

Incongruity-aware Tension Field Network for Multi-modal Sarcasm Detection

Jiecheng Zhang^{1,2,3} C.L. Philip Chen^{1,2,3} Shuzhen Li² Tong Zhang^{* 1,2,3}

¹Guangdong Provincial Key Laboratory of Computational AI Models and Cognitive Intelligence, School of Computer Science & Engineering, South China University of Technology

²Pazhou Lab, Guangzhou, China

³Engineering Research Center of the Ministry of Education on Health Intelligent Perception and Paralleled Digital-Human, Guangzhou, China

{jiechengzhang.alex, melody13szli}@gmail.com, {philipchen, tony}@scut.edu.cn

Abstract

Multi-modal sarcasm detection (MSD) identifies sarcasm and accurately understands users' real attitudes from text-image pairs. Most MSD researches explore the incongruity of text-image pairs as sarcasm information through consistency preference methods. However, these methods prioritize consistency over incongruity and blur incongruity information under their global feature aggregation mechanisms, leading to incongruity distortions and model misinterpretations. To address the above issues, this paper proposes a pioneering inconsistency preference method called incongruity-aware tension field network (ITFNet) for multi-modal sarcasm detection tasks. Specifically, ITFNet extracts effective text-image feature pairs in fact and sentiment perspectives. It then constructs a fact/sentiment tension field with discrepancy metrics to capture the contextual tone and polarized incongruity after the iterative learning of tension intensity, effectively highlighting incongruity information during such inconsistency preference learning. It further standardizes the polarized incongruity with reference to contextual tone to obtain standardized incongruity, effectively implementing instance standardization for unbiased decision-making in MSD. ITFNet performs well in extracting salient and standardized incongruity through an incongruity-aware tension field, significantly tackling incongruity distortions and cross-instance variance. Moreover, ITFNet achieves state-of-the-art performance surpassing LLaVA1.5-7B with only 17.3M trainable parameters, demonstrating its optimal performance-efficiency in multi-modal sarcasm detection tasks.

1 Introduction

Sarcasm is a widely used implicit expression in which the real attitude conflicts with the literal meaning (Gibbs, 1986). The incongruity between

* Corresponding authors.

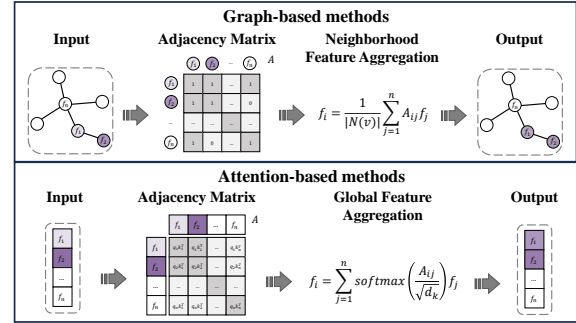


Figure 1: Schematic diagram of incongruity distortion. Here, the color difference reflects the degree of incongruity between features (f_1, f_2). Graph-based methods average aggregate neighbor features through their connected edge. Attention-based methods aggregate global features through their attention weights. The outputs illustrate the blurring of incongruity information compared to the inputs.

real attitude and literal meaning is a crucial clue for identifying sarcastic intent (Joshi et al., 2015). Multi-modal sarcasm detection has extensive applications in social media monitoring and management, intelligent interactive systems, news dissemination, public opinion analysis, etc. With the development of social media, Multi-modal sarcasm detection has garnered significant attention in academic research and application (Kolchinski and Potts, 2018; Desai et al., 2022).

The key challenge of multi-modal sarcasm detection is mining feature incongruities across various modalities to identify sarcastic intentions. Existing methods can be mainly divided into graph-based methods and attention-based methods. Graph-based methods use prior knowledge to build static graph structures and mine the semantics of graph nodes through information aggregation to model incongruities. For example, CMGCN (Liang et al., 2022) built cross-modal graphs to extract sentiment incongruities, and InCrossMGs (Liang et al., 2021) used a hierarchical graph to extract multi-modal incongruities. However, graph-based meth-

ods may distort features’ inherent incongruity due to the similarity graph and the neighbor node aggregation, as shown in Figure 1. Attention-based methods learn the dependencies between features and abstract high-level semantics to model incongruities. For example, Cai et al. (2019) utilized cross-attention to guide hierarchical modality fusion, and (Xu et al., 2020) modeled semantic associations by cross-attention. However, attention-based methods may harm features’ inherent incongruity due to the global feature aggregation based on the similarity attention weights, as shown in Figure 1. To sum up, graph-based and attention-based methods use similarity operations between feature points to model implicit incongruities, leading to incongruity distortion problems and misinterpretation performance in multi-modal sarcasm detection. Therefore, we collectively refer to them as consistency preference learning methods. This work focuses on an incongruity-specific learning method for multi-modal sarcasm detection tasks.

This paper proposes an Incongruity-aware Tension Field Network (ITFNet) to implement an inconsistency preference learning through an incongruity-aware tension field based on discrepancy metrics for addressing the incongruity distortion. Specifically, ITFNet utilizes a fact-sentiment multi-task feature learning module to extract fact and sentiment features from text-image data. In the fact/sentiment features branch, ITFNet obtains polarized incongruity and contextual tones through iterative learning with tension intensity. It then standardizes the incongruity with reference to contextual tone. ITFNet effectively addresses the incongruity distortion and cross-instance variance for efficient performance of multi-modal sarcasm detection tasks. The state-of-the-art performance on the widely-used dataset demonstrates the superiority of ITFNet in multi-modal sarcasm detection tasks.

