-modal information neglected the effect of inherent ambiguity within text and image modalities on multimodal fusion. To overcome these limitations, we propose a novel Ambiguity-aware Multi-level Incongruity Fusion Network (AMIF) for multi-modal sarcasm detection. Our method involves a multi-level incongruity learning module to capture the incongruity information simultaneously at the text-level, image-level and cross-modal-level. Additionally, an ambiguity-based fusion module is developed to dynamically learn reasonable weights and interpretably aggregate incongruity features from different levels. Comprehensive experiments conducted on a publicly available dataset demonstrate the superiority of our proposed model over state-of-the-art methods.

## 1 Introduction

Sarcasm represents a pervasive linguistic phenomenon denoting a discrepancy between literal meanings and implied intentions (Liu et al., 2022; Wu et al., 2025). Consequently, sarcasm detection holds the potential to unveil a person’s real emotions and attitudes, providing significant advantages for tasks like product review analysis and political opinion mining (Wen et al., 2023; Chen et al., 2024b,a). In this paper, sarcasm refers generally to all linguistic phenomena with irony or satire.

\*These authors contributed equally to this work.

†Corresponding author.



Figure 1: Examples of Twitter data with sarcasm.

Early research on sarcasm detection predominantly focused on unimodal approaches. These methods regarded the sarcastic contexts or the sentiments of sarcasm makers as valuable indicators for modeling the textual incongruity, which captured sarcastic features at the text-level (Tay et al., 2018; Xiong et al., 2019). Alternatively, other works explored the incongruity of different visual regions to mine sarcastic features at the image-level (Cai et al., 2019; Kumar and Garg, 2019). However, since most social media posts contain abundant multimodal information (e.g., image and text), relying solely on unimodal approaches is insufficient for accurate sarcasm/non-sarcasm classification.

Recently, multi-modal sarcasm detection has attracted significant attention. Previous methods sought to explore diverse multimodal strategies for fusing textual and visual features to improve multi-modal sarcasm detection performance (Liang et al., 2021, 2022; Song et al., 2023; Liu et al., 2024; Zhong et al., 2024). However, they neglected the importance of capturing unimodal text-level and image-level sarcastic features. Furthermore, it’s crucial to recognize that not all levels of sarcastic features contribute equally to the decision-making process. In Figure 1(a), the image portrays a beautiful night scene while the text describes it as "ugly". It is challenging to distinguish sarcasm/non-sarcasm based solely on text-level or image-level sarcastic features. Prediction performance can be improved by capturing and enhancing cross-modal-level incongruity representation. However, in cer-

tain instances, unimodal incongruities play a more crucial role, while cross-modal-level incongruity may not be pivotal and could potentially introduce noise to the classification task. As illustrated in Figure 1(b), the interplay among the words “pain”, “thanks” and “forcing” is rich in sarcasm, providing a basis for considering it a sarcastic tweet. Likewise, in Figure 1(c), we can observe that the sentence in the image contains a wealth of sarcastic incongruity information, while the text makes a limited contribution to sarcasm detection. Hence, the primary questions revolve around: 1) how to more effectively mine text-level, image-level and cross-modal-level incongruity information simultaneously; 2) how different levels of incongruity affect the decision-making process and how to weigh their importance.

Taking the consideration above, we propose a novel **Ambiguity-aware Multi-level Incongruity Fusion Network (AMIF)** for multi-modal sarcasm detection. Specifically, we design a **Multi-level Incongruity Learning module (MIL)** to learn text-level, image-level and cross-modal-level incongruity information simultaneously. In this module, we introduce **Unimodal Gating Incongruity Extracting module (UGIE)** to leverage the contextual information of the text and image, effectively mining text-level and image-level sarcastic features. Additionally, we introduce **Cross-modal Incongruity Graph Reasoning module (CIGR)**, which captures the vector-based incongruity relationship between local and global alignments to identify more complex cross-modal sarcastic features. Subsequently, we feed the learned multi-level incongruity information into **Ambiguity-based Incongruity Fusion module (AIF)** for adaptive weighted fusion. In this module, we initially utilize cross-modal ambiguity (Chen et al., 2022) to quantify the relationship between incongruity information at different levels, which guides the modality-wise attention mechanism to adaptively assign reasonable weights to different levels of incongruity information, facilitating the effective aggregation of sarcastic features across diverse levels.

The main contributions of this work are summarized as follows:

- We propose a novel MIL module to simultaneously mine incongruity information from different levels of unimodal and cross-modal, which utilizes UGIE to mine the text-level and image-level sarcastic features and adopts

CIGR to extract the complex cross-modal sarcastic features.

- We are the first to introduce cross-modal ambiguity to quantify the correlations among different levels of incongruity information for multi-modal sarcasm detection, based on which we propose an AIF module that includes a modality-wise attention mechanism and an ambiguity guidance to adaptively assign reasonable weights and interpretably aggregate incongruity features from different levels.
- We conduct numerous experiments on a publicly available benchmark dataset. Experimental results show the superiority of our method over state-of-the-art methods.

## 2 Related Work

### 2.1 Multi-modal Sarcasm Detection

Unlike previous text-only sarcasm or irony detection tasks in other source languages that distinguish between sarcasm and irony (Potamias et al., 2020; Tomás et al., 2023; Cervone et al., 2017). In this paper, "sarcasm" refers to a linguistic phenomenon in general that is not fundamentally different from irony or satire in English. The rise of multimedia platforms has led to an increase in posts presented as image-text pairs, attracting plentiful research on multimodal sarcasm detection. Cai et al. (2019) established a new dataset and demonstrated that cross-modal information can provide complementary advantages to improve the performance of multi-modal sarcasm detection. Liang et al. (2021); Liu et al. (2022, 2024); Ma et al. (2024) revealed that the key to achieving effective multi-modal sarcasm detection lies in accurately extracting incongruities from different modalities. Hence, Liang et al. (2022) adopted object extraction technology to integrate specific visual regions and local semantic information to capture cross-modal sarcastic features, but they ignored the global alignment between the image and text. Another approach (Zhong et al., 2024) modeled textual and visual information at the multi-scale and multi-span token level to address image-text incongruity. Fang et al. (2024) investigated a cross-modal multi-granularity alignment module to capture align context features. The latest LLMs-based work (Jia et al., 2024) defined the task of out-of-distribution to evaluate models' generalizability when the word distribution is different in training and testing settings. Although

multimodal approaches had also achieved notable performance, most of them tended to use multimodal strategies to fuse textual and visual features, ignoring the significance of capturing text-level and image-level sarcastic features. Different from these works, we aim to improve this task by simultaneously mining incongruity information from different levels of unimodal and cross-modal.

