Toward Generalized Multimodal Sarcasm Detection With Multi-View Learning

Diandian Guo, Hao Peng, *Senior Member, IEEE*, Cong Cao, Fangfang Yuan, Yanbing Liu, and Philip S. Yu, *Life Fellow, IEEE*

***Abstract*—Multimodal Sarcasm Detection (MSD) is crucial for understanding complex human communication and building intelligent emotional computing systems. However, existing MSD methods often over-rely on spurious correlations, causing the learned features to deviate from the true semantics of sarcasm. This bias significantly undermines the generalization ability of current models beyond the training environment. This paper proposes a Confict-based Disentangled multi-view incongruity learning framework (ConDi), aiming to effectively disentangle and interact with heterogeneous features in multimodal sarcasm. Given that multimodal embedding spaces are typically heterogeneous, direct fusion can disrupt the inherent structure of embeddings from different modalities. To address this issue, we employ the optimal transport algorithm to align embeddings from different modalities into a unified space. Subsequently, we jointly learn incongruities from three views: modality disentanglement, global sentiment, and local description. To achieve a debiased fusion of sarcasm features, we design a conflict-based fusion module to integrate features from these three views. Experimental results demonstrate the superiority of ConDi on multimodal sarcasm datasets, and further analysis shows that ConDi can effectively reduce reliance on spurious correlations. Additionally, out-of-distribution (OOD) experiments reveal that ConDi achieves better generalization.**

***Index Terms*—multimodal, sarcasm detection, spurious correlation, debiased learning, generalization.**

I. Introduction

SARCASM detection is an important research task in data mining and natural language processing. Its core challenge lies in identifying language phenomena where the surface meaning of a text contradicts or significantly differs from the deeper intended message [1]. With the widespread use of sarcasm on social media platforms, this task has gradually become a research hotspot. An effective sarcasm detection mechanism is crucial for uncovering users' emotional stances, especially in social media contexts, where genuine emotions are often masked by exaggerated rhetoric or misleading expressions. This technology has significant value in downstream applications such as online public opinion mining [2], [3] and sentiment analysis [4], [5].

Early research focuses mainly on text-based unimodal sarcasm detection [6], [7], [8]. Jonshi et al. [9] experimentally establish the decisive role of incongruity characteristics in sarcasm detection, sparking a surge of research based on incongruity theory. For example, typical methods involve constructing explicit positive and negative sentiment pairs for feature extraction [8], [10]. However, these approaches face significant limitations in capturing context-sensitive implicit negative emotions, especially when dealing with complex rhetorical phenomena such as metaphors and irony. This limitation prompts researchers to introduce external structured knowledge bases to enhance implicit sentiment reasoning. Sentiment dictionary-based methods, such as those utilizing SenticNet [11], construct sentiment correlation matrices to capture sarcasm clues by calculating emotional differences between words [12], [13]. The development of multimedia social platforms has led to increasingly diverse modes of user expression, with multimodal communication gradually becoming the mainstream. This shift has redirected research attention toward multimodal sarcasm detection (MSD), where modeling cross-modal incongruity features remains a core research focus. Recent advances have been made in designing feature fusion architectures [14], [15], [16]. Currently, mainstream approaches primarily derive deep representations for each modality through pre-trained models, followed by the application of cross-attention mechanisms [15], [17] or graph neural networks [18], [19] to capture semantic conflicts between modalities. Notably, debiasing learning has become a key technology for improving the robustness of multimodal methods [20], [21]. Although a previous study [22] has attempted to mitigate the modal bias in MSD through adversarial training, MICL [23] shows that existing methods generally rely on spurious correlations. Therefore, mitigating spurious correlations and enhancing model generalizability remain key challenges in the field today.

Spurious correlation refers to the phenomenon where machine learning models mistakenly capture non-generalizable features, rather than genuinely discriminative features that support

Received 23 April 2025; revised 10 October 2025; accepted 3 November 2025. Date of publication 17 November 2025; date of current version 3 December 2025. The work of Hao Peng was supported in part by NSFC under Grant 62322202, Grant 62441612, and Grant 62432006 and in part by Beijing Natural Science Foundation under Grant L253021. The work of Philip S. Yu was supported by NSF under Grant III-2106758 and Grant POSE-2346158. This work was supported by the National Key R&D Program of China under Grant 2023YFC3303800. The associate editor coordinating the review of this article and approving it for publication was Dr. Xiao-Lei Zhang. *(Corresponding authors: Cong Cao; Yanbing Liu.)*

Diandian Guo, Cong Cao, Fangfang Yuan, and Yanbing Liu are with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100045, China, and also with the School of Cyber Security, University of Chinese Academy of Sciences, Beijing 101408, China (e-mail: guodiandian@iie.ac.cn; caocong@iie.ac.cn; yuanfangfang@iie.ac.cn; liuyanbing@iie.ac.cn).

Hao Peng is with the State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China (e-mail: penghao@buaa.edu.cn).

Philip S. Yu is with the Department of Computer Science, University of Illinois Chicago, Chicago, IL 60607 USA (e-mail: psyu@uic.edu).

Digital Object Identifier 10.1109/TASLPRO.2025.3633050

(a) Image Modification (b) Sentiment Modification (c) Non-critical term modification

Fig. 1. Existing models suffer from deficiencies that lead to spurious correlations on MSD tasks.

generalization [24]. Based on empirical experimental analysis, we identify two typical types of spurious correlations in MSD: **(1) Cross-modal learning bias.** As shown in Fig. 1(a), when the text remains unchanged, replacing the image content (e.g., changing "rainy day" to "sunny day") does not significantly affect the model's prediction. Similarly, in Fig. 1(b), modifying the sentiment polarity of the text while keeping the visual content unchanged results in no change to the model's prediction. **(2) Intra-modal dependency misalignment.** As shown in Fig. 1(b), removing key sentiment words in the text modality does not alter the model's prediction. However, modifying non-essential modifiers in Fig. 1(c) causes the prediction to reverse. A similar phenomenon is observed in the visual modality as well. It is important to note that these spurious correlations have a random emergence characteristic, with their occurrence probability closely related to the training process, making it difficult for traditional regularization methods to alleviate them effectively. Additionally, overparameterization of the model has been proven to be a key factor that triggers spurious correlations [25]. Therefore, existing approaches are constrained by the following limitations: (1) They lack cross-dataset generalization capability due to spurious correlations; (2) They introduce excessive computational complexity during the feature learning stage; (3) Many methods require a large amount of additional labeled data for data augmentation. This paper aims to construct a more parameter-efficient and generalized solution.

To address these challenges, we introduce ConDi, a novel multi-view incongruity learning method for MSD. Existing methods usually adopt a heterogeneous embedding strategy, where multimodal contents are first mapped to independent representation spaces. These features are then fused by simple `concat` or `mean` operations as the unified representation for multimodal sarcasm detection. However, since the embeddings of different modalities are often distributed in heterogeneous geometric spaces, direct fusion may disrupt the inherent distributions of each modality, leading to semantic inconsistencies and loss of feature integrity in the unified space. To overcome this, we introduce the optimal transport (OT) theory [26] during the multimodal feature extraction stage to achieve cross-modal space alignment. OT minimizes the Wasserstein distance between the modalities to construct a joint distribution mapping. This process transfers heterogeneous embeddings into a unified semantic space while preserving the topological structures of individual modalities, effectively overcoming alignment bias caused by spatial heterogeneity. Since sarcasm often involves entities or objects in multimodal contexts with emotional polarity, ConDi constructs a multi-view incongruity learning module that captures robust features from three views: modality disentanglement, global sentiment, and local description. We propose a disentangled multimodal feature learning method that separates modality-agnostic consistent features from modality-specific heterogeneous features, thereby significantly reducing model parameters while mitigating spurious correlations. Notably, the quality of features from different views exhibits significant sample dependency — when specific views encounter noise interference or information loss, performance can be severely affected [27]. To address this, we design a confidence-weighted fusion module based on evidence conflict to guide the fusion of multi-view incongruity features.

We conduct comprehensive experiments on two benchmark datasets, and the results show that ConDi outperforms the existing state-of-the-art models. In addition, we perform extensive comparisons in scenarios involving spurious correlations and out-of-distribution (OOD) data. Analysis shows that ConDi effectively reduces reliance on spurious correlations and performs better generalization with fewer parameters. The main contributions of this paper can be summarized as follows:

- We propose a Confict-based Disentangled multi-view incongruitylearning framework (ConDi), an efficient and generalized MSD method.
- We propose a disentangled incongruity learning method that does not require any additional augmented data. Compared to mainstream approaches, it reduces training costs and cuts more than 50% of the core parameters.
- We conduct extensive experiments on various datasets to verify the SOTA sarcasm detection performance and superior generalization ability of ConDi.

The paper is organized as follows: Section II reviews a series of representative works on sarcasm detection and mitigating spurious correlations; Section III describes the task definition and proposed framework ConDi; Section IV describes the experimental setup and baseline, and shows extensive experimental results, fully demonstrating the superiority of ConDi; Section V discusses the contribution and future work.

II. Related Work

A. Sarcasm Detection

Early research primarily adopts rule-based and linguistic evidence-driven unimodal analysis methods, laying the

foundation for symbolic feature engineering in sarcasm detection. Foundational works explore linguistic patterns (e.g., lexical tokens, syntactic structures) to construct basic detection frameworks. For example, hashtag tokenization [4] pioneers the statistical validity of sentiment features in sarcasm prediction using topic tags. PBLGA [28] develops a framework for extracting prosodic features based on dependency syntax parsing. At the feature representation level, researchers focus on the incongruity in sarcastic texts through specific embeddings, such as word shape [29] and word extension [30]. Cognitive linguistic studies reveal that semantic incongruity is a key cue for sarcasm detection [9], driving innovation in sarcasm detection methods [8], [10]. For instance, LBPR [8] proposes a collaborative attention network that dynamically detects opposing sentiment pairs within the text. However, it still suffers from significant limitations in handling implicit negation scenarios due to its lack of deep reasoning capabilities. To overcome this bottleneck, emerging methods enhance implicit reasoning abilities through knowledge-augmented strategies. ADGCN [13] builds an affective dynamic graph convolutional network based on SenticNet [11], enabling spectral-domain modeling of sentiment relationships within sentences.

The multimodal trend in social media drives research paradigms toward cross-modal sarcasm detection. Early pioneering work [31] first verifies the effectiveness of joint modeling of text and images, opening a new direction for MSD. The fine-grained labeled dataset constructed by Cai et al. [3] becomes a domain benchmark, providing critical support for subsequent research, particularly the inclusion of object-level visual sentiment labels. CMGCN [32] builds cross-modal graphs to model sarcastic relationships between text and images. Recent research focuses on refining the modeling of cross-modal incongruity features: DMSD-CL [22] proposes counterfactual enhancement and contrastive learning frameworks to improve model generalization. MILNet [15] designs a multi-scale incongruity learning mechanism to capture local-global semantic conflicts. G^{2}SAM [19] introduces a fine-grained multimodal graph and embeds it into the semantic space, combines global semantic consistency to retrieve k-nearest neighbor instances for voting prediction. MICL [23], proposes a multi-view incongruity learning framework to interference from surface-level associative noise.

However, existing approaches struggle to strike a balance between incongruity learning and parameter quantity, typically achieving marginal improvements in feature learning capability at the expense of substantial parameter increases. The improvements of ConDi against MICL include but are not limited to: (1) ConDi uses a more efficient incongruity learning framework, reducing the number of parameters by 90%. (2) ConDi resolves the conflict problem among different views. (3) ConDi does not introduce any additional data.

B. Mitigating Spurious Correlations

Mitigating spurious correlations in multimodal learning is a critical challenge for enhancing the generalization ability of supervised models. Current research can be systematically categorized into two major technical approaches: invariant feature mining and data distribution reconstruction. Invariant feature learning aims to build cross-domain stable feature association mechanisms. IRM [33] optimizes via environmental constraints to establish invariant feature mapping functions in latent variable spaces, providing a theoretical framework for domain generalization research. Some methods apply distributed robust optimization to model risk sensitivity by minimizing worst-case group losses [34]. A recent breakthrough, DFR [35], proposes a two-stage training process. First, it uses empirical risk minimization to obtain base representations, and then fine-tunes on a balanced dataset to effectively decouple spurious correlations.