The main contributions of our paper can be summarized as follows:

- To our knowledge, ITFNet is a pioneering inconsistency preference method for multi-modal sarcasm detection tasks. It implements incongruity representation learning via the incongruity-aware tension field based on discrepancy metrics, tackling incongruity distortions inherent in consistency preference methods.
- This paper designs an innovative tension field for the salient and standardized incongruity ex-

traction. It iteratively extracts contextual tone and polarized incongruity and standardizes them into standardized incongruity, mitigating cross-instance variance in MSD tasks.

- ITFNet achieves superior accuracy 3.38% over LLaVA1.5-7B with only 17.3M trainable parameters. The SOTA performance demonstrates optimal performance-efficiency in multi-modal sarcasm detection tasks.

2 Related work

Multi-modal sarcasm detection.

Some multi-modal sarcasm detection methods distinguish sarcastic samples from non-sarcastic ones based on sample-level feature differences. DIP (Wen et al., 2023) leverages a Gaussian distribution to model uncertain correlations. KnowleNet (Yue et al., 2023) detects sarcasm by assessing semantic similarity at the sample level. DMSD-CL (Jia et al., 2024) constructs sarcastic–non-sarcastic sample pairs and employs contrastive learning to detect sarcasm. G2SAM (Wei et al., 2024) utilizes graph contrastive learning for sarcasm detection. MICL (Guo et al., 2025) integrates multi-view incongruities via contrastive learning for multi-modal sarcasm detection.

The main-stream multi-modal sarcasm detection methods mine fine-grained feature-level incongruities to detect and understand sarcasm. Graph-based methods can use prior knowledge to construct static similarity graphs to extract features. CMGCN (Liang et al., 2022) built cross-modal graphs to extract sentiment incongruities. InCrossMGs (Liang et al., 2021) used distinct GCNs to extract multi-modal incongruities. Attention-based methods can extract task-relevance features through similarity attention weights. HFM (Cai et al., 2019) utilized cross-attention to guide hierarchical modality fusion. D&RNet (Xu et al., 2020) modeled semantic associations by cross-attention. Att-Bert (Pan et al., 2020) and Multi-View CLIP (Qin et al., 2023) employed distinct attention to extract multi-modal incongruities. FSICN (Lu et al., 2024), AMIF (Li et al., 2025), DynRT-Net (Tian et al., 2023), and MuMu (Wang et al., 2024b) also use an attention variant mechanism to detect incongruity.

The differences and advantages of ITFNet and the above methods are as follows. (1) Most of the above methods model incongruity through consistency preference methods, leading to incongruity

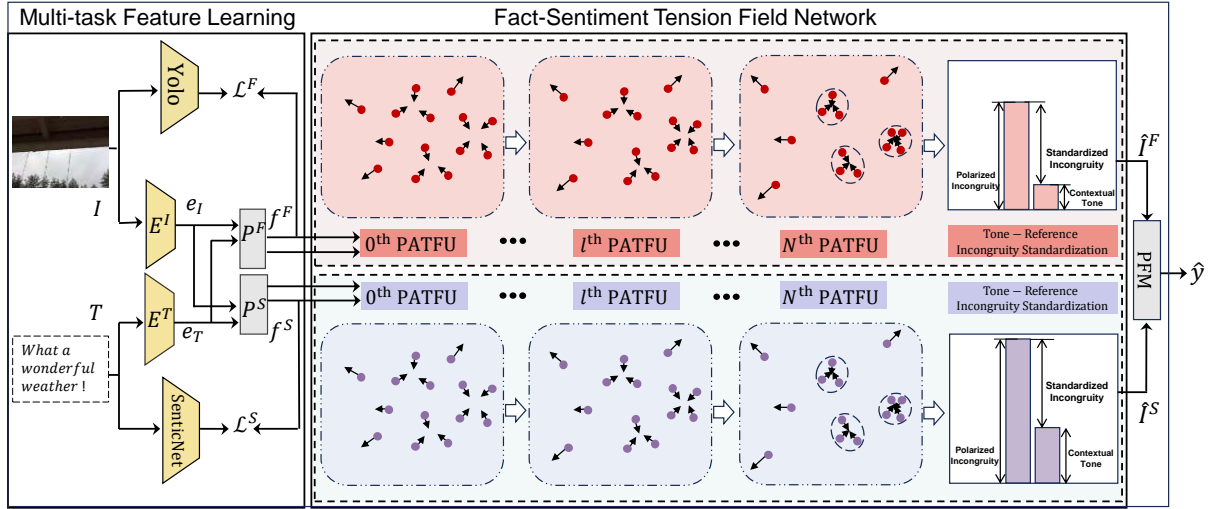


Figure 2: Overview of our framework. (I) Multi-task feature learning module extracts fact and sentiment features for the following fact-sentiment tension field network; (II) Fact/Sentiment tension field performs inconsistency preference learning to capture polarized incongruity and contextual tone. It then standardizes polarized incongruity with reference to contextual tone for each instance to obtain standardized incongruity. (III) Perspective fusion module (PFM) integrates the standardized incongruities of fact and sentiment perspectives for prediction.

