

Multi-View Incongruity Learning for Multimodal Sarcasm Detection

Diandian Guo^{1,2}, Cong Cao^{1,*}, Fangfang Yuan¹, Yanbing Liu^{1,2,*},
Guangjie Zeng³, Xiaoyan Yu⁴, Hao Peng³, Philip S. Yu⁵

¹Institute of Information Engineering, Chinese Academy of Sciences

²School of Cyber Security, University of Chinese Academy of Sciences

³State Key Laboratory of Software Development Environment, Beihang University

⁴School of Computer Science and Technology, Beijing Institute of Technology

⁵Department of Computer Science, University of Illinois at Chicago

{guodiandian, caocong, yuanfangfang, liuyanbing}@iie.ac.cn,
{zengguangjie, penghao}@buaa.edu.cn, xiaoyan.yu@bit.edu.cn, psyu@uic.edu

Abstract

Multimodal sarcasm detection (MSD) is essential for various downstream tasks. Existing MSD methods tend to rely on spurious correlations. These methods often mistakenly prioritize non-essential features yet still make correct predictions, demonstrating poor generalizability beyond training environments. Regarding this phenomenon, this paper undertakes several initiatives. Firstly, we identify two primary causes that lead to the reliance of spurious correlations. Secondly, we address these challenges by proposing a novel method that integrates Multimodal Incongruities via Contrastive Learning (MICL) for multimodal sarcasm detection. Specifically, we first leverage incongruity to drive multi-view learning from three views: token-patch, entity-object, and sentiment. Then, we introduce extensive data augmentation to mitigate the biased learning of the textual modality. Additionally, we construct a test set, SPMSD, which consists potential spurious correlations to evaluate the model's generalizability. Experimental results demonstrate the superiority of MICL on benchmark datasets, along with the analyses showcasing MICL's advancement in mitigating the effect of spurious correlation.

1 Introduction

Sarcasm, inherently metaphorical, seeks to convey meanings that diverge from literal interpretations. Its prevalence on social media platforms underscores the critical need for effective sarcasm detection, which is a tool pivotal for uncovering the genuine opinions and emotions of users. This capability supports essential applications such as public opinion mining (Cai et al., 2019; Prasanna et al., 2023) and sentiment analysis (Farias and Rosso, 2017; Khare et al., 2023).

Early attempts of sarcasm detection focus solely on textual modality (Davidov et al., 2010; Zhang

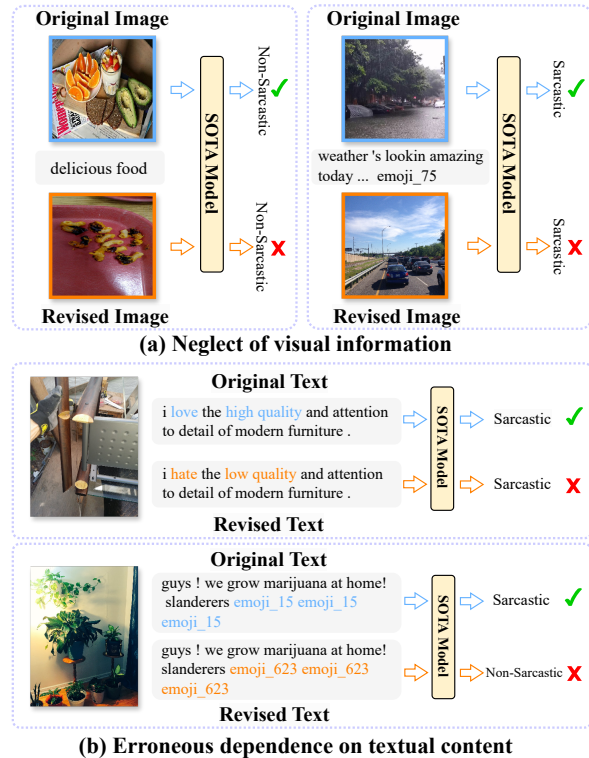


Figure 1: Existing models suffer from two deficiencies that lead to spurious correlations on MSD task.

et al., 2016; Xiong et al., 2019), modeling the incongruities within the text. However, the proliferation of multimedia social platforms enables users to convey opinions and emotions using multimodal information. As a result, MSD has recently attracted widespread attention. Joshi et al. (2015) demonstrate the incongruity as a pivotal factor for detecting sarcasm, which sparks a surge of research into learning incongruity using textual and visual cues, achieving outstanding results (Wen et al., 2023; Qiao et al., 2023).

Despite these efforts, existing models still suffer from reliance on spurious correlations. Spurious correlation is a phenomenon where models learn non-generalizable features, rather than core features truly related to the real labels, thus under-

*Corresponding authors

mining the model’s generalizability (Deng et al., 2024). We conduct experiments on the current SOTA model (Jia et al., 2024) and attribute the spurious correlations in MSD to two primary oversights: **1)** Overemphasis on the text encoder while underestimating visual information. For example shown in Figure 1(a), changing the image does not affect the model’s result when the textual input remains the same, revealing a biased dependence on the textual modality. **2)** Erroneously relying on non-critical textual features rather than critical emotional features. As the instance illustrated in Figure 1(b), changing key emotional words does not affect the model’s result. Conversely, the model makes an opposite judgment when non-critical descriptions that do not affect semantics are modified. In summary, the above findings reveal that existing models rely on spurious correlations, failing to capture the necessary task-related features.

To address the above issues, we introduce MICL, a novel multi-view incongruity learning method for MSD. This method is structured around three modules: multimodal feature encoding, multi-view incongruity learning, and multi-view fusion. Specifically, for multimodal feature encoding, in addition to the traditional textual and visual encoding, we introduce the OCR-texts for supplementary element to uncover the information contained within the image to a greater extent. Yang et al. (2024) demonstrate that multi-view learning can improve the effectiveness of models in social media. Considering that sarcastic content often involves an entity or object in a multimodal context and carries sentiment polarity, the multi-view incongruity learning module learns robust features from three aspects: token-patch, entity-object, and sentiment, to mitigate spurious correlations. However, the quality and importance of each view vary significantly across samples (Wu et al., 2022). Therefore, we propose using a beta distribution-based multi-view fusion module to perform confidence-weighted fusion of the learned embeddings, producing more reliable results. Furthermore, we extend beyond conventional text data augmentation techniques, which tend to perpetuate a bias towards textual information. Instead, MICL incorporates a dual augmentation strategy, enhancing both text and image data. Our contributions are as follows.

- We propose MICL, a novel multi-view learning method that comprehensively learns incongruities and integrates them credibly.
- We introduce robust data augmentation strate-

gies that enriches both textual and visual contents, mitigating biased learning of the textual modality.