Data distribution reconstruction strategies focus on improving model robustness through distribution balancing. Core methods involve generating augmented information within existing datasets [36] and employing other strategies to balance or diversify the data [37], [38]. For data augmentation, UV-Droid [36] generates augmented samples by injecting latent unmeasured variables. For data balancing, predefined concepts are often used to generate pseudo-labels to support training. For example, Discover and Cure (DISC) [37] expands training data using unstable predefined concepts across different environments to reduce spurious correlations. SSA [38] designs a pseudo-label propagation algorithm that estimates group distributions using a small number of labeled attribute samples.

While they achieve promising results, the overhead of introducing additional data during training proves too high, severely limiting its application potential and generalizability in real-time scenarios.

III. Methodology

As shown in Fig. 2, the architecture of ConDi mainly consists of three parts: multimodal feature extraction, multi-view incongruity learning, and conflict-based evidence fusion. This section first introduces the task definition and then describes the details of the architecture of our model.

A. Problem Definition

Given a sample $(\mathcal{X}_i, y_i)$ of training set, with $\mathcal{X}_i = (\mathcal{T}_i, \mathcal{V}_i)$, $\mathcal{T}_i$ represents the i-th text, $\mathcal{V}_i$ represents the corresponding image, and $y_i \in \{0, 1\}$ represents the true label. When the i-th sample contains sarcasm information, $y_i = 1$ otherwise $y_i = 0$. We aim to design a model $\mathcal{F}$ that uses visual and textual modalities for sarcasm detection and generates a predicted label $\hat{y}_i$. This process can be expressed as:

$$\mathcal{F}(\mathcal{X}_i|\Theta) \longrightarrow \hat{y}_i, \tag{1}$$

where Θ represents all parameters of model $\mathcal{F}$, $\hat{\mathcal{Y}}_i \in \{0, 1\}$ is the prediction result of model $\mathcal{F}$.

B. Multimodal Feature Extraction

1) Input Encoding: For each image $\mathcal{V}_i = \{v(i)_{cls}, v(i)_1, \ldots, v(i)_{n_{\mathcal{V}_i}}\}$, we define $v(i)_j$ as the j-th patch of $\mathcal{V}_i$, and $n_{\mathcal{V}_i}$ as the total number of patches that the image is divided into. We define the original text input

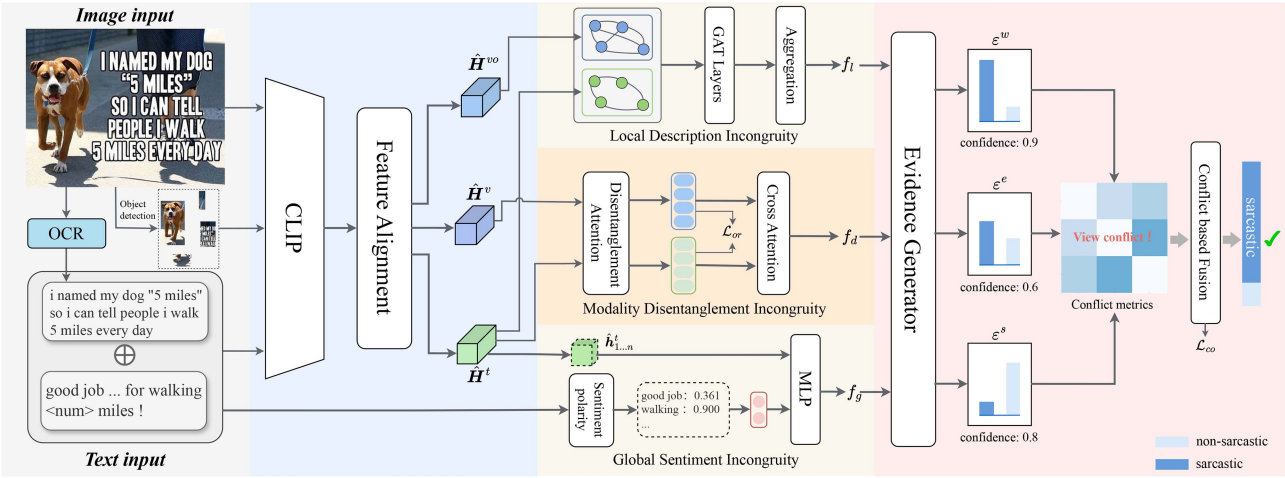
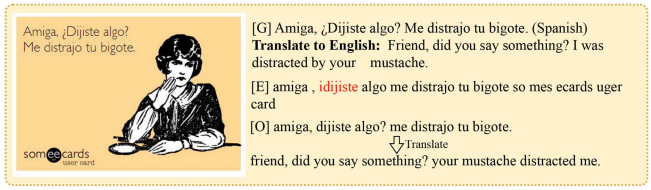
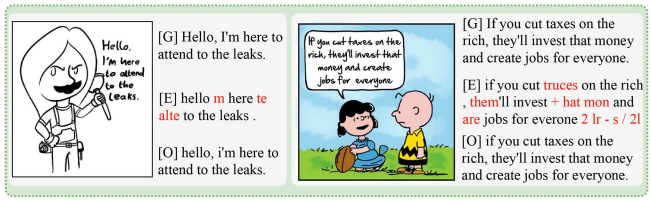


Fig. 2. The overall architecture of ConDi primarily comprises three key modules: (a) Multimodal Feature Extraction, (b) Multi-View Incongruity Learning, and (c) Conflict-based Evidence Fusion.



(a) OCR for non-English contents



(b) OCR for comics

Fig. 3. OCR comparison. [G] means ground truth, [E] means existing OCR result, and [O] means ours. 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.

$\mathcal{T}_i = \{t(i)_{cls}, t(i)_1, t(i)_2, \ldots, t(i)_{n_{\mathcal{T}_i}}\}$, where $t(i)_{cls}$ represents the [CLS] token, $t(i)_j$ represents the j-token of $\mathcal{T}_i$, and $n_{\mathcal{T}_i}$ is the sequence length of $\mathcal{T}_i$. Our observation indicates that the textual information embedded in images often complements the text modality. Based on this observation, we treat the Optical Character Recognition (OCR) text $\mathcal{O}_i$ extracted from images as an auxiliary input to the raw text input $\mathcal{T}_i$. Similarly, $\mathcal{O}_i = \{o(i)_1, o(i)_2, \ldots, o(i)_{n_{\mathcal{O}_i}}\}$, as an auxiliary input alongside $\mathcal{T}_i$, where $o(i)_j$ denotes the j-token of $\mathcal{O}_i$. However, existing work [10] reveals that OCR text suffers from issues such as low accuracy and ambiguous meaning, as shown in Fig. 3. Low-quality OCR text will negatively impact model performance [39]. Thus, we provide more precise OCR-text extraction and translation, complemented by meticulous manual proofreading to generate refined OCR-text. Then, we concatenate $\mathcal{T}_i$ and $\mathcal{O}_i$ as the final textual input.

In current multimodal learning methods, textual and visual information are typically encoded independently [15], [23]. Considering that using different pretrained feature encoders may introduce additional biases in multimodal features, we directly utilize the pre-trained CLIP [40] model to obtain multimodal feature representations:

$$\boldsymbol{H}_i^t, \boldsymbol{H}_i^v = \text{Self\_Att}(\text{CLIP}([\mathcal{T}_i; \mathcal{O}_i], \mathcal{V}_i)), \tag{2}$$

where $\boldsymbol{H}_i^t = [\boldsymbol{h}(i)_{cls}^t, \boldsymbol{h}(i)_1^t, \boldsymbol{h}(i)_2^t, \ldots, \boldsymbol{h}(i)_{n_i}^t] \in \mathbb{R}^{(n_i+1)\times d}$ is the textual representation of the input text, $\boldsymbol{h}(i)_j^t \in \mathbb{R}^d$ denotes the hidden state vector of j-token, d denotes the dimension of the hidden representations, $n_i = n_{\mathcal{T}_i} + n_{\mathcal{O}_i}$ is the total number of tokens after concatenating the original text and OCR-text, and $[\cdot;\cdot]$ refers to the concatenation operation. $\boldsymbol{H}_i^v = [\boldsymbol{h}(i)_{cls}^v, \boldsymbol{h}(i)_1^v, \boldsymbol{h}(i)_2^v, \ldots, \boldsymbol{h}(i)_{n_{\mathcal{V}}}^v] \in \mathbb{R}^{(n_{\mathcal{V}_i}+1)\times d}$ is the visual representation of the input image, $\boldsymbol{h}(i)_j^v \in \mathbb{R}^d$ represents the j-th patch embedding. Self_Att means a self-attention layer. For clarity and simplification, we use $\boldsymbol{e}_t^i$ and $\boldsymbol{e}_v^i$ to represent $\boldsymbol{h}(i)_{cls}^t$ and $\boldsymbol{h}(i)_{cls}^v$ in subsequent expressions. Similarly, we can also obtain $\boldsymbol{H}_i^{vo}$ partitioned by object detection for local description incongurity learning.

2) Feature Alignment: Considering the heterogeneity of multimodal features, we align the multimodal embeddings using OT. OT serves as a way to compare two probability distributions, which can provide a transport plan to transfer one point to another [26]. The process of feature alignment can be modeled as transferring the embeddings of different modalities into a unified space. Specifically, the procedure is as follows: (1) Estimate the distribution μ of the textual embeddings and the distribution ν of the visual embeddings. (2) Find the transport coupling $\mathbf{T}$ from μ to ν. (3) Compute the barycenter using $\mathbf{T}$, and then map the textual and visual embeddings into the unified space.

First, we define $\mathbf{u} = \{u_i\}_{i=1}^n$ and $\mathbf{v} = \{v_i\}_{i=1}^m$ as the probabilistic simplices for the textual and visual features, respectively.

Let $\mu = \sum_{i=1}^{n} u_i \delta_{\boldsymbol{e}_t^i}$ and $\nu = \sum_{j=1}^{m} v_j \delta_{\boldsymbol{e}_v^i}$ represent the discrete distributions of the textual and visual embeddings, where δ is the Dirac delta function. Following [26], we set $u_i = \frac{1}{m}$ and $v_j = \frac{1}{n}$, where m and n are the dimensions of the embeddings $\boldsymbol{e}_t$ and $\boldsymbol{e}_v$, respectively. We then define the joint distribution $\Pi(\mu, \nu)$ as follows:

$$\Pi(\mu, \nu) = \left\{ \mathbf{T} \in \mathbb{R}_+^{n \times m} \mid \mathbf{T}\mathbf{1}_m = \mathbf{u},\ \mathbf{T}^\top \mathbf{1}_n = \mathbf{v} \right\}, \tag{3}$$

where $\mathbf{1}$ denotes an all-one vector. Next, we compute the OT coupling $\mathbf{T}$ between μ and ν by minimizing the Wasserstein distance between the two distributions, which can be expressed as:

$$\mathbf{OT}(\mu, \nu, \mathbf{C}) = \min_{\mathbf{T} \in \Pi(\mu,\nu)} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{T}_{ij} \mathbf{C}_{ij}. \tag{4}$$

where $\mathbf{C}_{ij} = \|\boldsymbol{e}_t^i, \boldsymbol{e}_v^j\|_2^2$ is the 2-norm Wasserstein distance between e_t^i and e_v^j. After obtaining the transport matrix $\mathbf{T}$, the textual embedding $\boldsymbol{E}^t = [\boldsymbol{e}_t^i]_{i \in \{1,\ldots,batch\_size\}}$ can be transformed into the target aligned embedding $\hat{e}_t$ using a barycenter-based strategy [41]:

$$\hat{\boldsymbol{E}}^t = \mathrm{diag}(1/\mathbf{v}) \left(\mathbf{T}^\top + \Delta_{\mathbf{T}} \right) \boldsymbol{E}^t \tag{5}$$

where $\Delta_{\mathbf{T}}$ is a tunable parameter. Subsequently, we can obtain the aligned visual embedding $\hat{\boldsymbol{E}}^v$ in a similar manner. Then we get the updated $\hat{\boldsymbol{H}}_i^t$ and $\hat{\boldsymbol{H}}_i^v$ (also the $\hat{\boldsymbol{H}}_i^{vo}$).