distortion, as shown in Figure 1. ITFNet is an inconsistency preference method that specifically captures incongruities between features through an incongruity-aware tension field based on discrepancy metrics. (2) ITFNet uses tension intensity as a discrepancy metric, considering comprehensive discrepancy. The tension intensity between a feature pair is defined as follows:

$$\begin{aligned}
 \mathcal{T}_{f_i, f_j} &= (f_i - f_j)^2 \\
 &= \|f_i\|_2 + \|f_j\|_2 - 2\|f_i\|_2\|f_j\|_2\cos(f_i, f_j) \\
 &= (\|f_i\|_2 + \|f_j\|_2 - 2\|f_i\|_2\|f_j\|_2) + 2\|f_i\|_2\|f_j\|_2(1 - \cos(f_i, f_j)) \\
 &= \underbrace{(\|f_i\|_2 - \|f_j\|_2)^2}_{\textcircled{1}} + \underbrace{2\|f_i\|_2\|f_j\|_2(1 - \cos(f_i, f_j))}_{\textcircled{2}},
 \end{aligned} \tag{1}$$

where i, j represent the i -th and j -th features. The tension intensity measures the degree of incongruity between the feature points. The first term is the square of the feature modulus difference and reflects the sensitivity of feature scale discrepancy. Large-scale features often refer to extreme facts or intense sentiment, an important component of satirical expressions with a sense of absurdity. The second term reflects the sensitivity of feature angle discrepancy. Features with large angle differences may indicate a sarcastic expression of positive and negative contrast.

3 Proposed method

Incongruity-aware Tension Field Network (ITFNet) is proposed to implement inconsistency preference learning via the incongruity-aware tension field

based on discrepancy metrics for addressing inconsistency distortions and obtaining efficient performance in multi-modal sarcasm tasks. The framework of ITFNet is shown in Figure 2. ITFNet is mainly composed of three procedures: 1) a fact-sentiment multi-task feature learning module extracts fact and sentiment features for the following fact-sentiment tension field network; 2) in fact/sentiment feature branch, a tension field performs inconsistency preference learning to capture polarized incongruity and contextual tone; it then standardizes polarized incongruity with reference to contextual tone for each instance to obtain standardized incongruity; 3) a perspective fusion module (PFM) integrates the standardized incongruities of fact and sentiment perspectives for prediction. The details are as follows.

3.1 Fact-Sentiment Multi-task Feature Learning Module

Fact incongruity refers to the incongruity between factual semantic information in multi-modal data. For instance, the description "Bombay Dyeing just found an accessory for their 6x6 bedsheets collection" for a woman in a floor-length dress creates a fact incongruity. The metaphorical comparison to a bedsheet satirizes the impracticality and extravagance of the dress.

Sentiment incongruity refers to the incongruity between dissatisfaction and a positive manner. For

instance, the statement "What a wonderful day!" alongside an image of a rainy day produces a sentiment incongruity. The word "wonderful" expresses a positive sentiment, but the rain in the image shows a negative impression.

The fact-sentiment multi-task feature learning module leverages multi-task learning to extract fact and sentiment text-image features for the following fact-sentiment tension field network. It mainly consists of fact text-image feature extraction and sentiment text-image feature extraction.

Given a sample (T, I) includes a text and an image, the text encoder maps T into word-level embeddings $\mathbf{e}_T \in \mathcal{R}^{n \times d}$ and the image encoder transforms I into patch-level embeddings $\mathbf{e}_I \in \mathcal{R}^{m \times d}$, where d is the feature dimension. n and m stand for feature channels, and we denote $(n + m)$ as r .

Fact Text-Image Feature Extraction obtains refined fact text-image features via Yolo-task. First, the fact projection layer \mathcal{P}^F projects the image embeddings \mathbf{e}_I into fact image features \mathbf{f}_I^F and projects the text embeddings \mathbf{e}_T into fact text features \mathbf{f}_T^F , where the superscript capital letter F stands for fact perspective here and in the following context.

To highlight the factual properties of this multi-modal fact feature space $\mathcal{S}^F \in \mathcal{R}^{r \times d}$ composed of fact text-image features, we utilize the object detection model Yolo to annotate the image and construct a Yolo-task for the fact projection layer. Specifically, Yolo queries the input image and then gets the pseudo labels $\mathbf{y}^F \in [0, 1]^k$, where k is the number of object categories that Yolo can detect inherently. We use the fact image features to predict the pseudo labels, and the binary cross-entropy loss is formalized as follows:

$$\mathcal{L}^F = - \sum_{i=1}^k (\mathbf{y}_i^F \log g(\mathbf{f}_I^F)_i + (1 - \mathbf{y}_i^F) \log(1 - g(\mathbf{f}_I^F)_i)), \quad (2)$$

where $g(\cdot)$ represents the fully connected layer and i represents the i -th object category.

Sentiment Text-Image Feature Extraction obtains effective sentiment text-image features via SenticNet-task. First, the sentiment projection layer \mathcal{P}^S projects the text embeddings \mathbf{e}_T into sentiment text features \mathbf{f}_T^S and projects the image embeddings \mathbf{e}_I into sentiment image features \mathbf{f}_I^S , where the superscript capital letter S stands for sentiment perspective here and in the following context.