- Experimental results indicate that our approach outperforms existing methods on the MSD task and demonstrates stronger robustness against spurious correlations.

2 Related Work

2.1 Multimodal Sarcasm Detection

Multimodal sarcasm detection is a research task that identifies sarcasm through multimodal cues. Schifanella et al. (2016) first propose integrating textual features with visual features to solve the sarcasm detection task. Following this, Cai et al. (2019) construct an advanced MSD dataset based on tweets, providing a benchmark for subsequent research. InCrossMGs (Liang et al., 2021) is the first to model the interaction of information within and between modalities by graph neural networks. DMSD-CL (Jia et al., 2024) employs counterfactual augmentation and contrastive learning to study MSD in out-of-distribution scenarios. Recently, many works have dedicated efforts to model the incongruity in text-image pairs. For example, MIL-Net (Qiao et al., 2023) focuses on the combination of local incongruity and global incongruity. However, existing methods only focus on token-patch incongruity, which leads to erroneous reliance on non-critical features. Our model proposes to learn multi-view incongruity information to improve performance and enhance robustness.

2.2 Mitigating Spurious Correlations

Mitigating spurious correlations in multimodal scenarios has attracted increasing research interest. Existing methods for improving robustness against spurious correlations can be divided into two lines of research. One line focuses on effectively using multimodal information to enhance robustness (Yenamandra et al., 2023). Some methods use the distributed robust optimization (DRO) framework to dynamically increase the weight of minimizing the worst group loss (Wen and Li, 2021). Most recently, Kirichenko et al. (2023) propose methods that train a model using Empirical Risk Minimization (ERM) first and then only finetune the last layer on balanced data. Another line of research focuses on mitigating the bias in training data by creating additional data to balance the training dataset (Niu et al., 2021). Inspired by these methods, we comprehensively mitigate the reliance on spurious

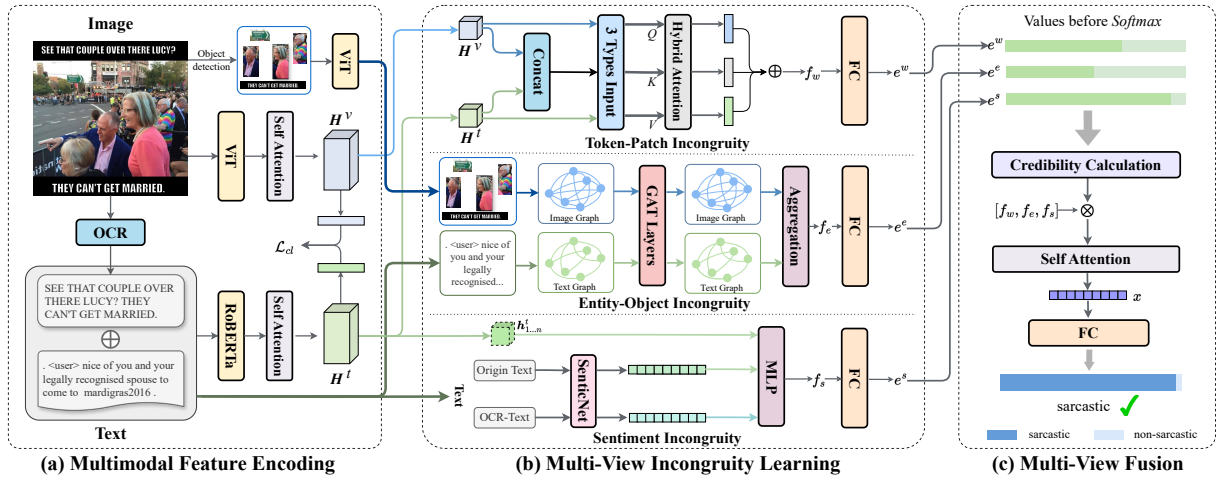


Figure 2: The overall architecture of MICL primarily comprises three key modules: (a) Multimodal Feature Encoding, (b) Multi-View Incongruity Learning, and (c) Multi-View Fusion. Additionally, we introduce data augmentation for each training data.

correlations in the MSD task from both the model and data perspectives.

3 Methodology

As shown in Figure 2, the architecture of MICL mainly consists of three parts: multimodal feature encoding, multi-view incongruity learning, and multi-view fusion. Additionally, to mitigate the modal bias problem from the data level, we introduce data augmentation for the training input.

3.1 Multimodal Feature Encoding

Given a text-image pair $\mathcal{X} = (\mathcal{T}, \mathcal{V})$, we first need to perform feature encoding, which is divided into two steps: text encoding and image encoding.

3.1.1 Text Encoding

In current multimodal learning approaches, textual and visual information are commonly encoded independently. However, our observation reveals that a number of images contain textual information that frequently complements the textual modality. Building upon this observation, we incorporate optical character recognition text (OCR-text) \mathcal{O} from images as an auxiliary input alongside the original text input \mathcal{T} . However, the OCR-text provided by existing work (Pan et al., 2020) has issues with low accuracy and ambiguous meaning, as shown in the Figure 3. Low-quality OCR-text may reduce model performance (Wang et al., 2024). Instead, we generate refined OCR-texts employing GLM-4V¹ (Wang et al., 2023) with more precision extraction and

translation, complemented by meticulous manual proofreading. Then, we concatenate \mathcal{T} and \mathcal{O} , and feed them into the text encoder. As shown in Figure 2(a), we apply the pre-trained language model RoBERTa (Liu et al., 2019) as the text encoder:

$$\mathbf{H}^t = \text{Self_Att}(\text{RoBERTa}(\mathcal{T} \oplus \mathcal{O})), \quad (1)$$

where $\mathbf{H}^t = [\mathbf{h}_{cls}^t, \mathbf{h}_1^t, \mathbf{h}_2^t, \dots, \mathbf{h}_n^t] \in \mathbb{R}^{(n+1) \times d}$ is the textual representation of the input text, $\mathbf{h}_i^t \in \mathbb{R}^d$ denotes the hidden state vector of i -token, d denotes the dimension of the hidden representations, n is the total number of tokens after concatenating the original text and OCR-text, Self_Att means a self-attention layer, and \oplus refers to the concatenation operation. For clarity and simplification, we use e_t^k to represent \mathbf{h}_{cls}^t of the k -th sample in subsequent expressions.

Original data	OCR by Pan et al.	Ours
	idhar 4 din ki girlfriend date pe ja rahi hai 2 saal ki dost khane le aane ko bahana maar rahi	here my girlfriend of 4 days is going on a date and my 2 year old friend keeps making excuses to come when asked
	o but its i'm going i only 4 : 50 ! to bed sorry what can't hear you cc nothing suspicio . us	i'm going to bed. but it's only 4:30! sorry what? can't hear you!