C. Multi-View Incongruity Learning

Existing MSD methods suffer from three fundamental limitations in incongruity modeling. First, they typically adopt a unified incongruity learning paradigm that inadequately separates modality-specific features. Second, they fail to model cross-modal contradictions emerging from global sentiment polarity shifts across entire instances. Third, they demonstrate weak capability in detecting fine-grained semantic mismatches between localized visual descriptions and their contextual verbal references. To address these challenges, we propose a multi-view incongruity learning framework through three complementary dimensions: **(1) Modality disentanglement incongruity**: A multimodal token disentangling Transformer architecture is designed to deeply decouple sarcasm-related features. **(2) Global sentiment incongruity**: A sentiment-aware reasoning mechanism is designed to explicitly model sentiment polarity conflicts. **(3) Local description incongruity**: A cross-modal semantic graph network is constructed to enable fine-grained semantic mismatch detection.

1) Modality Disentanglement Incongruity Learning: Existing cross-modal attention frameworks [15], [22] predominantly employ text as query with images as key/value pairs, a design susceptible to modality dominance bias. Our method introduces a novel multimodal disentangling attention mechanism that establishes balanced fusion pathways for heterogeneous features, preserving critical sarcasm features while enabling cross-modal complementary enhancement. Specifically, we randomly initialize three token vectors $\{\boldsymbol{Z}_i^f, \boldsymbol{Z}_i^t, \boldsymbol{Z}_i^v\}$, where $\boldsymbol{Z}_i^f \in \mathbb{R}^{4 \times d}$ is used to learn inherent sarcasm-related features between textual and visual modalities, $\boldsymbol{Z}_i^t$ and $\boldsymbol{Z}_i^v \in \mathbb{R}^{2 \times d}$ are used to learn modality-specific heterogeneous features for text and image, respectively. These tokens concatenate with multimodal features $\boldsymbol{H}_i = [\hat{\boldsymbol{H}}_i^t; \hat{\boldsymbol{H}}_i^v]$ for self-attention processing:

$$\hat{\boldsymbol{Z}}_i^f, \hat{\boldsymbol{Z}}_i^t, \hat{\boldsymbol{Z}}_i^v = \mathrm{Self\_Att}([\boldsymbol{Z}_i^f; \boldsymbol{Z}_i^t; \boldsymbol{Z}_i^v; \hat{\boldsymbol{H}}_i]). \tag{6}$$

We treat $\hat{\boldsymbol{Z}}_i^f$ as the inherent feature and $[\hat{\boldsymbol{Z}}_i^t; \hat{\boldsymbol{Z}}_i^v]$ as the heterogeneous feature. The cross-attention mechanism operates through dual pathways for disentangled incongruity learning:

$$\text{Path 1:}\quad \boldsymbol{Q}_f = \hat{\boldsymbol{Z}}_i^f,\ \boldsymbol{K}_f = \boldsymbol{V}_f = [\hat{\boldsymbol{Z}}_i^t; \hat{\boldsymbol{Z}}_i^v],$$

$$\text{Path 2:}\quad \boldsymbol{Q}_m = [\hat{\boldsymbol{Z}}_i^t; \hat{\boldsymbol{Z}}_i^v],\ \boldsymbol{K}_m = \boldsymbol{V}_m = \hat{\boldsymbol{Z}}_i^f. \tag{7}$$

$$\boldsymbol{F}_f = \mathrm{Cross\_Att}(\boldsymbol{Q}_f, \boldsymbol{K}_m, \boldsymbol{V}_m), \tag{8}$$

$$\boldsymbol{F}_m = \mathrm{Cross\_Att}(\boldsymbol{Q}_m, \boldsymbol{K}_f, \boldsymbol{V}_f), \tag{9}$$

$$\boldsymbol{f}_d = [\mathrm{Mean}(\boldsymbol{F}_f); \mathrm{Mean}(\boldsymbol{F}_m)], \tag{10}$$

where Cross_Att is a standard cross attention layer, Mean represents average pooling, and we use $\boldsymbol{f}_d$ as the final output.

To ensure the separation of feature $\hat{\boldsymbol{Z}}_i^f$ from the heterogeneous features $[\hat{\boldsymbol{Z}}_i^t; \hat{\boldsymbol{Z}}_i^v]$, we design an orthogonality loss:

$$\mathcal{L}_{or} = \frac{1}{N} \sum_{k=1}^{N} \left(\|\hat{\boldsymbol{Z}}_i^{f\top} \hat{\boldsymbol{Z}}_i^t\|_2^2 + \|\hat{\boldsymbol{Z}}_i^{f\top} \hat{\boldsymbol{Z}}_i^v\|_2^2 + \|\hat{\boldsymbol{Z}}_i^{t\top} \hat{\boldsymbol{Z}}_i^v\|_2^2 \right) \tag{11}$$

As illustrated in Fig. 4, our approach demonstrates two key advancements over MICL: (1) Enhanced parameter efficiency with cross-attention input dimensions reduced to $2 \times d$ (vs. $100 \times d$ in MICL); (2) Optimized architectural design for sarcasm feature preservation.

2) Global Sentiment Incongruity Learning: Considering the crucial role that emotional context plays in MSD as demonstrated in [9], our proposed model incorporates global sentiment incongruity. The primary objective is to identify and discern incongruities that exist between the original text and the OCR-text. More precisely, we utilize SenticNet [11] to extract the sentiment polarity of both the source text and the OCR-text. Then, we use a multi-layer perceptron to catch the sentiment incongruity:

$$\boldsymbol{s}_t = \mathrm{SenticNet}(\mathcal{T}),\ \boldsymbol{s}_o = \mathrm{SenticNet}(\mathcal{O}), \tag{12}$$

$$\boldsymbol{f}_g = \mathrm{MLP}([\boldsymbol{s}_t; \boldsymbol{s}_o; \hat{\boldsymbol{h}}_{1\ldots n}^t]), \tag{13}$$

where MLP is a muti-layer perceptron. If OCR-text is unavailable, $\boldsymbol{f}_g$ is assigned a value of 0.

3) Local Description Incongruity Learning: To enable local description incongruity learning, we construct dual semantic graph representations for textual and visual content. Both graphs adopt an undirected structure with self-loop connections to maintain bidirectional information flow and preserve intrinsic node features. The text semantic graph is built upon dependency parsing results from spaCy,[1] this graph uses syntactic entities as nodes and their dependency relationships as edges, enabling precise detection of semantic role contradictions. The

[1][Online]. Available: https://spacy.io/

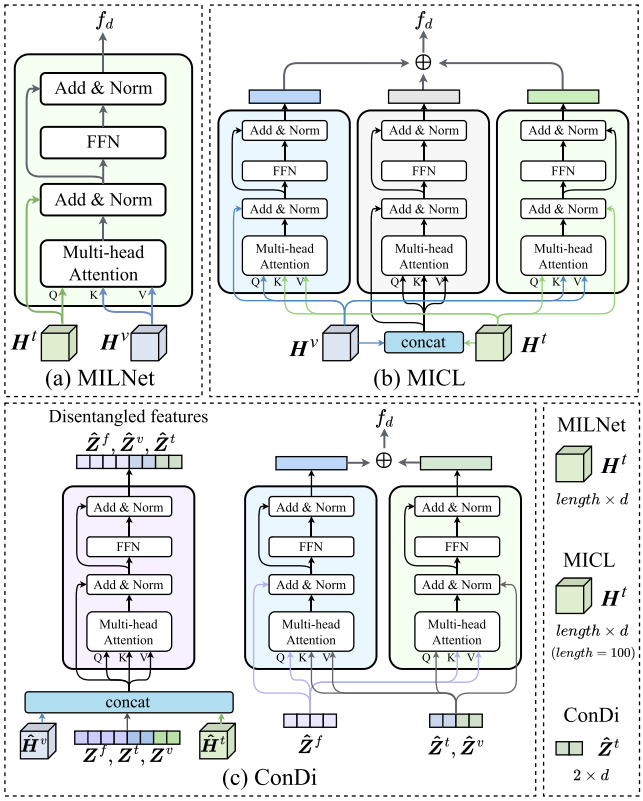


Fig. 4. Comparison of ConDi and other methods in token-patch incongruity learning.

visual semantic graph is generated using region segmentation results from a pretrained object detection model. Nodes are created through region features, and spatial-semantic edges are formed by calculating cosine similarity between region features. For graph processing, we employ Graph Attention Networks (GAT) [42] with dynamic feature propagation through attention mechanisms. The computation process is defined as follows:

$$\alpha_{i,j}^{l} = \frac{\exp\left(\text{LeakyReLU}\left(\boldsymbol{u}_l^{\top}[\boldsymbol{W}_l \boldsymbol{g}_i^{l} \| \boldsymbol{W}_l \boldsymbol{g}_j^{l}]\right)\right)}{\sum_{k \in \mathcal{N}(i) \cup i} \exp\left(\text{LeakyReLU}\left(\boldsymbol{u}_l^{\top}[\boldsymbol{W}_l \boldsymbol{g}_i^{l} \| \boldsymbol{W}_l \boldsymbol{g}_k^{l}]\right)\right)} \tag{14}$$

$$\boldsymbol{g}_i^{l+1} = \alpha_{i,i}^{l} \boldsymbol{W}_l \boldsymbol{g}_i^{l} + \sum_{j \in \mathcal{N}(i)} \alpha_{i,j}^{l} \boldsymbol{W}_l \boldsymbol{g}_j^{l} \tag{15}$$

where $\boldsymbol{g}_i^{l}$ denote the feature of node i at layer l, with learnable parameters $\boldsymbol{W}_l \in \mathbb{R}^{d \times d}$ and $\boldsymbol{u}_l$. Initial node features $\boldsymbol{g}_i^{0} = \hat{\boldsymbol{h}}(m)_i^{t}$ from input $\mathcal{T}_m$.

The final representations $\mathcal{G}^T = \{\boldsymbol{g}_0, \ldots, \boldsymbol{g}_n\}$ and $\mathcal{G}^V$ are concatenated as $\mathcal{G} = [\mathcal{G}^T; \mathcal{G}^V]$. Local description incongruity is learned through:

$$\boldsymbol{f}_l = \frac{1}{|\mathcal{G}|} \sum_{\boldsymbol{g}_i \in \mathcal{G}} \text{Softmax}\left(\boldsymbol{g}_i \boldsymbol{W}_g + b_g\right) \boldsymbol{g}_i \tag{16}$$

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

D. Conflict-Based Evidence Fusion

Our fusion mechanism addresses view-specific confidence variation in ConDi's three incongruity features (Fig. 3(c)), where adaptive feature selection enhances sensitivity to discriminative clues. Building upon TMC's [43] verification of Dirichlet distribution for confidence estimation, we adopt Subjective Logic (SL) for multi-class uncertainty modeling. SL provides a theoretical framework for deriving the probabilities of different categories (belief masses), as well as the overall uncertainty (uncertainty mass), in multi-class classification problems. This framework is based on the *evidence* gathered from the data. Specifically, for a multi-class classification problem with K classes, SL assigns a belief mass to each class label, along with an uncertainty mass to the entire classification frame, based on the available evidence. For the i-th view, all mass values are non-negative and sum to one. The belief masses for each class and the uncertainty mass must satisfy the following condition:

$$a^i + \sum_{k=1}^{K} b_k^i = 1, \tag{17}$$

where $a^i >= 0$ and $b_k^i >= 0$ indicate the overall uncertainty and the probability for the k-th class, respectively.

