International Journal of Scientific Engineering and Research (IJSER)
ISSN (Online): 2347-3878
SJIF (2025): 8.036

Hierarchical Feature Fusion with Differential Attention Enhancement for Multimodal Sarcasm Detection

Di He

North China Electric Power University, School of Control and Computer, No.2 Beinong Road, Huilongguan, Changping District, Beijing, 102206, China
Email: *1065385644[at]qq.com*

Abstract: ***Multimodal sarcasm detection requires effective modeling of semantic interactions between textual and visual information. Existing approaches often struggle to distinguish sarcasm-relevant cues from redundant noise during cross-modal feature learning. To address this issue, this paper proposes a multimodal sarcasm detection framework based on hierarchical feature fusion and differential attention enhancement. The proposed method incorporates differential attention into a vision-language dual-encoder architecture to suppress noisy attention patterns and emphasize salient sarcasm-related features. A bidirectional cross-modal interaction module is further introduced to capture semantic correspondences and contradictions between text and images. Finally, a hierarchical fusion strategy progressively integrates multimodal information at the feature, semantic, and decision levels. Experiments conducted on the MMSD2.0 dataset demonstrate that the proposed framework achieves 88.10% accuracy, 85.85% precision, 89.20% recall, and 87.29% F1-score, outperforming several existing state-of-the-art methods. The results confirm the effectiveness of differential attention and hierarchical fusion for multimodal sarcasm detection.***

Keywords: Multimodal Sarcasm Detection, Differential Attention, Vision-Language Model, Hierarchical Feature Fusion, Cross-Modal Learning, Vision-Language Pretraining

1. Introduction

With the vigorous development of social media platforms such as Twitter and Weibo, sarcasm, as a complex linguistic phenomenon, has become a common way for netizens to express criticism, humor, or irony. Sarcasm detection holds significant value for applications including sentiment analysis, public opinion monitoring, and human-computer interaction. Traditional sarcasm detection methods primarily rely on textual cues, analyzing linguistic structures, sentiment polarities, and contextual information to judge sarcastic intent. However, social media content is often presented in a multimodal format combining text and images. Relying solely on the textual modality fails to capture the full range of conflict between literal meaning and genuine intent in sarcastic expressions. For example, the text "love traffic" juxtaposed with an image of congested traffic creates a stark contrast- this contradiction between text and image is precisely the key indicator of sarcasm, yet neither text alone nor image alone can reliably identify such sarcastic expressions.

In recent years, multimodal sarcasm detection has gradually become an active research area. This task aims to identify whether a text-image pair contains sarcastic expression by fusing multimodal information. Unlike traditional unimodal sarcasm detection, the core challenge of multimodal sarcasm detection lies in effectively modeling the consistency and contradiction between different modalities. In this context, pretrained vision-language models provide a powerful feature extraction foundation for this task. Models represented by CLIP [15] (Contrastive Language-Image Pre-training) map images and texts into a unified embedding space through large-scale contrastive learning on image-text pairs, demonstrating excellent performance in various cross-modal tasks.

However, existing methods based on vision-language models still face several critical challenges. Recent works have explored various strategies within vision-language pretrained models for multimodal sarcasm detection. Qin et al. [8] proposed a multi-view CLIP framework that integrates features from text, image, and joint text-image encoders. Yu et al. [20] combined CLIP with counterfactual reasoning to construct dual-level indicative representations. Despite these advances, existing methods still rely on standard self-attention mechanisms in the feature extraction stage. Hammoud et al. [16] explicitly pointed out that standard attention mechanisms are susceptible to noise interference when processing complex multimodal inputs, resulting in diffuse attention weights over irrelevant regions. This issue is particularly prominent in sarcasm detection- the semantic contradiction between sarcastic text-image pairs often manifests locally rather than globally, requiring the model to possess fine-grained attention allocation capability. Moreover, existing methods mostly adopt simple concatenation or shallow interaction for multimodal feature fusion, failing to fully exploit deep semantic associations and conflicts across modalities, resulting in cross-modal sarcastic cues being submerged in redundant information. Furthermore, semantic information at different levels- from lexical to sentential, from local visual regions to global visual scenes- contributes differently to sarcasm judgment. How to design an effective multi-level fusion strategy to integrate multi-granularity multimodal features remains an open problem.