Figure 3: In the first example, since the text is in Hindi, it is difficult for a non-multilingual pre-trained RoBERTa to understand. Our method automatically translates the extracted text into English. In the second example, existing OCR result exhibits deficiencies in both recognition accuracy and sequential integrity, whereas our result performs better.

¹<https://open.bigmodel.cn>

3.1.2 Image Encoding

We use a pre-trained ViT (Dosovitskiy et al., 2020) as the image encoder. For each image $\mathcal{V} = \{v_{cls}, v_1, \dots, v_{n_{\mathcal{V}}}\}$, where v_{cls} means the [CLS] token, v_i represents the i -patch of \mathcal{V} , and $n_{\mathcal{V}}$ is the total patch number. We feed \mathcal{V} into ViT:

$$\mathbf{H}^v = \text{Self_Att}(\text{ViT}(\mathcal{V})), \quad (2)$$

where $\mathbf{H}^v = [\mathbf{h}_{cls}^v, \mathbf{h}_1^v, \mathbf{h}_2^v, \dots, \mathbf{h}_{n_{\mathcal{V}}}^v] \in \mathbb{R}^{(n_{\mathcal{V}}+1) \times d}$ is the visual representation of the input image, $\mathbf{h}_i^v \in \mathbb{R}^d$ represents the i -th patch embedding. For clarity and simplification, we use e_v^k to represent \mathbf{h}_{cls}^v of the k -th sample in subsequent expressions.

3.2 Multi-View Incongruity Learning

For the MSD task, cross-modal incongruity learning predominantly focuses on the token-patch levels. However, sarcastic contents are often closely related to specific entities or objects in multimodal contexts. Furthermore, sarcastic contents typically involve strong emotions that existing models overlook. To achieve a more comprehensive incongruity learning, we further incorporate incongruity learning from the entity-object and the explicit sentiment perspectives as shown in Figure 2(b).

3.2.1 Token-patch Incongruity Learning

A cross-attention mechanism is commonly used for modeling cross-modal interactions. Existing methods (Qiao et al., 2023; Jia et al., 2024) often use text as the query and images as the key and value, which may lead to modality bias. Instead, we design a hybrid attention interaction mechanism for unbiased token-patch incongruity learning, which can integrate text and image features in a balanced manner. Based on the input differences of the multi-head attention layer, it can be divided into the following parts:

$$\mathbf{Q}_{tv} = \mathbf{K}_{tv} = \mathbf{V}_{tv} = \mathbf{H}^{tv}, \quad (3)$$

$$\mathbf{Q}_t = \mathbf{H}^t, \mathbf{K}_t = \mathbf{V}_t = \mathbf{H}^v, \quad (4)$$

$$\mathbf{Q}_v = \mathbf{H}^v, \mathbf{K}_v = \mathbf{V}_v = \mathbf{H}^t, \quad (5)$$

where $\mathbf{H}^{tv} = \mathbf{H}^t \oplus \mathbf{H}^v$. Then, we feed different inputs into a standard cross-attention layer:

$$\mathbf{F} = \text{Cross_att}(\mathbf{Q}, \mathbf{K}, \mathbf{V}). \quad (6)$$

We define \mathbf{F}_{tv} , \mathbf{F}_t and \mathbf{F}_v as the outputs of the attention mechanisms from the input Eq. (3), (4) and (5), respectively. For \mathbf{F}_{tv} , \mathbf{F}_t and \mathbf{F}_v , we treat the encoding of their [CLS] tokens, \mathbf{f}_{tv} , \mathbf{f}_t and \mathbf{f}_v , as the final output:

$$\mathbf{f}_w = \mathbf{f}_{tv} \oplus \mathbf{f}_t \oplus \mathbf{f}_v. \quad (7)$$

3.2.2 Entity-object Incongruity Learning

To effectively capture entity-object incongruity, we construct semantic graphs for both text and images. Specifically, for the text semantic graph, we treat entities as nodes and use spaCy² to extract dependencies between entities as edges. If there is a dependency between two entities, an edge will be created between them in the text graph. For the visual semantic graph, we follow Anderson et al. (2018) to segment the image into object regions. We treat each region as a node, and create edges based on cosine similarity. Additionally, both graphs are undirected and contain self-loops.

Then, we model the graphs with Graph Attention Network (GAT) (Veličković et al., 2018). Taking the textual graph as an example, let $\alpha_{i,j}^l$ be the attention score between i and j , and \mathbf{g}_i^l denote the feature of node i in the l -th layer. We have:

$$\alpha_{i,j}^l = \frac{\exp\left(LR\left(\mathbf{u}_l^\top [\mathbf{W}_l \mathbf{g}_i^l \parallel \mathbf{W}_l \mathbf{g}_j^l]\right)\right)}{\sum_k \exp\left(LR\left(\mathbf{u}_l^\top [\mathbf{W}_l \mathbf{g}_i^l \parallel \mathbf{W}_l \mathbf{g}_k^l]\right)\right)}, \quad (8)$$

$$\mathbf{g}_i^{l+1} = \alpha_{i,i}^l \mathbf{W}_l \mathbf{g}_i^l + \sum_{j \in \mathcal{N}(i)} \alpha_{i,j}^l \mathbf{W}_l \mathbf{g}_j^l, \quad (9)$$

where $k \in \mathcal{N}(i) \cup i$ belongs to the neighbor nodes of i and i itself. LR denotes the LeakyReLU layer. $\mathbf{W}_l \in \mathbb{R}^{d \times d}$ and \mathbf{u}_l are learnable parameters of the l -th textual GAT layer. We initialize $\mathbf{g}_i^0 = \mathbf{h}_i^t$.

We denote the final textual representation as $\mathbf{G}^T = \{\mathbf{g}_0, \dots, \mathbf{g}_n\}$. Similarly, we can obtain \mathbf{G}^V . We define $\mathbf{G} = \mathbf{G}^T \oplus \mathbf{G}^V$, then we can learn the entity-object incongruity:

$$\mathbf{f}_e = \frac{1}{|\mathbf{G}|} \sum_{\mathbf{g}_i \in \mathbf{G}} \text{Softmax}(\mathbf{g}_i \mathbf{W}_g + b_g) \mathbf{g}_i, \quad (10)$$

where \mathbf{W}_g and b_g are learnable parameters.

