2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Hierarchical Incongruity-Aware Fusion Network with Adaptive Refinement for Multi-modal Sarcasm Detection

Fang Wang[1], Lei Chen[1,2] (✉), and Hao Pan[1]

[1] School of Computer Science and Technology, East China Normal University, Shanghai, China
[2] NPPA Key Laboratory of Publishing Integration Development, ECNUP, Shanghai, China
`51255901030@stu.ecnu.edu.cn, lchen@cs.ecnu.edu.cn`

Abstract. Multi-modal sarcasm detection (MSD) aims to identify sarcastic sentiment conveyed through textual and visual modalities. The key challenge lies in capturing underlying incongruity across modalities. However, many existing studies rely on shallow feature fusion strategies, resulting in limited interaction between textual and visual features. Moreover, they often overlook localized inconsistencies in sarcasm, leading to insufficient representation of fine-grained sarcastic cues. To address these challenges, we propose a hierarchical incongruity-aware fusion network with semantic adaptive refinement (HIAF). Specifically, we first introduce a hierarchical fusion module that progressively captures multi-level incongruity through iterative transformer layers, guided by a cross-modal locality-constrained attention mechanism. Second, we design a semantic adaptive refinement module that dynamically integrates unimodal and cross-modal features based on their contextual contributions. Experiments demonstrate consistent outperformance over strong baselines, validating its capability in capturing multi-modal incongruity.

Keywords: Multi-modal Sarcasm Detection · Multi-modal Fusion · Hierarchical Attention

1 Introduction

Sarcasm is a distinctive linguistic phenomenon characterized by a discrepancy between literal expressions and the speaker’s actual emotional intent [1]. It often manifests through humor, irony, self-deprecation, or mockery to subtly convey implicit attitudes. Accurately identifying sarcasm is essential for uncovering the underlying meaning of language, thereby enhancing semantic understanding in natural language processing. This task holds practical importance in applications such as opinion mining, public sentiment analysis, and customer service [2], and has garnered increasing research attention in recent years.

Early studies [3] on sarcasm detection primarily focused on unimodal data, particularly textual information. However, with the growing prevalence of multi-modal content on social media, users frequently express opinions and emotions through combined text and images. In such contexts, models that rely solely on textual cues while ignoring the visual modality risk missing the semantic and emotional interplay across modalities, ultimately compromising the accuracy of sarcasm recognition.

Multi-modal Sarcasm Detection (MSD) presents unique challenges that distinguish it from conventional multi-modal tasks. While typical multi-modal applications assume semantic consistency or complementarity between textual and visual modalities, the core of sarcasm detection lies in identifying semantic incongruity across modalities. Linguistic theories [4] have highlighted incongruity as a key indicator of sarcastic expression, particularly when the literal meaning of text sharply contrasts with the visual context. For example, the sentence “what a gorgeous day!!” paired with an image of a rainy street vividly illustrates the essence of sarcasm through a strong conflict between modalities (as shown in Fig. 1).

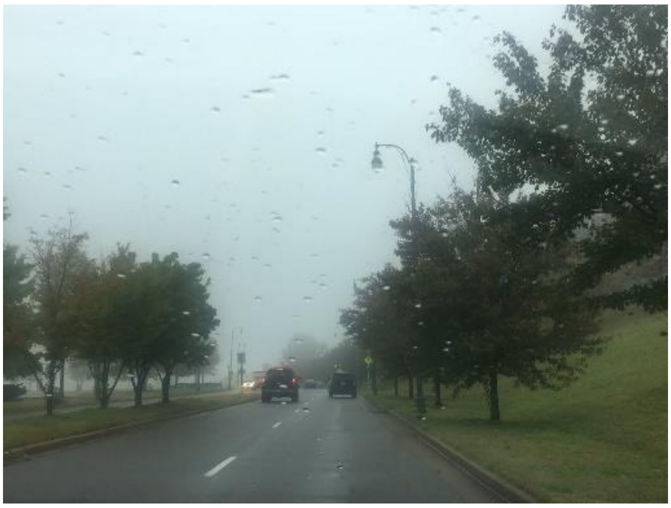


Fig. 1. Example of multi-modal sarcasm in a tweet: the textual expression “what a gorgeous day!!” is contrasted with a rainy street scene, illustrating cross-modal incongruity.

Sarcasm detection has undergone a paradigm shift toward multi-modal modeling, with numerous studies adopting early or late fusion frameworks to incorporate cross-modal information. MSD was first explored in early work [5], and subsequently extended with the release of a benchmark dataset and a hierarchical fusion model [1]. Subsequent research explored various strategies, including neural architectures for semantic alignment [6,7,8] and disentangling shared and specific features [9,10]. However, current methods mainly use attention mechanisms or graph structures to combine features [11,12], often relying on shallow fusion and neglecting deep inter-modal interactions. Moreover, they often overlook fine-grained cues that are essential for identifying the subtle semantic conflicts that define multi-modal sarcasm.

Prior research has explored a range of approaches including neural fusion frameworks, graph-based models, and inter-modal interaction mechanisms to enhance the understanding of cross-modal sarcasm [13,14,15,16]. However, many of these methods fail to explicitly address the central challenge of sarcasm detection: the identification of incongruity. While some recent efforts attempt to capture cross-modal incongruity by modeling global semantic mismatches between modalities [17,18], they tend to underestimate the context-specific nature of sarcasm, which frequently stems from localized inconsistencies between textual and visual fragments. To address this issue, we design a hierarchical incongruity-aware transformer.