To address the above challenges, this paper proposes a multimodal sarcasm detection method based on hierarchical feature fusion and differential attention enhancement. The

differential attention mechanism was originally proposed by Hammoud et al. to enhance cross-modal alignment capability of vision-language models. Its core idea is to cancel out common noise components by computing the difference between two complementary attention distributions, thereby amplifying truly relevant signals. This paper systematically introduces the differential attention mechanism into multimodal sarcasm detection to the best of our knowledge, embedding it into the dual-encoder architecture of vision-language models to enhance the model's perception of sarcasm-relevant features. On this basis, we design a cross-modal interaction module to explicitly model bidirectional semantic associations between text and images, and adopt a hierarchical feature fusion strategy to progressively integrate multimodal information at the feature, semantic, and decision levels, achieving multi-granularity feature representation from fine to coarse. The main contributions of this paper are summarized as follows:

- We introduce the differential attention mechanism into multimodal sarcasm detection. By integrating differential attention into the dual encoders of vision-language models, we effectively suppress attention noise and highlight sarcasm-relevant salient features, improving the model's perception of textual-visual semantic conflicts.
- We design a hierarchical feature fusion strategy that progressively fuses unimodal features and cross-modal interaction features at three levels- feature, semantic, and decision- achieving complementary enhancement of multi-granularity information.
- We conduct extensive experiments on the MMSD2.0 dataset, and the proposed method achieves superior performance over existing methods in accuracy, precision, recall, and F1-score.

2. Related Work

2.1 Multimodal Sarcasm Detection

Research on multimodal sarcasm detection originated from the prevalence of image-text paired content on social media. Early work mainly focused on simple fusion of textual and visual cues. Schifanella et al. [16] first attempted to concatenate textual features with low-level visual features for multimodal sarcasm detection, but the concatenation strategy failed to model deep semantic associations across modalities. Subsequently, researchers explored more sophisticated fusion strategies. Cai et al. [18] proposed the Hierarchical Fusion Model (HFM), which treated global image features, image attribute features, and text word embedding features as three heterogeneous modalities and fused them through a hierarchical gated mechanism, alleviating the alignment difficulty caused by modality heterogeneity to some extent.

With the rise of attention mechanisms, a series of studies adopted cross-modal attention to implicitly integrate image-text information. Xu et al. [19] utilized a co-attention network to simultaneously attend to key words in text and key regions in images, achieving information exchange across modalities through bidirectional flow of attention weights.

In recent years, researchers have gradually realized that the core of sarcasm detection lies in capturing the semantic gap between literal expressions and genuine intentions. Liang et al. [6] proposed a contrastive semantic inconsistency network that explicitly contrasts textual semantics and visual semantics to judge sarcasm. Qin et al. analyzed the issue of spurious cues in the MMSD dataset [8] (such as hashtag words and emojis leading to biased learning) and constructed a revised version, MMSD2.0, along with a multi-view learning framework.

With the advent of vision-language pretrained models (VLPMs), several works have explored their application in multimodal sarcasm detection. VLPMs such as CLIP [15], ALIGN [21], and BLIP [22] learn high-quality cross-modal representations through large-scale pretraining on image-text pairs, providing powerful foundation models for downstream multimodal tasks. CLIP [8] adopts a dual-encoder architecture with Transformer and ResNet/ViT as text and image encoders respectively, mapping image and text features into a unified embedding space via a contrastive learning objective. This architecture is efficient and flexible, allowing independent encoding of unimodal features and similarity measurement via cosine distance.