3.2.3 Sentiment Incongruity Learning

Given the pivotal role of emotional context in MSD (Joshi et al., 2015), our model integrates sentiment analysis to discern incongruity in the original text and OCR-text. Specifically, we extract the sentiment polarity of the source text and OCR-text through SenticNet (Cambria et al., 2024):

$$\mathbf{s}_t = \text{SenticNet}(\mathcal{T}), \mathbf{s}_o = \text{SenticNet}(\mathcal{O}), \quad (11)$$

$$\mathbf{f}_s = \text{MLP}(\mathbf{s}_t \oplus \mathbf{s}_o \oplus \mathbf{h}_{1 \dots n}^t), \quad (12)$$

where MLP is a multi-layer perceptron. If OCR-text is unavailable, \mathbf{f}_s is assigned a value of 0.

²<https://spacy.io/>


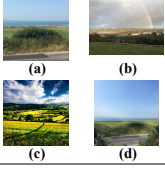
Original data	Text augmentation	Visual augmentation
 <p>i think i have got the worst ever weather of <num> enforced days off! scorchio freelance</p>	<p>sarcastic I believe I've experienced the most terrible weather during my enforced <num> days off!</p> <p>non-sarcastic I believe I've enjoyed the best possible weather during my mandatory <num> days off!</p>	 <p>(a) (b) (c) (d)</p>

Figure 4: Summary of text and visual augmentation methods. Text augmentation generates samples with the same or opposite labels. Visual augmentation methods include: (a) cropping, (b) swapping images, (c) image generation, and (d) image style transfer.

3.3 Multi-View Fusion

As shown in Figure 2(c), the credibility of the three incongruity features varies across different MSD scenarios. Measuring the confidence of different features helps improve detection performance. TMC (Han et al., 2021) has proved that the Dirichlet distribution can effectively estimate the credibility of a single view. In binary classification scenario, the Beta distribution shares the same mathematical significance. Following Ma et al. (2024), we use the output before the softmax operation of the m -th view classifier as evidence e^m , then the credibility c^m can be expressed as:

$$c^m = \frac{e_0^m + e_1^m}{S_m} = \frac{e_0^m + e_1^m}{(e_0^m + 1) + (e_1^m + 1)}, \quad (13)$$

where e_r^m represents the output of the final layer of the classifier for the m -th view regarding the r -th classification result. In binary classification tasks, $r \in \{0, 1\}$. The derivation process can be found in the Appendix B.

After obtaining the credibility, we use a self-attention network to obtain the fusion feature:

$$\mathbf{x} = \text{Self_Att}([\mathbf{f}_w, \mathbf{f}_e, \mathbf{f}_s] \cdot [c^w, c^e, c^s]^\top). \quad (14)$$

3.4 Data Augmentation and Contrastive Learning

3.4.1 Data Augmentation

Images serve as a vital source of incongruity clues, which is essential for comprehensive sarcasm analysis. However, previous MSD methods (Pan et al., 2020; Jia et al., 2024) focus on enhancing textual content and overlook the importance of image data augmentation. This inadequate data augmentation fails to enhance model performance and may even impede the performance (Wang et al., 2024). To address this issue, we adopt augmentation involving both textual and visual data, ensuring a balanced and effective enhancement.

As shown in Figure 4, for text augmentation, we employ two strategies: **1)** Replacing key entities or reversing sentiment words to obtain samples with opposite labels; **2)** Paraphrasing the original samples to keep the meaning unchanged, obtaining samples with the same labels. Text augmentation is performed by ChatGPT³. We apply the above strategies at a 1:1 ratio to generate augmented texts for all training samples.

For image augmentation, we use four strategies: **1)** Randomly cropping images and resizing them to 224×224 ; **2)** Randomly swapping images of samples with the same label; **3)** Employing stable diffusion for image style transfer; **4)** Prompting GLM-4V to generate image titles, and then using stable diffusion to generate new images based on those titles. We apply these four strategies at a 3:3:2:2 ratio to generate augmented images for all training samples.

3.4.2 Contrastive Learning Framework

For the generated sample $\tilde{\mathcal{X}} = (\tilde{\mathcal{T}}, \tilde{\mathcal{V}})$, we input $\{\mathcal{X}, \tilde{\mathcal{X}}\}$ into the training process together. We construct a contrastive learning framework based on whether the labels are the same to determine the positive and negative examples. Specifically, within a batch, samples with the same label as the anchor sample are considered positive samples, forming the positive sample set S_P ; otherwise, they belong to the negative sample set S_N . We define the sample set in one batch as $S = S_P + S_N$. In our entire model framework, the key is modeling the incongruity between text and image. Therefore, when constructing the contrastive learning framework, we use the text-image matching approach to obtain scores for positive and negative examples.

For k -th sample in the training set, $t \rightarrow v$ contrastive loss is:

$$\mathcal{L}_k^{t \rightarrow v} = \frac{1}{|S_P|} \sum_{i \in |S_P|} -\log \frac{\exp(\cos(\mathbf{e}_t^k, \mathbf{e}_v^i)/\tau)}{\sum_{j \in S} \exp(\cos(\mathbf{e}_t^k, \mathbf{e}_v^j)/\tau)}, \quad (15)$$

where $\tau \in \mathbb{R}^+$ is the temperature parameter. Similarly, we can obtain $v \rightarrow t$ contrastive loss $\mathcal{L}_k^{v \rightarrow t}$. The overall contrastive loss is as follows:

$$\mathcal{L}_{cl} = \frac{1}{N} \sum_{k=1}^N \left(\frac{1}{2} \mathcal{L}_k^{t \rightarrow v} + \frac{1}{2} \mathcal{L}_k^{v \rightarrow t} \right), \quad (16)$$

where N is the total number of samples in the training set.

³<https://chat.openai.com>

Modality	Method	Acc.(%)	Binary-Average			Macro-Average		
			Pre.(%)	Rec.(%)	F1(%)	Pre.(%)	Rec.(%)	F1(%)
Text	TextCNN	80.03	74.29	76.39	75.32	78.03	78.28	78.15
	Bi-LSTM	81.90	76.66	78.42	77.53	80.97	80.13	80.55
	BERT	83.85	78.72	82.27	80.22	81.31	80.87	81.09
	RoBERTa	85.51	78.24	88.11	82.88	84.83	85.95	85.16
Image	Image	64.76	54.41	70.80	61.53	60.12	73.08	65.97
	ViT	67.83	57.93	70.07	63.43	65.68	71.35	68.40
Text+Image	HFM	83.44	76.57	84.15	80.18	79.40	82.45	80.90
	D&RNet	84.02	77.97	83.42	80.60	-	-	-
	Res-BERT	84.80	77.80	84.15	80.85	78.87	84.46	81.57
	Att-BERT	86.05	78.63	83.31	80.90	80.87	85.08	82.92
	CMGCN	87.55	83.63	84.69	84.16	87.02	86.97	87.00
	Multi-View CLIP	88.33	82.66	88.65	85.55	-	-	-
	MILNet	89.50	85.16	89.16	87.11	88.88	89.44	89.12
	DMSD-CL	88.95	84.89	87.90	86.37	88.35	88.77	88.54
	G ² SAM*	91.07	88.27	90.09	89.17	90.67	90.92	90.78
	MICL (ours)	92.08	90.05	90.61	90.33	91.85	91.77	91.81

Table 1: Main results on MMSD dataset for sarcasm detection. We use * indicates the reproduced results by using RoBERTa as the textual backbone.