In this work, we propose a hierarchical incongruity-aware fusion (HIAF) model for MSD. First, we introduce a hierarchical fusion module to progressively model interactions between modalities. This module employs iterative transformer layers to capture increasingly complex incongruity patterns, guided by a cross-modal locality-constrained attention mechanism. Then, we design an adaptive refinement module that dynamically integrates modality-specific and cross-modal features based on their semantic contribution.

The main contributions can be summarized as follows:

- Our proposed HIAF model is developed to address the core challenge of modeling cross-modal incongruity in MSD. By introducing a hierarchical fusion architecture, our model enables progressive interaction between modalities, bridging the gap left by shallow fusion methods in previous work.
- We develop a locality-aware attention mechanism to highlight fine-grained conflicts between modalities, effectively capturing context-specific incongruity information that are essential for accurate sarcasm recognition.
- We further enhance multi-modal representation through an adaptive refinement module that dynamically aggregates modality-specific and cross-modal features. Extensive experiments demonstrate that our model achieves superior performance on benchmark datasets, validating the effectiveness of each component in addressing the task-specific challenges.

2 Methodology

Given a set of text-image pairs $D = \{(x_i^t, x_i^v) \mid 1 \leq i \leq N\}$, where N denotes the total number of pairs, each sample (x^t, x^v) consists of a textual modality t and a visual modality v. The goal of MSD is to determine the sarcasm label for each text-image pair. To achieve this, we propose a HIAF model that effectively captures and integrates cross-modal incongruity and semantic nuances.

The overall architecture of our proposed HIAF model is illustrated in Fig. 2 and summarized as follows: First, modality-specific encoders extract unimodal representations using pre-trained language and vision backbones. Second, we design a hierarchical fusion strategy that progressively captures local incongruity through iterative attention layers. Third, a semantic adaptive refinement module dynamically aggregates and recalibrates fused features based on their contextual significance. The final representation is then fed into a classifier to determine the presence of sarcasm.

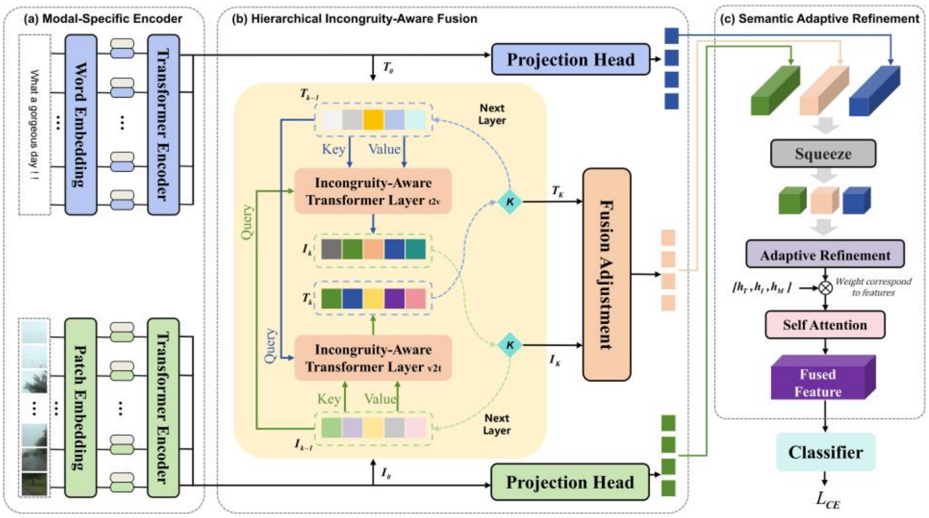


Fig. 2. Overview of the proposed HIAF model with three core components: modal-specific encoder, hierarchical incongruity-aware fusion, and semantic adaptive refinement.

In the following three subsections, we discuss in detail the specific implementation of the proposed innovation modules.

2.1 Modal-Specific Encoder

Text Encoder. To capture rich semantic and contextual information, we adopt a pre-trained transformer-based encoder, which leverages large-scale corpora to embed extensive world knowledge and provides robust feature representations for downstream tasks.

Given a text input $x^t = \{[\text{CLS}], w_1, w_2, \dots, w_{n-1}\}$, where n represents the number of tokens and [CLS] serves as the global token, we utilize BERT to obtain unimodal textual representations:

$$E^t = \text{BERT}(\{[\text{CLS}], w_1, w_2, \dots, w_{n-1}\}) = [e_1^t, e_2^t, \dots, e_n^t], \tag{1}$$

where $e_i^t \in \mathbb{R}^{d_t}$ denotes the token embeddings. The resulting representation E^t is then processed by a multi-layer perceptron to generate the final unimodal textual feature representations:

$$T = [t_1, t_2, \dots, t_n] \in \mathbb{R}^{n \times d}. \tag{2}$$

Image Encoder. Pre-trained vision transformers have demonstrated exceptional capability in various visual tasks. These models, trained on large-scale visual datasets, effectively encode extensive visual knowledge, facilitating superior image feature extraction.

