Enhancing Semantic Awareness by Sentimental Constraint With Automatic Outlier Masking for Multimodal Sarcasm Detection

Shaozu Yuan, Yiwei Wei, Hengyang Zhou, Qinfu Xu, Meng Chen, and Xiaodong He, *Fellow, IEEE*

***Abstract*—Multimodal sarcasm detection, aiming to uncover sarcastic sentiment behind multimodal data, has gained substantial attention in multimodal communities. Recent advancements in multimodal sarcasm detection (MSD) methods have primarily focused on modality alignment with pre-trained vision-language (V-L) model. However, text-image pairs often exhibit weak or even opposite semantic correlations in MSD tasks. Consequently, directly aligning these modalities can potentially result in feature shift and inter-class confusion, ultimately hindering the model's ability. To alleviate this issue, we propose the Enhancing Semantic Awareness Model (ESAM) for multimodal sarcasm detection. Specifically, we first devise a Modality-decoupled Framework (MDF) to separate the textual and visual features from the fused multimodal representation. This decoupling enables the parallel integration of the Sentimental Congruity Constraint (SCC) within both visual and textual latent spaces, thereby enhancing the semantic awareness of different modalities. Furthermore, given that certain outlier samples with ambiguous sentiments can mislead the training and weaken the performance of SCC, we further incorporate Automatic Outlier Masking. This mechanism automatically detects and masks the outliers, guiding the model to focus on more informative samples during training. Experimental results on two public MSD datasets validate the robustness and superiority of our proposed ESAM model.**

***Index Terms*—Multimodal sarcasm detection, sentimental congruity constraint, automatic outlier masking.**

I. Introduction

SARCASM detection, aiming to detect sarcastic sentiment within multimodal data, has gained substantial attention from both academia and industry. In past years, the relevant research primarily focused on textual modalities [1], [2]. With the development of social media platforms, there is an increasing need for people to express their emotions and opinions through multimodal data, including images and text. Thus, it is necessary to accurately detect the sarcastic sentiment behind the multimodal data. Compared to purely text-based sarcasm detection methods, multimodal sarcasm detection [3], [4], [5] is more challenging as it requires exploring sarcastic cues from different modalities.

To detect sarcastic sentiment in MSD, researchers [3], [4], [5], [6], [7] introduced various networks aimed at facilitating multimodal interaction between text and image. Recognizing the superior ability exhibited by pre-trained vision-language (V-L) model in tasks, such as visual captioning [8], [9], visual question answering [10], and multimodal sentiment detection [11], [12], [13], recent studies [14], [15] employ those pre-trained V-L models to align unimodal representations, enabling more exhibited understanding of sarcastic sentiment within multimodal contexts.

However, MSD poses a unique challenge: unlike aforementioned tasks [8], [10], [11] where modalities exhibit strong semantic consistency, text-image pairs in sarcastic contexts often exhibit weak or contradictory semantic relationships. Aligning such disparate modalities is challenging, sometimes leading to significant feature shifts where projected features deviate from their intended semantic distribution in the latent space, as depicted in Fig. 1. This shift features impairs the model's semantic awareness, causing inter-class confusion and undermining overall performance. Existing works [16][6] [7][17] utilize attention mechanisms or graph structures to align features from an atomic perspective, which are not insufficient to address the feature shift presented at the global level. Hence, this critical aspect has remained unresolved in MSD research.

To address this gap, we propose the Enhancing Semantic Awareness Model (ESAM) for multimodal sarcasm detection. ESAM's novelty lies in its three key components: the Modality-decoupled Framework (MDF), Sentimental Congruity Constraint (SCC), and Automatic Outlier Masking (AOM). In MDF, it first aligns the visual and textual features via unimodal encoders to capture initial semantic correlations and then decouples the textual and visual modalities with modality masking [18]. This can facilitate parallel processing at the decoupled feature level, which is more effective in mitigating feature shifts than fusing features. Subsequently, we introduce

Received 20 May 2024; revised 25 August 2024 and 13 November 2024; accepted 15 December 2024. Date of publication 17 February 2025; date of current version 27 August 2025. This work was supported in part by the Ke Campus Autonomous Region Basic Scientific Research Fund under Grant XQZX20240048 and in part by the Karamay City - Special Program for Innovative Talent under Grant XQZX20240069. The associate editor coordinating the review of this article and approving it for publication was Dr. Zhenzhen Hu. *(Corresponding author: Yiwei Wei.)*

Shaozu Yuan is with Meituan, Beijing 100102, China (e-mail: yuanshaozu@meituan.com).

Yiwei Wei and Hengyang Zhou are with the College of Petroleum Engineering, China University of Petroleum (Beijing) at Karamay, Karamay 834000, China (e-mail: 360976808@qq.com; hengyangzhou@outlook.com).

Qinfu Xu is with the Qingdao Institute of Software, College of Computer Science and Technology, China University of Petroleum (East China), Shandong 257099, China (e-mail: xqfupc@163.com).

Meng Chen and Xiaodong He are with JD AI Research, Beijing 100176, China (e-mail: chenmengdx@gmail.com; hexiaodong@jd.com).

Digital Object Identifier 10.1109/TMM.2025.3543074

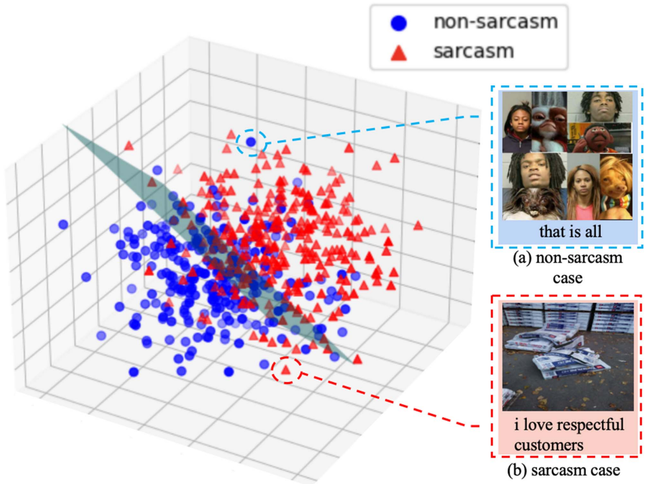

Fig. 1. The aligned multimodal feature suffers from semantic-level feature shift in the latent space, for the weak or contradictory correlations between modalities in MSD.

sentimental congruity constraints in both visual and textual latent spaces to enhance sentiment awareness, finally alleviating feature shifts. Nevertheless, MSD scenarios often contain sentimentally ambiguous outlier samples that will mislead the training and weaken the performance of SCC. To counter this, we propose the outlier detection algorithm [19] to automatically detect and mask these outliers during training, eliminating the interruption of these cases. Finally, we harness multi-view fusion [20] to yield the final prediction. Comprehensive experiments validate ESAM's robustness and superiority on two public Multimodal Sarcasm Detection datasets.

In summary, the main contributions of this paper are as follows:

- This paper is the first work to specifically tackle and mitigate the problem of feature shift that restricted the performance of existing models in the MSD task.
- We introduce three novel modules including a modality-decoupled framework, sentimental congruity constraint, and an automatic outlier masking, which can effectively alleviate feature shift in the latent semantic space for MSD.
- Experimental results on two public MSD datasets [5], [20] validate the robustness and superiority of our proposed ESAM model. We further verify its effectiveness through comprehensive analyses.