3.5 Training and Inference

We obtain the final results based on the fused features:

$$\hat{y} = \mathbf{W}\mathbf{x} + b, \quad (17)$$

where \mathbf{W} and b are learnable parameters. The binary cross-entropy loss is calculated as:

$$\mathcal{L}_{ce} = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})). \quad (18)$$

The final loss function for MICL is defined as the combination of the contrastive learning loss in Eq. (16) and the cross-entropy loss in Eq. (18):

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda \mathcal{L}_{cl}, \quad (19)$$

where λ is hyperparameter.

4 Experiments

4.1 Datasets and Metrics

Our experiments are conducted on the public Multimodal Sarcasm Detection Dataset (MMSD) (Cai et al., 2019). Each entry in this dataset is a text-image pair, categorized into either sarcastic or non-sarcastic examples based on the specific hashtags. The dataset is divided into a training set, a validating set, and a test set, which includes 19,816, 2,410, and 2,409 samples, respectively. Following previous works (Jia et al., 2024), we report the accuracy, precision, recall, F1-score, and macro-average results to measure the model performance.

To further investigate the models’ capability to generalize and their susceptibility to spurious correlations, we meticulously design a small test set, SPMSD. It is refined and expanded from the MMSD dataset, comprising a total of 1,000 samples, including 573 sarcastic items and 427 non-sarcastic items. Detailed information of this dataset can be found in the Appendix A.

4.2 Baseline Models

We compare our proposed model MICL with several baselines, which can be broadly categorized into two groups:

Unimodal Baselines. These methods simply take textual or visual information as input, including: TextCNN (Kim, 2014), Bi-LSTM (Graves and Schmidhuber, 2005), BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) for textual, Image (Cai et al., 2019) and ViT (Dosovitskiy et al., 2020) for visual.

Multimodal Baselines. These methods exploit both visual and textual information as input, including: HFM (Cai et al., 2019), D&RNet (Xu et al., 2020), Res-BERT (Pan et al., 2020), Att-BERT (Pan et al., 2020), CMGCN (Liang et al., 2022), Multi-View CLIP (Qin et al., 2023), MILNet (Qiao et al., 2023), DMSD-CL (Jia et al., 2024) and G²SAM (Wei et al., 2024).

Method	Acc.	Binary-Average			Macro-Average		
		P.	R.	F1	P.	R.	F1
BERT	55.50	66.41	45.20	53.79	57.47	57.26	55.44
RoBERTa	51.30	60.84	22.33	32.66	65.29	22.33	33.28
ResNet	52.30	59.75	51.30	55.21	52.42	52.47	52.10
ViT	53.60	60.42	55.14	57.66	53.27	53.33	53.17
Res-BERT	58.10	66.17	54.97	60.05	58.47	58.63	57.99
Att-BERT	58.30	67.56	52.35	58.99	59.23	59.31	58.29
MILNet	56.20	66.83	46.42	54.79	57.96	57.79	56.10
DMSD-CL	60.60	64.09	71.02	67.38	59.30	58.81	58.82
MICL	68.70	70.70	77.48	73.94	68.01	67.20	67.38

Table 2: Comparison results on SPMSD dataset (%).

4.3 Main Results

The main results are shown in Table 1. Our analysis yields the following insights: **1)** The proposed MICL emerges as the most effective model, outperforming all baseline models. It records improvements ranging from 2.71% to 5.16% over the latest DMSD-CL model across various metrics and consistently surpasses the state-of-the-art model G²SAM in all metrics. **2)** Text-based models demonstrate superior performance over image-based models, with the RoBERTa model achieving an accuracy of 85.51%, compared to only 67.83% by the ViT model. This indicates that text carries a higher information density than images in the multimodal sarcasm detection task. The substantial disparity in performance causes multimodal models to rely excessively on textual data, potentially compromising their ability to generalize. These insights underscore MICL’s proficiency in leveraging multimodal data to achieve exceptional results in the multimodal sarcasm detection task.

4.4 Analysis on SPMSD

We design a comparative experiment on the spurious correlation test set SPMSD, as shown in Table 2. The analysis reveals that, unlike the main experimental results with high recall, most baseline models exhibit lower recall compared to precision. This discrepancy in performance metrics highlights the significant impact of varying data distributions on the decision-making processes of existing models, tentatively affirming the presence of the spurious correlation issue. Notably, the proposed MICL significantly outperforms all baselines, achieving a 68.7% accuracy rate. Specifically, against DMSD-CL, MICL displays a more significant 6.46% to 8.71% improvement across various metrics, which is more significant than that on MMSD. These results demonstrate that MICL can effectively mitigate reliance on spurious correlations, showing better generalization ability on new data.

Base	f_w	f_e	f_s	c	MMSD		SPMSD	
					Acc.(%)	F1(%)	Acc.(%)	F1(%)
✓					88.54	85.73	61.60	63.77
✓	✓				89.97	87.20	62.80	65.46
✓	✓	✓			91.32	89.33	63.70	65.79
✓	✓		✓		90.77	88.84	66.10	69.29
✓	✓	✓	✓		91.45	89.51	67.90	71.23
✓	✓	✓	✓	✓	92.08	90.33	68.70	73.94

Table 3: Experiment results of ablation study.

Method	MMSD	SPMSD	Method	MMSD	SPMSD
MILNet*	89.54	56.70	MILNet	89.50	56.20
+ OCR'	89.50	56.20	+ aug'	89.41	59.00
+ ours	89.66	57.80	+ ours	89.58	65.40
DMSD-CL	88.95	60.60	DMSD-CL*	89.08	57.20
+ OCR'	88.62	59.10	+ aug'	88.95	60.60
+ ours	89.04	60.90	+ ours	89.29	65.30
MICL	91.40	67.40	MICL	91.91	56.90
+ OCR'	90.27	64.80	+ aug'	91.07	59.20
+ ours	92.08	68.70	+ ours	92.08	68.70

Table 4: Results of using different extra data (Acc %). * MILNet removes the OCR module, DMSD-CL removes the data augmentation module.