The input image x^v is first resized to 224×224 pixels following standard preprocessing procedures. It is then partitioned into m two-dimensional patches, denoted as $x^v = \{p_1, p_2, \dots, p_m\}$. We employ a pre-trained ViT model as the image encoder, which produces visual feature embeddings:

$$E^v = \mathrm{ViT}(\{p_1, p_2, \dots, p_m\}) = [e_1^v, e_2^v, \dots, e_m^v], \tag{3}$$

where $e_j^v \in \mathbb{R}^{d_v}$ represents the embedding of the j-th image patch. These embeddings are subsequently processed by a multi-layer perceptron to derive the final unimodal visual feature representations:

$$I = [v_1, v_2, \dots, v_m] \in \mathbb{R}^{m \times d}. \tag{4}$$

2.2 Hierarchical Incongruity-Aware Fusion

We present a hierarchical incongruity-aware fusion network to iteratively model local cross-modality incongruity. To this end, we extend the standard transformer architecture into a cross-modality variant that explicitly captures inter-modal dependencies. Each layer integrates a cross-modality attention mechanism guided by specifically designed incongruity-aware constraints.

Hierarchical Iterative Fusion. Existing methods often adopt shallow fusion and fail to model deep inter-modal interactions, limiting their ability to capture fine-grained incongruity. To address this, we propose a hierarchical iterative fusion mechanism that progressively captures cross-modal dependencies at multiple levels. Specifically, we employ iterative incongruity-aware transformer layers to extract increasingly enriched cross-modal representations. Within the cross-modal transformer constrained by local incongruity, textual and visual embeddings are processed through IATL, formulated as:

$$T_k, I_k = \mathrm{IATL}_k(T_{k-1}, I_{k-1}), \quad k \in [1, K], \tag{5}$$

where T_k and I_k represent the outputs of the k-th IATL layer, with initial inputs $T_0 = T$ and $I_0 = I$. K denotes the total number of IATL layers, and the final outputs T_K and I_K serve as the fully refined multi-modal features.

The fusion adjustment layer refines modality interactions by applying stacked transformer layers over concatenated embeddings, followed by attention-based weighting of averaged modality-specific features, enabling adaptive integration of final textual and visual representations.

Incongruity-Aware Transformer Layer. To enable fine-grained information exchange between textual and visual modalities, we design the incongruity-aware transformer layer (IATL) as a bidirectional structure. As shown in Fig. 3, each IATL layer consists of a multi-head co-attention LIGA (MHCALIGA) module followed by a feed-forward network (FFN), with both components equipped with residual connections and layer normalization (LN).

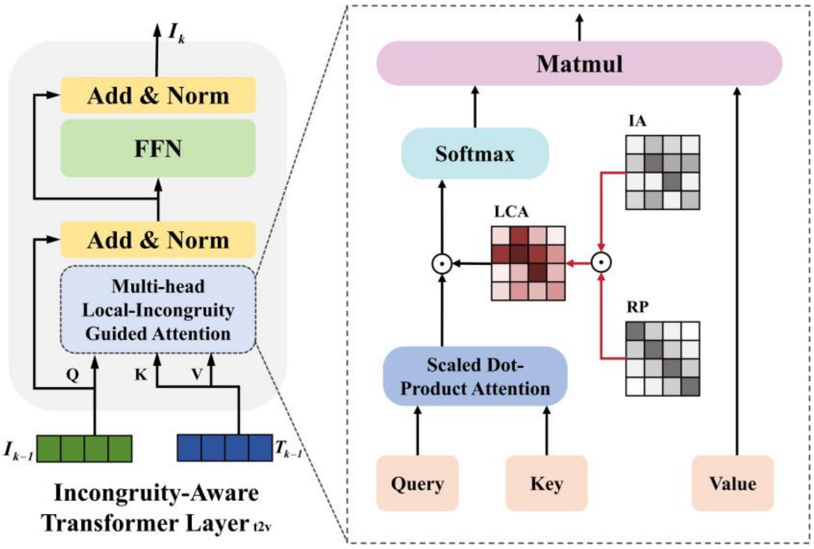

I_k
Add & Norm
FFN
Add & Norm
Multi-head Local-Incongruity Guided Attention
Q
K
V
I_{k-1}
T_{k-1}
Incongruity-Aware Transformer Layer t2v
Matmul
Softmax
IA
LCA
RP
Scaled Dot-Product Attention
Query
Key
Value

Fig. 3. The architecture of the IATL.

The transformation at the k-th IATL layer is defined as:

$$T_{k-1}^{r} = \mathrm{LN}(\mathrm{MHCALIGA}_{k}^{v2t}(T_{k-1}, I_{k-1}) + T_{k-1}), \tag{6}$$

$$I_{k-1}^{r} = \mathrm{LN}(\mathrm{MHCALIGA}_{k}^{t2v}(I_{k-1}, T_{k-1}) + I_{k-1}), \tag{7}$$

$$T_{k} = \mathrm{LN}(\mathrm{FFN}_{k}(T_{k-1}^{r}) + T_{k-1}^{r}), \tag{8}$$

$$I_{k} = \mathrm{LN}(\mathrm{FFN}_{k}(I_{k-1}^{r}) + I_{k-1}^{r}), \tag{9}$$

where $k \in [1, K]$ denotes the layer index. $T_k \in \mathbb{R}^{n \times d_t}$ and $I_k \in \mathbb{R}^{m \times d_v}$ are the outputs of the k-th IATL layer. T_{k-1}^{r} and I_{k-1}^{r} are the intermediate representations produced by the MHCALIGA module.