## 2.2 Multimodal Fusion

Multimodal fusion is the core content of multimodal deep learning technology, which aims to fuse the information from distinct modalities to derive abundant features (Mai et al., 2024; Zhang et al., 2023, 2024a). Wang et al. (2019) constructed a deep neural network to embed inputs from different modalities and computed the similarity of different semantic vectors for late multimodal fusion. Ben-Younes et al. (2019) proposed Block, a new multimodal fusion model based on the block-superdiagonal tensor decomposition. Nagrani et al. (2021) modeled a architecture that used ‘‘fusion bottlenecks’’ for modality fusion at multiple layers to improve multimodal fusion performance. Other studies started from the common feature and specific feature, aiming to explore the redundancy and complementarity of diverse modalities for the final fusion (Lu et al., 2020; Wu et al., 2021). And Chen et al. (2022) considered the inherent ambiguity between different contents and proposed the cross-modal ambiguity learning from the perspective of information theory to measure the importance of different modalities in multi-modal classification tasks. Cross-modal ambiguity learning achieved more efficient multimodal fusion. Nevertheless, existing works for multi-modal sarcasm detection utilized attention mechanism, GCN and GAT to fuse sarcastic features from different modalities (Liu et al., 2022; Song et al., 2023; Fang et al., 2024; Liu et al., 2024; Wu et al., 2025). They neglected the effect of inherent ambiguity for multi-modal sarcasm detection.

## 3 Methodology

### 3.1 Modal-specific Encoder

Given the input  $x = (x^t, x^v) \in \mathcal{D}$ , where  $x^t$ ,  $x^v$  and  $\mathcal{D}$  denote the text, image and dataset, respectively. In this paper,  $x^t = \{w_i\}_{i=1}^n$ ,  $n$  refers to the length of the text, we utilize the pre-trained BERT (Kenton and Toutanova, 2019) to embed each word of the text. The final word embedding denotes

as  $e_i^t \in \mathbb{R}^d$ ,  $i$  means the  $i$ -th word. For each image, we first divide the image into  $K$  regions, and then choose the pre-trained ViT (Dosovitskiy et al., 2020) as our image encoder to embed each visual region of the image. The final region embedding denotes as  $e_j^v \in \mathbb{R}^d$ ,  $j$  means the  $j$ -th region.

### 3.2 Multi-level Incongruity Learning

**Unimodal Gating Incongruity Extracting.** To comprehensively exploit the hidden intra-modal contextual sarcastic cues, we propose the UGIE, which contains the multi-head self-attention and gate mechanism.

The multi-head self-attention with  $h$  heads can capture intra-modal contextual sarcastic cues from different subspaces. This can be calculated as:

$$H_i = \text{Attention}(Q_i, K_i, V_i) \quad (1)$$

where  $H_i$  denotes the output of the  $i$ -th head,  $Q_i = XW_i^Q$ ,  $K_i = XW_i^K$ ,  $V_i = XW_i^V$  denote the query, key and value of the  $i$ -th head, respectively.  $X \in \mathbb{R}^{n \times d}$  refers to the input,  $n$  and  $d$  separately denote the length and hidden dimension of  $X$ .  $W_i^Q$ ,  $W_i^K \in \mathbb{R}^{n \times d_k}$  and  $W_i^V \in \mathbb{R}^{n \times d_v}$  are learnable parameter matrices.  $d_k = d_v = d/h$ .

The multi-head self-attention mechanism attempts to lead the model to concentrate on correlations among various segments of the complete unimodal input. Nevertheless, the projected queries and keys may incorporate noise or sarcasm-unrelated information. To effectively convey the captured intra-modal useful sarcastic cues and suppress the irrelevant ones, we design a novel fusion strategy that integrates the multi-head self-attention mechanism with the gate mechanism. Specifically, for the  $i$ -th head, we initially map  $Q_i$  and  $K_i$  into a common space and fuse them. Then, we adopt two fully-connected layers to get the gating masks:

$$G_i = (Q_i W_Q^G) \odot (K_i W_K^G) \quad (2)$$

$$M_Q^i = \sigma(G_i W_Q^M); M_K^i = \sigma(G_i W_K^M) \quad (3)$$

where  $G_i \in \mathbb{R}^{n \times d_k}$  is the fusion result.  $W_Q^G$ ,  $W_K^G$ ,  $W_Q^M$ ,  $W_K^M \in \mathbb{R}^{d_k \times d_k}$  are the learnable projection matrices.  $M_Q^i$ ,  $M_K^i \in \mathbb{R}^{n \times d_k}$  denote the gating masks.  $\sigma$  refers to the sigmoid function.

Finally, we use these obtained gating masks to filter sarcasm-unrelated contextual information of original  $Q_i$  and  $K_i$ , as follows:

$$\tilde{H}_i = \text{Attention}(Q_i \odot M_Q^i, K_i \odot M_K^i, V_i) \quad (4)$$