4.5 Ablation Study

Analysis of components. To probe the effectiveness of each component in MICL, we conduct ablation experiments. The experimental results are shown in Table 3, where *Base* represents the direct concatenation of H_v and H_t for prediction. f_w , f_e , and f_s correspond to the token-patch, entity-object, and sentiment incongruity learning modules, respectively. c represents multi-view fusion using credibility. According to the results, we have the following findings: **1)** All incongruity learning modules can improve performance compared to the base model. **2)** f_s effectively improves the model’s performance on the SPMSD dataset, reducing erroneous dependence on the text. **3)** f_e significantly improves performance on the MMSD dataset, proving that entity-object incongruity is crucial in the MSD task. **4)** c can effectively integrate features from different views and improve performance.

Analysis of extra data. From a data perspective, we conduct another set of ablation experiments to validate the efficacy of our OCR-text and data augmentation. The results are shown in Tables 4, where *ours* refer to the OCR-text and data augmentation proposed in this paper, *OCR'* represents the OCR-text extracted by Pan et al. (2020), and *aug'* refers to the data augmentation of DMSD-CL. The analysis yields several key insights: **1)** Additional data does not necessarily enhance model performance. In some instances, it may even impair the model’s effectiveness due to distributional

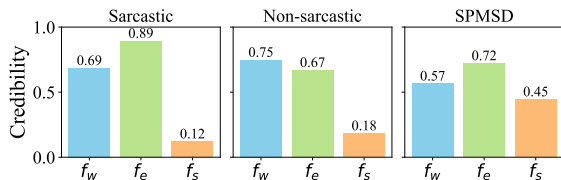


Figure 5: Credibility study.

differences from the original data. For example, MILNet+aug’ slightly improves performance on the SPMSD dataset but causes a decrease on the MMSD dataset. 2) Our OCR-text can enhance the models’ performance. All methods show better results on both benchmark datasets. 3) Our novel data augmentation approach improves model robustness in spurious correlation scenarios without compromising the baseline performance. These findings collectively affirm the effectiveness of the OCR-text and data augmentation devised in our study.

4.6 Credibility Study

To investigate the credibility of incongruity features from different perspectives in various scenarios, we conduct a credibility study, with the results shown in Figure 5. We divide the study into three scenarios: the sarcastic and non-sarcastic scenarios of the MMSD dataset, and the SPMSD scenario. We calculate and display the average credibility of each feature. The experimental results show that entity-object incongruity exhibits high credibility for sarcastic samples, indicating that this view is effective in capturing sarcastic entity information. Conversely, traditional token-patch incongruity effectively detects non-sarcastic samples. Moreover, sentiment incongruity is beneficial in reducing the model’s dependence on spurious correlations. In addition, the credibility of each view is relatively balanced on SPMSD. Therefore, the components of our multi-view incongruity learning method complement each other across different scenarios, demonstrating effective mitigation of spurious correlation issues.

4.7 Case Study

To provide an intuitive comprehension of MICL on spuriously correlated samples, we design a case study. Based on empirical summaries, we present four types of spuriously correlated samples and compare the results of MILNet, DMSD-CL and MICL, as shown in Figure 6. In case 1, the focus is






Text-image Pair	MILNet	DMSD-CL	MICL
 i just love setting an example by getting to work before everyone else . waking up at 3:30 am is so exciting . i just hate setting an example by getting to work before everyone else . waking up at 3:30 am is so frustrating .	Sarcastic ✓ Sarcastic ✗	Sarcastic ✓ Non-Sarcastic ✓	Sarcastic ✓ Non-Sarcastic ✓
 proud to say i've received an offer from my hometown team , indiana university of pennsylvania ! emoji_53 thrilled to announce i've been approached by the local legend, indiana university of pennsylvania !!	Non-Sarcastic ✓ Non-Sarcastic ✓	Non-Sarcastic ✓ Sarcastic ✗	Non-Sarcastic ✓ Non-Sarcastic ✓
iterally loving the weather today . can't wait to get out there and start filming . really . let's go . now .  	Sarcastic ✓ Sarcastic ✗	Sarcastic ✓ Sarcastic ✗	Sarcastic ✓ Non-Sarcastic ✓
 happy new year , everyone ! Legend: Text + Image Image Only Text Only	Sarcastic ✓ Non-Sarcastic ✗ Non-Sarcastic ✓	Sarcastic ✓ Non-Sarcastic ✗ Sarcastic ✗	Sarcastic ✓ Sarcastic ✓ Non-Sarcastic ✓

Figure 6: Case studies on spuriously correlated samples.

mainly on the particular emotional words in the text. Case 2 investigates the impact of modifying non-critical information. Case 3 examines whether models can handle situations where the image and text are congruent. Case 4 examines whether models can correctly handle unimodal inputs. The results show that MILNet struggles with most spurious correlation scenarios (case 1, 3, and 4), showing obvious over-focusing on the text modality. DMSD-CL can handle scenarios involving emotive words (case 1), but it also has modality learning bias (case 3 and 4). In addition, DMSD-CL makes mistakes in learning key textual content (case 2). Therefore, the problem of spurious correlations strongly affects the model’s generalizability. Meanwhile, the proposed MICL, through data augmentation and multi-view incongruity learning, can detect sarcasm properly in various scenarios, emphasizing its generalizability and superiority in MSD.

5 Conclusion

In this paper, we introduce MICL, an innovative approach that leverages contrastive learning to learn multi-view incongruities. This method is designed to counteract the prevalent issue of spurious correlations observed in current MSD models. Furthermore, we tackle the challenge of models’ excessive dependence on textual data by integrating a comprehensive text-image data augmentation scheme. To empirically highlight the problem of spurious correlations, we introduce a test set, SPMSD, built upon the foundational MMSD dataset. Experimental results show that MICL not only achieves state-of-the-art performance on the MSD task but also effectively mitigates spurious correlations.

6 Limitation

Although MICL can reduce the dependence on spurious correlations, it achieves only a 68% accuracy rate on the SPMSD dataset, indicating still substantial scope for further enhancement. Our empirical experiments and existing literature (Wang et al., 2024) show that some spurious correlations can improve model performance, which is a point not discussed in this paper. Additionally, MICL’s complexity, particularly with integrating hybrid attention and graph attention networks, may pose challenges in scalability and efficiency.

Acknowledgments

This research was supported by the National Key R&D Program of China (No. 2023YFC3303800). Hao Peng was supported by NSFC through grant 62322202. Prof. Philip S. Yu was supported in part by NSF under grants III-2106758, and POSE-2346158.