Within each IATL layer, MHCALIGA employs h parallel attention heads of dimension d_h (with $d_h = d_t/h$). The outputs of all heads are concatenated and linearly projected to yield the final attention result:

$$\mathrm{MHCALIGA}_{k}^{v2t}(T_{k-1}, I_{k-1}) = \mathrm{concat}\left(\left[\mathrm{head}_{i}^{k}\right]_{i=1}^{h}\right) O_{T}^{k}, \tag{10}$$

where $O_T^k \in \mathbb{R}^{d_t \times d_t}$ is a learnable projection matrix, and $\mathrm{head}_i^k \in \mathbb{R}^{n \times d_h}$ is computed via the co-attention LIGA (CALIGA) mechanism with the following formulation:

$$\begin{aligned} \mathrm{head}_{i}^{k} &= \mathrm{CALIGA}_{i}^{k}(T_{k-1}, I_{k-1}) \\ &= \mathrm{CA}_{i,j}^{k}\left(Q_{i,j,k}, K_{i,j,k}, V_{i,j,k}, \mathrm{LIGA}\right) \\ &= \sigma\left(\frac{Q_{i,j,k}K_{i,j,k}^{\mathrm{T}}}{\sqrt{d_h}} \odot \mathrm{LIGA}\right) V_{i,j,k}, \end{aligned} \tag{11}$$

where $\sigma(\cdot)$ denotes the softmax function and $\odot$represents element-wise (Hadamard) product. The attention matrix is defined as $M_{i,j,k} = Q_{i,j,k}K_{i,j,k}^{\top} \in \mathbb{R}^{n\times m}$. The query, key, and value matrices are given by $Q_{i,j,k} = T_{k-1}W_{i,j,k}^{Q}$, $K_{i,j,k} = I_{k-1}W_{i,j,k}^{K}$, $V_{i,j,k} = I_{k-1}W_{i,j,k}^{V}$, with $W_{i,j,k}^{Q} \in \mathbb{R}^{d_t\times d_h}$, $W_{i,j,k}^{K}, W_{i,j,k}^{V} \in \mathbb{R}^{d_v\times d_h}$ as learnable projection weights. The conventional attention weights are further modulated by the LIGA mask to emphasize local incongruity.

Local-Incongruity Guided Attention. Previous studies have employed contrastive learning or consistency-based approaches to align modalities [19,20], primarily focusing on highly correlated regions to infer semantic coherence across modalities. However, such methods often overlook local incongruity, which has been theoretically shown to be a critical indicator of sarcasm [2]. In these approaches, segments with low inter-modal relevance tend to receive lower attention scores during fusion, potentially leading to the loss of crucial sarcastic cues, such as subtle inconsistencies. To address this issue, we explicitly model local incongruity to enhance the model's sensitivity to sarcastic information.

We construct incongruity activation masks (IA) to emphasize inter-modal contrastive information and model the local discrepancy patterns inherent in sarcasm. We employ cosine similarity to quantify the correlation between textual and visual segments. The similarity score is defined as:

$$s_{ij} = \frac{t_i v_j^{\top}}{\|t_i\|\|v_j\|}, \quad i \in [1, N],\ j \in [1, M], \tag{12}$$

where $s_{i,j} \in [-1,1]$ is the relevance score of the textual feature t_i and visual region feature v_j. Inspired by prior work, we construct the incongruity activation masks (IA) as:

$$IA_{i,j} = \exp\left(-\frac{(s_{i,j}-\tilde{a})^2}{2\lambda^2}\right), \tag{13}$$

where $\lambda \in [0,1)$ is a threshold parameter introduced to suppress dominant high-correlation signals and amplify subtle incongruent cues; $\tilde{a}$ is defined as $\mathrm{Median}\big(\{s_{i,j} \mid s_{i,j} \neq 0\}\big)$.

Since both textual and visual features tend to exhibit strong local dependencies within their respective modalities, we introduce a relative position weighting (RP) strategy that considers segment-level positional relationships. The final local-incongruity guided attention (LIGA) is computed as the element-wise product of the sigmoid-activated relative position weighting and the modal-incongruity mask $\mathrm{LIGA} = \mathrm{sigmoid}(\mathrm{RP} \times \mathrm{IA})$, where RP is defined as:

$$\mathrm{RP}_{m,n} = \begin{cases} \mathrm{M} - C(n-m)^2 & \text{if}\, n \leq N, m \leq M, \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

2.3 Semantic Adaptive Refinement

To enhance the final representation by dynamically integrating modality-specific and cross-modal features, we introduce a semantic adaptive refinement (SAR) module. Let $\mathbf{h}_T, \mathbf{h}_I, \mathbf{h}_M \in \mathbb{R}^{d_h}$ denote the encoded representations of the text, visual and cross-modal branches, respectively. These are first concatenated along the modality dimension as $\mathbf{H}_{\text{cat}} = \text{Stack}(\mathbf{h}_T, \mathbf{h}_I, \mathbf{h}_M) \in \mathbb{R}^{d_h \times 3}$.