To highlight the sentimental properties of this multi-modal sentiment feature space $\mathcal{S}^S \in \mathcal{R}^{r \times d}$

composed of sentiment text-image features, we similarly utilize the widely used sentiment dictionary SenticNet to construct a SenticNet-task for the sentiment projection layer. Specifically, it queries the input text for a continuous sentiment polarity vectors $\mathbf{y}^S \in [-1, 1]^n$, where n denotes the number of tokens and $y_i^S = 0$ indicates no sentiment match. We use the sentiment text features to predict the sentiment polarity vectors, and the mean squared error is formalized as follows:

$$\mathcal{L}^S = \frac{1}{n} \sum_{i=1}^n (y_i^S - g(\mathbf{f}_T^S)_i)^2, \quad (3)$$

where $g(\cdot)$ represents the fully connected layer and i represents the i -th word.

3.2 Fact-Sentiment Tension Field Network

Most multi-modal sarcasm detection methods prioritize consistency over incongruity and suffer from incongruity distortion and misjudgment of incongruity. Meanwhile, variations in sarcasm basic strength across instances due to the context further hinder reliable multi-modal sarcasm detection with incongruity information.

In the gravitational field's differential force mechanism, weaker tidal gradients facilitate clustering, while strong gradients induce repulsion (Brill and Wheeler, 1957). Inspired by this concept, this paper designs an incongruity-aware tension field based on discrepancy metrics to implement inconsistency preference learning for addressing incongruity distortion and obtain standardized incongruity for multi-modal sarcasm detection. The tension field comprises a polarization-aggregation tension field block and a tone-reference incongruity standardization module.

For incongruity learning of fact and sentiment perspectives, we show the details of the tension field in the following with a multi-modal feature space $\mathcal{S} \in \mathcal{R}^{r \times d}$ is either multi-modal fact feature space \mathcal{S}^F or multi-modal sentiment feature space \mathcal{S}^S . Any pair of features in multi-modal feature space \mathcal{S} is denoted as (f_i, f_j) , where i and j represent the i -th and j -th feature.

3.2.1 Polarization-Aggregation Tension Field Block

Polarization-aggregation tension field block performs inconsistency preference learning via iteration of N polarization-aggregation tension field units (PATFUs) to capture significant sarcasm infor-

mation for the following tone-reference incongruity standardization.

Polarization-Aggregation Tension Field Unit.

To quantify the incongruity information of sarcasm inherent across all features, it is essential to construct a tension field. PATFU calculates the tension intensity between all pairs of features using the discrepancy metrics. The tension intensity between features f_i and f_j in the tension intensity matrix T is formalized as follows:

$$T_{i,j} = (f_i - f_j)^2. \quad (4)$$

This tension intensity forms the foundation of the tension field. A higher tension intensity indicates a greater degree of incongruity between features f_i and f_j , while a lower tension intensity implies that the features are relatively consistent.

The degree of attraction and repulsion between features needs to be determined based on their tension intensity. PATFU maps the tension intensity matrix into an interaction weight matrix W . The interaction weight between features f_i and f_j is calculated by:

$$W_{ij} = \frac{e^{-T_{ij}}}{\sum_{k=1}^r e^{-T_{ik}}}. \quad (5)$$

At this stage, PATFU negatively modulates the tension intensity. Hence, high weights indicate that the feature pair is relatively consistent and has a strong attraction, while low weights do the opposite.

Next, the composite effect can be calculated according to the interactions across all features. The resulting representation \hat{f}_i for feature f_i under the composite effect is :

$$\hat{f}_i = f_i + \sum_{j=0}^r W_{ij} f_j. \quad (6)$$

The second term calculates the composite effect of all features on feature f_i . It reflects the deviation direction and intensity of feature f_i compared to the original. Thus, the interactions across all features lead to the following representation set:

$$F = \{\hat{f}_1, \dots, \hat{f}_i, \dots, \hat{f}_r\}. \quad (7)$$

The tension field forms a new distribution under the composite effect of each feature. Consistent features tend to cluster together, with each cluster's prototype representing contextual tone information. In contrast, incongruous features are polarized to

reveal significant incongruity information and alleviate the incongruity distortion.

Given that the tension field may exhibit complex incongruity patterns that are crucial to capturing the subtle nuances of sarcasm, PATFU uses a nonlinear field space transformation to effectively model and resolve these complex feature relationships, which is formalized as:

$$\hat{F} = g(F), \quad (8)$$

where $g(\cdot)$ represents the fully connected layer.

Iterations. Iteration of N PATFUs perform inconsistency preference learning to obtain polarized incongruity \mathbf{I} and contextual tone \mathbf{C} . The explicit capture is formalized as:

$$\mathbf{I} = h(f_i^*, f_j^*) \Big|_{\arg \max_{(f_i^*, f_j^*) \in \mathcal{S} \times \mathcal{S}} T_{ij}} \quad (9)$$

$$\mathbf{C} = \sum_{i=1}^r f_i / r, \quad (10)$$

where (f_i^*, f_j^*) is the feature pair with the highest tension intensity in the multi-modal feature space \mathcal{S} after the iteration. i stands for the i -th feature. The polarized incongruity is the fusion result of the feature pair through the star operator $h(\cdot, \cdot)$ as a key local sarcasm clue (Ma et al., 2024). It measures the most significant degree of incongruity at the feature level. The contextual tone represents the instance-specific global factual semantics or sentiment. It measures the differences in sarcasm's basic strength across instances and provides a reference for incongruity standardization.