References

- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.
- 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.
- Erik Cambria, Xulang Zhang, Rui Mao, Melvin Chen, and Kenneth Kwok. 2024. Senticnet 8: Fusing emotion ai and commonsense ai for interpretable, trustworthy, and explainable affective computing. In *International Conference on Human-Computer Interaction (HCII)*.
- Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcasm in twitter and amazon. In *Proceedings of the fourteenth conference on computational natural language learning*, pages 107–116.
- Yihe Deng, Yu Yang, Baharan Mirzasoleiman, and Quanquan Gu. 2024. Robust learning with progressive data expansion against spurious correlation. *Advances in Neural Information Processing Systems*, 36.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.
- DI Hernández Farias and Paolo Rosso. 2017. Irony, sarcasm, and sentiment analysis. In *Sentiment Analysis in Social Networks*, pages 113–128. Elsevier.
- Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.
- Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. 2021. Trusted multi-view classification. In *International Conference on Learning Representations*.
- 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.
- Aditya Joshi, Vinita Sharma, and Pushpak Bhat-tacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762.
- Arpit Khare, Amisha Gangwar, Sudhakar Singh, and Shiv Prakash. 2023. Sentiment analysis and sarcasm detection in indian general election tweets. In *Research Advances in Intelligent Computing*, pages 253–268. CRC Press.
- Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.
- Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. 2023. Last layer re-training is sufficient for robustness to spurious correlations. *ICLR 2023*.
- 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)*, volume 1, pages 1767–1777. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Zihan Ma, Minnan Luo, Hao Guo, Zhi Zeng, Yiran Hao, and Xiang Zhao. 2024. Event-radar: Event-driven multi-view learning for multimodal fake news detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5809–5821.
- Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12700–12710.
- Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1383–1392.
- MSM Prasanna, SG Shaila, and A Vadivel. 2023. Polarity classification on twitter data for classifying sarcasm using clause pattern for sentiment analysis. *Multimedia Tools and Applications*, 82(21):32789–32825.
- Yang Qiao, Liqiang Jing, Xuemeng Song, Xiaolin Chen, Lei Zhu, and Liqiang Nie. 2023. Mutual-enhanced incongruity learning network for multi-modal sarcasm detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9507–9515.
- Libo Qin, Shijue Huang, Qiguang Chen, Chenran Cai, Yudi Zhang, Bin Liang, Wanxiang Che, and Ruifeng Xu. 2023. Mmsd2. 0: Towards a reliable multi-modal sarcasm detection system. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10834–10845.
- Rossano Schifanella, Paloma De Juan, Joel Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1136–1145.
- P Veličković, A Casanova, P Liò, G Cucurull, A Romero, and Y Bengio. 2018. Graph attention networks.
- Wei Han Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogvlm: Visual expert for pretrained language models. *Preprint*, arXiv:2311.03079.
- Yifei Wang, Jizhe Zhang, and Yisen Wang. 2024. Do generated data always help contrastive learning? *arXiv preprint arXiv:2403.12448*.
- Yiwei Wei, Shaozu Yuan, Hengyang Zhou, Longbiao Wang, Zhiling Yan, Ruosong Yang, and Meng Chen. 2024. G²sam: Graph-based global semantic awareness method for multimodal sarcasm detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 9151–9159.
- 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.
- Zixin Wen and Yuanzhi Li. 2021. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pages 11112–11122. PMLR.
- Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. 2022. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *International Conference on Machine Learning*, pages 24043–24055. PMLR.
- Tao Xiong, Peiran Zhang, Hongbo Zhu, and Yihui Yang. 2019. Sarcasm detection with self-matching networks and low-rank bilinear pooling. In *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.
- Yingguang Yang, Qi Wu, Buyun He, Hao Peng, Renyu Yang, Zhifeng Hao, and Yong Liao. 2024. Sebot: Structural entropy guided multi-view contrastive learning for social bot detection. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3841–3852.
- Sriram Yenamandra, Pratik Ramesh, Viraj Prabhu, and Judy Hoffman. 2023. Facts: First amplify correlations and then slice to discover bias. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4794–4804.
- Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Tweet sarcasm detection using deep neural network. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: technical papers*, pages 2449–2460.

A Dataset

The statistics of MMSD dataset is as shown in Table 6.

The SPMSD dataset is derived and expanded from the MMSD dataset, specifically designed to evaluate the models’ reliance on spurious correlations. To ensure the fairness of the dataset, we randomly select 1,000 samples from the MMSD dataset and use these samples as the basis for constructing the SPMSD dataset. We employ various strategies to construct SPMSD, aiming to obtain a wide range of potential spurious correlations. These strategies include transforming the sentiment of the text, only describing the content of the image in the text, replacing entities in the text with entities appearing in the image, regenerating sarcastic text based on the image, swapping text-image pairs, and using only image/text.

B Estimating Credibility

In the context of multi-class classification, Subjective logic (SL) associates the parameters of the Dirichlet distribution. Subjective logic defines a theoretical framework for obtaining the probabilities of different categories (belief masses) and the overall uncertainty (uncertainty mass) of multi-classification problems based on *evidence* collected from the data. Specifically, for the K classification problems, subjective logic tries to assign a belief mass to each class label and an overall uncertainty mass to the whole frame based on the *evidence*. Accordingly, for the v -th view, the $K + 1$ mass values are all non-negative and their sum is one:

$$u^v + \sum_{k=1}^K b_k^v = 1, \quad (20)$$

where $u_k^v \geq 0$ and $b_k^v \geq 0$ indicate the overall uncertainty and the probability for the k -th class, respectively.

For the v -th view, subjective logic connects the *evidence* $e^v = [e_1^v, \dots, e_K^v]$ to the parameters of the Dirichlet distribution $\alpha^v = [\alpha_1^v, \dots, \alpha_K^v]$. Specifically, the parameter α_k^v of the Dirichlet distribution is induced from e_k^v , *i.e.*, $\alpha_k^v = e_k^v + 1$. Then, the belief mass b_k^v and the uncertainty u^v are computed as:

$$b_k^v = \frac{e_k^v}{S_v} = \frac{\alpha_k^v - 1}{S_v}, \quad u^v = \frac{K}{S_v}, \quad (21)$$

where $S_v = \sum_{i=1}^K (e_i^v + 1) = \sum_{i=1}^K \alpha_i^v$ is the Dirichlet strength. We follow the work of Ma et al.

(2024) and simply use 1 minus the uncertainty u^v to estimate the credibility of each view, that is:

$$\begin{aligned} c^v &= 1 - u^v \\ &= \sum_{k=1}^K b_k^v \\ &= b_0^v + b_1^v \\ &= \frac{e_0^v}{S_v} + \frac{e_1^v}{S_v} \\ &= \frac{e_0^v + e_1^v}{(e_0^v + 1) + (e_1^v + 1)}. \end{aligned} \quad (22)$$

C Experiments Compared with LVLMs

Large Vision-Language Models (LVLMs) have demonstrated remarkable results across various multimodal tasks. We compare the performance of MICL with existing LVLMs on the MSD task, and the results are presented in Table 5. It can be seen that without fine-tuning most LVLMs do not reach the performance of mainstream methods on the MMSD and SPMSD datasets. However, ChatGPT-4’s accuracy on the SPMSD dataset is slightly higher than that of MICL.