II. Related Work

A. Multimodal Sarcasm Detection

Multimodal sarcasm detection has become an increasingly attentive task, driven by the growing need to analyze multimodal content on social media platforms. Schifanella et al. [3] was the first to tackle this challenge, approaching it as a multimodal classification problem by combining manually engineered multimodal features. Subsequently, researchers [3], [4], [5] introduced various multimodal networks to explore the incongruity between text and image data. In addition, InCrossMGs [16], CMGCN [6], HKEmodel [7] and MILNet [21] have leveraged transformers [17] and graph neural networks to model intra-modality and inter-modality incongruities. Recent study [20], [22] has achieved progress by utilizing pre-trained vision-language (V-L) models to first extract unimodal representations and then align them in a latent space. However, a critical challenge remains unresolved in multimodal sarcasm detection (MSD): directly aligning text and image modalities can lead to feature shift, due to weak or opposing semantic correlations.

B. Vision-Language Pretraining

In recent years, sustained progress has been achieved in the field of vision-language pretraining. Early efforts in this domain mainly focused on tasks like image classification and caption generation [23], [24]. More recently, researchers have proposed methods that jointly pretrain image and text representations, enabling more effective transfer learning across a diverse array of tasks. Notable advancements include VisualBERT [25], which integrates visual and textual inputs to craft a pre-trained model capable of zero-shot image classification. Similarly, ALIGN [26] introduces a dual-encoder framework for image-text alignment. Other relevant contributions include methods for multimodal learning, such as CLIP [27], which jointly optimizes image and text classification tasks, and UNITER [23], which introduces a multimodal decoder for caption generation. Inspired by the success of large language models (LLMs), the scaling of vision-language pretraining models (VLM) to tens and even hundreds of billions of parameters has shown consistent performance [8], [28], [29]. Despite these significant advancements, it remains a challenging task to apply these models to multimodal sarcasm detection (MSD) due to the weak correlations of different modalities in MSD scenarios.

III. Methodology

The architecture for our Enhancing Semantic Awareness Model is depicted in Fig. 2. Details of each constituent module are provided in the following sections.

A. Modality-Decoupled Framework

Modality-decoupled Framework comprises three modules: unimodal encoder, modality-aligned module, and modality-decoupled module.

Unimodal Encoder: Let $\{\boldsymbol{x}, \boldsymbol{y}\}$ indicate the text-image pair. For text encoder $\mathbb{T}$, we use the pre-trained BERT model [30] to encode the text $\boldsymbol{x}$ and obtain the textual representation $\boldsymbol{T}$:

$$\boldsymbol{T} = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n) = \mathbb{T}(\boldsymbol{x}) \in \mathbb{R}^{\{n, d^t\}} \tag{1}$$

where n stands for the sequence length of $\boldsymbol{x}$. For image encoder $\mathbb{V}$, we first use a pre-trained toolkit [31] to extract m regions, denoted as $\boldsymbol{R} = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_m\}$, then we leverage the VIT model [32] to generate visual $[\boldsymbol{cls}]$ token for each region. Finally, we concatenate all the $[\boldsymbol{cls}]$ tokens as final visual representation $\boldsymbol{I}$ for the given image:

$$\boldsymbol{I} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_m) = \mathbb{V}(\boldsymbol{y}) \in \mathbb{R}^{\{m, d^v\}} \tag{2}$$

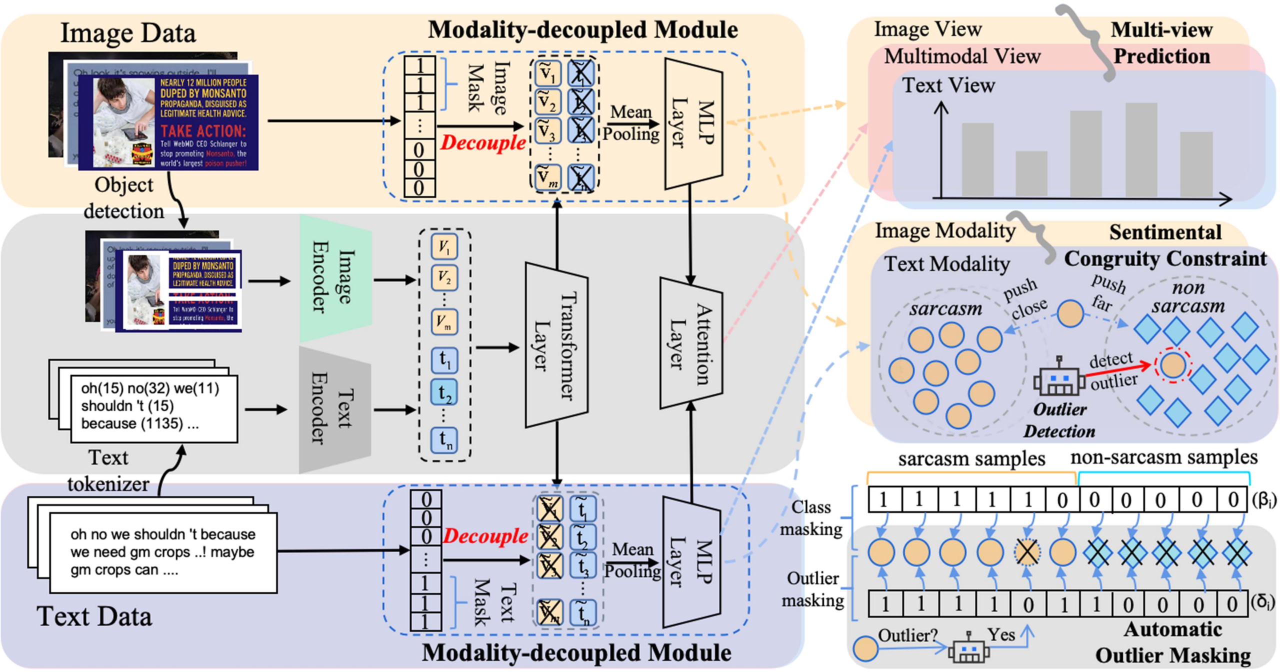

Fig. 2. The overall architecture of the proposed method ESAM. Outlier Detection indicates the proposed outlier detection algorithm in the automatic outlier masking module.

Modality-aligned Module: To align text and image features, we first employ a fully connected layer to map them into a common dimension d. Subsequently, we concatenate the textual representation $\boldsymbol{T}$ and visual representation $\boldsymbol{I}$ as $\boldsymbol{F} = Concat\{\boldsymbol{T}, \boldsymbol{I}\}$, where $\boldsymbol{F} \in \mathbb{R}^{\{n+m,d\}}$. We then leverage a Transformer-Encoder [17] to align multimodal features. For $l-th$ layer of the Transformer, the output can be computed as follows:

$$\tilde{\boldsymbol{F}_l} = \text{softmax}\left(\frac{(\boldsymbol{F}\boldsymbol{W_q})(\boldsymbol{F}\boldsymbol{W_k})^T}{\sqrt{d}}\right)(\boldsymbol{F}\boldsymbol{W_v}) \tag{3}$$

where $\boldsymbol{W_q} \in \mathbb{R}^{d\times d}$, $\boldsymbol{W_k} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{W_v} \in \mathbb{R}^{d\times d}$ are query, key, and value projection matrices, respectively. For simplicity, we omit the residual connection and layer normalization of each layer. Here, we denote the multimodal aligned representations for the last layer as $\tilde{\boldsymbol{F}} = (\tilde{\boldsymbol{v}}_1, \ldots, \tilde{\boldsymbol{v}}_m, \tilde{\boldsymbol{t}}_1, \ldots, \tilde{\boldsymbol{t}}_n)$.

Modality-decoupled Module: To decouple the text feature and image feature from the multimodal aligned representation $\tilde{\boldsymbol{F}}$, we generate mask vectors [18] based on the respective sequence lengths of the image and text data. These masks are then applied to their corresponding modality features, effectively decoupling them from the fused representation. Here we emphasize the decoupling of image and text features, but they are not entirely isolated from each other. This module ensures the relationship between text and images can be captured, which benefits the effectiveness of feature embedding.