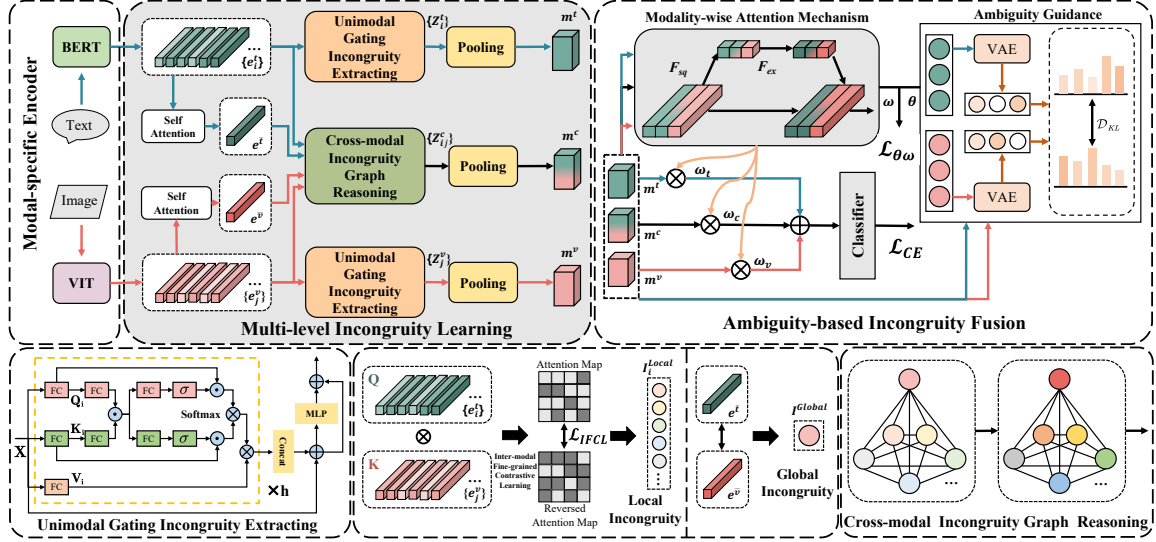


Figure 2: The architecture of the proposed AMIF.

where  $\tilde{H}_i \in \mathbb{R}^{n \times d_v}$  is the gating multi-head self-attention result of the  $i$ -th head.

We concatenate all heads to obtain the hidden intra-modal contextual sarcastic cues:

$$f(X) = \text{Concat}(\tilde{H}_1, \tilde{H}_2, \dots, \tilde{H}_h) + X \quad (5)$$

We take the residual connection operation on the output of an additional MLP and the original  $f(X)$  to obtain the final output of UGIE, as follows:

$$F_{UGIE}(X) = f(X) + \text{MLP}(f(X)) \quad (6)$$

Finally, we adapt UGIE for both the text and image, denoted as:  $Z_i^t = F_{UGIE}(e_i^t)$  and  $Z_i^v = F_{UGIE}(e_i^v)$ , where  $Z_i^t$  and  $Z_i^v$  separately represent the output of UGIE for the text and image.

### Cross-modal Incongruity Graph Reasoning.

To extract intricate cross-modal sarcastic cues, we perform cross-modal incongruity graph reasoning on graph constructed with aligned local and global incongruity representations as graph nodes.

Existing multi-modal sarcasm detection models have employed scalar-based methods to represent the incongruity information between the feature vectors of the text and image, which is unable to capture comprehensive correspondences (Liu et al., 2022; Song et al., 2023; Ma et al., 2024; Zhong et al., 2024; Zhang et al., 2024b). Therefore, in this section, we adopt a vector-based approach to compute both local and global correspondences for cross-modal sarcastic features alignment, which can capture rich incongruity information between feature representations from different modalities.

This can be computed as:

$$I(a, b; W_I) = \frac{W_I |a - b|^2}{\|W_I |a - b|^2\|_2} \quad (7)$$

where  $a, b \in \mathbb{R}^d$  are two different vectors,  $|\cdot|^2$  and  $\|\cdot\|_2$  separately represent the element-wise square and  $\ell_2$ -norm.  $W_I \in \mathbb{R}^{m \times d}$  is a parameter matrix.

**Local incongruity.** To explore the correspondence between local features of the image and text, we apply cross attention mechanism to attend on each region with respect to each word. Then, we obtain the attended visual representation  $a_i^v$  with respect to  $i$ -th word, as follows:

$$c_{i,j} = \frac{e_i^t (e_j^v)^\top}{\sqrt{d_k}} \quad (8)$$

$$\alpha_i^j = \frac{\exp(\beta c_{i,j})}{\sum_{j=1}^K \exp(\beta c_{i,j})}; a_i^v = \sum_{j=1}^K \alpha_i^j e_j^v \quad (9)$$

where  $c_{i,j}$  indicates the attention score for the intermediate transition between  $i$ -th word feature and  $j$ -th region feature,  $\alpha_i^j$  denotes the final attention weight between  $i$ -th word feature and  $j$ -th region feature.  $\beta$  is the inversed temperature factor.

We also design Inter-modal Fine-grained Contrastive Learning (IFCL) to refine the attention mechanism for acquiring more precise text-guided visual representation. Specifically, the attended visual representation  $a_i^v$  prioritizes the congruity between the image and text, while the reversed attention visual representation  $\hat{a}_i^v$  emphasizes the incongruity between the image and text:

$$\hat{\alpha}_i^j = \frac{\exp(\beta(1 - c_{i,j}))}{\sum_{j=1}^K \exp(\beta(1 - c_{i,j}))}; \hat{a}_i^v = \sum_{j=1}^K \hat{\alpha}_i^j e_j^v \quad (10)$$

The loss function of IFCL can be defined as:

$$\mathcal{L}_{IFCL} = [Sim(e_i^t, \hat{a}_i^v) - Sim(e_i^t, a_i^v) + \gamma]_+ \quad (11)$$

where  $\gamma$  controls the similarity difference margin.  $Sim$  is the similarity function.

Then, we calculate the vector-based cross-modal local incongruity between  $a_i^v$  and  $e_i^t$  with Equation 7:

$$I_i^{Local} = I(e_i^t, a_i^v; W_I^l) \quad (12)$$

where  $W_I^l \in \mathbb{R}^{m \times d}$  is a parameter matrix.

**Global incongruity.** To explore effective and in-depth correspondence between the global features of the entire text and image, we initially execute self-attention over all the obtained word and region embeddings respectively to yield the global text feature representation  $e^{\bar{t}} \in \mathbb{R}^d$  and the global image feature representation  $e^{\bar{v}} \in \mathbb{R}^d$ .

Likewise, we calculate the vector-based cross-modal global incongruity between  $e^{\bar{t}}$  and  $e^{\bar{v}}$  with Equation 7:

$$I^{Global} = I(e^{\bar{t}}, e^{\bar{v}}; W_I^g) \quad (13)$$

where  $W_I^g \in \mathbb{R}^{m \times d}$  is a parameter matrix.