3.2.2 Tone-Reference Incongruity Standardization

Tone-reference incongruity standardization standardizes significant incongruity to solve the problem of sarcasm basic strength differences between instances and enable a fair comparison of incongruity information for generalized multi-modal sarcasm detection. Specifically, we standardize polarized incongruity with reference to contextual tone for each instance through adaptive learning by the neural network, which is formalized as:

$$\hat{\mathbf{I}} = g((\mathbf{I}, \mathbf{C})), \quad (11)$$

where $\hat{\mathbf{I}}$ is the standardized incongruity, (\cdot, \cdot) represents the concat operator and $g(\cdot)$ represents the fully connected layer. We obtain standardized incongruity information to improve model performance for reliable multi-modal sarcasm detection.

3.3 Perspective Fusion Module and Prediction

PFM aims to integrate the standardized incongruities of fact and sentiment perspectives as the total evaluation E , which is formalized as:

$$E = (\hat{\mathbf{I}}^F, \hat{\mathbf{I}}^S), \quad (12)$$

where (\cdot, \cdot) represents the concat operator. $\hat{\mathbf{I}}^F$ and $\hat{\mathbf{I}}^S$ stand for fact standardized incongruity and sentiment standardized incongruity, respectively. ITFNet then inputs the total evaluation and predicts the sarcasm label. The total loss function balances the contributions of the fact-related loss, sentiment-related loss, and prediction loss:

$$\mathcal{L} = \alpha\mathcal{L}^F + \alpha\mathcal{L}^S + (1 - 2\alpha)\mathcal{L}^{\text{pred}}. \quad (13)$$

4 Experiment

4.1 Dataset and evaluation metrics

We conduct experiments on the publicly available MMSD1.0 dataset (Cai et al., 2019), MMSD2.0 dataset (Qin et al., 2023), DMSD dataset (Jia et al., 2024). Each sample in the dataset consists of text-image pairs. Samples expressing sarcasm are labeled as positive, and those without sarcasm are labeled as negative. Following previous works (Cai et al., 2019; Xu et al., 2020; Liang et al., 2022), we report accuracy, precision, recall, and F1-score results for evaluation.

4.2 Implementation details

We used the CLIP ViT-B/32 model (Radford et al., 2021) with frozen parameters for unified token-level image and text feature extraction. We employed YOLO v10-s with frozen parameters (Wang et al., 2024a) for the fact feature extraction. The ITFNet was trained using AdamW with a learning rate set to $1e-4$, weight decay at $1e-4$, and α at 7.5%, over ten epochs.

4.3 Baseline models

To evaluate the performance of ITFNet, we compare it against several state-of-the-art baselines, categorized into image-modality, text-modality, and multi-modal methods.

Image-modality methods: ResNet (Cai et al., 2019), ViT (Dosovitskiy, 2020).

Text-modality methods: Bi-LSTM (Graves and Schmidhuber, 2005), SIARN (Tay et al., 2018), SMSD (Xiong et al., 2019), BERT (Kenton and Toutanova, 2019), RoBERTa (Qin et al., 2023).

Multi-modal methods: HFM (Cai et al., 2019), InCrossMGs (Liang et al., 2021), CMGCN (Liang et al., 2022), Att-BERT (Pan et al., 2020), DIP (Wen et al., 2023), KnowleNet (Yue et al., 2023), FSICN (Lu et al., 2024), Mumu (Wang et al., 2024b), AMIF (Li et al., 2025), Multi-view CLIP (Qin et al., 2023), DMSD-CL (Jia et al., 2024), G2SAM (Wei et al., 2024), MICL (Guo et al., 2025), DynRT-Net (Tian et al., 2023).

The details of these methods have been described in the related work section. Furthermore, we cite the performance of MLLM models (LLaVA1.5 and LLaVA1.5-VIDR) as reported by Tang et al. (2024), who performed LoRA-based PTFT on the training set of MMSD1.0 and MMSD2.0 dataset.

4.4 Main Result

MMSD1.0 Dataset. Table 1 presents the result on MMSD1.0. We have the following observations: (1) Compared with uni-modal methods, multi-modal methods perform better due to more comprehensive sarcasm information from multiple modalities. (2) ITFNet with frozen feature extractor (17.3M parameters) achieves the best performance and defeats the dominant multi-modal large language models (Tang et al., 2024). ITFNet exceeds LLaVA1.5-7B by 3.8% in accuracy and has an extremely low number of trainable parameters. This demonstrates that ITFNet captures significant and standardized incongruities as sarcasm information and effectively addresses the incongruity distortion through the iterative inconsistency preference learning to obtain efficient performance in multi-modal sarcasm detection tasks.

MMSD2.0 Dataset. Table 2 presents the result on MMSD2.0. We have the following observations: (1) Compared with the general model, ITFNet with frozen feature extractor (17.3M parameters) achieves the best performance. This illustrates that ITFNet is a reliable model. (2) ITFNet can surpasses the 7B multi-modal large language model (Tang et al., 2024) through full fine-tuning, which shows the large potential of ITFNet.