Qin et al.'s multi-view CLIP [15] framework extracted features from three perspectives—text encoder, image encoder, and joint text-image encoder—and integrated multi-granularity cues for sarcasm judgment, achieving significantly better performance than traditional methods on MMSD2.0. Yu et al. [19] combined CLIP with counterfactual reasoning to construct dual-level indicative representations that capture intra- and inter-modal semantic contrasts.

Despite these advances, existing methods still rely on standard self-attention and cross-attention mechanisms in the feature extraction stage. Hammoud et al. [16] explicitly pointed out that standard attention mechanisms are susceptible to noise interference when processing complex multimodal inputs, resulting in diffuse attention weights over irrelevant regions and reduced cross-modal alignment accuracy. Sarcasm detection is extremely sensitive to cross-modal semantic conflicts- the diffuse attention problem is particularly detrimental because key inconsistency cues may only exist in a local text fragment or a local image region, and diffuse attention dilutes these critical signals with background noise.

To summarize, while prior studies have made notable progress in multimodal sarcasm detection using vision-language models, they predominantly adopt standard self-attention and cross-attention mechanisms. The attention diffusion problem identified by Hammoud et al. [16] remains insufficiently addressed in the context of sarcasm detection, where subtle cross-modal inconsistencies require precise attention allocation. This motivates our introduction of differential attention to mitigate attention noise and enhance the perception of sarcasm-relevant features.

2.2 Differential Attention Mechanism

The differential attention mechanism was originally proposed by Hammoud et al. [16] to enhance the robustness of cross-modal representations in vision-language models. Its core idea stems from an in-depth analysis of the limitations of standard attention mechanisms. The standard scaled

dot-product attention produces a single probability distribution via softmax normalization, which is susceptible to interference from noise components in the input, causing non-zero attention weights to be assigned to semantically irrelevant positions. Differential attention addresses this limitation by introducing two independent linear projections, computing two separate attention distributions, and subtracting them to cancel out common noise components.

Hammoud et al. integrated differential attention into the dual encoders of CLIP [15], replacing the standard self-attention modules. Experiments showed that DiffCLIP [16] consistently outperformed standard CLIP [15] on tasks including zero-shot classification, and image-text retrieval, proving the effectiveness of differential attention in suppressing noise and highlighting salient features. This mechanism aligns well with the needs of sarcasm detection—the semantic contradiction between sarcastic text-image pairs often manifests as local inconsistencies, and differential attention can precisely suppress sarcasm-irrelevant visual backgrounds and textual redundancies, allowing the model to focus on cross-modal conflict regions that truly reflect sarcastic intent. However, existing work has not systematically introduced this mechanism into multimodal sarcasm detection.

3. Method

3.1 Problem Formulation

The multimodal sarcasm detection task is formally defined as follows: given a text-image pair (T, I), where $T = \{w_1, w_2, \ldots, w_n\}$ is a text sequence of length n and I is the corresponding image, the model predicts a binary label $y \in \{0,1\}$, where $y = 1$ indicates the pair contains sarcastic expression and $y = 0$ otherwise. The core objective is to effectively capture the semantic consistency and contradiction between text and image, thereby accurately inferring the author's sarcastic intent.

3.2 Overall Model Architecture

The overall framework of the proposed model, illustrated in **Figure 1**, consists of three cascaded core modules: the differential attention-enhanced dual-encoder module, the cross-modal interaction module, and the hierarchical feature fusion module.

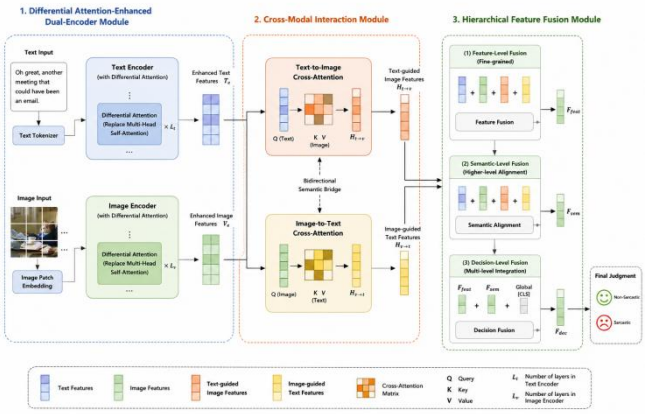

Figure 1: Overall Framework of the Proposed Model.