We then apply a modality-wise attention mechanism to compute the importance of each modality. Specifically, the attention weights $\boldsymbol{\alpha} = [\alpha_T, \alpha_I, \alpha_M] \in \mathbb{R}^3$ are computed as:

$$\boldsymbol{\alpha} = \text{Softmax}\Big(W_2\, \delta\big(W_1\, \text{GAP}(\mathcal{F}(\mathbf{H}_{\text{cat}}))\big)\Big), \tag{15}$$

where $\mathcal{F}(\cdot)$ denotes a transformation we refer to as the squeeze-and-excitation mechanism, which consists of: an initial 1×1 convolutional layer to project input features into a higher-dimensional space, followed by multiple residual blocks, each incorporating convolutional layers and channel-wise recalibration via excitation gates. The final convolutional layer projects the output into a fixed space. GAP($\cdot$) denotes global average pooling, W_1 and W_2 are learnable parameter matrices, and $\delta(\cdot)$ is the ReLU activation. The softmax ensures normalized attention weights across modalities.

The aggregated modality-integrated feature can be equivalently computed via weighted matrix multiplication over the modality axis as $\mathbf{h}_{\text{agg}} = \mathbf{H}_{\text{cat}} \cdot \boldsymbol{\alpha}^{\top}$, where $\mathbf{h}_{\text{agg}} \in \mathbb{R}^{d_h}$ is the aggregated representation, which is further refined by a self-attention layer to produce the final representation $\mathbf{h}_{\text{fused}}$.

2.4 Sarcasm Classifier

Finally, the representation h_{fused} is passed through the classifier F_{agg} to predict the sarcasm label $\hat{y}$. The classifier is implemented as a two-layer fully connected network. The prediction is optimized using cross-entropy loss, defined as:

$$\mathcal{L}_{\text{CE}} = -(y\log(\hat{y}) + (1-y)\log(1-\hat{y})). \tag{16}$$

3 Experiments

This section details the experimental setup and compares the performance of GCLCP with different baseline methods. Additionally, we compare training dynamics and efficiency, perform an ablation study, and analyze the sensitivity to hyperparameters.

3.1 Experimental Configurations

Datasets. We evaluate the effectiveness of the proposed HIAF model on a widely used MSD benchmark dataset [1]. The dataset consists of user posts from Twitter, containing both textual content and corresponding images. A summary of the dataset statistics is shown in Table 1.

Table 1. The statistics of the MSD dataset.

	Train	**Val**	**Test**
Sarcastic	8642	959	959
Non-sarcastic	11174	1451	1450
Total	19816	2410	2409

Implementation Details and Evaluation Metrics. We use ViT-base-patch32-224[2] with 7×7 grids for visual encoding after resizing images to 224×224 pixels, and the first layer of RoBERTa-base for text encoding. The maximum text length and number of image patches are set to $n = 100$ and $m = 49$, respectively. The model includes $K = 3$ IATL layers and $h = 2$ heads in the IATL module, with hidden dimensions dv=dt=d=768. The classifier applies a dropout rate of 0.5. We adopt Adam with a learning rate of 1×10^{-6}, weight decay of 0.01, and a batch size of 32. Following prior work, we report accuracy, precision, recall, and F1-score. The best checkpoint on the val set is used for testing, and results are averaged over five random seeds. All experiments are conducted on GeForce RTX 3080 Ti GPUs.

Baseline. We compare our method with the following baseline methods:

Image-based methods. We consider **ResNet** [22] and **ViT** [23] as image-only baselines, which extract visual features using convolutional or transformer-based backbones but lack the capacity to model semantic incongruity or contextual alignment across modalities.

Text-based methods. Text-only baselines include **Bi-LSTM** [24], **SMSD** [25], and **BERT** [26], which model sequential, self-matching, or global semantic dependencies within text, but fail to capture cross-modal contrasts essential for sarcasm detection.

Multi-modal methods. The following models represent advanced approaches that leverage both textual and visual modalities for MSD.

- **HFM** [1] adopts a hierarchical fusion mechanism to integrate textual and visual features at multiple semantic levels.
- **InCrossMGs** [14] builds intra- and inter-modal graphs to capture fine-grained interactions between image and text.
- **CMGCN** [12] constructs dynamic cross-modal graphs and applies graph convolutional networks to identify modality inconsistencies indicative of sarcasm.
- **HKEmodel** [15] incorporates external commonsense knowledge to enhance both atomic- and composition-level congruity reasoning.
- **GAAN** [17] introduces a global-aware attention mechanism to model sarcasm-related features across different granularities.
- **Multi-View CLIP** [11] leverages CLIP-based encoders and multi-view contrastive learning to align cross-modal semantic spaces.

- **MCEF** [18] employs a multi-channel fusion network to extract complementary cues from different modalities via channel-wise enhancement.
- **SEF** [19] focuses on aligning semantically divergent content between modalities using multi-scale contrastive learning.
- **SAHFN** [16] models sentiment-aware hierarchical fusion by capturing object-attribute relations and multi-level alignment of sarcastic cues.
- **MICL** [27] mitigates spurious correlations by leveraging multi-view incongruity contrastive learning across token-patch, entity-object, and sentiment dimensions.
- **KnowleNet** [28] incorporates external commonsense knowledge and cross-modal semantic similarity to enhance sarcasm detection via ConceptNet and contrastive learning.