Next, to achieve comprehensive reasoning for captured local and global incongruity information, we construct an incongruity graph to transmit the cross-modal incongruity. Specifically, we denote the local incongruity representations and global incongruity representation as nodes  $\mathcal{N} = \{I_1^{Local}, \dots, I_n^{Local}, I^{Global}\}$ , and the edge from node  $I_v \in \mathcal{N}$  to  $I_u \in \mathcal{N}$  can be computed as:

$$E(I_u, I_v) = \frac{\exp((P_I^{in} I_u)(P_I^{out} I_v))}{\sum_v \exp((P_I^{in} I_u)(P_I^{out} I_v))} \quad (14)$$

where  $P_I^{in}, P_I^{out} \in \mathbb{R}^{d \times d}$  are the linear transformations for incoming and outgoing nodes, respectively.

Subsequently, we perform cross-modal incongruity graph reasoning by iteratively updating the nodes and edges within the graph, as follows:

$$\hat{I}_u^t = \sum_v E(I_u^t, I_v^t) \cdot I_v^t; I_u^{t+1} = ReLU(P_I^t \hat{I}_u^t) \quad (15)$$

where  $I_u^0$  and  $I_v^0$  are taken from  $\mathcal{N}$  at step  $t = 0$ ,  $P_I^t$  is a learnable parameter in each step. We iterate

$T$  steps of incongruity reasoning and converge the information of the global node and all local nodes at the last step as the final reasoned incongruity representation  $Z_{ij}^c$ .

Finally, we take average-pooling on the output of UGIE and CIGR respectively to obtain the final multi-level incongruity representations, namely, text-level incongruity  $m^t$ , cross-modal-level incongruity  $m^c$  and image-level incongruity  $m^v$ .

### 3.3 Ambiguity-based Incongruity Fusion

**Modality-wise Attention Mechanism (MAM).** Inspired by the excellent performance of channel attention in computer vision (Zhang et al., 2022; Zhao et al., 2023), we design a modality-wise attention mechanism module, which help us to reweight the different levels of incongruity information before fusing them. Concretely, we first concatenate  $m^t, m^c$  and  $m^v \in \mathbb{R}^{d \times 1}$  as:  $m^f = m^t \oplus m^c \oplus m^v$ . Then, we take the squeeze operation  $F_{sq}(m^f) = GlobalAveragePooling(m^f)$  to aggregate global modality-wise incongruity information into a  $\mathbb{R}^{1 \times 3}$  vector. Subsequently, we adopt a gating mechanism  $F_{ex}(m^f) = \sigma(W_2 \delta(W_1 F_{sq}(m^f)))$  to obtain the modality-wise attention weight  $\omega = \{\omega_t, \omega_c, \omega_v\}$ , where  $\sigma$  refers to the sigmoid activation,  $\delta$  is the ReLU function,  $W_1 \in \mathbb{R}^{3 \times 1}$  and  $W_2 \in \mathbb{R}^{1 \times 3}$  are learnable parameter matrices.

**Ambiguity Guidance (AG).** When the cross-modal information gap is small, unimodal incongruity features alone are adequate for accurate sarcasm detection. Conversely, when there exists a significant information gap between unimodalities, relying solely on unimodal incongruity features becomes insufficient and additional attention should be paid to the cross-modal incongruity features. Therefore, inspired by Chen et al. (2022), we introduce the cross-modal ambiguity to measure the information gap between the text-level incongruity and the image-level incongruity.

Specifically, we use the Variational Autoencoder (VAE) (Khattar et al., 2019) to model the divergence over feature space to approximate the ambiguity between  $m^t$  and  $m^v$ . The variational posterior can be denoted as:  $q(z^{t/v} | m^{t/v}) = \mathcal{N}(z^{t/v} | \mu(m^{t/v}), \sigma(m^{t/v}))$ , where the mean  $\mu$  and variance  $\sigma$  can be obtained from the modal-specific encoder,  $t/v$  denotes text or image-modality. Considering the distribution over the mini-batch:

$$q(z^{t/v}) = \frac{1}{A} \sum_{i=1}^A q(z_i^{t/v} | m_i^{t/v}) \quad (16)$$

where  $A$  denotes the size of mini-batch,  $i$  refers to the  $i$ -th sample  $x_i$  in mini-batch. We quantify the ambiguity between the image-level incongruity distributions and the text-level incongruity distributions of sample  $x_i$  by the averaged KL divergence:

$$\theta_i^{t \rightarrow v} = \left( \frac{D_{KL}(q(z_i^t || m_i^t) || q(z_i^v || m_i^v))}{D_{KL}(q(z^t) || q(z^v))} \right) \quad (17)$$

where  $D_{KL}(\cdot || \cdot)$  stands for the KL divergence.

Likewise, we can get the  $\theta_i^{v \rightarrow t}$  with Equation 17, and compute the final ambiguity:

$$\theta_i = \text{sigmoid}\left(\frac{\theta_i^{t \rightarrow v} + \theta_i^{v \rightarrow t}}{2}\right) \quad (18)$$

Then, we obtain the cross-modal ambiguity scores  $\theta = \{1 - \theta_i, \theta_i, 1 - \theta_i\}_{i=1}^A$ .

We also exploit a loss function as:  $\mathcal{L}_{\theta\omega} = JS(\theta || \omega)$  to calculate the logarithmic difference between the attention weight  $\omega = \{\omega_t, \omega_c, \omega_v\}$  and the cross-modal ambiguity scores  $\theta = \{1 - \theta_i, \theta_i, 1 - \theta_i\}_{i=1}^A$ .  $JS$  stands for the JS divergence.

By minimizing  $\mathcal{L}_{\theta\omega}$ , AIF can assign more reasonable attention scores to different levels of incongruity representations with ambiguity guidance.

**Sarcasm Classifier.** For each sample in mini-batch, given the different levels of incongruity representations  $m^t$ ,  $m^c$  and  $m^v$ , the attention weights  $\omega = \{\omega_t, \omega_c, \omega_v\}$ . We can compute the final multi-level incongruity fusion representation as follows:

$$\tilde{x}_f = (\omega_t \times m^t) \oplus (\omega_c \times m^c) \oplus (\omega_v \times m^v) \quad (19)$$

where  $\oplus$  denoted the concatenation operation.