The differential attention-enhanced dual-encoder module builds upon the dual-encoder architecture of vision-language pretrained models, replacing the standard multi-head self-attention layers in both text and image encoders with differential attention. This module outputs enhanced textual and visual semantic features, where noise components have been effectively suppressed and sarcasm-relevant salient features amplified. The cross-modal interaction module builds upon the differential attention-enhanced unimodal features, establishing bidirectional semantic bridges from text to image and from image to text via bidirectional cross-attention, explicitly capturing cross-modal semantic correspondences and conflicts. The hierarchical feature fusion module progressively integrates unimodal features and cross-modal interaction features at three levels: feature-level fusion integrates fine-grained intra-modal and inter-modal information; semantic-level fusion aligns cross-modal semantic representations at a higher abstraction level; and decision-level fusion integrates multi-level semantic information for final judgment.

3.3 Differential Attention-Enhanced Dual Encoders

Standard multi-head self-attention computes attention via softmax-normalized dot product, which inevitably distributes probability mass to semantically irrelevant positions when the input contains noise or task-irrelevant features. In sarcasm detection, this diffusion is especially harmful because subtle cross-modal inconsistencies may be overshadowed. To address this, we introduce differential attention into both text and image encoders. The mechanism generates two independent attention distributions from separate learned projections and subtracts them, canceling common noise while preserving discriminative signals.

Given an input sequence, we compute two scaled dot-product attentions A_1 and A_2 from two sets of projections. The differential attention is defined as $A_{diff} = A_1 - \lambda A_2$, where λ is a learnable parameter controlling suppression strength, and the output is $\mathrm{DiffAttention}(Q, K, V) = A_{diff} V$. This operation eliminates shared background patterns that both distributions attend to, while retaining their differences which correspond to task-relevant features. We extend to multi-head by applying this differential attention in each head independently.

In the text encoder, differential attention emphasizes sentiment-laden words and intensifiers while suppressing function words and generic descriptors, thereby capturing cues that may contradict the visual semantics. In the image encoder, it focuses on regions semantically mismatched with the text and suppresses irrelevant backgrounds and textures. Through this dual enhancement, both unimodal representations become more discriminative for sarcasm, providing a robust foundation for subsequent cross-modal interaction and fusion.

3.4 Cross-Modal Interaction Module

After obtaining the differential attention-enhanced text features $F_T \in \mathbb{R}^{n \times d}$ and image features $F_I \in \mathbb{R}^{m \times d}$ (where n is the text sequence length, m is the number of image patches, and d is the feature dimension), the cross-modal

interaction module further mines deep semantic associations and conflicts between the two modalities. This module employs bidirectional cross-attention for information flow.

The text-to-image cross-attention is computed as:

$$F_{T\to I} = \text{CrossAttn}(Q_T, K_I, V_I) = \text{softmax}\left(\frac{Q_T K_I^T}{\sqrt{d_k}}\right)V_I \quad (1)$$

where $Q_T = F_T W_Q^{TI}$ originates from text features, and $K_I = F_I W_K^{TI}$, $V_I = F_I W_V^{TI}$ come from image features. This operation allows each text position to retrieve visually relevant information from the entire image. Similarly, the image-to-text cross-attention is:

$$F_{I\to T} = \text{CrossAttn}(Q_I, K_T, V_T) = \text{softmax}\left(\frac{Q_I K_T^T}{\sqrt{d_k}}\right)V_T \quad (2)$$

where $Q_I = F_I W_Q^{IT}$ comes from image features, and $K_T = F_T W_K^{IT}$, $V_T = F_T W_V^{IT}$ come from text features.