Concretely, assuming n and m denote the sequence lengths for the text and image representations, respectively. We construct the modality mask by a mask vector of size $m+n$ binary values (0 or 1) according to modality. As such, the text mask is $\boldsymbol{M_v} = \left[\underbrace{1, 1, \ldots, 1}_{m}, \underbrace{0, 0, \ldots, 0}_{n}\right]$, while the image mask is $\boldsymbol{M_t} = \left[\underbrace{0, 0, \ldots, 0}_{m}, \underbrace{1, 1, \ldots, 1}_{n}\right]$.

Given the multimodal aligned representation $\tilde{\boldsymbol{F}} \in \mathbb{R}^{\{n+m,d\}}$, along with the image mask $\boldsymbol{M_v} \in \mathbb{R}^{\{n+m\}}$ and text mask $\boldsymbol{M_t} \in \mathbb{R}^{\{n+m\}}$, we extract the decoupled text features $\tilde{\boldsymbol{T}} = (\tilde{\boldsymbol{t}}_1, \ldots, \tilde{\boldsymbol{t}}_n)$ and image features $\tilde{\boldsymbol{I}} = (\tilde{\boldsymbol{v}}_1, \ldots, \tilde{\boldsymbol{v}}_m)$ by selecting features from $\tilde{\boldsymbol{F}}$ at the indices where the $\boldsymbol{M_t}$ and $\boldsymbol{M_v}$ have a value of 1, respectively. Moreover, considering the decoupled features $\tilde{\boldsymbol{T}}$ and $\tilde{\boldsymbol{I}}$ are not directly suitable for feature constraint and sentiment classification, we employ the mean-pooling operation followed by a multi-layer perceptron (MLP) to perform dimensionality reduction, thereby obtaining the final decoupled features $\bar{\boldsymbol{t}}$ and $\bar{\boldsymbol{v}}$.

B. Sentimental Congruity Constraint

In this section, we introduce sentimental congruity constraints to effectively mitigate the feature shift. In particular, feature shift manifests as the projected features deviate from their corresponding semantic distribution in the latent space. Our key insight is to enforce congruity among instances sharing the same sentimental label, thereby reducing inter-class dispersions.

Here, we take the textual modality $\bar{\boldsymbol{t}}$ for illustration, and the same principles apply to the visual modality. For a given batch of size b, we define $g(\bar{\boldsymbol{t}}_i) = \{\bar{\boldsymbol{t}}_k \mid \bar{\boldsymbol{t}}_k \in \{\bar{\boldsymbol{t}}_1, \bar{\boldsymbol{t}}_2, \ldots, \bar{\boldsymbol{t}}_b\}, \bar{\boldsymbol{t}}_k \neq \bar{\boldsymbol{t}}_i\}$ as a set of instance excluding $\bar{\boldsymbol{t}}_i$. We introduce loss function $\boldsymbol{L}_{\boldsymbol{scc}}^{ij}$ aimed at minimizing the distance between pairs of sample $\bar{\boldsymbol{t}}_i$ and

Algorithm 1: Outlier Detection Algorithm.

```
Result: Anomaly scores for each instance in the dataset
Input : dataset $S$, number of trees $N$, subsample size $\psi$
Output: Anomaly scores $\Omega$
Initialize an empty set $F$
for $i \leftarrow 1$ to $N$ do
    Randomly select $\psi$ instances from $S$ to form $S'$
    Build a tree $\boldsymbol{T}_i$ using ConstructTree($S'$) described in algorithm 2
    Add $\boldsymbol{T}_i$ to $F$
end
Initialize an empty list of scores $\Omega$
for each instance $s$ in $S$ do
    Initialize path length $h(s) \leftarrow 0$
    for each tree $\boldsymbol{T}$ in $F$ do
        Compute the path length $h_t(s)$ of $s$ in $T$ using ComputePathLength($s, T$) described in algorithm 3
        Update $h(s) \leftarrow h(s) + h_t(s)$
    end
    Normalize $h(s)$
    Add the anomaly score to the list $\Omega$
end
return Anomaly scores $\Omega$;
```

$\bar{\boldsymbol{t}}_j$ within the same sentimental class. This loss is formulated as:

$$\boldsymbol{L}_{scc}^{ij} = -\boldsymbol{\beta}_{ij} log \frac{e^{-d(\bar{\boldsymbol{t}}_{\boldsymbol{i}},\bar{\boldsymbol{t}}_{\boldsymbol{j}})}/\tau}{\sum_{k\in g(\bar{\boldsymbol{t}}_{\boldsymbol{i}})} e^{-d(\bar{\boldsymbol{t}}_{\boldsymbol{i}},\bar{\boldsymbol{t}}_{\boldsymbol{k}})/\tau}} \tag{4}$$

where $d(.,.)$ is the Euclidean distance, τ indicates learning temperature, and $\boldsymbol{\beta}_{ij}$ denotes the class indicator to ensure that the SCC loss only applies to samples of the same category. Specifically, if $\bar{\boldsymbol{t}}_i$ and $\bar{\boldsymbol{t}}_j$ share the same sentimental label, $\boldsymbol{\beta}_{ij}$ is set as 1, encouraging their closer proximity in the latent space. Conversely, if their labels differ, $\boldsymbol{\beta}_{ij}$ will be set as 0 to avoid unnecessary clustering. Mathematically, $\boldsymbol{\beta}_{ij}$ can be defined as follows:

$$\begin{cases} \boldsymbol{\beta}_{ij} = 1, \text{ if } \boldsymbol{S}_i == \boldsymbol{S}_j \\ \boldsymbol{\beta}_{ij} = 0, \text{ else} \end{cases} \tag{5}$$

where $\boldsymbol{S}_i$ and $\boldsymbol{S}_j$ represents the sentimental labels for $\bar{\boldsymbol{t}}_i$ and $\bar{\boldsymbol{t}}_j$ respectively. Thus, for a given sample $\bar{\boldsymbol{t}}_i$, we can formalize the overall loss as $L_{scc}^i = \sum_{j\in g(i)} L_{scc}^{ij}$. This formulation encourages samples from the same class as $\bar{\boldsymbol{t}}_i$ to cluster together. To efficiently implement this loss, we introduce class masking vector $\boldsymbol{\beta}_i$ defined as $\left[\underbrace{1,1,\ldots,1}_{K}, \underbrace{0,0,\ldots,0}_{b-1-K}\right]$, where K denotes the number of samples belonging to the same class and b denotes the batch size. This vector allows us to apply the SCC loss selectively within each batch. Finally, the total loss of SCC can be calculated as $L_{scc} = \sum_i L_{scc}^i/2$.