For the i-th view, subjective logic connects the *evidence* $\boldsymbol{\varepsilon}^i = [\varepsilon_1^i, \ldots, \varepsilon_K^i]$ to the parameters of the Dirichlet distribution $\boldsymbol{\beta}^i = [\beta_1^i, \ldots, \beta_K^i]$. Specifically, the parameter β_k^i of the Dirichlet distribution is induced from ε_k^i, *i.e.*, $\beta_k^i = \varepsilon_k^i + 1$. Then, the belief mass b_k^i and the uncertainty a^i are computed as:

$$b_k^i = \frac{\varepsilon_k^i}{S_i} = \frac{\beta_k^i - 1}{S_i}, \; a^i = \frac{K}{S_i}, \tag{18}$$

where $S_i = \sum_{k=1}^{K} (\varepsilon_k^i + 1) = \sum_{k=1}^{K} \beta_k^i$ is the Dirichlet strength.

The core challenge in multi-view fusion lies in decision conflicts between high-confidence views, as it is difficult to determine which view is of higher quality. Ideally, the uncertainty of multi-view learning results should not decrease with the increase in the number of views, especially when the learning results of two views conflict. To solve this problem, we use a conflict opinion aggregation method to complete multi-view fusion. MSD is a binary classification task, so the Dirichlet distribution can be simplified to a mathematically equivalent Beta distribution. Thus, we use ε_k^i represents the output of the final layer of the classifier for the i-th view regarding the k-th classification result. In binary classification tasks, $k \in \{0, 1\}$. Specifically, our ConDi structure includes three views $\{w, e, s\}$. We use the output before the softmax operation of the i-th view classifier as evidence ε^i. Taking views w (token-patch incongruity) and e (entity-object incongruity) as an example, we use the following combination rule to ensure that the quality of the new opinion is proportional to the combined opinion:

$$b_k^{w,e} = \frac{b_k^w a^e + b_k^e a^w}{a^w + a^e}, \; a^{w,e} = \frac{2 a^w a^e}{a^w + a^e}. \tag{19}$$

This combination ensures that the combined evidence $\varepsilon^{w,e}$ can be easily calculated:

$$\varepsilon^{w,e} = b_k^{w,e} S = \frac{b^{w,e} K}{a^{w,e}} = \frac{b_k^w a^e + b_k^e a^w}{a^w + a^e} \cdot \frac{a^w + a^e}{2a^w a^e} \cdot K = \frac{b_k^w K}{2a^w} + \frac{b_k^e K}{2a^e} = \frac{\varepsilon_k^w + \varepsilon_k^e}{2}. \tag{20}$$

In this way, the final detection result can be simply written as $\hat{y} = \varepsilon^{w,e,s}$. In addition, in order to ensure that the various perspectives can be consistent during the training process and thus reduce perspective conflicts, we also need to constrain it through the loss function $\mathcal{L}_{co}$:

$$\mathcal{L}_{co} = \sum_{p \in \{w,e,s\}} \left(\sum_{q \neq p} c(p,q) \right), \tag{21}$$

$$c(p,q) = (1 - a^p)(1 - a^q) \cdot \sum_{k \in \{0,1\}} \frac{|b_k^p - b_k^q|}{2}, \tag{22}$$

when $c(p,q) = 0$, it indicates that there is no conflict between views p and q; when $c(p,q) = 1$, views p and q absolutely believe in their judgments and are in conflict.

E. Optimization Objective

Contrastive learning enables the effective alignment of features across different modalities. In this work, we construct a contrastive learning framework in which positive and negative examples are determined based on whether their labels match. Specifically, within a batch, samples that share the same label as the anchor sample are considered positive, forming the positive sample set S_P, while samples with different labels belong to the negative sample set S_N. The complete sample set within a batch is defined as $S = S_P + S_N$. In our model, the core focus is on capturing the incongruity between textual and visual modalities. Therefore, when constructing the contrastive learning framework, we adopt a text-image matching approach to derive similarity scores for both positive and negative pairs. For k-th sample in the training set, the $t \to v$ contrastive loss is as follows:

$$\mathcal{L}_k^{t \to v} = \frac{1}{S_P} \sum_{i \in |S_P|} -\log \frac{\exp\left(\cos(\hat{\boldsymbol{e}}_t^k, \hat{\boldsymbol{e}}_v^i)/\tau\right)}{\sum_{j \in S} \exp\left(\cos(\hat{\boldsymbol{e}}_t^k, \hat{\boldsymbol{e}}_v^j)/\tau\right)}, \tag{23}$$

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

$$\mathcal{L}_k^{v \to t} = \frac{1}{S_P} \sum_{i \in |S_P|} -\log \frac{\exp\left(\cos(\hat{\boldsymbol{e}}_v^k, \hat{\boldsymbol{e}}_t^i)/\tau\right)}{\sum_{j \in S} \exp\left(\cos(\hat{\boldsymbol{e}}_v^k, \hat{\boldsymbol{e}}_t^j)/\tau\right)}. \tag{24}$$

The overall contrastive loss is:

$$\mathcal{L}_{cl} = \frac{1}{N} \sum_{k=1}^{N} \left(\frac{1}{2} \mathcal{L}_k^{t \to v} + \frac{1}{2} \mathcal{L}_k^{v \to t} \right). \tag{25}$$

The final optimization objective for ConDi is defined as the combination of the cross-entropy loss $\mathcal{L}_{ce}$, the orthogonality loss in (11), the credibility loss in (21), and the contrastive learning loss in (25):

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_1 \mathcal{L}_{or} + \lambda_2 \mathcal{L}_{co} + \lambda_3 \mathcal{L}_{cl}, \tag{26}$$

where λ_1, λ_2 and λ_3 are hyperparameters.

TABLE I
STATISTICS OF DATASETS

Datasets	Split	Positive	Negative	Total	Avg Len.
MMSD	Training	8,642	11,174	19,816	15.71
	Validation	959	1,451	2,410	15.72
	Test	959	1,450	2,409	15.89
MMSD 2.0	Training	9,576	10,240	19,816	13.42
	Validation	1,042	1,368	2,410	13.64
	Test	1,037	1,372	2,409	13.52
SPMSD	-	573	427	1,000	12.77
RedEval	-	395	609	1,004	7.35
FBHM	-	510	490	1,000	10.42

IV. Experiments

A. Datasets and Experiment Setting

We conduct experiments using the publicly available Multimodal Sarcasm Detection Dataset (MMSD) [3] and MMSD 2.0 [17]. Each dataset entry comprises a text-image pair, labeled as either sarcastic or non-sarcastic based on specific hashtags. To further evaluate the models' generalization capability and their susceptibility to spurious correlations, we use SPMSD [23] and RedEval [44] as additional test sets. The SPMSD dataset, derived and extended from MMSD, is specifically designed to assess models' reliance on spurious correlations. In contrast, the RedEval dataset, collected from the Reddit platform, serves as out-of-distribution data for robustness evaluation. In addition, we also use the out-of-domain task dataset FBHM [45] to test the generalization ability in hateful meme detection. The dataset statistics are summarized in Table I.

We employ the pre-trained CLIP[2] for multimodal encoding. For constructing textual graphs, we use the `en_core_web_trf` model in SpaCy to extract dependencies between entities. For visual graphs, edges are added between regions with cosine similarity exceeding 0.6. The feature dimension d is set to 768. The hyperparameters τ, λ_1, λ_2 and λ_3 are set to 0.07, 0.2, 0.2 and 1, respectively. The batchsize is 128. The maximum number of OT iterations is set to 200. Optimization is performed using the AdamW optimizer, with a learning rate of 5e-4. All experiments are conducted on a single NVIDIA Tesla A100 GPU (80 GB). Data and code are publicly accessible on GitHub.[3]

B. Baselines and Evaluation Metrics

We compare our proposed model ConDi with several baselines, which are broadly categorized into two groups.

Unimodal Baselines: These methods use either textual or visual information as input: **(1) Textual models**: This group includes traditional models like TextCNN [46], Bi-LSTM [47],

[2][Online]. Available: https://huggingface.co/openai/clip-vit-base-patch32
[3][Online]. Available: https://github.com/Remwlp/ConDi

TABLE II
MAIN RESULTS ON MMSD DATASET FOR SARCASM DETECTION. † MEANS LORA FINE-TUNED ON THE CORRESPONDING DATASET. ‡ MEANS ZERO-SHOT SETTING, TESTING THE MODEL DIRECTLY ON THE DATASET WITHOUT FINE-TUNING. **BOLD** MEANS THE BEST PERFORMANCE OF ALL METHODS

Method	MMSD							MMSD 2.0						
	Acc.	Binary-Average			Macro-Average			Acc.	Binary-Average			Macro-Average		
		P.	R.	F1	P.	R.	F1		P.	R.	F1	P.	R.	F1
Text-only														
TextCNN	80.03	74.29	76.39	75.32	78.03	78.28	78.15	71.61	64.62	75.22	69.52	67.59	70.39	68.96
Bi-LSTM	81.90	76.66	78.42	77.53	80.97	80.13	80.55	72.48	68.02	68.08	68.05	71.36	66.34	68.76
BERT	83.85	78.72	82.27	80.22	81.31	80.87	81.09	75.09	68.50	78.01	72.94	74.96	75.44	74.93
RoBERTa	85.51	78.24	88.11	82.88	84.83	85.95	85.16	79.66	76.74	75.70	76.21	79.78	80.36	79.71
Qwen 2.5-7B‡	69.27	61.39	57.08	59.15	67.57	67.07	67.26	67.20	64.82	52.07	57.75	66.64	65.35	65.47
Llama 3-8B‡	63.33	51.97	78.16	62.43	65.70	66.01	63.31	65.17	57.14	76.37	65.37	66.59	66.53	65.17
Phi-4-mini-3.8B‡	63.84	67.73	21.87	33.07	65.49	57.34	54.15	61.72	69.89	19.47	30.46	65.25	56.56	52.03
Qwen 2.5-7B†	90.64	89.94	86.79	88.34	90.52	90.04	90.26	77.45	71.40	79.45	75.21	77.21	77.70	77.27
Llama 3-8B†	93.17	91.90	91.33	91.61	92.97	92.88	92.92	82.31	76.35	85.34	80.60	82.09	82.68	82.17
Image-only														
Image	64.76	54.41	70.80	61.53	60.12	73.08	65.97	65.50	61.17	54.39	57.58	65.29	65.27	63.96
ViT	67.83	57.93	70.07	63.43	65.68	71.35	68.40	72.02	65.26	74.83	69.72	73.20	73.65	72.96
Multimodal														
HFM	83.44	76.57	84.15	80.18	79.40	82.45	80.90	70.57	64.84	69.05	66.88	70.04	70.23	69.14
Res-BERT	84.80	77.80	84.15	80.85	78.87	84.46	81.57	81.65	75.75	75.82	75.79	81.41	81.98	81.50
Att-BERT	86.05	78.63	83.31	80.90	80.87	85.08	82.92	80.03	76.28	77.82	77.04	79.97	80.49	80.06
CMGCN	87.55	83.63	84.69	84.16	87.02	86.97	87.00	79.83	75.82	78.01	76.90	78.88	79.45	78.83
Multi-View CLIP	88.33	82.66	88.65	85.55	86.82	87.85	87.21	85.64	80.33	**88.24**	84.10	84.72	84.86	84.79
MILNet	89.50	85.16	89.16	87.11	88.88	89.44	89.12	82.77	76.76	86.01	81.12	82.56	83.16	82.64
DMSD-CL	88.95	84.89	87.90	86.37	88.35	88.77	88.54	78.33	71.80	81.77	76.47	78.75	78.21	78.19
G^2SAM	90.48	87.95	89.02	88.48	89.44	89.79	89.65	79.43	72.04	85.20	78.07	78.68	79.13	78.78
MICL	92.08	90.05	90.61	90.33	91.85	91.77	91.81	85.05	82.08	83.51	82.79	84.42	85.06	84.48
Qwen 2.5-VL-7B‡	60.42	49.35	57.72	53.21	59.53	59.94	59.46	57.70	50.86	51.20	51.03	56.89	56.90	56.90
LLaVA 1.6-7B‡	65.10	58.20	37.18	45.38	62.80	60.06	59.87	63.30	63.68	34.32	44.61	63.43	59.76	58.58
MiniCPM-V 2.6-8B‡	61.18	50.16	66.81	57.30	61.62	62.20	60.86	53.25	45.55	43.97	44.74	52.14	52.12	52.12
Gemma 3-12B‡	52.97	44.58	91.43	59.94	61.54	56.50	46.90	51.72	46.65	87.46	60.85	59.34	55.94	48.66
Qwen 2.5-VL-7B†	90.30	90.73	84.93	87.73	90.38	89.47	89.86	77.54	71.86	78.59	75.08	77.22	77.67	77.32
LLaVA 1.6-7B†	93.21	**91.47**	91.95	91.71	92.95	93.01	92.98	81.11	73.85	86.88	79.84	81.20	81.81	81.03
GPT 5‡	70.37	61.12	75.43	67.53	70.44	71.16	70.14	71.36	63.93	76.75	69.76	71.61	72.02	71.27
Gemini 2.5 Flash‡	71.51	63.62	70.58	66.92	70.83	71.36	70.95	71.64	67.21	66.63	66.92	71.08	71.03	71.05
Claude Sonnet 4‡	68.31	57.06	90.40	69.97	72.98	71.73	68.21	70.73	60.69	90.83	72.76	74.80	73.18	70.57
ConDi (ours)	**93.84**	90.62	**94.73**	**92.63**	**93.43**	**93.98**	**93.67**	**88.67**	**86.24**	87.65	**86.94**	**88.39**	**88.54**	**88.46**