DMSD Dataset. Table 3 presents the result on DMSD. We have the following observations: (1) DMSD is an OOD test set, and DMSD-CL is a method based on positive and negative sample contrast learning, which is very robust (Jia et al., 2024). (2) ITFNet addresses the cross-instance variance through the standardization of incongruity with reference of contextual tone and achieves the best

Modality	Model	Acc(%)	Binary-Average			Macro-Average		
			P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
Image	Resnet	64.76	54.41	70.80	61.53	60.12	73.08	65.97
	ViT	67.83	57.93	70.07	63.43	65.69	71.35	68.40
Text	Bi-LSTM	81.90	76.66	78.42	77.53	80.97	80.13	80.55
	SIARN	80.57	75.55	75.70	75.63	80.34	78.81	79.57
	SMSD	80.90	76.46	75.18	75.82	80.87	78.20	79.51
	BERT	83.85	78.72	82.27	80.22	81.31	80.87	81.09
	RoBERTa	93.97	90.39	94.59	92.45	-	-	-
Multi-modal	HFM	86.63	83.84	84.18	84.01	86.24	86.28	86.26
	InCrossMGs	86.10	81.38	84.36	82.84	85.39	85.80	85.60
	CMGCN	87.55	83.63	84.69	84.16	87.02	86.97	87.00
	Att-Bert	86.05	78.63	83.31	80.90	80.87	85.08	82.92
	DIP	89.59	87.76	86.58	87.17	88.46	89.13	89.01
	KnowleNet	88.87	88.59	84.18	86.33	88.83	88.21	88.51
	DMSD-CL	88.95	84.89	87.90	86.37	88.35	88.77	88.54
	AMIF	90.10	86.55	89.68	88.09	-	-	-
	G2SAM	90.48	87.95	89.02	88.48	89.44	89.79	89.65
	FSICN	90.55	89.93	89.51	89.72	90.16	90.42	90.29
	Multi-view CLIP	88.33	82.66	88.65	85.55	-	-	-
	MuMu	90.73	88.81	88.44	88.62	90.43	90.37	90.40
	MIL-Net (RoBERTa+ViT)	89.50	85.16	89.16	87.11	88.88	89.44	89.12
	MICL (RoBERTa+ViT)	92.08	90.05	90.61	90.33	<u>91.85</u>	<u>91.77</u>	<u>91.81</u>
	DynRT-Net (RoBERTa+ViT)	93.59	93.06	93.60	93.31	-	-	-
	LLaVA1.5 (MLLM-7B)	<u>93.67</u>	<u>93.70</u>	93.14	<u>93.40</u>	-	-	-
	LLaVA1.5-VIDR (MLLM-7B)	89.97	89.26	89.58	89.42	-	-	-
	ITFNet	92.04	90.21	90.30	90.25	91.75	<u>91.77</u>	91.76
	ITFNet (RoBERTa+ViT)	97.05*	99.45*	<u>93.29</u>	96.27*	97.51*	96.47*	96.92*

Table 1: Model performance on MMSD1.0 dataset. Results with * denote the significance tests of our ITFNet over the baseline models in the same area at p-value < 0.01. The best results are highlighted in boldface, while the second-best results are underlined.

performance on DMSD-CL. This shows the adaptability and robustness of ITFNet.

4.5 Ablation Study

We conducted an ablation study to assess the impact of each module of the ITFNet with frozen CLIP as a feature extractor on MMSD1.0 and MMSD2.0, as shown in Table 4. First, the absence of FISN and SISN also led to performance degradation. It supports the motivation that detecting sarcasm from both fact and sentiment perspectives is more comprehensive.

Ablating the tension field greatly decreased performance, which illustrates that the incongruity-aware tension field plays a critical role in ITFNet. The fact/sentiment tension fields capture significant incongruity information through inconsistency preference learning and obtain standardized incongruity for unbiased decision-making in multi-modal sarcasm detection.

ITFNet without \mathbf{I}^F , \mathbf{I}^S means only the consensus extraction channel is enabled in the tension field of

both the FISN and SISN. It performs well since the contextual tone is the mean of the global feature space. However, it is still lower than the complete model by 1.56% in accuracy and 1.69% in F1 score on the MMSD1.0 dataset, and by 2.22% in accuracy and 3.18% in F1 score on the MMSD2.0 dataset.

ITFNet without \mathbf{C}^F , \mathbf{C}^S means only the incongruities extraction channel is enabled in the tension field of both the FISN and SISN. Its performance is slightly worse than the previous one because the polarized incongruity is a local clue. Moreover, this supports the motivation that cross-instance variance about sarcasm basic strength harms reliable multi-modal sarcasm detection and demonstrates the necessity of the reference function provided by contextual tone for incongruities.

4.6 Analysis

Optimal Settings Exploring. In this section, we experiment to determine the optimal settings for ITFNet. Figure-3a presents the performance

Modality	Model	Acc(%)	P(%)	R(%)	F1(%)
Text	TextCNN	71.61	64.62	75.22	69.52
	BiLSTM	72.48	68.02	68.08	68.05
	SMSD	73.56	68.45	71.55	69.97
Image	ResNet	65.50	61.17	54.39	57.58
	ViT	72.02	65.26	74.83	69.72
Multi-modal	HFM	70.57	64.84	69.05	66.88
	Att-Bert	80.03	76.28	77.82	77.04
	CMGCN	79.83	75.82	78.01	76.90
	HKE	76.50	73.48	72.07	72.25
	DynRT-Net	71.40	71.80	72.17	71.34
	Multi-view CLIP (Frozen)	84.72	-	-	83.64
	Multi-view CLIP (Full Finetuned)	85.64	80.33	88.24	84.10
	LLaVA1.5	85.18	<u>85.89</u>	85.20	<u>85.11</u>
	LLaVA1.5-VIDR	<u>86.43</u>	87.00	86.30	86.34
	ITFNet (Frozen)	85.83	80.58	<u>88.30</u>	84.26
	ITFNet (Full Finetuned)	86.73*	81.08	88.03*	84.41

Table 2: Model performance on MMSD2.0 dataset. Results with * indicate the statistical significance of our ITFNet over the baseline models in the same area at p-value < 0.05. The best results are highlighted in boldface, while the second-best results are underlined.