Algorithm 2: Construct Tree Algorithm.

```
ConstructTree($S'$) {
if $|S'| \leq 1$ then
    return Leaf Node
Randomly select a feature $f$ from $S'$
Determine a split point $p$ based on $f$
Split $S'$ into two subsets: $S'_L$ (samples with $f < \rho$) and $S'_R$ (samples with $f \geq \rho$)
Create a node with split point $\rho$ and feature $f$
Set the left child as ConstructTree($S'_L$)
Set the right child as ConstructTree($S'_R$)
return the constructed node
}
```

Algorithm 3: Compute Path Length Algorithm.

```
ComputePathLength($s, T$) {
Initialize path length $h \leftarrow 0$
Set the current node as the root of $T$
while current node is not a leaf do
    Increment $h$ by 1
    if $s$ falls in the left subtree based on the split feature and point then
        Move to the left child node
    else
        Move to the right child node
return $h$
}
```

C. Automatic Outlier Masking

In multimodal sarcasm detection scenarios, the existence of sentimentally ambiguous samples (outliers) poses a new challenge, which will mislead the training and weaken the performance of SCC. To mitigate the impact of these outliers, we design an automatic outlier masking $\boldsymbol{\delta}_i$ to guide the training. Generally, we first set $\boldsymbol{\delta}_i$ as a vector of size $b-1$ with all elements initialized to 1 and then dynamically adjust the values of $\boldsymbol{\delta}_i = \left[\underbrace{1,1,\ldots,1}_{b-1}\right]$ based on the position of the outliers. To detect outliers efficiently, we propose an outlier detection algorithm that can search for outliers in high-dimensional features. Specifically, to separately handle outliers in non-sarcasm and sarcasm cases in MSD, we first divide the case in a batch into two subsets: $\{S^+, S^-\}$. For each subset, we construct N random binary trees $\{\boldsymbol{T}^i\}_{i=1}^N$, where N is determined by the algorithm to enhance model robustness. The details of the outlier detection algorithm are shown in Algorithms 1, 2, and 3.

The tree construction process in Algorithm 2 involves recursively partitioning the data through a random splitting method. First, a feature $\boldsymbol{f}$ is randomly selected from the dataset S', and a split point ρ is determined within the range of that feature. Samples with values less than ρ are assigned to the left subtree, while samples with values greater than or equal to ρ are assigned to the right subtree. This process is recursively repeated until each leaf node contains only one sample, resulting in a binary tree structure. Based on the constructed binary tree, the path length of a data point can be calculated using algorithm 3 to determine whether it is an outlier. Specifically, the path length h refers to the distance from the root node to the leaf node where the data point is located. A shorter path length indicates that the data point is more easily isolated and thus more likely to be an outlier. The anomaly score is computed based on this path length and typically ranges between [0, 1]. A score closer to 1 suggests a higher likelihood of the data point being an outlier, while a score closer to 0 indicates that the data point is more likely to be normal.

During the inference stage, outliers are identified according to the anomaly scores Ω in the outlier detection algorithm, with the top γ of samples exhibiting the highest scores being detected as outliers. Suppose the j^{th} case is outlier, we update $\boldsymbol{\delta}_i$ by replacing the corresponding element with 0, as $\left[\underbrace{1,1,\ldots,1}_{j-1},\underbrace{0}_{1},\underbrace{1,1,\ldots,1}_{b-1-j}\right]$. This updated $\boldsymbol{\delta}_i$ is then used to modify the SCC loss as:

$$\tilde{L}_{scc} = \sum_i \delta_i L^i_{scc} \tag{6}$$

D. Multi-View Prediction

Inspired by [33], our final results are obtained from three views. For the decoupled textual feature $\bar{\boldsymbol{t}}$ and visual features $\bar{\boldsymbol{v}}$, we apply simple attention along the last dimension to obtain the multimodal feature m:

$$\boldsymbol{p}_t, \boldsymbol{p}_v = softmax(\boldsymbol{W_p}(\bar{\boldsymbol{t}}, \bar{\boldsymbol{v}})) \tag{7}$$

$$\boldsymbol{m} = \boldsymbol{p}_t \bar{\boldsymbol{t}} + \boldsymbol{p}_v \bar{\boldsymbol{v}} \tag{8}$$

where $\boldsymbol{W_p}$ is the learned parameters, and $\boldsymbol{p}_t$, $\boldsymbol{p}_v$ are attention weights. Then, we generate the predicting distributions $\boldsymbol{y}^{\{v,t,m\}}$ text view $\boldsymbol{y}^t$, image view $\boldsymbol{y}^v$, and multimodal view $\boldsymbol{y}^m$ by employing the softmax function along the last dimension:

$$\boldsymbol{y}^{\{v,t,m\}} = softmax(\boldsymbol{W}^{\{v,t,m\}}(\bar{\boldsymbol{v}}, \bar{\boldsymbol{t}}, \boldsymbol{m})) \tag{9}$$

where $\boldsymbol{W}^{\{v,t,m\}}$ denote the learned parameters. Finally, we generate the prediction $\boldsymbol{y}_o$ as follows:

$$\boldsymbol{y}^o = \boldsymbol{y}^t + \boldsymbol{y}^v + \boldsymbol{y}^m \tag{10}$$

Additionally, for the single multimodal view, the prediction $\boldsymbol{y}_o$ can be generated as follows:

$$\boldsymbol{y}^o = \boldsymbol{y}^m \tag{11}$$

where $\boldsymbol{y}^m$ denotes the multimodal output.

E. Training Loss

During training, we minimize the cross-entropy loss for image v, text t, and multimodal m views:

$$L_{msd} = \sum_{i \in \{t,v,m\}} (\boldsymbol{y}^i log(\hat{\boldsymbol{y}}^i)) + (1 - \boldsymbol{y}^i) log(1 - \hat{\boldsymbol{y}}^i) \tag{12}$$

where $\boldsymbol{y}^i$ and $\hat{\boldsymbol{y}}^i$ denotes the label and prediction respectively. Thus, the total loss L_t includes two parts: cross-entropy loss L_{msd} and the constraint losses $\{L^t_{scc}, L^v_{scc}\}$ for textual and visual modality:

$$L_t = L_{msd} + \lambda(\tilde{L}^t_{scc} + \tilde{L}^v_{scc}) \tag{13}$$

where λ is a balance coefficient.

TABLE I
STATISTICS OF THE EXPERIMENTAL DATA

Dataset	Label	Train	Val	Test
MMSD	Sarcasm	8642	959	959
	N-sarcasm	11174	1451	1450
	All	19816	2410	2409
MMSD2.0	Sarcasm	9572	1042	1037
	N-sarcasm	10240	1368	1372
	All	19816	2410	2409

IV. Experimental Setup

In this section, we introduce the experimental settings and the baseline models.

A. Datasets

We demonstrate the effectiveness of our method on two public datasets which are MMSD [5] and MMSD2.0 [20] Both datasets collect data from Twitter, and each text-image pair is labeled by a single sentiment. For a fair comparison, we follow the experimental settings of prior work [5], which divides the data into training, validation, and test sets in a ratio of 80%:10%:10%. To better understand the dataset details, we provide the statistics for MMSD and MMSD2.0 datasets in Table I. It shows that MMSD and MMSD2.0 are binary multimodal classification tasks and have a relatively balanced category distribution. The difference between those two datasets is that MMSD2.0 conducts data optimization to address the issues in MMSD by removing the spurious cues and fixing unreasonable annotation, for multimodal sarcasm detection.

B. Implementation Details

In our experiments, we utilized the pre-trained BERT model to extract textual features, while the extraction of visual features was accomplished via a pre-trained Vision Transformer. A 6-layer transformer was employed for modality alignment. During training, the model is trained for 10 epochs, with a batch size of 64. And we use AdamW [34] as an optimizer with the learning rate of $2e-5$. Regarding the hyper-parameters, we set λ in (13) to 1, the learning temperature τ in SCC loss to 20 and 15 for the MMSD and MMSD2.0 datasets, and the threshold to 0.2 and 0.25 in AOM. To evaluate the performance of the model, we follow the previous work [5] and adopt Accuracy, F1 score, Precision, and Recall as the evaluation metrics. All experiments are conducted with an NVIDIA 4090 GPU.