3.2 Overall Performance

We evaluate our proposed model against representative baseline approaches on the MSD dataset. The experimental results are reported in Table 2, clearly demonstrating that our model achieves consistent performance improvements across various metrics compared to existing approaches.

Table 2. Comparison of accuracy and F1-score results between the proposed HIAF model and other strong existing models.

Modality	Model	Acc. (%)	Binary-Average		
			Pre. (%)	Rec.(%)	F1.(%)
Text-Only	Bi-LSTM	81.90	76.66	78.42	77.53
	SMSD	80.90	76.46	75.18	75.82
	BERT	83.85	78.72	82.27	80.22
Image-Only	ResNet	64.76	54.41	70.80	61.53
	VIT	67.83	57.93	70.07	63.43
Multi-Modal	HFM	83.44	76.57	84.15	80.18
	D&R Net	84.02	77.97	83.42	80.60
	InCrossMGs	86.10	81.38	84.36	82.84
	CMGCN	87.55	83.63	81.69	84.16
	HKEmodel	87.36	81.84	86.48	84.09
	GAAN	87.42	82.91	86.62	84.72
	Multi-View CLIP	88.33	82.66	88.65	85.55
	MCEF	87.80	84.10	85.50	84.80
	SEF	88.45	**85.35**	86.58	85.96
	SAHFN	87.22	82.71	87.33	84.95
	HIAF (Ours)	**89.65**	**84.36**	**90.39**	**87.27**

- The results show that text-based approaches outperform image-based approaches, indicating inherent challenges in visual modality due to noise and semantic sparsity. For instance, the BERT text-only baseline achieves 83.85% accuracy, significantly surpassing the best image-only baseline ViT (67.83%). By effectively combining visual and textual modalities, our multi-modal model further improves the accuracy to 89.65% and the F1-score to 87.27%, confirming the necessity and effectiveness of multi-modal integration for sarcasm detection.
- Models employing deep cross-modality interactions (e.g., MCEF, SAHFN) generally surpass simpler fusion-based methods (e.g., HFM, D&R Net), highlighting the importance of capturing nuanced multi-modal semantics.
- Additionally, compared with the strong baseline SEF, our model achieves a further accuracy improvement of 1.20% and an F1-score improvement of 1.31%, further validating the advantages of explicitly modeling hierarchical incongruity across modalities.
- We also evaluate the effectiveness of different transformer backbones by comparing BERT and RoBERTa. Results in Table 3 indicate that employing RoBERTa significantly boosts model performance, increasing accuracy from 89.65% to 93.85% and F1-score from 87.27% to 92.31%. This improvement underscores RoBERTa's superior capacity for capturing contextual semantics, further validating the robustness of our modal.

Table 3. Comparison of accuracy and F1-score results between the proposed HIAF model and other transformer-based models.

Modality	**Model**	**Acc. (%)**	**Binary-Average**		
			Pre. (%)	**Rec.(%)**	**F1.(%)**
Text	BERT	83.85	78.72	82.27	80.22
	RoBERTa	85.51	78.24	88.11	82.88
Image + Text	VisualBERT	83.51	76.66	82.94	79.68
	ViLBERT	84.68	77.52	86.37	81.71
	KnowleNet + ALBERT	92.69	91.57	90.85	91.21
	MICL + RoBERTa	92.08	90.05	90.61	90.33
Ours	HIAF + BERT	89.65	84.36	90.39	87.27
	HIAF + RoBERTa	**93.85**	**90.03**	**94.70**	**92.31**

3.3 Ablation Study

To investigate the contribution of each core component in our model, we conduct a comprehensive ablation study, including the removal of key modules and replacement of our proposed model. Specifically, we evaluate the following five variants: (i) w/o LIGA: removing the LIGA mechanism; (ii) w/o IATL: removing the entire IATL; (iii) w/o SAR: removing the SAR module and directly concatenating features; (iv) w/o IATL+Image2Text: replacing IATL with unidirectional image-to-text attention; and (v) w/o IATL+Text2Image: replacing IATL with text-to-image attention. Table 4 shows the corresponding results.

Table 4. Results of ablation experiments on HIAF model.

Method	Acc. (%)	F1. (%)
w/o LIGA	88.58	85.83
w/o IATL	84.23	82.74
w/o SAR	88.62	85.98
w/o IATL+Image2Text	87.65	84.93
w/o IATL+Text2Image	85.35	83.93
HIAF (Ours)	**89.65**	**87.27**

It is evident that removing or altering any component consistently leads to performance degradation, highlighting the effectiveness of each design. Detailed analyses are as follows:

- Removing the LIGA module (w/o LIGA) leads to a 1.07% drop in accuracy and a 1.44% decrease in F1-score, demonstrating its effectiveness in capturing context-specific incongruity essential for accurate sarcasm recognition.
- The w/o IATL variant leads to a substantial performance drop of 5.42% in accuracy and 4.53% in F1-score, confirming the critical role of the hierarchical iterative fusion in enabling the model to capture deep cross-modal interactions.
- Excluding the SAR module (w/o SAR) also leads to a notable performance drop, suggesting that direct feature concatenation fails to capture the relative importance of unimodal and cross-modal features.
- Replacing the IATL module with a unidirectional image-to-text or text-to-image attention mechanism leads to moderate performance degradation. While the image-to-text variant enables the alignment of visual features with textual anchors, it lacks bidirectional interaction and thus fails to capture mutual incongruity comprehensively.