Then, we feed  $\tilde{x}_f$  into a softmax layer with a fully-connected network to perform the prediction:

$$\hat{y} = \text{softmax}(MLP(\tilde{x}_f)) \quad (20)$$

Thereafter, we take the cross-entropy loss function to calculate the loss:

$$\mathcal{L}_{CE} = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})) \quad (21)$$

where  $y$  refers the ground-truth label,  $\hat{y}$  is the prediction label.

### 3.4 Training Objective

The overall loss function for AMIF is as follows:

$$\mathcal{L}_{All} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{\theta\omega} + \lambda_{IFCL} \mathcal{L}_{IFCL} \quad (22)$$

where  $\lambda$  and  $\lambda_{IFCL}$  control the ratio of  $\mathcal{L}_{\theta\omega}$  and  $\mathcal{L}_{IFCL}$ , respectively.

Model	Acc(%)	Pre(%)	Rec(%)	F1(%)
<b>Image-Only</b>				
Image	64.76	54.41	70.80	61.53
VIT	67.83	57.93	70.07	63.43
<b>Text-Only</b>				
SIARN	80.57	75.55	75.70	75.63
SMSD	80.90	76.46	75.18	75.82
BERT	83.85	78.72	82.27	80.22
<b>Multi-Modal</b>				
HFM	83.44	76.57	84.15	80.18
D&R Net	84.02	77.97	83.42	80.60
InCrossMGs	86.10	81.38	84.36	82.84
HKE	87.36	81.84	86.48	84.09
CMGCN	87.55	83.63	84.69	84.16
GAAN	87.42	82.91	86.62	84.72
MCEF	87.80	84.10	85.50	84.80
SAHFN	87.22	82.71	87.33	84.95
SEF	88.45	85.35	86.58	85.96
DMSD-CL	88.95	84.89	87.90	86.37
DGP	87.21	<b>87.10</b>	86.48	86.75
<b>AMIF(Ours)</b>	<b>90.10</b>	86.55	<b>89.68</b>	<b>88.09</b>

Table 1: Comparison results for sarcasm detection.

## 4 Experiments

### 4.1 Experiment Settings

**Dataset.** We assess our model by conducting experiments on a publicly available multi-modal sarcasm detection benchmark dataset constructed by Cai et al. (2019). The statistics of the dataset are shown in Appendix A.1.

**Baselines.** We compare the proposed AMIF with sixteen baselines shown in Table 1. More details on baselines are provided in Appendix A.2.

**Implementation.** The details of parameter implementations are listed in Appendix A.3.

### 4.2 Experimental Results and Analysis

We evaluate the effectiveness of our proposed framework through comparative analyses with baseline models, as presented in Table 1. Additionally, we derive the following observations: 1) It is evident that both the text and image play crucial roles in sarcasm detection. Therefore, it is imperative to fully excavate the incongruity information at the image-level and text-level. Our AMIF successfully addresses this problem by designing the UGIE module. Multi-modal models consistently outperform text-only and image-only models in terms of the performance, which suggests that simultaneously exploiting the textual and visual contents to capture complex cross-modal-level incongruity information can improve the performance of multi-

Model	Acc(%)	F1(%)	Model	Acc(%)	F1(%)
<b>AMIF</b>	<b>90.10</b>	<b>88.09</b>	w/o IFCL	89.42	87.01
w/o UGIE	89.42	87.18	w/o AG	89.21	87.28
w/o CIGR	89.16	86.93	w/o AIF	88.53	86.29

Table 2: Experimental results of ablation study.

modal sarcasm detection. The AMIF model utilizes the CIGR module to achieve this objective. 2) Moreover, AMIF consistently outperforms image-only, text-only and multi-modal models on all evaluation metrics, which indicates that AMIF significantly boosts the performance of multi-modal sarcasm detection compared to existing works. Specifically, compared with the classical methods in recent years that do not consider the distribution gap between unimodal sarcastic information, AMIF is 4.00%, 3.93%, 3.37%, 3.29%, 2.13% and 1.34% higher than HKE, CMGCN, GAAN, MCEF, SEF and DGP on F1-score. This indicates that AMIF achieves more advanced performance by capturing multi-level incongruity information and introducing cross-modal ambiguity in AIF module to measure the incongruity information gap. Compared with the recent LLMs-based model DMSD-CL, AMIF achieves considerable improvements in all metrics, outperforming *Acc*, *Pre*, *Rec* and *F1-score* by 1.15%, 1.66%, 1.78% and 1.72%, respectively. Furthermore, AMIF improves 2.89% on accuracy and 1.34% on F1-score over the latest multi-modal state-of-the-art model DGP. The above experimental results validate the effectiveness and superiority of our proposed method.

### 4.3 Ablation Study

To further investigate the effectiveness of each component in AMIF, we conduct a series of ablation studies: **1) w/o UGIE**: we remove the unimodal gating incongruity extracting module at the text-level and image-level; **2) w/o CIGR**: we remove the cross-modal incongruity graph reasoning module at the cross-modal-level; **3) w/o IFCL**: we remove the inter-modal fine-grained contrastive learning that assist cross-modal local incongruity alignment; **4) w/o AG**: we remove the ambiguity guidance module; **5) w/o AIF**: we remove the entire ambiguity-based incongruity fusion module and directly concatenate the three incongruity representations for the final classification.

Table 2 shows the results of ablation study. It is evident that the performance after removing any components is worse than the original AMIF, which

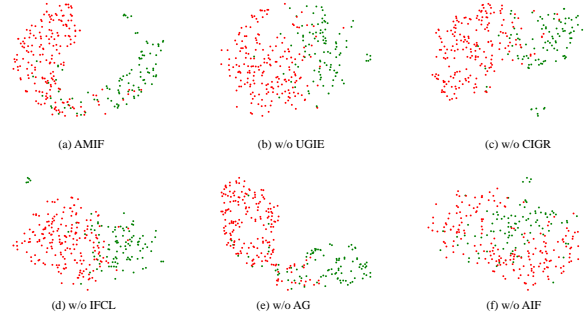


Figure 3: T-SNE visualizations of the sarcastic feature vectors before classification that are learned by AMIF and its five variants.

demonstrates the effectiveness of each component. And detailed analyses are as follows: 1) Both the AMIF w/o UGIE and AMIF w/o CIGR exhibit significantly lower performance, demonstrating that simultaneously and efficiently mining the hidden incongruity information at the image-level, text-level and cross-modal-level is necessary for multi-modal sarcasm detection. 2) The performance of AMIF w/o IFCL decreases obviously, which verifies that IFCL is essential for mining complex and deep-seated cross-modal sarcastic features by comparing fine-grained inter-modal fragments. 3) AMIF w/o AG exhibits a 0.81% decrease in F1-score and a 0.89% decrease in accuracy compared to the full AMIF, suggesting that accounting for the inherent cross-modal ambiguity for multi-modal sarcasm detection task can be advantageous for improving performance. 4) Notably, the performance of AMIF w/o AIF experiences a further decline of nearly 1% on F1 and 0.68% on Acc compared to AMIF w/o AG, which validates the importance of adaptively fusing the different levels of incongruity.