Modality	Model	Acc(%)	P(%)	R(%)	F1(%)
Text	TextCNN	37.25	37.30	36.71	36.58
	BiLSTM	34.50	33.20	32.77	32.94
	BERT	21.25	22.22	22.28	21.25
	RoBERTa	29.50	28.07	27.34	27.64
Image	ResNet	28.25	27.87	27.04	27.36
	ViT	22.00	22.53	21.36	21.55
Multi-modal	Res-BERT	20.75	21.62	20.77	20.60
	Att-Bert	28.25	27.50	26.46	26.69
	HKE	37.50	37.90	37.36	37.04
	CMGCN	34.25	35.52	35.22	34.20
	DMSD-CL	<u>70.25</u>	<u>70.41</u>	<u>71.34</u>	<u>69.96</u>
	ITFNet	75.75*	76.81*	71.72*	72.53*

Table 3: Model performance on DMSD dataset. Results with * indicate the statistical significance of our ITFNet over the baseline models at p-value < 0.05. The best results are highlighted in boldface, while the second-best results are underlined.

of ITFNet with a frozen feature extractor on MMSD1.0, varying the iteration count N . When N equals 0, ITFNet’s incongruity-aware tension field remains inactive. Increasing N from 0 to 1 yields a dramatic performance boost, with accuracy rising by 4.15% and F1 score by 4.23%. This jump confirms that inconsistency preference learning works effectively. Further increases in N continue to enhance both accuracy and F1 score, demonstrating that iterative learning based on discrepancy metrics gradually captures significant incongruity information for sarcasm detection. However, when N exceeds 4, accuracy begins to decline, likely due to instability from over-aggregation of consistent features and over-polarization of incongruous features.

ITFNet with different backbones. Since ITFNet

Ablation settings	MMSD1.0		MMSD2.0	
	Acc(%)	F1(%)	Acc(%)	F1(%)
w/o FISN	91.23	90.87	84.75	82.44
w/o SISN	91.28	90.89	84.37	82.12
w/o Tension fields	85.34	84.51	74.94	72.76
w/o I^F, I^S	90.48	90.07	83.61	81.08
w/o C^F, C^S	90.29	89.73	82.98	81.02
ITFNet (Complete)	92.04	91.76	85.83	84.26

Table 4: Ablation study results.

leverages an existing advanced model to construct a Yolo-task for extracting multi-modal fact features, we conducted an experiment to investigate the impact of different backbones on ITFNet’s performance. Figure-3b shows the performance of ITFNet with a frozen feature extractor on MMSD1.0, varying the backbone. According to public information, Fast R-CNN with Resnet 50 has parameters of 41.3M, Yolo v5-X has parameters of 87.7M, Yolo v10-N has parameters of 2.3M, Yolo v10-X has parameters of 29.5M, and Yolo v10 has parameters of 7.2M. Overall, models from the Yolo v10 series serve as stronger backbones for ITFNet. However, the performance range across different backbones does not exceed 0.49%. Additionally, ITFNet using Yolo v10-S as the backbone has 75.59% fewer parameters than Yolo v10-X, yet achieves 0.21% and 0.24% higher accuracy and F1 score, respectively. This experiment demonstrates that ITFNet generalizes effectively, with its performance remaining independent of any specific backbone architecture.

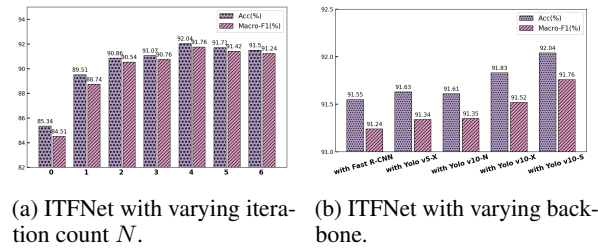


Figure 3: Performance under different settings.

Case study. We provide a case study to illustrate the effectiveness of inconsistency preference learning, as shown in Figure 4. These cases obscurely convey sarcastic intentions, which hinder the model from identifying the true label. Despite this, ITFNet correctly captures the sarcastic clues in three cases. In contrast, our reproduced AttBERT model (Pan et al., 2020), which is a consistency preference method for modeling implicit incongruities, failed to detect sarcasm in the sec-

ond and third cases. This proves our effectiveness in capturing explicit incongruities through the inconsistency preference method. The ITFNet of the ablation tension field failed to detect sarcasm in none of the three cases, highlighting the critical role of the incongruity-aware tension field in capturing significant incongruity information.