BERT [48] and RoBERTa [49]. Additionally, we include state-of-the-art large language models (LLMs) such as Llama 3-8B [50], Qwen 2.5-7B [51], and Phi-4-mini-3.8B [52], which excel at text-only tasks. **(2) Visual models**: For visual inputs, we evaluate baseline models like Image [3] and ViT [53].

Multimodal Baselines: These methods combine both textual and visual information: **(1) Traditional multimodal models**: These methods include HFM [3], D&RNet [54], Res-BERT [10], Att-BERT [10], Att-BERT [10], ViLBERT [55], CMGCN [32], Multi-View CLIP [17], MILNet [15], DMSD-CL [22], G^2SAM [19], and MICL [23]. **(2) Multimodal large language models**: To further evaluate against cutting-edge methods, we include MLLMs such as LLaVA 1.6-7B [56], MiniCPM-V 2.6-8B [57], Qwen 2.5-VL-7B [58], and Gemma 3-12B [59] which are pre-trained on extensive multimodal datasets and demonstrate strong generalization capabilities. To provide a more comprehensive comparison, we also compare with closed-source MLLMs such as GPT5,[4] Claude Sonnet 4,[5] and Gemini 2.5-flash.[6]

Following prior studies [15], [19], [22], we report accuracy, precision, recall, F1-score, and macro-average metrics to comprehensively evaluate model performance.

[4][Online]. Available: https://chatgpt.com/
[5][Online]. Available: https://claude.ai/
[6][Online]. Available: https://gemini.google.com/

C. Main Results

To validate the effectiveness of the proposed ConDi method, we evaluate their performance on the MMSD dataset and MMSD 2.0 dataset. Table II shows that ConDi significantly outperforms existing baseline methods across all metrics, achieving state-of-the-art performance. A deeper analysis of the experimental results leads to the following key insights: **(1) ConDi exhibits outstanding performance**, achieving the best results across most metrics. Additionally, compared to multimodal baselines, the performance difference of ConDi's accuracy on MMSD and MMSD 2.0 is only 5.08%, second only to Multi-View CLIP. This verifies the effectiveness of the multi-view incongruity disentangled learning mechanism in resisting spurious correlations. **(2) Text-only methods outperform image-only methods in sarcasm detection tasks.** Taking RoBERTa as an example, its accuracy on MMSD reaches 85.51%, while the image-only method ViT achieves only 67.83%, with a performance gap of 17.68%. This confirms the irreplaceability of the text modality in capturing subtle semantic patterns like irony and puns. However, this advantage may lead to modality dependence bias, resulting in the complementary information of the visual modality not being fully mined. **(3) Multimodal methods outperform single-modal methods.** Multi-View CLIP achieves F1 scores of 85.55% and 84.10% on MMSD and MMSD 2.0, respectively, showing significant improvement over single-modal methods. This phenomenon reveals the bidirectional compensation

TABLE III
COMPARISON RESULTS ON REDEVAL DATASET (%). † MEANS LORA FINE-TUNED ON THE CORRESPONDING DATASET. ‡ MEANS ZERO-SHOT SETTING, TESTING THE MODEL DIRECTLY ON THE DATASET WITHOUT FINE-TUNING

Method	Trained on MMSD							Trained on MMSD 2.0						
	Acc.	Binary-Average			Macro-Average			Acc.	Binary-Average			Macro-Average		
		P.	R.	F1	P.	R.	F1		P.	R.	F1	P.	R.	F1
RoBERTa	64.34	72.34	13.92	23.45	69.41	55.48	50.13	68.53	77.62	28.10	41.26	72.32	61.42	59.88
ViT	63.75	53.29	63.54	57.97	63.71	63.14	63.05	65.44	54.53	73.16	62.49	66.80	66.08	65.22
Att-BERT	67.53	73.79	27.09	39.63	60.42	70.13	58.71	70.12	68.48	44.56	53.99	65.93	69.58	65.63
Multi-view CLIP	76.29	**75.67**	73.70	74.30	-	-	-	80.98	**80.85**	82.62	80.73	-	-	-
MILNet	67.72	72.04	29.36	41.72	69.47	60.98	59.70	76.09	69.52	69.87	69.69	74.95	75.00	71.38
DMSD-CL	73.30	68.19	60.25	63.97	72.11	71.01	71.38	78.38	75.42	66.83	70.87	77.69	76.35	76.84
MICL	51.10	25.76	12.91	17.20	44.39	41.54	41.25	45.82	35.07	44.30	39.15	45.55	45.75	45.16
Llama 3-8B‡	53.48	43.54	61.51	50.99	54.73	54.89	53.36	53.48	43.54	61.51	50.99	54.73	54.89	53.36
Qwen 2.5-7B‡	50.09	41.89	69.36	52.24	53.66	53.48	49.99	50.09	41.89	69.36	52.24	53.66	53.48	49.99
LLaVA 1.6-7B‡	72.50	72.79	48.10	57.92	72.60	68.22	68.75	72.50	72.79	48.10	57.92	72.60	68.22	68.75
MiniCPM-V 2.6-8B‡	69.92	68.07	44.81	54.04	69.38	65.59	65.89	69.92	68.07	44.81	54.04	69.38	65.59	65.89
Qwen 2.5-VL-7B‡	58.36	47.45	54.17	50.59	57.35	57.63	57.30	58.36	47.45	54.17	50.59	57.35	57.63	57.30
Qwen 2.5-7B†	67.52	69.49	31.14	43.00	68.30	61.13	60.15	75.49	72.37	61.01	66.21	74.71	72.95	73.49
Llama 3-7B†	70.41	73.52	36.70	49.10	72.36	64.49	64.25	76.69	72.67	65.31	68.80	75.78	74.69	75.09
Qwen 2.5-VL-7B†	66.43	70.14	25.56	37.47	67.97	59.25	57.26	75.19	70.62	63.29	66.75	74.15	73.10	73.49
LLaVA 1.6-7B†	72.41	64.25	67.34	65.76	71.19	71.51	71.32	73.50	65.17	70.12	67.56	72.39	72.91	72.58
ConDi	**84.36**	74.18	**92.40**	**82.29**	**84.16**	**85.77**	**84.14**	**87.64**	78.40	**94.68**	**85.77**	**87.21**	**88.88**	**87.43**

mechanism between the text and image modalities. The visual modality can provide context for ambiguous or incomplete text information, while linguistic cues can correct visual misjudgments. **(4) The performance gap between MMSD and MMSD 2.0 highlights the impact of spurious correlations.** Taking G^{2}SAM as an example, its accuracy on MMSD is as high as 90.48%, but it drops sharply to 79.43% on MMSD 2.0, a decrease of 11.05% , indicating that its learning process is seriously disturbed by the spurious correlations of the original dataset. Interestingly, the image-only method ViT shows a counter-trend improvement on MMSD 2.0, which indirectly confirms that the MMSD 2.0 dataset enhances multimodal data by filtering noisy visual patterns. These observations further emphasize that while multimodal methods generally outperform single-modal methods, they are more prone to spurious correlations due to their reliance on interactions between modalities, making them susceptible to noise or irrelevant information. **(5) LLMs and MLLMs perform poorly under zero-shot conditions**, achieving performance comparable to ViT. This clearly indicates that the MSD task is a challenging task. Even powerful closed-source models still fall far short of fine-tuning methods in a zero-shot setting. Moreover, the fine-tuned LLMs (e.g. Qwen 2.5 and Llama 3) and MLLMs (e.g. Qwen 2.5-VL and LLaVA 1.6) have far more parameters than other baselines to achieve comparable performance with ConDi.

The above findings comprehensively reveal the limitations of single-modal methods, the necessity of multimodal fusion, and the decisive impact of data bias on model generalization. These limitations are comprehensively solved through the innovative architecture of ConDi.

D. Performance in OOD Scenario

To validate the robustness and practical applicability of the model in handling out-of-distribution (OOD) data, we design a generalization experiment on the RedEval dataset and FBHM dataset, as shown in Tables III and IV.

TABLE IV
COMPARISON RESULTS ON FBHM DATASET IN OOD SCENARIO.(%). † MEANS LORA FINE-TUNED ON THE CORRESPONDING DATASET

Method	Trained on MMSD		Trained on MMSD 2.0		Trained on FBHM	
	Acc.	F1	Acc.	F1	Acc.	F1
ViLBERT	-	-	-	-	64.70	55.78
DMSD-CL	48.20	43.15	51.70	37.32	61.30	58.33
Multi-view CLIP	50.90	39.95	48.70	48.64	60.20	57.02
MICL	54.00	45.44	53.50	51.46	62.90	59.52
Qwen 2.5 VL-7B†	55.80	**54.71**	56.00	49.85	65.20	62.09
LLaVA 1.6-7B†	**58.10**	52.73	60.20	55.40	66.90	64.39
Condi	56.60	50.62	**60.90**	**57.75**	**68.70**	**66.45**

OOD settings for the same task: As shown in Table III, ConDi performs optimally across most metrics, particularly excelling in recall, demonstrating its strong robustness and out-of-distribution scenarios. The experimental results indicate that models trained on the MMSD dataset generally perform significantly worse compared to those trained on the MMSD 2.0 dataset. This demonstrates that the MMSD dataset tends to make the model focus on domain-specific data features (particularly text modality features), leading it to overly rely on spurious correlations. In contrast, the MMSD 2.0 dataset, after being corrected and optimized, facilitates the model's learning of cross-modal interaction features, resulting in more accurate judgments. Moreover, most baseline methods show significant performance drops compared to main experiments. For instance, MICL suffers over 40% metric decline when trained on MMSD. Despite being designed to reduce spurious correlations, MICL struggles with real OOD data. Although the performance of LLMs and MLLMs is not outstanding under the zero-shot setting, its performance does not decline significantly like other methods, which proves its versatility in OOD scenarios. However, fine-tuned LLM and MLLM do not perform well in OOD scenarios, and in some metrics, their performance is even worse than the original models. This suggests that the prior knowledge

TABLE V
COMPARISON RESULTS ON SPMSD DATASET (%). † MEANS LORA FINE-TUNED ON THE CORRESPONDING DATASET. ‡ MEANS ZERO-SHOT SETTING, TESTING THE MODEL DIRECTLY ON THE DATASET WITHOUT FINE-TUNING