### 4.4 Visualization

We further analyze the proposed model using the t-SNE (Van der Maaten and Hinton, 2008) algorithm to map the features before classification into a 2-dimensional Euclid space and visualize the feature vectors in Figure 3. These feature vectors are learned by AMIF and its five variants on the test dataset of Twitter. It is evident that the points corresponding to different labels in AMIF have a more distinct boundary than its all variants, demonstrating that the captured sarcastic features in AMIF are more discriminative. As shown Figure 3(d), the learned sarcastic features in AMIF w/o IFCL are easily misclassified, which reveals that

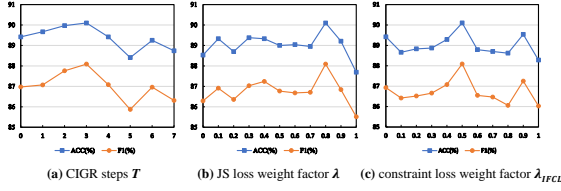


Figure 4: The influence of hyper-parameters.

IFCL can learn more accurate cross-modal correlations and deeply distinguish sarcasm/non-sarcasm posts. Comparing Figure 3(a), Figure 3(b) and Figure 3(c), we can see that simultaneously mining the unimodal and cross-modal hidden incongruity information can obtain multi-level sarcastic features from different modalities, thus effectively generating distinguished feature representations. Additionally, comparing Figure 3(a), Figure 3(e) and Figure 3(f), we can observe that considering the cross-modal ambiguity and adaptively fusing the different levels of incongruity information can significantly improve the representation ability of the final features.

#### 4.5 Parametric Analysis

To explore the influence of different hyper-parameters on the final prediction, we conduct extensive experimental studies for the number of CIGR steps  $T$ , JS loss weight factor  $\lambda$  and constraint loss weight factor  $\lambda_{IFCL}$ . Figure 4 records the metrics ACC and F1-score for different  $T$ ,  $\lambda$  and  $\lambda_{IFCL}$ . Specifically, as depicted in Figure 4(a), performing 3 step of cross-modal incongruity graph reasoning yields optimal performance, and the model exhibits poor performance when the number of steps exceeds 3. We speculate the reason may be that an excess of reasoning steps leads to overfitting of local and global cross-modal incongruity information. Moreover, as illustrated in Figure 4(b), the highest F1-score and ACC are achieved when  $\lambda$  is set to 0.8, AG and MAM can play a maximum role. When  $\lambda$  exceeds 0.8, the model’s performance steadily diminishes as  $\lambda$  increases. we speculate that the reason is assigning excessive weight to  $\mathcal{L}_{\theta\omega}$  introduces unexpected noise to the fusion module, disrupting the final prediction. Finally, as shown in Figure 4(c), the model achieves optimal performance when  $\lambda_{IFCL}$  is 0.5, whereas deviation from this value results in performance degradation. The reason behind this may be that excessively large or small values don’t enhance the accuracy of local incongruity computation, thereby affecting the sub-




	Text-Level	Image-Level	Cross-modal-Level
Image			
Text	(a) she must not realize she looks like the bigger idiot # funny	(b) this is reality ... lol # reality # poorpeoples # richpeople # society # blogsbar	(c) i think i may be the greatest fisherman who has ever lived . just look at the size of my largemouth bass .
$\omega_v$	0.195	0.707	0.361
$\omega_c$	0.323	0.264	0.752
$\omega_t$	0.786	0.359	0.257
HKE	Non-sarcasm(×)	Non-sarcasm(×)	Sarcasm(✓)
SEF	Sarcasm(✓)	Sarcasm(✓)	Non-sarcasm(×)
AMIF(ours)	Sarcasm(✓)	Sarcasm(✓)	Sarcasm(✓)

Figure 5: Different level sarcastic samples of case study on HKE, SEF and AMIF.

sequent cross-modal incongruity graph reasoning.

#### 4.6 Case Study

To qualitatively analyze the advantage of our model, we visualize the attention weights in AIF module and present the predictions of open-sourced methods HKE, SEF and AMIF on three representative examples at different levels in Figure 5. Specifically, HKE achieves accurate detection for cross-modal-level example by capturing cross-modal atomic-level and composition-level incongruities, but they ignore the unimodal sarcastic features, leading to false predictions for the text-level and image-level examples. Moreover, SEF uses multiple contrastive learning to enhance unimodal semantic features, thus correct predictions are made in the image-level and text-level example. For the cross-modal-level example, SEF makes wrong prediction because it only adapts multi-scale fusion strategy to align the image and text, which neglects both the cross-modal coarse- and fine-grained features contribute to multi-modal sarcasm detection and is insufficient to capture complex cross-modal incongruity information. Conversely, our AMIF exploits the UGIE module to capture subtle unimodal incongruity at the image-level and text-level, implements more comprehensive vector-based local and global cross-modal incongruity graph reasoning by designing the CIGR module simultaneously, and then adaptively assigns reasonable weights using a modality-wise attention mechanism with ambiguity guidance to ensure accurate predictions for the three examples.

### 5 Conclusion

In this paper, we propose a novel Ambiguity-aware Multi-level Incongruity Fusion Network (AMIF) for multi-modal sarcasm detection. Specifically,



we first suggest a multi-level incongruity learning module which can effectively and simultaneously mine both the latent unimodal sarcastic cues and complex cross-modal sarcastic cues. Moreover, we also design an ambiguity-based fusion module which can adaptively fuse the learned incongruity information from diverse levels. In this module, cross-modal ambiguity can guide the specific attention mechanism to assign reasonable weights for three different level incongruity information. Evaluation results demonstrate our AMIF significantly outperforms state-of-the-art methods on the benchmark dataset.