Case	Result
 <p>me reflecting on the weekend .</p>	<div>Label: $p(\text{sarcasm})=1$</div> <div>ITFNet: $p(\text{sarcasm})=0.86$</div> <div>Att-BERT: $p(\text{sarcasm})=0.62$</div> <div>w/o Tension Field: $p(\text{sarcasm})=0.39$</div>
 <p>thanks for the math # cocacola smh ... # advertisinggenius</p>	<div>Label: $p(\text{sarcasm})=1$</div> <div>ITFNet: $p(\text{sarcasm})=0.91$</div> <div>Att-BERT: $p(\text{sarcasm})=0.25$</div> <div>w/o Tension Field: $p(\text{sarcasm})=0.11$</div>
 <p>Is this you? #Monday #SameOld</p>	<div>Label: $p(\text{sarcasm})=1$</div> <div>ITFNet: $p(\text{sarcasm})=0.76$</div> <div>Att-BERT: $p(\text{sarcasm})=0.27$</div> <div>w/o Tension Field: $p(\text{sarcasm})=0.13$</div>

Figure 4: User study for sampled instances. Label stands for ground truth. ITFNet is our model.

5 Conclusion

This paper proposes that ITFNet perform inconsistency preference learning through an incongruity-aware tension field based on discrepancy metrics to address incongruity distortion and cross-instance variance in multi-modal sarcasm tasks. ITFNet utilizes a fact-sentiment multitask feature learning module to extract fact and sentiment text-image features. In the fact/sentiment feature branch, ITFNet presents a tension field to capture significant and standardized incongruity information. Experimental results show ITFNet attains state-of-the-art performance.

Limitations

This work demonstrates the effectiveness of our inconsistent preference methods in multi-modal sarcasm detection. Moreover, inconsistency is critical in areas such as rumor detection, where some conventional approaches that model inconsistency through consistency extraction may lead to distortion. Although our approach shows significant potential, it has not yet been used across different applications. In future work, we aim to extend our

inconsistent preference methods to a broader range of domains.

Acknowledgments

This work was funded in part by the National Natural Science Foundation of China grant under number 62222603, in part by the STI2030-Major Projects grant from the Ministry of Science and Technology of the People’s Republic of China under number 2021ZD0200700, in part by the Key-Area Research and Development Program of Guangdong Province under number 2023B0303030001, in part by the Program for Guangdong Introducing Innovative and Entrepreneurial Teams (2019ZT08X214), and in part by the Science and Technology Program of Guangzhou under number 2024A04J6310.

References

- Dieter R Brill and John A Wheeler. 1957. Interaction of neutrinos and gravitational fields. *Reviews of Modern Physics*, 29(3):465.
- Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multimodal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2506–2515.
- Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. 2022. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10563–10571.
- Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Raymond W Gibbs. 1986. On the psycholinguistics of sarcasm. *Journal of experimental psychology: general*, 115(1):3.
- Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.
- Diandian Guo, Cong Cao, Fangfang Yuan, Yanbing Liu, Guangjie Zeng, Xiaoyan Yu, Hao Peng, and Philip S. Yu. 2025. *Multi-view incongruity learning for multimodal sarcasm detection*. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1754–1766, Abu Dhabi, UAE. Association for Computational Linguistics.
- 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.
- Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of naacL-HLT, volume 1, page 2.
- Y Alex Kolchinski and Christopher Potts. 2018. Representing social media users for sarcasm detection. arXiv preprint arXiv:1808.08470.
- Kuntao Li, Yifan Chen, Qiaofeng Wu, Weixing Mai, Fenghuan Li, and Yun Xue. 2025. Ambiguity-aware multi-level incongruity fusion network for multi-modal sarcasm detection. In Proceedings of the 31st International Conference on Computational Linguistics, pages 380–391, Abu Dhabi, UAE. Association for Computational Linguistics.
- 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.
- Qiang Lu, Yunfei Long, Xia Sun, Jun Feng, and Hao Zhang. 2024. Fact-sentiment incongruity combination network for multimodal sarcasm detection. Information Fusion, 104:102203.
- Xu Ma, Xiyang Dai, Yue Bai, Yizhou Wang, and Yun Fu. 2024. Rewrite the stars. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5694–5703.
- 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.
- 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. arXiv preprint arXiv:2307.07135.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sasttry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR.
- Binghao Tang, Boda Lin, Haolong Yan, and Si Li. 2024. Leveraging generative large language models with visual instruction and demonstration retrieval for multi-modal sarcasm detection. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 1732–1742.
- Yi Tay, Luu Anh Tuan, Siu Cheung Hui, and Jian Su. 2018. Reasoning with sarcasm by reading in-between. arXiv preprint arXiv:1805.02856.
- Yuan Tian, Nan Xu, Ruikang Zhang, and Wenji Mao. 2023. Dynamic routing transformer network for multimodal sarcasm detection. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2468–2480.
- Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. 2024a. Yolov10: Real-time end-to-end object detection. arXiv preprint arXiv:2405.14458.
- Jie Wang, Yan Yang, Yongquan Jiang, Minbo Ma, Zhuyang Xie, and Tianrui Li. 2024b. Cross-modal incongruity aligning and collaborating for multi-modal sarcasm detection. Information Fusion, 103:102132.
- 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.
- 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.
- Tan Yue, Rui Mao, Heng Wang, Zonghai Hu, and Erik Cambria. 2023. Knowlenet: Knowledge fusion network for multimodal sarcasm detection. Information Fusion, 100:101921.