Intent-Oriented Hierarchical Incongruity Guidance for Multimodal Sarcasm Detection

1st Siyuan Li
School of Computer Science and Technology
Soochow University
Suzhou, China
20245227058@stu.suda.edu.cn

2nd Bangjun Wang*
School of Computer Science and Technology
Soochow University
Suzhou, China
wangbangjun@suda.edu.cn

***Abstract*—Multimodal sarcasm detection aims to identify whether an image–text pair conveys sarcastic sentiment. Many existing methods have achieved strong performance by modeling cross-modal incongruity. However, these methods typically ignore the important role of implicit visual intent semantics related to sarcasm, thus limiting the further improvement. To address this limitation, we propose an Intent-oriented Hierarchical Incongruity Guidance framework(IHIG). We first leverage a large visual-language model to generate deep visual intent semantics. Then, we design a Hierarchical Incongruity Guidance Module(HIGM), which first models global relationships in each modality to obtain shallow inconsistency signals between text and image, and deep inconsistency signals between text and visual intent semantics. These signals further guide a three-stage interaction and fusion process between the text graph and the visual graph, thus sarcasm-related cross-modal incongruity is strengthened in the graph structure, mitigating the dilution of key cues. Finally, in the graph contrastive regularized representation space, an adaptive Gaussian kernel-weighted prediction strategy is employed to perform multimodal sarcasm detection. Extensive experiments on benchmark datasets demonstrate our approach outperforms previous state-of-the-art baselines.**

***Index Terms*—multimodal sarcasm detection, hierarchical incongruity guidance, visual intent semantics**

I. Introduction

Sarcasm is a rhetorical device in which the literal meaning of an expression contradicts its sarcastic sentiment. In social media, sarcasm is increasingly expressed through multimodal content, where text and images jointly convey meaning. Multimodal Sarcasm Detection (MSD) seeks to identify sarcastic expressions by jointly analyzing text and images from social media posts [1]–[3]. Early studies on MSD primarily concentrated on graph-based modeling, InCrossMGs [4]leveraged Graph Convolutional Networks (GCNs) to capture both intra- and inter-modal relationships. HKEmodel [5] used GCNs to concatenates both atomic-level and composition-level congruity features. G^{2}SAM [6] designed a framework from the perspective of global semantic consistency based on graphs.

Beyond these graph-based approaches, many existing studies also focus on capturing cross-modal incongruity as a key cue for MSD. DIP model [7] emphasized fine-grained inconsistency and distinguished between emotional and factual incongruities. MIL-Net [8] combined local and global incongruity. MICL [9] and AMIF-Net [10] utilized multi-view, ambiguity-aware multi-level incongruity to enhance sarcasm detection. MMFN [11] employed cross-modal attention to amplify incongruity cues and aggregated unimodal with fused predictions. While these studies have recognized and exploited cross-modal incongruity for sarcasm detection, they often lack an explicit understanding of the visual intention embedded in multimodal content, which is crucial for deciphering sarcastic meaning. In such circumstances, the sarcasm information related to visual intent semantics between modalities can easily be obscured during the direct feature fusion process, which may lead to the dilution of subtle sarcasm-relevant cues.

To address these limitations, we propose a novel framework named Intent-oriented Hierarchical Incongruity Guidance framework (IHIG), which integrates three complementary modules, as illustrated in Fig. 1. First, the Intent-oriented Visual Semantic Interpretations (IVSI) module queries a frozen vision–language model with carefully designed intent-oriented prompts to extract high-level visual intent semantics. Its core module, the Hierarchical Incongruity Guidance Module(HIGM), integrates intent-oriented visual semantics from IVSI into a cross-modal incongruity graph and operates in three stages. In the first stage, shallow incongruity signals are derived from text–image interactions and deep incongruity signals from text–intent interactions. In the second stage, to capture fine-grained structural within each modality, HIGM constructs a text graph and an image graph, and applies the shallow and deep incongruities to emphasize sarcasm-relevant words and regions, avoiding diluting sarcastic cues. In the final stage, the enhanced text and image graphs are fused into a cross-modal incongruity graph that serves as basis for prediction. Finally, the Adaptive Gaussion Weighting Module (AGWM) introduces graph-based contrastive learning to encourage samples with similar labels to form tighter clusters in the representation space and makes predictions through adaptive Gaussian-weighted aggregation to achieve MSD.

In summary our main contributions are as follows:

1) We propose the IHIG framework and design a noval IVSI module to extract implict deep visual intent semantics relevant for sarcasm understanding.

2) We propose the HIGM module to exploit shallow and deep incongruities at three stages: early detection to con-

* Corresponding author.