## Limitations

At this stage, we concentrate on two limitations of this work, aiming to inspire future potential research directions.

- For multi-modal sarcasm detection of social media posts, incorporating more external knowledge to enrich sentiment semantic information could improve the model's predictive performance. Our model ignores the importance of external knowledge for this task.
- Although our cross-modal incongruity graph reasoning module improves the model's performance, it primarily focuses on mining semantic-level information while neglecting syntactic dependencies between text nodes.

## Acknowledgement

This work was supported in part by Guangdong Basic and Applied Basic Research Foundation(2023A1515011370), National Natural Science Foundation of China (32371114), Characteristic Innovation Projects of Guangdong Colleges and Universities (2018KTSCX049), International Scientific and Technological Cooperation Project of Huangpu and Development Districts in Guangzhou (2022GH01, 2023GH05).

## References

Hedi Ben-Younes, Remi Cadene, Nicolas Thome, and Matthieu Cord. 2019. [Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection](#). In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8102–8109.

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*, pages 2506–2515.

- Alessandra Cervone, Evgeny A Stepanov, Fabio Celli, and Giuseppe Riccardi. 2017. [Irony detection: from the twittersphere to the news space](#). In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, volume 2006.
- Yifan Chen, Kuntao Li, Weixing Mai, Qiaofeng Wu, Yun Xue, and Fenghuan Li. 2024a. [D2r: Dual-branch dynamic routing network for multimodal sentiment detection](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3536–3547.
- Yifan Chen, Haoliang Xiong, Kuntao Li, Weixing Mai, Yun Xue, Qianhua Cai, and Fenghuan Li. 2024b. [Relevance-aware visual entity filter network for multimodal aspect-based sentiment analysis](#). *International Journal of Machine Learning and Cybernetics*, pages 1–14.
- Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. [Cross-modal ambiguity learning for multimodal fake news detection](#). In *Proceedings of the ACM Web Conference 2022*, pages 2897–2905.
- 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 *Proceedings of the International Conference on Learning Representations*.
- Hong Fang, Dahao Liang, and Weiyu Xiang. 2024. [Multi-modal sarcasm detection based on multi-channel enhanced fusion model](#). *Neurocomputing*, 578:127440.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. [Deep residual learning for image recognition](#). In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Mengzhao Jia, Can Xie, and Liqiang Jing. 2024. [Debiasing multimodal sarcasm detection with contrastive learning](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18354–18362.
- Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the naacL-HLT*, volume 1, page 2.
- Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. [Mvae: Multimodal variational autoencoder for fake news detection](#). In *Proceedings of the world wide web conference*, pages 2915–2921.

- Akshi Kumar and Geetanjali Garg. 2019. [Sarc-m: sarcasm detection in typo-graphic memes](#). In *Proceedings of the ICAESMT-2019, Uttarakhand University, Dehradun, India*.
- 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*, pages 4707–4715.
- 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)*, pages 1767–1777.
- Hao Liu, Runguo Wei, Geng Tu, Jiali Lin, Cheng Liu, and Dazhi Jiang. 2024. [Sarcasm driven by sentiment: A sentiment-aware hierarchical fusion network for multimodal sarcasm detection](#). *Information Fusion*, 108:102353.
- Hui Liu, Wenya Wang, and Haoliang Li. 2022. [Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement](#). In *Proceedings of the 2022 Conference on EMNLP*, pages 4995–5006.
- Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. 2020. [Cross-modality person re-identification with shared-specific feature transfer](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13379–13389.
- Huiying Ma, Dongxiao He, Xiaobao Wang, Di Jin, Meng Ge, and Longbiao Wang. 2024. [Multi-modal sarcasm detection based on dual generative processes](#). In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 2279–2287. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- Weixing Mai, Zhengxuan Zhang, Yifan Chen, Kuntao Li, and Yun Xue. 2024. [Geda: Improving training data with large language models for aspect sentiment triplet extraction](#). *Knowledge-Based Systems*, 301:112289.
- Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. [Attention bottlenecks for multimodal fusion](#). *Advances in Neural Information Processing Systems*, 34:14200–14213.
- Rolandos Alexandros Potamias, Georgios Siolas, and Andreas-Georgios Stafylopatis. 2020. [A transformer-based approach to irony and sarcasm detection](#). *Neural Computing and Applications*, 32(23):17309–17320.
- Liuqing Song, Zefang Zhao, Yuxiang Ma, Yuyang Liu, and Jun Li. 2023. [Global-aware attention network for multi-modal sarcasm detection](#). In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2409–2414. IEEE.
- Yi Tay, Anh Tuan Luu, Siu Cheung Hui, and Jian Su. 2018. [Reasoning with sarcasm by reading in-between](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1010–1020.
- David Tomás, Reynier Ortega-Bueno, Guobiao Zhang, Paolo Rosso, and Rossano Schifanella. 2023. [Transformer-based models for multimodal irony detection](#). *Journal of Ambient Intelligence and Humanized Computing*, 14(6):7399–7410.
- Laurens Van der Maaten and Geoffrey Hinton. 2008. [Visualizing data using t-sne](#). *Journal of machine learning research*, 9(11).
- Shuohang Wang, Sheng Zhang, Yelong Shen, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, and Jing Jiang. 2019. [Unsupervised deep structured semantic models for commonsense reasoning](#). In *Proceedings of the NAACL-HLT*, pages 882–891.
- Changsong Wen, Guoli Jia, and Jufeng Yang. 2023. [Dip: Dual incongruity perceiving network for sarcasm detection](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2540–2550.
- Qiaofeng Wu, Wenlong Fang, Weiyu Zhong, Fenghuan Li, Yun Xue, and Bo Chen. 2025. [Dual-level adaptive incongruity-enhanced model for multimodal sarcasm detection](#). *Neurocomputing*, 612:128689.
- Yang Wu, Zijie Lin, Yanyan Zhao, Bing Qin, and Li-Nan Zhu. 2021. [A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis](#). In *Findings of the ACL-IJCNLP 2021*, pages 4730–4738.
- Tao Xiong, Peiran Zhang, Hongbo Zhu, and Yihui Yang. 2019. [Sarcasm detection with self-matching networks and low-rank bilinear pooling](#). In *Proceedings of the world wide web conference*, pages 2115–2124.
- 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*, pages 3777–3786.
- Hu Zhang, Keke Zu, Jian Lu, Yuru Zou, and Deyu Meng. 2022. [Epsanet: An efficient pyramid squeeze attention block on convolutional neural network](#). In *Proceedings of the Asian Conference on Computer Vision*, pages 1161–1177.
- Zhengxuan Zhang, Jianying Chen, Xuejie Liu, Weixing Mai, and Qianhua Cai. 2024a. [‘what’ and ‘where’ both matter: dual cross-modal graph convolutional networks for multimodal named entity recognition](#). *International Journal of Machine Learning and Cybernetics*, 15(6):2399–2409.