Method	Trained on MMSD							Trained on MMSD 2.0						
	Acc.	Binary-Average			Macro-Average			Acc.	Binary-Average			Macro-Average		
		P.	R.	F1	P.	R.	F1		P.	R.	F1	P.	R.	F1
Att-BERT	58.30	67.56	52.35	58.99	59.23	59.31	58.29	60.10	64.89	66.14	65.51	59.13	59.06	59.09
Multi-view CLIP	61.50	66.37	66.49	66.43	60.65	60.64	60.64	60.60	63.37	**73.99**	68.27	59.18	58.30	58.14
MILNet	56.20	66.83	46.42	54.79	57.96	57.79	56.10	57.80	63.45	62.12	62.78	57.01	57.05	57.02
DMSD-CL	60.60	64.09	71.02	67.38	59.30	58.81	58.82	61.10	66.91	63.52	65.17	60.53	60.68	60.56
MICL	68.70	70.70	**77.48**	**73.94**	68.01	67.20	67.38	62.40	68.55	63.53	65.94	61.99	62.21	61.99
Qwen 2.5-VL-7B‡	57.10	**73.68**	39.09	51.08	61.77	60.17	56.44	57.10	**73.68**	39.09	51.08	61.77	60.17	56.44
LLaVA 1.6-7B‡	49.80	70.48	27.92	40.00	58.52	56.11	50.00	49.80	70.48	27.92	40.00	58.52	56.11	50.00
MiniCPM-V 2.6-8B‡	52.10	67.80	31.23	42.77	57.13	55.66	50.79	52.10	67.80	31.23	42.77	57.13	55.66	50.79
Qwen 2.5-VL-7B†	60.60	71.56	51.83	60.12	62.19	62.09	60.59	64.80	68.08	72.60	70.27	63.86	63.46	63.56
LLaVA 1.6-7B†	63.60	69.80	67.68	68.72	62.69	62.53	62.59	59.10	67.82	54.45	60.40	59.74	59.89	59.05
ConDi	**69.20**	73.29	72.77	73.03	**68.53**	**68.58**	**68.55**	**65.70**	68.91	73.12	**70.95**	**64.81**	**64.43**	**64.53**

gained from large-scale pretraining possesses an antidistribution shift characteristic, while task fine-tuning may disrupt this inherent robustness.

OOD settings for different tasks: As shown in Table IV, to further demonstrate the generalization capability of ConDi, we conduct transfer tests on hateful meme detection task. Specifically, we evaluate the models by training them on the MMSD and MMSD 2.0 datasets and subsequently testing them on the FBHM dataset. For comparison, we also include the in-distribution results, where models are trained and tested on the FBHM dataset. The key conclusions drawn from the results are as follows: (1) ConDi demonstrates the strongest overall performance and generalization ability. In the ideal in-distribution scenario, ConDi achieves the highest accuracy and F1-score, which proves its strong foundational capabilities on this task. In the OOD settings, ConDi also exhibits outstanding performance. When trained on MMSD 2.0, ConDi's accuracy and F1-score are significantly superior to those of other models, showcasing its exceptional ability to learn from a source domain and generalize to a target domain. When trained on MMSD, although LLaVA 1.6-7B achieves a slightly higher accuracy (58.10% vs. 56.60%), ConDi's overall performance remains highly competitive and ranks among the top. This indicates that the generalization performance of ConDi is robust. (2) All models exhibit a noticeable performance degradation in OOD scenarios compared to the in-distribution setting. This performance decay is observed across all compared methods, highlighting the inherent difficulty of generalizing from one data distribution to another and underscoring the significant impact of the domain shift problem on model performance. (3) The quality and relevance of training data are crucial. By comparing the two sets of OOD results from *Trained on MMSD* and *Trained on MMSD 2.0*, we observe that using MMSD 2.0 as the training set generally leads to better performance on the FBHM dataset. This may suggest that the data distribution of MMSD 2.0 is closer to that of the target dataset FBHM, or that its data quality and diversity are higher, thereby facilitating the model's ability to learn more generalizable features.

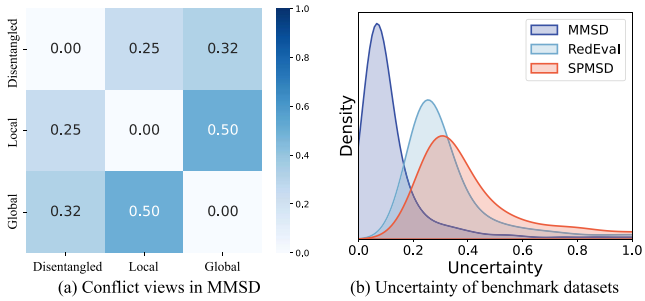
These experiments strongly demonstrate that the ConDi model achieves state-of-the-art performance in multi-modal tasks, both in in-distribution and challenging cross-domain (OOD) scenarios.

E. Performance Against Spurious Correlations

To intuitively evaluate the model's ability to handle spurious correlations, we conduct comparative experiments on the SPMSD dataset, and the results are shown in Table V. Since there are unimodal samples in the SPMSD data, all baselines are multimodal models to ensure valid comparison. The experiments reveal that most baseline methods experience catastrophic performance degradation under this challenging setting, exposing their excessive reliance on spurious features. Additionally, unlike the phenomenon observed in the MMSD dataset, where most methods have higher recall than precision, most baselines on the SPMSD dataset exhibit higher precision than recall. This highlights a general difference in performance indicators, emphasizing the significant impact of different data distributions on the decision-making process of existing models. MICL stands out from other baseline models with its complex data augmentation and can better combat spurious correlations. However, our ConDi performs better on most metrics, indicating that ConDi is robust to spurious correlations while maintaining the best OOD performance. It is worth noting that MLLMs without fine-tuning also struggle in the spurious correlation scenario, and even perform worse than Att-BERT. These results demonstrate that ConDi can significantly improve the reliability of dealing with spurious correlation problems.

F. Credibility Study

To evaluate the effectiveness and credibility of the multi-view incongruity features in different scenarios, we conduct a credibility study. Taking the model trained on the MMSD dataset as an example, we visualize the view conflicts of three incongruity learning, as shown in Fig. 5(a). The results indicate that conflicts of opinions indeed exist between different views, especially the global emotion view, where conflicts are more prominent. Fig. 5(b) shows the uncertainty of the modality disentanglement view varies across different datasets. While the uncertainty is low on the MMSD dataset, it increases 20% on the spurious correlation dataset SPMSD and the ood dataset RedEval. This further shows that focusing solely on one view is insufficient, and information from other views is needed to obtain credible results. In addition, we also conduct a comparative analysis of the accuracy of different views in different data scenarios, as



(a) Conflict views in MMSD (b) Uncertainty of benchmark datasets

Fig. 5. Credibility study, mainly analyzes the conflict and uncertainty of different views.

TABLE VI
EXPERIMENT RESULTS OF ABLATION STUDY (%)

Base	f_d	f_g	f_l	c	MMSD		MMSD 2.0		SPMSD	
					Acc.	F1	Acc.	F1	Acc.	F1
✓					86.97	85.85	83.52	80.08	61.10	66.49
✓	✓				90.22	88.11	85.76	84.34	63.60	68.72
✓	✓	✓			90.73	89.08	86.55	85.54	66.10	72.14
✓	✓		✓		91.61	90.28	87.50	86.07	64.90	70.22
✓	✓	✓	✓		92.07	90.11	87.71	85.38	67.30	71.78
✓	✓	✓	✓	✓	**93.84**	**92.63**	**88.67**	**86.94**	**69.20**	**73.03**

shown in Fig. 6. The study includes the four datasets used in the experiment. The results show that the modality disentanglement view has high accuracy across all datasets, indicating its effectiveness in capturing the contextual information inherent to sarcastic samples. The global sentiment view plays a pivotal role in detecting sarcastic samples and plays a key role in reducing the model's reliance on spurious correlations. Additionally, the local description view can serve as a powerful supplement to support or correct the judgment of the modality disentanglement view, which is particularly evident in SPMSD and RedEval scenarios.

G. Further Analysis

Ablation study of components: To explore the effectiveness of each component in ConDi, we conduct ablation experiments. The experimental results are summarized in Table VI, where *Base* represents the direct connection of visual features $\hat{\boldsymbol{H}}_v$ and text features $\hat{\boldsymbol{H}}_t$. $\boldsymbol{f}_d$, $\boldsymbol{f}_g$ and $\boldsymbol{f}_l$ correspond to the modality disentanglement, global sentiment, and local description incongruity learning modules, respectively. c represents the multi-view feature fusion mechanism based on view conflict. The key findings can be summarized as follows: (1) All incongruity learning modules significantly outperform the base model, verifying the effectiveness of multi-view incongruity modeling. (2) The sentiment module $\boldsymbol{f}_g$ shows specificity advantages on the SPMSD dataset, effectively reducing the model's erroneous dependence on text. (3) $\boldsymbol{f}_l$ achieves significant performance improvement on the MMSD dataset, further verifying the key role of local description incongruity in multimodal tasks. (4) The credibility fusion mechanism c based on view contradiction achieves complementary advantages of multi-view features and improves performance.

TABLE VII
EXPERIMENTAL RESULTS OF USING DIFFERENT BACKBONE(%)

Method	**MMSD**		**MMSD 2.0**		**SPMSD**	
	Acc.	F1	Acc.	F1	Acc.	F1
BERT						
↪ DMSD-CL	88.24	85.43	68.70	61.97	61.10	66.49
↪ G^{2}SAM	90.48	88.48	79.43	78.07	52.90	59.07
↪ **ConDi**	**90.89**	**88.13**	**86.79**	**84.45**	**64.20**	**67.86**
RoBERTa						
↪ DMSD-CL	88.95	86.37	78.33	76.47	60.60	67.38
↪ G^{2}SAM	91.07	89.17	77.96	76.33	55.80	60.81
↪ **ConDi**	**92.28**	**90.07**	**86.34**	**83.95**	**65.70**	**71.67**

TABLE VIII
EXPERIMENTAL RESULTS OF USING EXTRA OCR DATA (%). *OCR'* MEANS USING THE OCR-TEXT EXTRACTED BY [10]

Method	MMSD		MMSD 2.0		SPMSD	
	Acc.	F1	Acc.	F1	Acc.	F1
DMSD-CL	88.95	86.37	78.33	76.47	60.60	**67.38**
+*ocr'*	88.62	86.23	71.65	57.23	59.10	62.64
+ours	**89.04**	**86.88**	78.78	78.62	**60.90**	66.48
G^{2}SAM	90.48	88.48	79.43	78.07	**52.90**	59.07
+*ocr'*	89.67	87.35	72.31	64.00	50.40	59.80
+ours	**90.72**	**89.07**	**80.07**	**79.00**	52.50	**61.96**
ConDi	92.83	91.48	86.71	84.99	67.20	**74.88**
+*ocr'*	91.48	89.87	85.59	83.60	65.40	73.66
+ours	**93.84**	**92.63**	**88.67**	**86.94**	**69.20**	73.03

TABLE IX
EXPERIMENTAL RESULTS OF USING DIFFERENT FEATURE ALIGNMENT METHODS (%). TRAIN TIME MEANS TRAINING TIME PER TERATION, EXTRA #PARA MEANS THE NUMBER OF EXTRA PARAMETERS

Method	**MMSD**		**MMSD2.0**		**Train time**	**Extra #para**
	Acc.	F1	Acc.	F1		
Linear	89.67	87.35	83.52	81.05	**7min 29s**	**0.8M**
Self Attention+Linear	90.36	88.18	84.43	81.66	7min 44s	6.3M
Cross Attention	90.98	88.59	84.56	82.61	7min 36s	2.9M
OT (100 iteration)	92.37	90.21	88.04	86.57	7min 33s	1.2M
OT (200 iteration)	**93.84**	**92.63**	**88.67**	**86.94**	7min 49s	1.2M
OT (400 iteration)	93.80	92.27	86.63	84.40	9min 18s	1.2M
OT (1000 iteration)	91.27	88.99	85.01	82.93	13min 2s	1.2M

Different backbones: In order to verify the compatibility of ConDi with different text encoders and ensure the fairness of the experimental results, we replace ConDi's text encoder with BERT and RoBERTa respectively. We compare them with the baseline model under the same experimental environment. The experimental results are shown in Table VII. ConDi achieves optimal performance when using different text encoders, which fully demonstrate the robustness of the architecture in learning from diverse features.