3.4 Hyperparameter analysis

The results in Fig. 4 illustrate that the model performance improves consistently with an increasing number of IATL layers from 1 to 3, reaching an optimal accuracy of 89.65% and an F1-score of 87.28% at three layers. However, further increasing the layers beyond three causes a gradual decline in performance, indicating a saturation in the model's learning capacity. We attribute this phenomenon to the hierarchical incongruity interactions becoming effectively captured by three cascaded layers of the IATL, beyond which additional layers introduce redundancy and limit the model's capacity to accurately encode cross-modal incongruity.

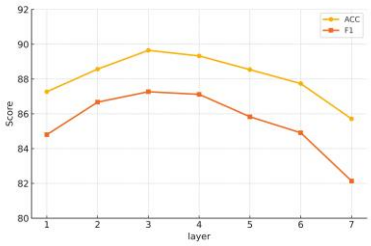


Fig. 4. Performance of using different IATL layers.

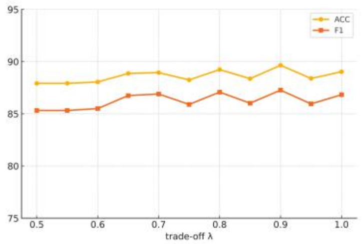


Fig. 5. Performance variations when altering the value of trade-off λ.

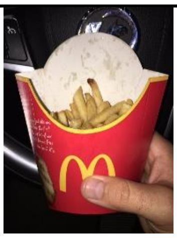
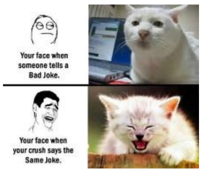
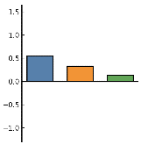
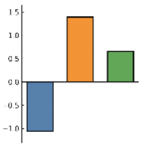
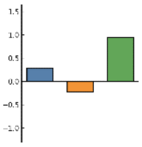
We also analyze the effect of hyperparameter λ in the LIGA module. As shown in Fig. 5, decreasing the value of λ initially improves the model's accuracy, suggesting that focusing on less similar semantic regions enhances the detection of subtle cross-modal incongruities. However, further reducing λ negatively impacts performance, as overly restrictive filtering may exclude essential contextual information from cross-modal interactions. The best performance is achieved at $\lambda = 0.9$, demonstrating an optimal balance between incongruity awareness and semantic completeness.

3.5 Case study

To qualitatively evaluate the effectiveness of our proposed model, we present a case study in Table 5, comparing the predictions of HKEmodal [15], Multi-View CLIP [11], and our HIAF model across three representative examples at the text, image, and multi-modal levels.

HKEmodal successfully identifies sarcasm in the multi-modal example by modeling atomic- and composition-level incongruities, but fails on the unimodal levels due to limited textual and visual representation capabilities. Multi-View CLIP performs well in the image and multi-modal settings by leveraging multi-grained cues from text, image, and interaction views, yet struggles with localized incongruity in text. In contrast, HIAF employs modality-specific encoders for unimodal feature extraction, a hierarchical iterative fusion strategy for cross-modal incongruity modeling, and a semantic adaptive refinement module for context-aware integration. Guided by learned attention weights from SAR, HIAF consistently achieves accurate predictions across all levels, demonstrating its robustness in capturing both unimodal and cross-modal incongruity.

Table 5. Comparison of text-level, image-level, and multi-modal-level predictions with learned weights.

	Text-Level	**Image-Level**	**Multi-modal-Level**
Image		 Your face when someone tells a Bad Joke. Your face when your crush says the Same Joke.	
Text	(a) apparently we have a potato shortage in rotherham this is what i received in a large fries box tonight <user> # valueformoney	(b) lmao ! - pctto	(c) when you want yo hold bae's hands but need to keep it as halaal as possible. # achasorry
● **Text** ● **Image** ● **Multi-modal**			
HKEmodal	Sarcasm	Non-sarcasm	Non-sarcasm
Multi-View CLIP	Non-sarcasm	Sarcasm	Sarcasm
HIAF(Ours)	Sarcasm	Sarcasm	Sarcasm

4 Conclusion

In this paper, we present a novel hierarchical model for MSD, termed HIAF, which comprises two integral components. The hierarchical fusion module progressively models cross-modal interactions by employing iterative incongruity-aware transformer layers, which capture increasingly complex incongruity patterns under the guidance of a cross-modal locality-constrained attention mechanism. The semantic adaptive refinement module dynamically integrates modality-specific and cross-modal representations by weighting their semantic contributions, thereby enabling discriminative feature fusion.

Extensive experiments on the benchmark MSD dataset demonstrate that the proposed HIAF model consistently outperforms existing advanced methods. Ablation studies further validate the effectiveness of each component, confirming their essential roles in enhancing sarcasm detection. In addition, qualitative analyses highlight the model's strong generalization ability and interpretability. These results suggest that HIAF is a promising and effective solution for nuanced multi-modal sarcasm understanding, with significant potential for real-world sentiment-aware applications.