TABLE II
EXPERIMENTAL RESULTS OF DIFFERENT TYPES OF BASELINE MODELS ON MMSD AND MMSD2.0 DATASETS

Modality	Model	MMSD				MMSD2.0			
		Acc(%)	Pre(%)	Recall(%)	F1(%)	Acc(%)	Pre(%)	Recall(%)	F1(%)
Text	TextCNN	80.03	74.29	76.39	75.32	71.61	64.62	75.22	69.52
	Bi-LSTM	81.9	76.66	78.42	77.53	72.48	68.02	68.08	68.05
	SMSD	80.9	76.46	75.18	75.82	73.56	68.45	71.55	69.97
	BERT	83.85	78.72	82.27	80.22	75.88	71.07	74.69	72.83
	GPT-3.5†	59.13	-	-	55.41	53.05	-	-	52.92
Image	ResNet	64.76	54.41	70.8	61.53	65.5	61.17	54.39	57.58
	ViT	67.83	57.93	70.07	63.4	72.02	65.26	74.83	69.72
Multimodal	HFM	83.44	76.57	84.15	80.18	70.57	64.84	69.05	66.88
	D&R Net	84.02	77.97	83.42	80.6	-	-	-	-
	Res-BERT	84.80	77.80	84.15	80.85	-	-	-	-
	Att-BERT	86.05	80.87	85.08	82.92	80.03	76.28	77.82	77.04
	InCrossMGs	86.1	81.38	84.36	82.84	-	-	-	-
	CMGCN	86.54	-	-	82.73	79.83	75.82	78.01	76.9
	HKEModel	87.36	81.84	86.48	84.09	76.5	73.48	71.07	72.25
	Multi-view CLIP	88.33	82.66	88.65	85.55	85.64	80.33	**88.24**	84.10
	MILNet	89.50	85.16	89.16	87.11	80.27	-	-	78.02
	DIP	89.59	**87.76**	86.58	87.17	80.59	-	-	78.23
	GPT-4.0†	66.66	-	-	63.19	57.87	-	-	55.52
	ESAM	**90.11**	86.87	**89.54**	**88.19**	**85.87**	**83.12**	86.05	**84.56**

† indicates the large pre-trained model. note that for the MMSD2.0 dataset, we directly use CLIP as text and image encoders for fine-tuning, without incorporating additional object detection methods to refine the image features.

C. Baselines

The baseline models are generally divided into three categories: text-based methods, image-based methods and multimodal-based methods.

Text-based models: For the text-based approaches, we adopt several well-known models, such as **TextCNN** [35], **Bi-LSTM** [36], and **BERT** [37]. Additionally, we also adopt the large language model (LLM) **GPT-3.5-turbo** [38] for this task, as it shows remarkable ability across various NLP tasks. Specifically, we first select a prompt "The sentiment of sentence is [sarcastic/not sarcastic], because", and then ask the GPT-3.5 model to perform a zero-shot prediction based on the prompt.

Image-based models: Given the inherent challenges in capturing ironic cues solely from images, the research community has only proposed a limited number of image-based baselines for this task. In our experiments, we utilize the feature representations from the pooling layer of **ResNet** [39] and [CLS] token from each image patch sequence produced by **ViT** [32] to generate predictions, respectively.

Multimodal based models: We consider diverse multimodal baselines for holistic comparison. Among these, **HFM** [5] proposed a hierarchical fusion model for multimodal sarcasm detection. **D&R Net** [40] proposes a decomposition and relation network to model cross-modality features. **Res-Att** [4] directly concatenated visual and textual features to capture ironic cues. Additionally, **Att-BERT** [4] proposed refined attention mechanisms to detect sarcasm. Furthermore, **InCrossMGs** [16] employed a heterogeneous graph structure to capture ironic features from different perspectives. **CMGCN** [6] constructed a cross-modal graph for each instance to explicitly draw the ironic relations between different modalities. **HKEmodel** [7] proposed a hierarchical framework for sarcasm detection by exploring atomic-level and composition-level congruities based on graph neural networks. **MILNet** [21] designed three graphs to capture multimodal incongruities. **Multi-view CLIP** [20] introduced a correction dataset called MMSD2.0, and they also presented a novel framework to leverage multi-grained cues from multiple perspectives. **DIP** [14], the current state-of-the-art model, effectively modeled the incongruity from factual and affective levels. We also report the performance of **GPT-4 V**. Specifically, the prompt for zero-shot prediction is "Examine the provided Twitter post, comprising both textual content and an image, please predict whether it expresses "sarcastic" or "non-sarcastic" emotions. If it is "sarcasm", answer 0. If it is "non-sarcasm", answer 1".

V. RESULTS AND ANALYSIS

In this section, we present the analysis for the results of the main experiment, ablation study, case study, and visualization respectively. For further discussions, please refer to the Appendix.

A. Main Results

In Table II, we present a comprehensive comparison between our proposed ESAM and various baseline models. Firstly, text-based models exhibit superior performance compared to image-based methods, primarily due to redundant visual information in the image, restricting its capacity to convey sentiment directly. Additionally, multimodal approaches outperform

TABLE III
ABLATION RESULTS OF OUR ESAM

Model	MMSD		MMSD2.0	
	ACC	F1	ACC	F1
MDF	87.77	86.39	83.01	82.39
MDF+SCC^{v}	88.05	86.81	83.45	82.72
MDF+SCC^{t}	88.29	86.99	83.99	83.05
MDF+SCC$^{v\&t}$	88.67	87.12	84.44	83.46
MDF+SCC+AOM	89.55	87.83	85.12	84.03
MDF$_m$+SCC+AOM	**90.11**	**88.19**	**85.87**	**84.56**
AFM	87.30	85.75	82.42	81.97
AFM+SCC+AOM	88.77	86.76	84.65	83.68

Here, AFM refers to directly aligning textual and visual features, while “MDF” denotes our foundational model equipped with a modality decoupling module. additionally, “SCC” signifies sentimental congruity constraint loss, where “v” and “t” denote visual and textual modality. Moreover, “AOM” indicates automatic outlier masking. And “m” represents the model augmented with multi-view predictions.

unimodal models, as the integration of multiple modalities facilitates a more comprehensive understanding of the data. Furthermore, we also evaluate the performance of large pre-trained models, demonstrating the limitations of both GPT-3.5 and GPT-4.0 on MSD tasks. We conjecture that the complexity and subtlety of sarcasm make it difficult for general-purpose large models to accurately capture and interpret ironic sentiment. Finally, our ESAM achieves remarkable performance, surpassing all the compared multimodal baselines. The main reason is that ESAM effectively alleviates feature shifts, further emphasizing the importance of addressing this issue in multimodal sentiment analysis.

B. Ablation Study

Table III presents an ablation study to investigate the contribution of each component in the ESAM. Initially, we evaluate the impact of sentimental congruity constraint within the Modality-decoupled Framework by constraining different modalities. It indicates that separately constraining visual or textual modalities can yield improvement, and the most substantial improvement stems from constraining both two modalities. Moreover, it can be observed that automatic outlier masking (AOM) can further boost performance, demonstrating its effectiveness in reducing the impact of the outliers. Although multi-view prediction, as presented in [20], yielded only marginal improvements, it significantly boosts the performance of our model. This is because SCC enhances the semantic awareness of different modalities, thereby improving the overall performance achieved by multi-view fusion. To highlight the advantages of MDF, we also conduct experiments with the Align and Fusion Model (AFM), which directly aligns and fuses text and image representations using a traditional transformer encoder, subsequently predicting the results based on the aligned representations. The AFM has been widely adopted by previous research [41], [42]. Overall, our approach manifests its universality by improving the performance of both AFM and MDF. Compared to AFM, our MDF only shows slight strength, While MDF only exhibits slight superiority compared to AFM under an unconstrained state, the significance of its improvement becomes evident when both models are equipped with SCC. This demonstrates the challenge of constraining fused features, implying the importance of mitigating feature shifts within individual latent spaces.