The outputs of the bidirectional cross-attentions are concatenated and passed through a two-layer feedforward network with residual connection:

$$F_{cross} = \text{FFN}([F_{T\to I}; F_{I\to T}]) + [F_{T\to I}; F_{I\to T}] \quad (3)$$

The cross-modal interaction feature F_{cross} encodes fine-grained semantic correspondences between text and image. For sarcastic samples, these correspondences often contain local contradictions or inconsistencies- for instance, the word "love" in the text conflicting with the congested traffic scene in the image in terms of sentiment polarity. The bidirectional design enables conflict capture from both perspectives, comprehensively uncovering cross-modal sarcastic cues.

3.5 Hierarchical Feature Fusion Module

The hierarchical feature fusion module progressively integrates multimodal information at three levels, achieving multi-granularity feature representation from fine to coarse.

3.5.1 Feature-Level Fusion

We fuse unimodal features with cross-modal interaction features via learnable gates. For the text modality:

$$F_T^{fuse} = \alpha_T \cdot F_T + (1-\alpha_T)\cdot \text{Proj}(F_{I\to T}) \quad (4)$$

$$\alpha_T = \sigma(W_\alpha[\text{Avg}(F_T); \text{Avg}(F_{I\to T})]) \quad (5)$$

where $\text{Avg}(\cdot)$ denotes global average pooling, σ is the sigmoid function, and W_α is a learnable parameter. The gate α_T adaptively controls cross-modal supplementation.

Similarly, for the image modality, the feature-level fusion is formulated as:

$$F_I^{fuse} = \alpha_I \cdot F_I + (1-\alpha_I)\cdot \text{Proj}(F_{T\to I}) \quad (6)$$

This level preserves intra-modal discriminative information while incorporating targeted cross-modal cues.

3.5.2 Semantic-Level Fusion.

We map the fused features into a unified semantic space:

$$H_T = \text{SemEnc}_T(F_T^{fuse}), \quad H_I = \text{SemEnc}_I(F_I^{fuse}) \quad (7)$$

where $SemEnc$ comprises stacked Transformer layers. Semantic-level cross-attention then aligns the modalities:

$$H_{align} = \text{CrossAttn}(H_T, H_I, H_I) + \text{CrossAttn}(H_I, H_T, H_T) \quad (8)$$

H_{align} encodes semantic alignment degree between text and image. Low alignment regions indicate potential sarcastic contradictions, while high alignment reflects consistency.

3.5.3 Decision-Level Fusion

The feature-level outputs F_T^{fuse} and F_I^{fuse} are concatenated and passed through classifier Cls_1 to obtain s_1, while H_{align} is passed through Cls_2 to obtain s_2:

$$s_1 = \text{Cls}_1([\text{Avg}(F_T^{fuse}); \text{Avg}(F_I^{fuse})]) \quad (9)$$

$$s_2 = \text{Cls}_2(\text{Avg}(H_{align})) \quad (10)$$

A learnable weight β integrates both scores:

$$\hat{y} = \sigma(\beta \cdot s_1 + (1-\beta)\cdot s_2) \quad (11)$$

The three fusion levels complement each other: feature-level fusion preserves fine-grained signals, semantic-level fusion achieves deep cross-modal alignment and conflict measurement, and decision-level fusion integrates multi-level information for comprehensive judgment. This progressive strategy effectively avoids information loss that may result from single-stage fusion.

3.6 Training Objective

The model is trained with binary cross-entropy loss:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)] \quad (12)$$

where N is the batch size, y_i is the ground-truth label of the i-th sample, and $\hat{y}_i$ is the predicted sarcasm probability. During training, the dual-encoder parameters of the vision-language pretrained model are fine-tuned end-to-end, and the learnable parameters in differential attention (λ) and gating parameters ($\alpha_T, \alpha_I, \beta$) are optimized together with other network parameters.