OCR data: From a data perspective, we conduct ablation experiments to verify the effectiveness of our OCR text. The experimental results are shown in Table VIII, from which we draw the following key conclusions: (1) Additional data does not always improve model performance. In some cases, it may even lead to performance degradation due to differences in data distribution. Taking G^{2}SAM+*ocr'* as an example, its performance on the SPMSD dataset is slightly improved, but its performance on the MMSD dataset is degraded. (2) Our OCR text can improve model performance, and all methods achieve better results on the benchmark datasets.

Parameter scale and efficiency: To comprehensively evaluate the parameter efficiency and training efficiency of ConDi, we

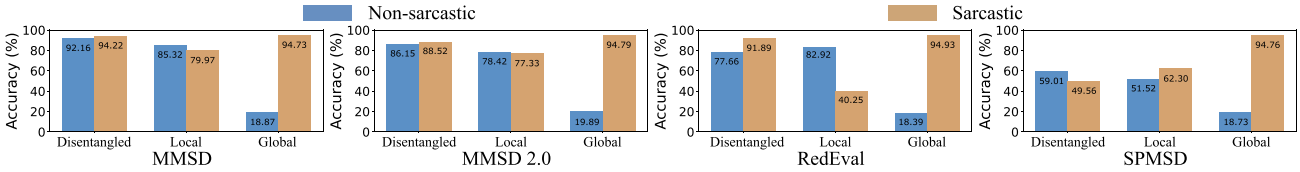
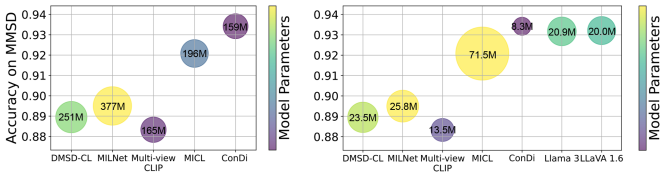


Fig. 6. Comparison of accuracy of different views.



(a) Comparison of whole model parameters (b) Comparison of core parameters

Fig. 7. Comparison of performance and model size of multimodal baselines.

compare the overall parameter scale and core network parameters (excluding the pre-trained encoder) of ConDi with those of baseline models in Fig. 7. For LLM models, only the parameters introduced during the LoRA fine-tuning stage are considered. The experimental results indicate that ConDi not only has the smallest overall parameter scale, but its core network structure is also extremely streamlined, with only 8.3 M parameters. Compared to MICL, ConDi's core network parameters are reduced by nearly 90%. This demonstrates that ConDi can maintain excellent performance while significantly reducing the model space overhead by proposing a more efficient incongruity learning structure.

Feature alignment methods: To evaluate the impact of different feature alignment methods on model performance, we conduct a series of comprehensive ablation studies in Table IX. The results first establish the performance of baseline methods: a simple linear projection achieves an accuracy of 89.67% on the MMSD dataset, which improves to 90.98% with a cross-attention mechanism. Although these attention-based methods show potential for dynamic alignment, their performance gains are moderate. In stark contrast, OT exhibits overwhelming superiority, with its accuracy leaping to 92.37% after just 100 iterations, significantly outperforming all baseline models. An in-depth analysis reveals that increasing the iterations from 100 to 200 pushes the model to its optimal state. It achieves a peak accuracy of 93.84% and an F1-score of 92.63% on the MMSD dataset, while also securing the best results on the MMSD2.0 dataset. This peak performance is attained with remarkable efficiency, as the training time per iteration only marginally increases from 7 m 33 s to 7 m 49 s. However, when the iteration count is further increased to 400 and 1000, not only does the model's performance decline, but the training time also escalates sharply. This finding provides compelling evidence that configuring OT with 200 iterations represents the optimal trade-off between maximizing model performance and controlling computational costs, making it the ideal approach for effective feature alignment.

Disentangled feature dimension: To investigate the impact of the disentangled feature dimension in the cross-attention mechanism on the overall model performance, we conduct a comprehensive evaluation to examine how varying the feature dimension influences performance. A basic principle is that the size of $\hat{\boldsymbol{Z}}_i^f$ is twice that of $\hat{\boldsymbol{Z}}_i^t$, so we use the dimension of $\hat{\boldsymbol{Z}}_i^t$ as the basic unit for explanation. As shown in Fig. 8, when the feature dimension increases from $1 \times d$ to $2 \times d$, both accuracy and F1 score improve substantially, indicating that very low-dimensional representations are insufficient to capture all the critical information required for the task. However, further enlarging the dimension from $2 \times d$ to $64 \times d$ yields only marginal gains, with the performance curve flattening out. Notably, the model already reaches its peak performance at $2 \times d$, demonstrating that relatively low-dimensional features are sufficiently powerful to capture most of the essential semantic information for downstream tasks. This trend is consistently observed across all four datasets, highlighting the robustness and generality of our conclusion.

H. Case Study

To provide an intuitive comprehension of ConDi on spurious correlation samples, we design a case study. We use three replacement strategies to study the impact of spurious correlation on the models, and the experimental results are shown in Fig. 9. All methods can correctly judge whether the original sample is sarcastic, but when the image is replaced with an image with the opposite meaning of the original image, G^{2}SAM and MICL do not make the correct judgment. When replacing the entire text or even just keywords, only MICL among the baselines can correctly detect sarcasm. This shows that the performance of existing methods is indeed constrained by spurious correlation dependence, which seriously affects the generalization ability of the model. While MICL shows robustness in text replacement scenarios, it remains insensitive to image modifications. DMSD-CL and G^{2}SAM can correctly judge in some scenarios, but still have the problem of modality bias learning. ConDi can correctly detect sarcasm in all the above scenarios, emphasizing its generalization ability in MSD.

I. Visualization

To intuitively demonstrate the concerns of modality disentanglement incongruity and local description incongruity learning, we conduct attention visualization experiments. For modality disentanglement incongruity, we extract the attention value of the disentanglement attention stage, and for local description incongruity, we focus on the attention in last layer of the GAT

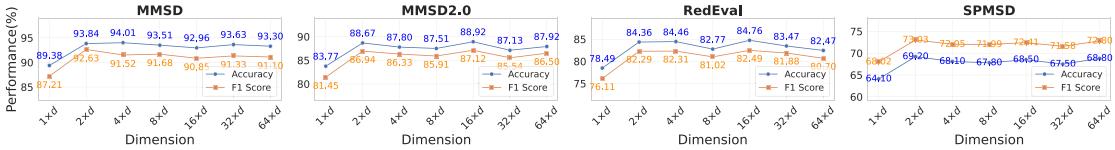


Fig. 8. Impact of disentangled feature dimension on model performance across datasets.

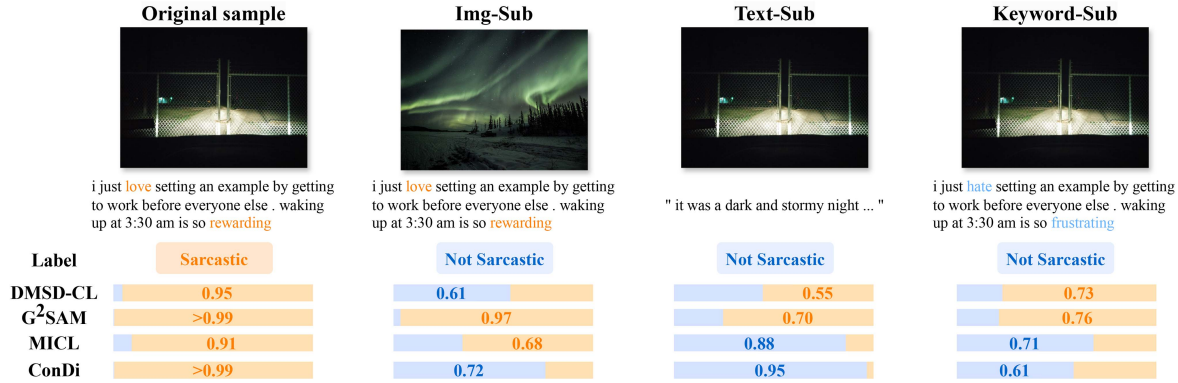


Fig. 9. Case studies. Img-Sub means replacing the image, Text-sub means replacing the text, and Keyword-sub means replacing the keywords.

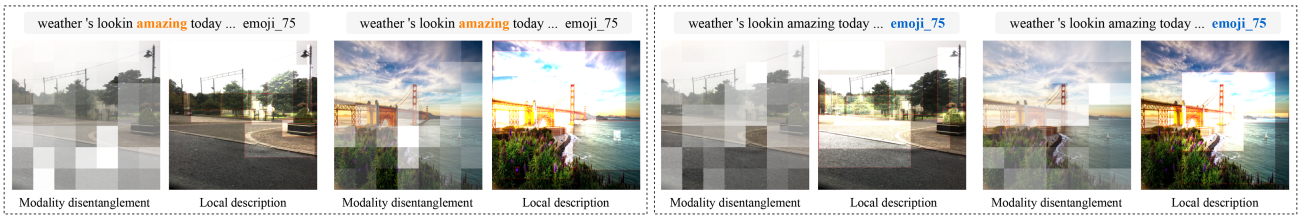
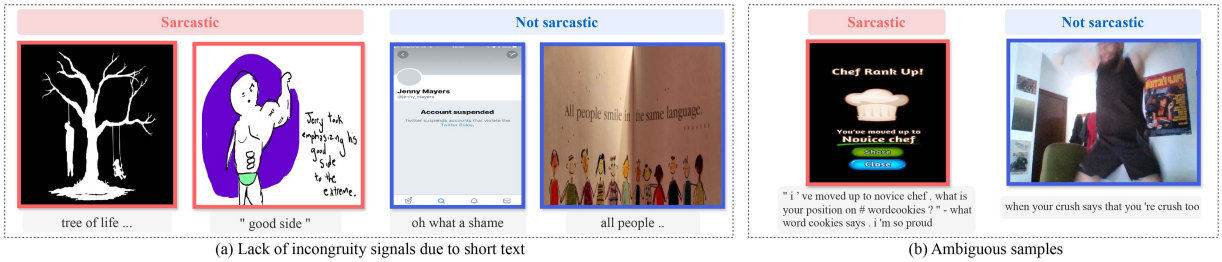


Fig. 10. Attention Visualization. In the token-patch view, the more opaque the area, the more important it is. In the entity-object view, we only show the top 10 important boxes, and the brighter the area, the more important it is.



(a) Lack of incongruity signals due to short text (b) Ambiguous samples

Fig. 11. Failure cases. The main failure cases are concentrated in the recognition of short text sarcasm samples, accounting for nearly 40% of all failure cases.

network. Fig. 10 shows that in sarcastic examples, both methods are able to focus on the inconsistent parts of the image corresponding to the key text. In non-sarcastic examples, the two types of incongruity learning are complementary and enable more comprehensive feature learning. For non-key text, neither method identifies a specific focus area in the image.The visualization results demonstrate that complementary learning from different perspectives can be achieved, highlighting the effectiveness of our ConDi.