C. Visualization Discussion

To visually demonstrate the superiority of sentimental congruity constraint(SCC) and Automatic Outilier Masking(AOM), we present the influence of different modules on feature distribution in Fig. 4. Here, we utilize T-SNE[1] for dimensionality reduction of features to visualize their distribution. The visual representation in Fig. 4(b) reveals that the application of SCC promotes the clustering of samples belonging to the same class. This observation sharply contrasts with Fig. 4(a), which depicts the class distributions in the absence of SCC—noticeably more diffuse and dispersed. The reason behind this contrast is that the SCC consciously integrates sentimental congruity to constrain the feature distribution, effectively alleviating the feature shift. However, as highlighted in Fig. 4(b), there still exist cases of semantic confusion. To explore those cases, in Fig. 5, two randomly selected cases(outliers) from the highlighted part. Those samples exhibit ambiguous sentiment, leading to misguidance during training and subsequently diminishing the performance of SCC. Finally, Fig. 4(c), demonstrates that AOM can effectively address this issue by masking these outliers, which confirms the necessity of AOM to guide the training.

D. Case Study

To further evaluate the effectiveness of sentiment congruity constraint (SCC), Fig. 3 showcases a comparative case study of multi-view predictions, both with and without the application of SCC. It is observed that detecting sarcastic sentiments in the absence of SCC is challenging because of the feature shift. Illustratively, in the first case, the scene depicted in the image comprises “newspapers scattered on the ground”, which contradicts the textual content “I love respectful customers”. In the second case, the text and image exhibit only weak semantic relevance. This mismatch leads to feature shift and inter-class confusion when the text-image pair is directly aligned. Consequently, it results in incorrect prediction across all three views. In contrast, the MDF equipped with SCC can accurately identify corresponding ironic sentiments, showing its critical role in enhancing prediction accuracy.

E. Optimal Parameter Exploring

In this section, we explore the optimal parameter configurations. Firstly, we examine the performance variations of SCC with increasing learning temperature τ of SCC, as formalized in (4). The associated performance curves are graphically represented in the left panel of Fig. 6. We find that the performance is improved at first and subsequently decreases as the

[1] https://github.com/mxl1990/tsne-pytorch

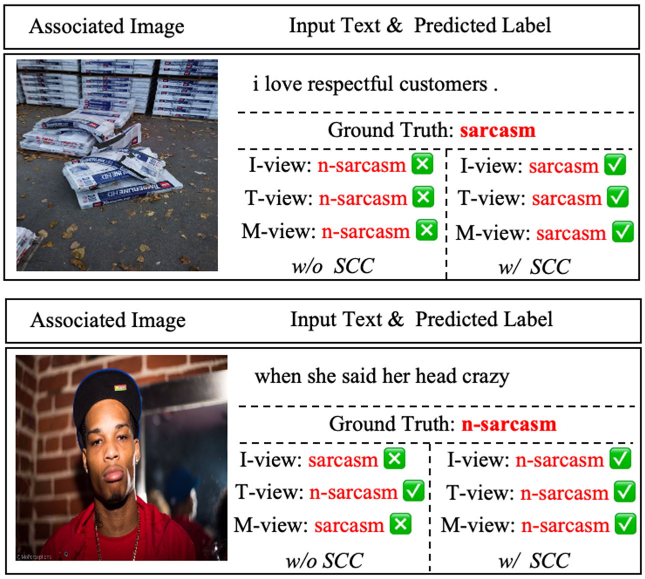

Fig. 3. Case study of multi-view prediction with and without(w and w/o) SCC, where "n-sarcasm" represents non-sarcasm prediction.

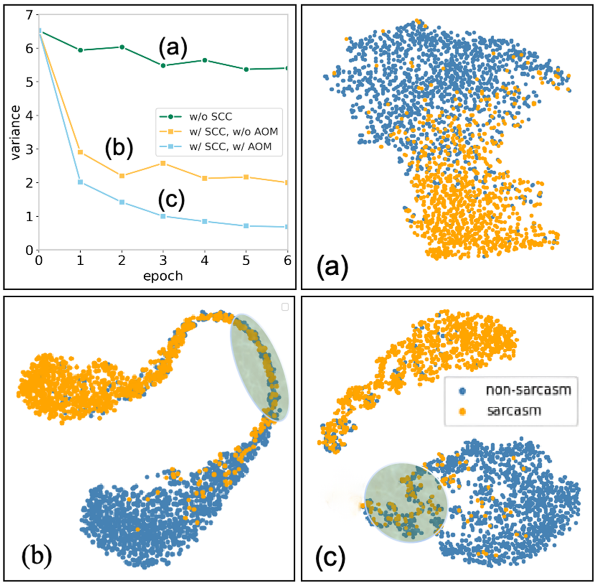

Fig. 4. The first figure demonstrates the training variance curves of the model equipped with different modules. "variance" denotes the metric that reflects the dispersion of a data distribution. The subsequent three figures visualize the representation of the corresponding curve.

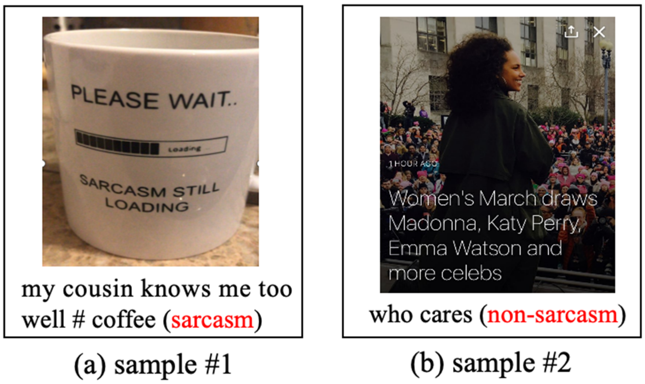
PLEASE WAIT..
SARCASM STILL LOADING
my cousin knows me too well # coffee (sarcasm)
(a) sample #1
Women's March draws Madonna, Katy Perry, Emma Watson and more celebs
who cares (non-sarcasm)
(b) sample #2

Fig. 5. Two randomly selected cases from the highlighted part of the second visualization distribution in Fig. 4.

TABLE IV
Ablation Study of Multi-View Prediction on MMSD and MMSD2.0 Datasets

Model	w/o SCC		w/ SCC	
	ACC	F1	ACC	F1
MMSD				
MDF$_m$	88.82	86.61	89.77	87.94
w/o L^t_{msd}	88.15	86.33	89.16	87.09
w/o L^v_{msd}	88.42	86.39	89.38	87.57
w/o L^m_{msd}	87.01	85.07	87.68	85.66
MMSD2.0				
MDF$_m$	84.65	83.51	85.27	84.16
w/o L^t_{msd}	83.52	82.67	84.39	83.41
w/o L^v_{msd}	83.43	82.85	84.52	83.48
w/o L^m_{msd}	82.29	81.95	82.58	82.39

τ increases. This trend can be attributed to the pivotal role played by the temperature coefficient in adjusting SCC's sensitivity to inter-sample distances, thereby influencing the model's constraint intensity on intra-class distribution. Both excessively stringent and lenient constraints on intra-class distribution have the potential to deteriorate the model's performance. In addition, we extend our experiment to investigate the effects of different threshold settings γ in the outlier detection algorithm of automatic outlier masking. The corresponding results are shown in the middle part of Fig. 6. We can see that the model attains peak performance when γ is set to 0.2 in the MMSD dataset and 0.25 in the MMSD2.0 dataset. A lower γ results in a more permissive outlier filtration, which causes some outliers to remain undetected. Conversely, an increasing γ leads to overly strict filtration, raising the likelihood of normal samples being erroneously identified as outliers, thereby negatively impacting the model's performance.