4. Experiments

4.1 Dataset

MMSD2.0 [8] is a revised version of the MMSD dataset proposed by Qin et al. The original version contained spurious cues (e.g., "#sarcasm" hashtags and emojis) and inappropriately labeled negative samples. MMSD2.0 removes these spurious cues and re-annotates unreasonable samples. It comprises approximately 19,000 Twitter image-text pairs, serving as one of the largest public benchmarks for multimodal sarcasm detection.

4.2 Experimental Settings

We adopt the CLIP ViT-B/32 [15] as the vision-language backbone, which consists of a 12-layer Transformer text encoder and a 12-layer Vision Transformer image encoder. Both encoders are initialized with the official pretrained weights provided by OpenAI. The feature dimension d is set to 512 for both modalities. The differential attention mechanism introduces a learnable suppression parameter λ, which is initialized to 0.5 and constrained to the range [0, 1]

via sigmoid activation. In each differential attention head, the two independent projection subspaces share the same feature dimension d_k=64, with the number of attention heads set to 8. For the cross-modal interaction module, we employ a 6-layer Transformer decoder with multi-head cross-attention, where each layer contains 8 attention heads and a feedforward network with a hidden size of 2048. The semantic encoders $SemEnc_T$ and $SemEnc_I$ in the semantic-level fusion stage each consist of 2 stacked Transformer layers with a hidden dimension of 512. The classifiers Cls_1 and Cls_2 are implemented as two-layer multilayer perceptrons (MLPs) with a hidden dimension of 256 and ReLU activation, followed by a dropout rate of 0.1 to prevent overfitting. During training, the AdamW optimizer is used with an initial learning rate of 1×10^{-4}, which is decayed using a cosine annealing schedule. The weight decay coefficient is set to 0.01, and the gradient clipping threshold is set to 1.0 to stabilize training. The batch size is configured as 32, and all input images are resized to 224×224 pixels with standard ImageNet normalization. The text sequence length is truncated to a maximum of 77 tokens. The model is trained for 50 epochs with an early stopping patience of 5 on the validation loss. All reported results are averaged over three independent runs with different random seeds, and all experiments are conducted on a single NVIDIA A100 GPU (40GB).

4.3 Main Results

To validate the effectiveness of our method, we compare against a comprehensive set of baselines spanning text-only models (BiLSTM 0, SMSD [2], BERT [2]), image-only models (ResNet [4], ViT[5]), and state-of-the-art multi-modal approaches including InCrossMGs [6], HKE [6], Multi-view CLIP [8], DIP [9], TFCD[10], MoBA[10], CofiPara [12], ADs [12], and GDCNet [14]. Among these, Multi-view CLIP leverages multi-perspective feature extraction from CLIP encoders, and GDCNet represents a recently proposed approach with competitive performance. These baselines cover a broad spectrum of methodological paradigms, providing a thorough and fair comparison for evaluating the effectiveness of our proposed method. In our experimental evaluation, we adopt accuracy (Acc.), precision (P), recall (R), and F1-score (F1) as the primary evaluation metrics.

Table 1 presents the comparative results on the MMSD2.0 benchmark. Multimodal approaches consistently outperform text-only and image-only models, confirming that integrating both visual and textual modalities is essential for effective sarcasm detection. Our proposed method achieves the highest overall performance with 88.10% accuracy and 87.29% F1-score, surpassing all competing approaches. Compared with GDCNet (87.38% accuracy), our method yields improvements of 0.72 and 0.95 percentage points in accuracy and F1, respectively. Notably, the substantial F1 gain demonstrates superior balance between precision and recall. When compared with CLIP-based methods such as Multi-view CLIP and CofiPara, our approach consistently outperforms them by significant margins. This advantage stems from the differential attention mechanism, which suppresses attention noise and amplifies sarcasm-relevant salient features, enabling more precise capture of subtle cross-modal semantic conflicts. The hierarchical feature fusion strategy further contributes by progressively integrating multi-granularity information. These results collectively validate the effectiveness of our proposed framework.