Zhengxuan Zhang, Weixing Mai, Haoliang Xiong, Chuhan Wu, and Yun Xue. 2023. [A token-wise graph-based framework for multimodal named entity recognition](#). In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2153–2158. IEEE.

Zhengxuan Zhang, Yin Wu, Yuyu Luo, and Nan Tang. 2024b. [Mar: Matching-augmented reasoning for enhancing visual-based entity question answering](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1520–1530.

Zhicai Zhao, Na Lv, Runquan Xiao, Qiang Liu, and Shanben Chen. 2023. [Recognition of penetration states based on arc sound of interest using vgg-se network during pulsed gtau process](#). *Journal of Manufacturing Processes*, 87:81–96.

Weiyu Zhong, Zhengxuan Zhang, Qiaofeng Wu, Yun Xue, and Qianhua Cai. 2024. [A semantic enhancement framework for multimodal sarcasm detection](#). *Mathematics*, 12(2):317.

## A Appendix

### A.1 Dataset

We assess our model by conducting experiments on a publicly available multi-modal sarcasm detection benchmark dataset constructed by [Cai et al. \(2019\)](#). This dataset comprises English tweets expressing sarcasm as Positive examples and those expressing non-sarcasm as Negative examples. Each example in the dataset consists of a text and an associated image. The statistical information of the dataset is shown in Table 3.

### A.2 Baseline Models

We compare the proposed AMIF with eleven baselines, categorized into three groups: 1) Image-Only models: **Image** and **ViT**. 2) Text-Only models: **SIARN**, **SMSD** and **BERT**. 3) Multi-Modal models: **HFM**, **D&R Net**, **InCrossMGs**, **HKE**, **CMGCN**, **GAAN**, **MCEF**, **SAHFN**, **SEF**, **DMSD-CL** and **DGP**.

#### A.2.1 Image-Only models

1. Image ([Cai et al., 2019](#)): a method that employs ResNet ([He et al., 2016](#)) to train a sarcasm classifier.
2. ViT ([Dosovitskiy et al., 2020](#)): a baseline that utilizes the [CLS] token representation of the pre-trained ViT for sarcasm detection.

Label	Training	Development	Testing
Positive	8642	959	959
Negative	11174	1451	1450
All	19816	2410	2409

Table 3: Statistics of the dataset.

#### A.2.2 Text-Only models

1. SIARN ([Tay et al., 2018](#)): a model that adopts inner-attention for textual sarcasm detection.
2. SMSD ([Xiong et al., 2019](#)): a network that explores a self-matching network to capture textual incongruity information
3. BERT ([Kenton and Toutanova, 2019](#)): a baseline that utilizes the vanilla pre-trained uncased BERT-base, taking [CLS] token as the input.

#### A.2.3 Text and Image models

1. HFM ([Cai et al., 2019](#)): a hierarchical multi-modal features fusion model for multi-modal sarcasm detection.
2. D&R Net ([Xu et al., 2020](#)): a Decomposition and Relation Network modeling both cross-modality contrast and semantic association.
3. InCrossMGs ([Liang et al., 2021](#)): a graph-based model for leveraging the sarcastic relations from intra and inter-modal perspectives.
4. HKE ([Liu et al., 2022](#)): a baseline that combines external knowledge and considers atomic and composition-level congruities.
5. CMGCN ([Liang et al., 2022](#)): a method that constructs a cross-modal graph for each instance to explicitly draw the sarcastic relations between textual and visual modalities.
6. GAAN ([Song et al., 2023](#)): a attention-based cross-modal multi-granularity alignment model.
7. MCEF ([Fang et al., 2024](#)): a multi-channel enhanced fusion model to maximize the information extraction between different modalities.
8. SAHFN ([Liu et al., 2024](#)): a hierarchical fusion model that uses attribute-object matching method to integrate sentiment information.

Hyper-parameters	Value
learning rate	1e-5
warm up proportion	0.2
number of training epochs	20
training batch size	32
maximum text length	64
number of CIGR steps $T$	3
JS loss weight factor $\lambda$	0.8
constraint loss weight factor $\lambda_{IFCL}$	0.5
constraint similarity margin $\gamma$	0.1
inversed temperature factor $\beta$	100

Table 4: Hyper-parameter setting of our AMIF model.

9. SEF (Zhong et al., 2024): a method modeling textual and visual information at the multi-scale and multi-span token level using contrastive learning.
10. DMSD-CL (Jia et al., 2024): a baseline using large language models to define the task of out-of-distribution aiming to evaluate models’ generalizability. It proposes a novel debiasing multimodal sarcasm detection framework with contrastive learning to mitigate the harmful effect of biased textual factors for robust out-of-distribution generalization.
11. DGP (Ma et al., 2024): a model based on dual generative processes to deeply explore emotional inconsistencies between modalities.

### A.3 Implementation Details

For a fair comparison, following the processing in Liang et al. (2021, 2022); Liu et al. (2022); Zhong et al. (2024); Ma et al. (2024), we utilize the pre-trained BERT-base-uncased model (Kenton and Toutanova, 2019) and ViT model (Dosovitskiy et al., 2020) as the text-encoder and image-encoder to embed each word and region. Following Liu et al. (2022); Fang et al. (2024); Zhong et al. (2024); Ma et al. (2024); Liu et al. (2024), we perform Accuracy, Precision, Recall, and F1-score to evaluate the model performance. The experimental results of our model are averaged over 10 runs using different random seeds to ensure statistical stability in the final reported results. The specific hyper-parameter settings are detailed in Table 4.