D Experiments on Different Backbones

To ensure a fair comparison of results, we standardize the text encoder of all models to BERT and conduct experiments on the MMSD dataset. The results are presented in Table 7. As shown in the table, our MICL model continues to achieve the best performance.

E Attention Visualization

To intuitively demonstrate the concerns of token-patch incongruity and entity-object incongruity learning, we conduct attention visualization experiments, using sub-modules with text as *Query* and images as *Key* and *Value*. Figure 7 shows that in the sarcastic examples, both methods can focus on the key parts. In non-sarcastic examples, the two methods are complementary properties to learn features more comprehensively.

F Implementation Details

We use the pre-trained RoBERTa-base⁴ model for text encoding and the pre-trained vit-base-patch32-

⁴<https://huggingface.co/FacebookAI/roberta-base>

Method	MMSD			SPMSD		
	Acc(%)	Binary-F1(%)	Macro-F1(%)	Acc(%)	Binary-F1(%)	Macro-F1(%)
MiniCPM-V 2.0	55.95	43.59	53.73	53.30	46.75	52.58
LLaVA 1.6	60.23	46.42	57.40	48.80	44.59	48.50
VisualGLM	60.81	44.66	41.03	60.80	58.41	48.33
Qwen-VL-Chat	45.08	27.01	43.27	44.20	38.83	45.08
mPLUG-Owl 2	59.40	34.62	52.59	47.90	33.16	45.30
ChatGPT 4	76.11	74.75	76.01	70.20	66.37	64.21
MICL(ours)	92.08	90.33	91.81	68.70	73.94	67.38

Table 5: Additional experimental results with LVLMs.

Label	Train	Val	Test
Positive	8642	959	959
Negative	11174	1451	1450
All	19816	2410	2409

Table 6: Statistics of MMSD.

Method	Acc	Binary-Average			Macro-Average		
		P	R	F1	P	R	F1
BERT	83.85	78.72	82.27	80.22	81.31	80.87	81.09
Res-BERT	84.80	77.80	84.15	80.85	78.87	84.46	81.57
Att-BERT	86.05	78.63	83.31	80.90	80.87	85.08	82.92
MILNet	88.72	84.97	87.79	86.37	87.75	88.29	88.04
DMSD-CL	88.24	86.47	84.42	85.43	87.65	87.94	87.79
G ² SAM	90.48	87.95	89.02	88.48	89.44	89.79	89.65
MICL(ours)	91.36	89.48	88.84	89.16	90.90	90.80	90.85

Table 7: Additional experimental results with BERT text encoder.

224⁵ model for image encoding. For textual graph, we use the `en_core_web_trf` model in `spacy` to extract dependencies between entities. For visual graphs, we add an edge between regions with cosine similarity > 0.6 . We use `gpt3.5-turbo` for text data augmentation. For image data augmentation, we extract the original image content with GLM-4V and complete the text-to-image and image-to-image steps using `stable diffusion`⁶. We set the feature dimension d to 768, and set the hyperparameters τ and λ to 0.07 and 1, respectively. We use the Adam optimizer to optimize our model. The learning rate is set to $1e-5$ for all components. The learning rate is reduced to 0 in the line schedule. All experiments are completed under a single Nvidia RTX 4090 (24 G).

⁵<https://huggingface.co/google/vit-base-patch32-224-in21k>

⁶<https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0>

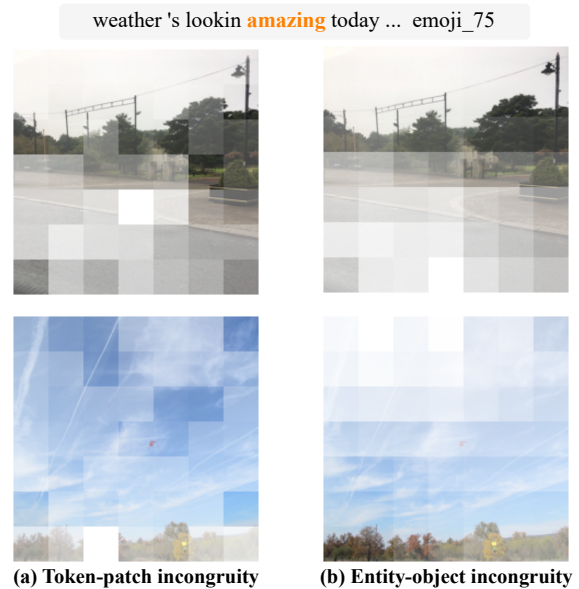


Figure 7: Attention Visualization.

G Prompts

Prompts for OCR. Please perform OCR on this image and translate any non-English text into English.

Prompts for Text Augmentation. Please rewrite these data from three aspects: 1. Reverse the meaning of sarcasm: that is, if the sarcasm item of the original sarcasm data is yes, please rewrite the original text into a sentence that does not contain sarcasm at all; if the sarcasm item of the original sarcasm data is no, please use a strong sarcasm emotion rewrite text; 2. Keep the sarcasm meaning: keep the sarcasm items of the original data unchanged, introduce some new concepts, and rewrite them.

Prompts for Image Captioning. Please describe the main content of this image in one sentence.

H OCR-text Examples

We give more OCR-text examples, as shown in Figure 8. Our approach can handle handwriting, comics, non-English, and photos.

Original data	OCR-text
	fighting the good fight. finally a rally i can get behind. stop premature christmas decorating.
	what's the wifi password here? respect the dead. all small letters?
	con el perro peluchon bolinha mi amigo ↓ Translate with the stuffed dog bolinha my friend
	revealing india's true history, hidden so far by pseudo-secular anti-nationals

Figure 8: OCR-text examples.

I Data Augmentation Examples

We give more data augmentation examples, as shown in Figure 9.

Original data	Augmented data
blocking out the haters	blocking out the haters
first "real" food of the day . yum . what a treat .	The first "actual" meal of the day. Delicious! What a delight!
that Vs it , trump . call the pope "disgraceful ." you 're guaranteed to make people happy with that remark .	That's it, Trump. Compliment the pope. You're bound to upset people with that comment.
pretty warm out ... but rainy and dark . gotta love this buffalo weather ... rain vapeclouds darkskys vape	It's quite warm outside, though wet and gloomy. One has to appreciate this Buffalo-style weather with its rainy vape clouds and overcast skies.

Figure 9: Data augmentation examples.