References

1. Cai, Y., Cai, H., Wan, X.: Multi-modal sarcasm detection in twitter with hierarchical fusion model. In: ACL. pp. 2506–2515 (2019)
2. Wen, C., Jia, G., Yang, J.: Dip: Dual incongruity perceiving network for sarcasm detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2540–2550 (2023)
3. Tay, Y., Luu, A.T., Hui, S.C., Su, J.: Reasoning with sarcasm by reading in-between. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1010–1020. Association for Computational Linguistics, Melbourne, Australia (jul 2018). https://doi.org/10.18653/v1/P18-1093, https://aclanthology.org/P18-1093/
4. Joshi, A., Sharma, V., Bhattacharyya, P.: 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). pp. 757–762 (2015)
5. Schifanella, R., De Juan, P., Tetreault, J., Cao, L.: Detecting sarcasm in multimodal social platforms. In: Proceedings of the 24th ACM International Conference on Multimedia. pp. 1136–1145 (2016)
6. Wang, S., Zhang, S., Shen, Y., Liu, X., Liu, J., Gao, J., Jiang, J.: Unsupervised deep structured semantic models for commonsense reasoning. In: Proceedings of the NAACL-HLT. pp. 882–891. Association for Computational Linguistics, Minneapolis, Minnesota (jun 2019). https://doi.org/10.18653/v1/N19-1094, https://aclanthology.org/N19-1094/
7. Ben-Younes, H., Cadene, R., Thome, N., Cord, M.: Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 8102–8109 (2019)
8. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. Advances in Neural Information Processing Systems 34, 14200–14213 (2021)
9. Lu, Y., Wu, Y., Liu, B., Zhang, T., Li, B., Chu, Q., Yu, N.: Cross-modality person re-identification with shared-specific feature transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13379–13389 (2020)
10. Wu, Y., Lin, Z., Zhao, Y., Qin, B., Zhu, L.: A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis. In: Findings of the ACL-IJCNLP 2021. pp. 4730–4738 (2021)
11. Qin, L., et al.: Mmsd2.0: Towards a reliable multi-modal sarcasm detection system. In: Findings of ACL. pp. 10834–10845 (2023)
12. Liang, B., et al.: Multi-modal sarcasm detection via cross-modal graph convolutional network. In: ACL. pp. 1767–1777 (2022)
13. Wang, X., Sun, X., Yang, T., Wang, H.: Building a bridge: a method for image-text sarcasm detection without pretraining on image-text data. In: Proceedings of the 1st International Workshop on Natural Language Processing beyond Text. pp. 19–29 (2020)
14. Liang, B., et al.: Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In: ACM MM. pp. 4707–4715 (2021)
15. Liu, H., Wang, W., Li, H.: Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. In: EMNLP. pp. 4995–5006 (2022)
16. Liu, H., Wei, R., Tu, G., Lin, J., Liu, C., Jiang, D.: Sarcasm driven by sentiment: A sentiment-aware hierarchical fusion network for multimodal sarcasm detection. Information Fusion 108, 102353 (2024)

17. Song, L., Zhao, Z., Ma, Y., Liu, Y., Li, J.: Global-aware attention network for multi-modal sarcasm detection. In: Proceedings of the 2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 2409–2414. IEEE (2023)
18. Fang, H., Liang, D., Xiang, W.: Multi-modal sarcasm detection based on multi-channel enhanced fusion model. Neurocomputing 578, 127440 (2024)
19. Zhong, W., Zhang, Z., Wu, Q., Xue, Y., Cai, Q.: A semantic enhancement framework for multimodal sarcasm detection. Mathematics 12(2), 317 (2024)
20. Wang, J., Yang, Y., Jiang, Y., Ma, M., Xie, Z., Li, T.: Cross-modal incongruity aligning and collaborating for multi-modal sarcasm detection. Information Fusion 103, 102132 (2024)
21. Kim, J., El-Khamy, M., Lee, J.: T-gsa: Transformer with gaussian-weighted self-attention for speech enhancement. In: ICASSP. pp. 6649–6653. IEEE (2020)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
23. Dosovitskiy, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
24. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural networks 18(5-6), 602–610 (2005)
25. Xiong, T., Zhang, P., Zhu, H., Yang, Y.: Sarcasm detection with self-matching networks and low-rank bilinear pooling. In: The World Wide Web Conference. pp. 2115–2124 (2019)
26. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT. pp. 4171–4186 (2019)
27. Guo, D., Cao, C., Yuan, F., Liu, Y., Zeng, G., Yu, X., Peng, H., Yu, P.S.: Multi-view incongruity learning for multimodal sarcasm detection. In: Proceedings of the 31st International Conference on Computational Linguistics. pp. 1754–1766. Association for Computational Linguistics, Abu Dhabi, UAE (jan 2025), https://aclanthology.org/2025.coling-main.119/
28. Yue, T., Mao, R., Wang, H., Hu, Z., Cambria, E.: Knowlenet: Knowledge fusion network for multimodal sarcasm detection. Information Fusion 100, 101921 (2023)