Table 1: Performance comparison on MMSD2.0.

Method	Acc.	P	R	F1
BiLSTM	72.48	68.02	68.08	68.05
SMSD	73.56	68.45	71.55	69.97
Bert	76.52	74.48	73.09	73.78
ResNet	65.50	61.17	54.39	57.58
ViT	72.02	65.26	74.83	69.72
InCrossMGs	79.83	75.82	78.01	76.90
HKE	76.50	73.48	71.07	72.25
Multi-view CLIP	85.14	80.33	88.24	84.09
DIP	84.63	84.17	85.20	84.68
TFCD	86.54	82.46	87.95	84.31
MoBA	85.01	80.46	87.67	83.64
CofiPara	85.66	85.79	85.43	85.61
ADs	85.60	85.28	85.60	85.41
GDCNet	87.38	83.39	**89.51**	86.34
Ours	**88.10**	**85.85**	89.20	**87.29**

4.4 Ablation Study

To quantitatively assess the independent contributions of each component, we conduct systematic ablation experiments on MMSD2.0. Six configurations are set: (1) w/o differential attention (replacing differential attention with standard attention in encoders); (2) w/o cross-modal interaction (removing bidirectional cross-attention and using only concatenation); (3) w/o hierarchical fusion (using only feature-level fusion, without semantic-level and decision-level fusion); (4) w/o feature-level fusion (replacing feature-level fusion with direct concatenation while keeping semantic and decision levels); (5) w/o semantic-level fusion (keeping feature and decision levels but removing semantic-level fusion); (6)full model. As shown in **Table 2**, removing any component leads to a performance decline, confirming that all modules contribute positively to the final results. The largest drop occurs when differential attention is removed, with accuracy falling from 88.10% to 87.56%, highlighting its critical role in suppressing attention noise and amplifying discriminative features. Removing hierarchical fusion also causes clear degradation, with accuracy dropping from 88.10% to 87.24%, demonstrating the necessity of progressive multi-level integration. Interestingly, feature-level fusion has the greatest influence within the hierarchical fusion strategy, as its removal results in the lowest accuracy of 86.98%. Overall, the ablation results validate that differential attention, cross-modal interaction, and hierarchical fusion together contribute to the superior performance of our proposed method.

Table 2: Systematic ablation experiments on MMSD2.0.

Configuration	Acc.	F1
w/o differential attention	87.56	86.31
w/o cross-modal interaction	87.92	86.67
w/o hierarchical fusion	87.24	87.01
w/o feature-level fusion	86.98	87.18
w/o semantic-level fusion	87.71	86.82
full model	**88.10**	**87.29**

5. Conclusion

This study presented a multimodal sarcasm detection framework that combines differential attention enhancement, bidirectional cross-modal interaction, and hierarchical feature fusion. By suppressing irrelevant attention patterns and strengthening sarcasm-related semantic cues, the proposed method effectively captures textual and visual inconsistencies. Experimental evaluation on MMSD2.0 demonstrated superior performance compared with multiple baseline and state-of-the-art approaches. Ablation studies further verified the contribution of each proposed component. Future work may explore multi-image scenarios, larger vision-language models, and improved model interpretability to further enhance multimodal sarcasm understanding.

References

[1] J. Li, Y. Xu and H. Shi, "Bidirectional LSTM with Hierarchical Attention for Text Classification," *2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, Chengdu, China, 2019, pp. 456-459, doi: 10.1109/IAEAC47372.2019.8997969.

[2] T. Xiong, P. Zhang, H. Zhu, and Y. Yang, "Sarcasm Detection with Self-matching Networks and Low-rank Bilinear Pooling," *in Proceedings of the World Wide Web Conference (WWW '19)*, New York, NY, USA, 2019, pp. 2115-2124.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, Minneapolis, MN, USA, 2019, pp. 4171-4186.

[4] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[5] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv preprint arXiv:2010.11929, 2020.