J. Failure Cases

We conduct an analysis of failure cases of Condi for uncovering the model's limitations and guiding future research. A core finding is that the model struggles with short-text sarcasm instances as shown in Fig. 11(a). Nearly 40% of all mispredictions are associated with inputs whose text length is substantially shorter than the dataset average. The robustness of ConDi stems from its ability to integrate incongruity

features from three complementary perspectives. However, when the input text is extremely short all three perspectives suffer severe limitations. Local description view relies on syntactic dependency structures to construct semantic graphs. For short texts, meaningful syntactic structures are virtually absent, making graph construction infeasible. Global sentiment view estimates sentiment polarity to assess potential conflicts. Without sufficient context, sentiment polarity extracted at face value is unreliable. Besides, sparse textual tokens hinder the formation of stable alignments in modality disentanglement incongruity. As a result, when textual information is extremely sparse, the three views fail to provide sufficiently informative signals for decision-making, and inter-view complementarity breaks down. Beyond short-text challenges, we found that ConDi sometimes makes mistakes with ambiguous examples. As shown in Fig. 11(b), many such examples exhibit inherently ambiguous sarcasm that even human annotators may find difficult to agree on, making ConDi's errors unsurprising. In summary, despite ConDi's strong overall performance, it continues to face substantial challenges in handling short-text sarcasm. Addressing this issue is an important avenue for future work. Promising directions include: (1) Developing adaptive inference strategies that shift toward vision-centric reasoning when textual signals are weak. (2) Investigating how to transfer complex incongruity patterns learned from long-text or high-signal samples to the short-text setting.

V. Conclusion

This paper proposes ConDi, an innovative approach to learn multi-view incongruity using disentangled learning. The approach aims to enhance model generalization while reducing the number of model parameters, counteracting the widespread spurious correlation problem observed in current MSD models. Specifically, we align modal features through an optimal transport strategy and design a disentangled learning module for learning incongruity accurately and efficiently. We consider the global sentiment view as a key sarcastic supplementary feature. Additionally, we analyze incongruity features for local description in sarcastic contexts by building a multimodal graph. To resolve the contradiction problem generated when merging multiple views, we propose a conflict opinion aggregation method to improve the credibility and accuracy of model judgments. Experimental results show that ConDi not only achieves the best performance on the MSD task with the least parameters, but also effectively alleviates spurious correlation.

References

[1] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert, and R. Huang, "Sarcasm as contrast between a positive sentiment and negative situation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 704–714.

[2] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou, "Low-quality product review detection in opinion summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.-Comput. Natural Lang. Learn.*, 2007, pp. 334–342.

[3] Y. Cai, H. Cai, and X. Wan, "Multi-modal sarcasm detection in twitter with hierarchical fusion model," in *Proc. Assoc. Comput. Linguistics*, 2019, pp. 2506–2515.

[4] D. G. Maynard and M. A. Greenwood, "Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis," in *Proc. Int. Conf. Lang. Resour. Eval.*, 2014, pp. 4238–4243.

[5] J. Fang, W. Wang, G. Lin, and F. Lv, "Sentiment-oriented sarcasm integration for video sentiment analysis enhancement with sarcasm assistance," in *Proc. 32nd ACM Int. Conf. Multimedia*, 2024, pp. 5810–5819.

[6] D. Davidov, O. Tsur, and A. Rappoport, "Semi-supervised recognition of sarcasm in Twitter and Amazon," in *Proc. Conf. Comput. Natural Lang. Learn.*, 2010, pp. 107–116.

[7] M. Zhang, Y. Zhang, and G. Fu, "Tweet sarcasm detection using deep neural network," in *Proc. Int. Conf. Comput. Linguistics*, 2016, pp. 2449–2460.

[8] T. Xiong, P. Zhang, H. Zhu, and Y. Yang, "Sarcasm detection with self-matching networks and low-rank bilinear pooling," in *Proc. World Wide Web Conf.*, 2019, pp. 2115–2124.

[9] A. Joshi, V. Sharma, and P. Bhattacharyya, "Harnessing context incongruity for sarcasm detection," in *Proc. Assoc. Comput. Linguistics*, 2015, pp. 757–762.

[10] H. Pan, Z. Lin, P. Fu, Y. Qi, and W. Wang, "Modeling intra and inter-modality incongruity for multi-modal sarcasm detection," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 1383–1392.

[11] E. Cambria, X. Zhang, R. Mao, M. Chen, and K. Kwok, "SenticNet 8: Fusing emotion AI and commonsense AI for interpretable, trustworthy, and explainable affective computing," in *Proc. Int. Conf. Hum.-Comput. Interaction*, 2024, pp. 197–216.

[12] A. Agrawal, A. An, and M. Papagelis, "Leveraging transitions of emotions for sarcasm detection," in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 1505–1508.

[13] C. Lou, B. Liang, L. Gui, Y. He, Y. Dang, and R. Xu, "Affective dependency graph for sarcasm detection," in *Proc. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 1844–1849.

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

[15] Y. Qiao, L. Jing, X. Song, X. Chen, L. Zhu, and L. Nie, "Mutual-enhanced incongruity learning network for multi-modal sarcasm detection," in *Proc. Conf. Assoc. Advance. Artif. Intell.*, 2023, pp. 9507–9515.

[16] Y. Wei et al., "DeepMSD: Advancing multimodal sarcasm detection through knowledge-augmented graph reasoning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 35, no. 7, pp. 6413–6423, Jul. 2025.

[17] L. Qin et al., "MMSD2.0: Towards a reliable multi-modal sarcasm detection system," in *Proc. Assoc. Comput. Linguistics*, 2023, pp. 10834–10845.

[18] B. Liang, C. Lou, X. Li, L. Gui, M. Yang, and R. Xu, "Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs," in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 4707–4715.

[19] Y. Wei et al., "G ^{2} sam: Graph-based global semantic awareness method for multimodal sarcasm detection," in *Proc. Conf. Assoc. Advance. Artif. Intell.*, 2024, pp. 9151–9159.

[20] D. Yang et al., "Towards multimodal sentiment analysis debiasing via bias purification," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 464–481.

[21] Q. Liu, J. Wu, S. Wu, and L. Wang, "Out-of-distribution evidence-aware fake news detection via dual adversarial debiasing," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 11, pp. 6801–6813, Nov. 2024.

[22] M. Jia, C. Xie, and L. Jing, "Debiasing multimodal sarcasm detection with contrastive learning," in *Proc. Conf. Assoc. Advance. Artif. Intell.*, 2024, pp. 18354–18362.

[23] D. Guo et al., "Multi-view incongruity learning for multimodal sarcasm detection," in *Proc. 31st Int. Conf. Comput. Linguistics*, 2025, pp. 1754–1766.

[24] Y. Deng, Y. Yang, B. Mirzasoleiman, and Q. Gu, "Robust learning with progressive data expansion against spurious correlation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2023, pp. 1390–1402.

[25] G. Sreekumar and V. N. Boddeti, "Spurious correlations and where to find them," in *Proc. 2nd Workshop Spurious Correlations Invariance Stability*, 2023, pp. 1–9.

[26] Z. Cao, Q. Xu, Z. Yang, Y. He, X. Cao, and Q. Huang, "OTKGE: Multi-modal knowledge graph embeddings via optimal transport," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 39090–39102.

[27] N. Wu, S. Jastrzebski, K. Cho, and K. J. Geras, "Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 24043–24055.

[28] S. K. Bharti, K. S. Babu, and S. K. Jena, "Parsing-based sarcasm sentiment recognition in twitter data," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2015, pp. 1373–1380.

[29] T. Ptáek, I. Habernal, and J. Hong, "Sarcasm detection on Czech and english Twitter," in *Proc. 25th Int. Conf. Comput. Linguistics*, 2014, pp. 213–223.
[30] A. Rajadesingan, R. Zafarani, and H. Liu, "Sarcasm detection on tTwitter: A behavioral modeling approach," in *Proc. Int. Conf. Web Search Data Mining*, 2015, pp. 97–106.
[31] R. Schifanella, P. De Juan, J. Tetreault, and L. Cao, "Detecting sarcasm in multimodal social platforms," in *Proc. ACM Multimedia*, 2016, pp. 1136–1145.
[32] B. Liang et al., "Multi-modal sarcasm detection via cross-modal graph convolutional network," in *Proc. Assoc. Comput. Linguistics*, 2022, pp. 1767–1777.
[33] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," 2019, *arXiv:1907.02893*.
[34] Z. Wen and Y. Li, "Toward understanding the feature learning process of self-supervised contrastive learning," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 11112–11122.
[35] P. Kirichenko, P. Izmailov, and A. G. Wilson, "Last layer re-training is sufficient for robustness to spurious correlations," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 29411–29447.
[36] M. Srivastava, T. Hashimoto, and P. Liang, "Robustness to spurious correlations via human annotations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 9109–9119.
[37] S. Wu, M. Yuksekgonul, L. Zhang, and J. Zou, "Discover and cure: Concept-aware mitigation of spurious correlation," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 37765–37786.
[38] J. Nam, J. Kim, J. Lee, and J. Shin, "Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 18222–18237.
[39] Y. Wang, J. Zhang, and Y. Wang, "Do generated data always help contrastive learning," in *Proc. Int. Conf. Learn. Representations*, 2024, pp. 10231–10249.
[40] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.
[41] S. P. Singh and M. Jaggi, "Model fusion via optimal transport," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 22045–22055.
[42] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 2920–2931.
[43] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 16471–16482.
[44] B. Tang, B. Lin, H. Yan, and S. Li, "Leveraging generative large language models with visual instruction and demonstration retrieval for multimodal sarcasm detection," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2024, pp. 1732–1742.
[45] D. Kiela et al., "The hateful memes challenge: Detecting hate speech in multimodal memes," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 2611–2624.
[46] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1746–1751.
[47] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, no. 5-6, pp. 602–610, 2005.
[48] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2019, pp. 4171–4186.
[49] Y. Liu et al., "RoBERTa: A robustly optimized bert pretraining approach," 2019, *arXiv:1907.11692*.
[50] A. Dubey et al., "The llama 3 herd of models," 2024, *arXiv:2407.21783*.
[51] A. Yang et al., "Qwen2.5 technical report," 2024, *arXiv:2412.15115*.
[52] M. Abdin et al., "Phi-4 technical report," 2024, *arXiv:2412.08905*.
[53] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 611–631.
[54] N. Xu, Z. Zeng, and W. Mao, "Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association," in *Proc. Assoc. Comput. Linguistics*, 2020, pp. 3777–3786.
[55] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 13–23.
[56] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 26296–26306.
[57] T. Yu et al., "RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 13807–13816.
[58] Q. Team, "Qwen2.5-vl," Jan. 2025. [Online]. Available: https://qwenlm.github.io/blog/qwen2.5-vl/
[59] G. Team et al., "Gemma 3 technical report," 2025, *arXiv:2503.19786*.

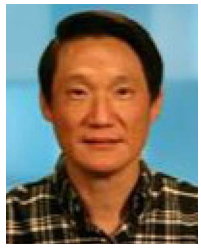
Diandian Guo is currently working toward the Ph.D. degree with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. His research interests include data mining and harmful speech detection.

Hao Peng (Senior Member, IEEE) is currently a Professor with the School of Cyber Science and Technology, Beihang University, Beijing, China. His current research interests include machine learning, deep learning, and reinforcement learning. He is an Associate Editor for *International Journal of Machine Learning and Cybernetics*, *Neural Networks*, and IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING.

Cong Cao is currently an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. His research interests include natural language processes and data mining.

Fangfang Yuan received the Ph.D. degree in computer system architecture from the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China, in 2020. She is currently an Associate Professor with the Institute of Information Engineering, Chinese Academy of Sciences. Her current research interests include information content security, artificial intelligence security, and data mining.

Yanbing Liu is currently a Professor with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China. His research interests include multimodal information processing, data mining, and content security.



Philip S. Yu (Life Fellow, IEEE) is currently a Distinguished Professor and the Wexler Chair of information technology with the Department of Computer Science, University of Illinois Chicago, Chicago, IL, USA. He is a Fellow of the ACM. Dr. Yu has authored or coauthored more than 1,200 referred conference and journal papers cited more than 220,000 times with an H-index of 204. He has applied for more than 300 patents. He was an Editor-in Chief of IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING from 2001 to 2004 and ACM TKDD from 2011 to 2017.