In addition, we also analyze the impact of the balance coefficient in the loss function and report the Acc score in the right part of Fig. 6. The results show that the best performance can be achieved when the balance coefficient is set to 1 in both the two datasets and the performance further drops when increasing the value of the balance coefficient. We conjecture the reason to be the excessive suppression of feature distribution caused by a larger balance coefficient, which disrupts the feature space structure and ultimately leads to under-fitting during model training.

F. Performance of Multi-View Prediction

To further analyze the performance of different modality views within our method, we conducted a series of experiments by selectively omitting the training loss of text view L^t_{msd}, image view L^v_{msd} and multimodal view L^m_{msd} both in the absence and presence of SCC. As illustrated in Table IV, the removal

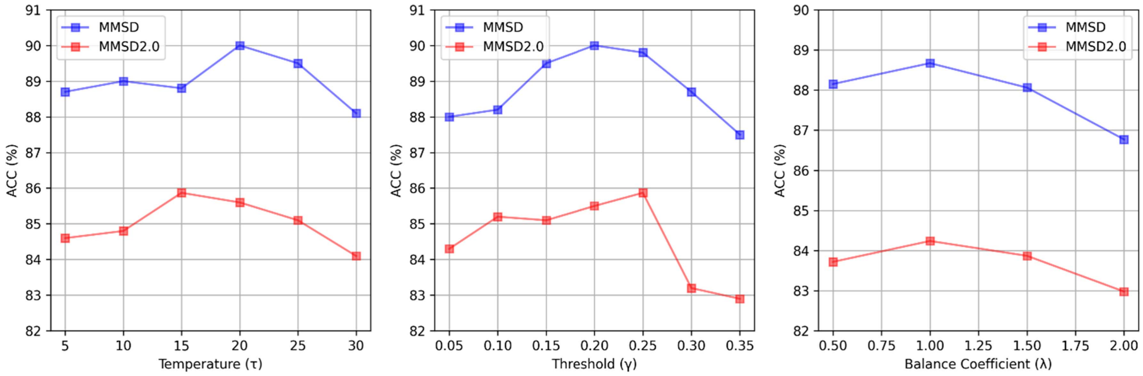

Fig. 6. The curve of performance for multimodal sarcasm detection with different hyperparameters on MMSD and MMSD2.0 datasets.

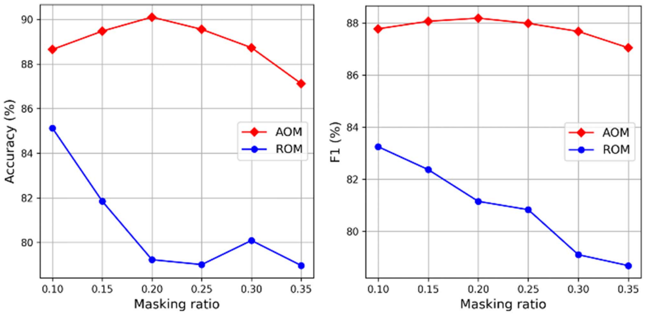

Fig. 7. The performance of different outlier masking strategies.

of L_{msd}^t, L_{msd}^v and L_{msd}^m individually led to notable decreases in accuracy, emphasizing their importance. The optimal performance of our framework is achieved by integrating all these views, indicating the pivotal role each view plays in bolstering the effectiveness of our model. This observation indirectly validates the efficacy of our proposed modality-decoupled framework, which is the prerequisite for multi-perspective prediction. Moreover, the introduction of SCC demonstrably enhances the performance across all views, underscoring its potent role in mitigating feature shifts within each modality.

G. The Effect of Automatic Outlier Masking

To analyze the impact of the proposed Automatic Outlier Masking, we compare it with Random Outlier Masking (ROM), which randomly selects the outlier from the dataset. The results are shown in Fig. 7. In the figure, the horizontal axis indicates the masking ratio, while the vertical axis indicates Accuracy and F1 score, respectively. It can be observed that the proposed AOM outperforms ROM across all masking ratios. This is primarily because ROM randomly selects outlier samples, resulting in the selection of many normal samples and consequently a significant performance degradation. In contrast, AOM dynamically selects outlier samples through the proposed outlier masking algorithm, thereby achieving superior performance.

VI. Limitation

While our method shows promising results on multimodal sarcasm detection, there are notable limitations in its performance when applied to datasets with varying characteristics. Specifically, as observed in our experiments, the proposed approach achieves a more substantial improvement on the MMSD dataset than MMSD2.0. This discrepancy highlights the dependency of our model on explicit sarcastic cues within the data, which, when absent, reduces its effectiveness in distinguishing outliers. In future work, we plan to address these limitations by developing a more robust approach for sarcasm detection that can adapt to different data distributions and identify outliers in datasets with more implicit or nuanced sarcastic features. Additionally, we aim to investigate the impact of dataset size and diversity on our model's performance to better understand its applicability across a wider range of scenarios.

VII. Conclusion

This paper is the first work to mitigate the feature shift issue that restricted the performance of existing models in the MSD task. Concretely, we propose a Modality-decoupled Framework (MDF) to separate the textual and visual modalities, as this parallel processing is particularly effective in mitigating feature shifts. Additionally, we incorporate sentimental congruity constraints in visual and textual latent spaces to enhance semantic awareness, further alleviating feature shifts. Moreover, we introduce Automatic Outlier Masking to identify and mask outliers during training automatically, which eliminates the interference of outliers and guides the training process. Our proposed model has been extensively evaluated on public datasets, consistently demonstrating superior performance.

References

[1] A. Joshi, V. Sharma, and P. Bhattacharyya, "Harnessing context incongruity for sarcasm detection," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguist., 7th Int. Joint Conf. Natural Lang. Process.*, 2015, pp. 757–762.

[2] Y. Tay, A. T. Luu, S. C. Hui, and J. Su, "Reasoning with sarcasm by reading in-between," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguist.*, 2018, pp. 1010–1020.

[3] R. Schifanella, P. De Juan, J. Tetreault, and L. Cao, "Detecting sarcasm in multimodal social platforms," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 1136–1145.

[4] H. Pan, Z. Lin, P. Fu, Y. Qi, and W. Wang, "Modeling intra and inter-modality incongruity for multi-modal sarcasm detection," in *Findings Assoc. Comput. Linguist.*, 2020, pp. 1383–1392.

[5] Y. Cai, H. Cai, and X. Wan, “Multi-modal sarcasm detection in twitter with hierarchical fusion model,” in *Proc. 57th Annu. Meeting Assoc. Comput. Linguist.*, 2019, pp. 2506–2515.

[6] B. Liang et al., “Multi-modal sarcasm detection via cross-modal graph convolutional network,” in *Proc. 60th Annu. Meeting Assoc. Comput. Linguist.*, 2022, pp. 1767–1777.

[7] H. Liu, W. Wang, and H. Li, “Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement,” in *Proc. 2022 Conf. Emp. Methods Natural Lang. Process.*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Abu Dhabi, United Arab Emirates, Dec. 2022, pp. 4995–5006. [Online]. Available: https://aclanthology.org/2022.emnlp-main.333/

[8] J. Li, D. Li, S. Savarese, and S. Hoi, “BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” 2023, *arXiv:2301.12597*.

[9] P. Zhu et al., “Prompt-based learning for unpaired image captioning,” *IEEE Trans. Multimedia*, vol. 26, pp. 379–393, 2024.

[10] L. Zhou et al., “Unified vision-language pre-training for image captioning and VQA,” in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 13041–13049.