[6] 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 Proceedings of the 29th ACM International Conference on Multimedia (MM)*, New York, NY, USA, 2021, pp. 4707-4715.

[7] H. Liu, W. Wang, and H. Li, "Towards Multi-Modal Sarcasm Detection via Hierarchical Congruity Modeling with Knowledge Enhancement," *in Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Abu Dhabi, United Arab Emirates, 2022, pp. 4995-5006.

[8] L. Qin, S. Huang, Q. Chen, C. Cai, Y. Zhang, B. Liang, W. Che, and R. Xu, "MMSD2.0: Towards a Reliable Multi-modal Sarcasm Detection System," in Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, 2023, pp. 10834-10845.

[9] C. Wen, G. Jia and J. Yang, "DIP: Dual Incongruity Perceiving Network for Sarcasm Detection," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, 2023, pp. 2540-2550, doi: 10.1109/CVPR52729.2023.00250.

[10] Z. Zhu, X. Zhuang, Y. Zhang, D. Xu, G. Hu, X. Wu, and Y. Zheng, "TFCD: Towards Multi-Modal Sarcasm Detection via Training-Free Counterfactual Debiasing," *in Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, 2024, pp. 6687-6695, doi: 10.24963/ijcai.2024/739.

[11] Y. Xie, Z. Zhu, X. Chen, Z. Chen, and Z. Huang, "MoBA: Mixture of Bi-directional Adapter for Multi-modal Sarcasm Detection," *in Proceedings of the 32nd ACM International Conference on Multimedia (MM)*, New York, NY, USA, 2024, pp. 4264-4272.

[12] H. Lin, Z. Chen, Z. Luo, M. Cheng, J. Ma, and G. Chen, "CofiPara: A Coarse-to-Fine Paradigm for Multimodal Sarcasm Target Identification with Large Multimodal Models," arXiv preprint arXiv:2405.00390, 2024.

[13] S. Jana, S. Danayak, and S.R. Singh, "Adapter-state Sharing CLIP for Parameter-efficient Multimodal Sarcasm Detection," arXiv preprint arXiv:2507.04508, 2025.

[14] S. Zhang, J. Lian, G. Yu, B. Xu, and X. Ao, "GDCNet: Generative Discrepancy Comparison Network for Multimodal Sarcasm Detection," arXiv preprint arXiv:2601.20618, 2026.

[15] A. Radford et al., "Learning Transferable Visual Models from Natural Language Supervision," *in Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021, pp. 8748-8763.

[16] H.A.A.K. Hammoud and B. Ghanem, "DiffCLIP: Differential Attention Meets CLIP," arXiv preprint arXiv:2503.06626, 2025.

[17] R. Schifanella, P.D. Juan, J.R. Tetreault, and L. Cao, "Detecting Sarcasm in Multimodal Social Platforms," *in Proceedings of the 24th ACM International Conference on Multimedia (MM)*, Amsterdam, Netherlands, 2016, pp. 1136-1145.

[18] Y. Cai, H. Cai, and X. Wan, "Multi-Modal Sarcasm Detection in Twitter with Hierarchical Fusion Model," *in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, Florence, Italy, 2019, pp. 2506-2515.

[19] B. Liang, C. Lou, X. Li, M. Yang, L. Gui, Y. He, W. Pei, and R. Xu, "Multi-Modal Sarcasm Detection via Cross-Modal Graph Convolutional Network," *in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, Dublin, Ireland, 2022, pp. 1767-1777, doi: 10.18653/v1/2022.acl-long.124.

[20] Y. Zhang, G. Su, T. Wang, M. Wu, and X. Wei, "A Dual-Level Indicative Representation Learning Method for Multimodal Sarcasm Detection," Information Processing & Management, vol. 63, p. 104839, 2026.

[21] C. Jia et al., "Scaling Up Visual and Vision-Language Representation Learning with Noisy Text Supervision," *in Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021, pp. 4904-4916.

[22] J. Li, D. Li, C. Xiong, and S.C. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," *in Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022, pp. 12888-12900.