[11] Z. Li, B. Xu, C. Zhu, and T. Zhao, “CLMLF: A contrastive learning and multi-layer fusion method for multimodal sentiment detection,” in in *Proc. Findings Assoc. Comput. Linguist.: NAACL*, ACL, 2022, pp. 2282–2294.

[12] D. Wang et al., “Cross-modal enhancement network for multimodal sentiment analysis,” *IEEE Trans. Multimedia*, vol. 25, pp. 4909–4921, 2022.

[13] X. Yang, S. Feng, D. Wang, and Y. Zhang, “Image-text multimodal emotion classification via multi-view attentional network,” *IEEE Trans. Multimedia*, vol. 23, pp. 4014–4026, 2021.

[14] C. Wen, G. Jia, and J. Yang, “DIP: Dual incongruity perceiving network for sarcasm detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 2540–2550.

[15] Y. Tian, N. Xu, R. Zhang, and W. Mao, “Dynamic routing transformer network for multimodal sarcasm detection,” in *Proc. 61st Annu. Meeting Assoc. Comput. Linguist.*, 2023, pp. 2468–2480.

[16] B. Liang et al., “Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs,” in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 4707–4715.

[17] A. Vaswani et al., “Attention is all you need,” in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[18] K. He et al., “Masked autoencoders are scalable vision learners,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16000–16009.

[19] F. Liu, K. Ting, and Z. Zhou, “Isolation forest,” in *Proc. 8th IEEE Int. Conf. Data Mining*, 2008, pp. 413–422.

[20] L. Qin et al., “MMSD2.0: Towards a reliable multi-modal sarcasm detection system,” in *Proc. Findings Assoc. Comput. Linguist.*, 2023, pp. 10834–10845.

[21] Y. Qiao et al., “Mutual-enhanced incongruity learning network for multi-modal sarcasm detection,” in *Proc. AAAI Conf. Artif. Intell.*, 2023, pp. 9507–9515.

[22] Y. Tian, N. Xu, R. Zhang, and W. Mao, “Dynamic routing transformer network for multimodal sarcasm detection,” in *Proc. 61th Annu. Meeting Assoc. Comput. Linguist.*, 2023, pp. 2468–2480.

[23] Y. Wang et al., “Unified vision-language pretraining with image-to-text captioning,” 2019, *arXiv:2019.09.18*.

[24] F. Chen et al., “VLP: A survey on vision-language pre-training,” *Mach. Intell. Res.*, vol. 20, pp. 38–56, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:246996617

[25] X. Li et al., “VisualBERT: A simple and performant baseline for vision and language learning,” 2020, *arXiv:2004.10934*.

[26] B. Zhou et al., “Aligning images and texts in the wild with deep bidirectional attention networks,” 2019, *arXiv:2019.08.14*.

[27] A. Radford et al., “Learning transferable visual models from natural language supervision,” in *Proc. 38th Int. Conf. Mach. Learn.*, M. Meila and T. Zhang, Jul. 2021, vol. 139, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[28] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, “Minigpt-4: Enhancing vision-language understanding with advanced large language models,” 2023, *arXiv:2304.10592*.

[29] OpenAI J. Achiam et al., “GPT-4 technical report,” 2024, *arXiv:2303.08774*.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” 2018, *arXiv:1810.04805*.

[31] P. Anderson et al., “Bottom-up and top-down attention for image captioning and visual question answering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086.

[32] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” 2020, *arXiv:2010.11929*.

[33] L. Qin et al., “MMSD2.0: Towards a reliable multi-modal sarcasm detection system,” 2023.

[34] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” 2017, *arXiv:1711.05101*.

[35] Y. Kim, “Convolutional neural networks for sentence classification,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1746–1751.

[36] P. Zhou et al., “Attention-based bidirectional long short-term memory networks for relation classification,” in *Proc. 54th Annu. Meeting Assoc. Comput. Linguist.*, 2016, pp. 207–212.

[37] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[38] L. Ouyang et al., “Training language models to follow instructions with human feedback,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 27730–27744.

[39] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[40] N. Xu, Z. Zeng, and W. Mao, “Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association,” in *Proc. 58th Annu. Meeting Assoc. Comput. Linguist.*, 2020, pp. 3777–3786.

[41] Z. Li, B. Xu, C. Zhu, and T. Zhao, “CLMLF: A contrastive learning and multi-layer fusion method for multimodal sentiment detection,” in *Proc. Find. Assoc. Comput. Linguistics*, M. Carpuat, M.-C. de Marneffe, R. Meza, and V. Ivan, Seattle, USA, Jul. 2022, pp. 2282–2294. [Online]. Available: https://aclanthology.org/2022.findings-naacl.175/

[42] Y. Wei et al., “Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection,” in *Proc. 61st Annu. Meeting Assoc. Comput. Linguist.*, 2023, pp. 5240–5252.

Shaozu Yuan is currently an Algorithm Engineer with JD.com, China. His research interests include multimodal emotion analysis, natural language processing, and large language models. He was a Reviewer for top-tier conferences including ACL, NeurIPS, and ACM MM.

Yiwei Wei is currently working toward the Ph.D. degree with the School of Intelligence and Computing, Tianjin University, Tianjin, China. He is also a Faculty Member with the China University of Petroleum (Beijing), Karamay Campus, Karamay, China. He has authored or coauthored more than 15 papers in ACL, AAAI, NeurIPS, and other venues. His research interests include multimodal sentiment analysis, natural language processing, and multimodal information extraction.

Hengyang Zhou is currently working toward the B.E. degree in data science and Big Data technology with the China University of Petroleum (Beijing) at Karamay, Karamay, China. His research interests include multimodal sentiment analysis and large language models.

Qinfu Xu received the B.S. degree in software engineering from China University of Petroleum (Beijing), Karamay, China, in 2022. He is currently working toward the M.S. degree in software engineering with the China University of Petroleum (East China), Shandong, China. His research interests include multimodal sentiment analysis, vision-language understanding, and machine learning.

Meng Chen is currently the AI Director of JD.COM. He was a Research Scientist with Nuance Communications. He has authored or Coauthored more than 40 papers in prestigious academic conferences, such as AAAI, IJCAI, ACM Multimedia, ACL, NAACL, ICASSP, Interspeech, and CIKM. His research interests include natural language processing, speech recognition, and dialogue systems. He is also a Program Committee Member for several top-tier conferences including AAAI, ACL, EMNLP, NAACL, EACL, ICASSP, Interspeech, and ACM Multimedia. He was the recipient of the Wuwenjun Artificial Intelligence Science and Technology Progress Award in 2023.

Xiaodong He (Fellow, IEEE) received the bachelor's degree from Tsinghua University, Beijing, China, in 1996, the M.S. degree from the Chinese Academy of Sciences, Beijing, China, in 1999, and the Ph.D. degree from the University of Missouri-Columbia, Columbia, MO, USA, in 2003. He was with Microsoft for about 15 years, and was a Principal Researcher and the Research Manager with DLTC at Microsoft Research, Redmond, WA USA. He is currently the Vice President with JD.COM and the Director with JD AI Research. He is also an Affiliate Professor with the University of Washington, Seattle, WA, USA, and works in Doctoral Supervisory Committees. He has authored or coauthored more than 200 papers in ACL, EMNLP, NAACL, CVPR, SIGIR, WWW, CIKM, NIPS, ICLR, ICASSP, PROCEEDINGS OF THE IEEE, IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, *IEEE Signal Processing Magazine*, and other venues. His research interests mainly include artificial intelligence areas focusing on deep learning, natural language, computer vision, speech, information retrieval, and knowledge representation. He was the recipient of the several awards including the Outstanding Paper Award at ACL 2015.