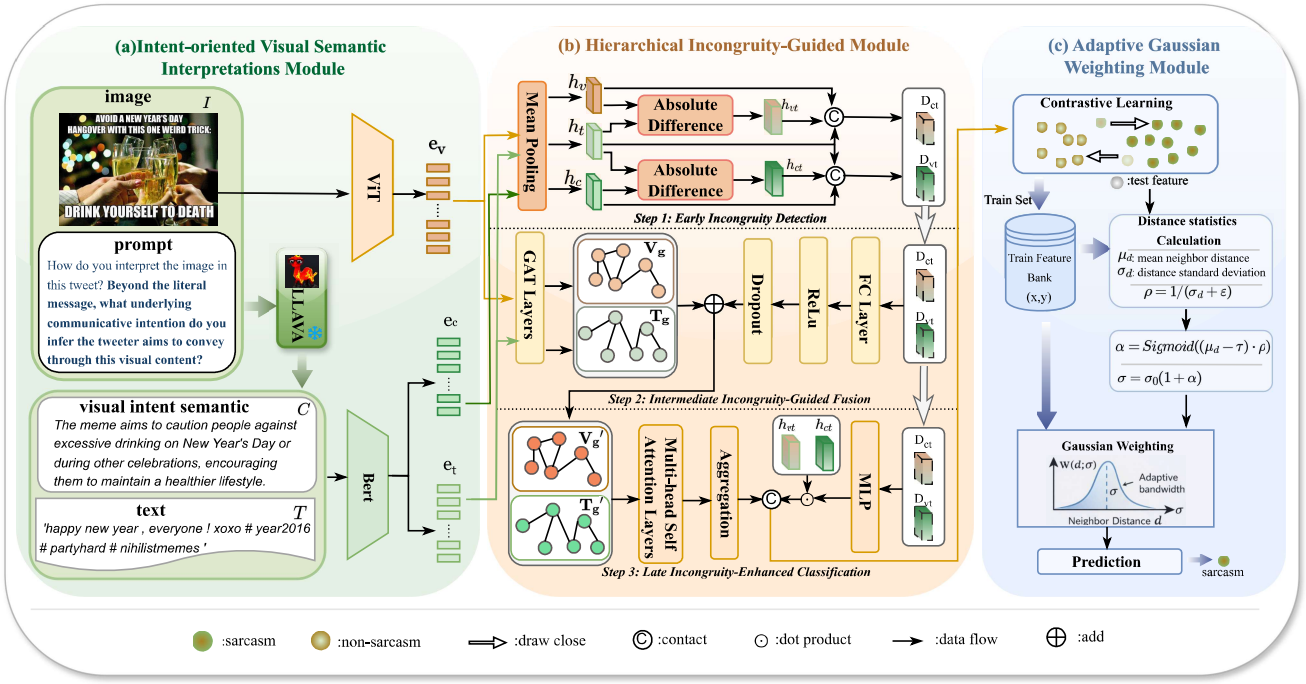

Fig. 1: The Overall Architecture of Our Proposed IHIG

struct incongruity-aware features, intermediate guidance to steer cross-modal fusion, and late enhancement to emphasize sarcasm-relevant cues in classification.

3) Experiments conducted on public datasets demonstrate that IHIG outperforms previous state-of-the-art baselines for MSD tasks.

Our code will be available at https://github.com/msdcnn/IHIG.

II. METHODOLOGY

A. Intent-oriented Visual Semantic Interpretations Module

To obtain deeper intent semantic information from images, we design a set of prompt templates that guide the model to infer implict intent from multiple semantic perspectives rather than relying on purely descriptive outputs. Specifically, we approach visual intention inference from four perspectives: critical analysis, subjective summary, affective atmosphere perception and objective entity overview as visualized in Table I.

Each input image I is paired with a prompt P_k, where $k \in \{1, 2, 3, 4\}$ denotes one of four distinct semantic perspectives. The prompt-image pair is then fed into a frozen vision-language model LLAVA [12] to generate an intent-oriented visual semantic interpretation C_k:

$$C_k = LLM\left(I, P_k\right), \tag{1}$$

where C_k represents the generated deeper semantic information reflecting the inferred visual intent under perspective k. Then, we evaluate all four prompts and name the visual intent semantics that yields the strongest performance as C in the multimodal sarcasm detection task.

B. Hierarchical Incongruity-Guided Module

For a given original text T and the generated visual intent semantics C, we first use a pre-trained BERT-base model [13] to obtain contextualized token-level embeddings $e_t \in \mathbb{R}^{L_t \times d}$ and $e_c \in \mathbb{R}^{L_c \times d}$, respectively, where L_t and L_c are sequence length, d is hidden size. For image I, we utilize a pre-trained ViT [14] to obtain patch-level feature $e_v \in \mathbb{R}^{N_p \times d}$,where N_p is the number of image patches.

TABLE I: Prompt Templates from Different Perspectives

Prompt ID	Perspective	Prompt Template
P1:	Critical Analysis	Examine this image for any contrasts, contradictions, or unusual elements.**Based on these visual inconsistencies, what intent or implied stance do you infer the author is deliberately expressing through the image?**
P2:	Subjective Summary	How do you interpret the image in this tweet? **Beyond the literal message, what underlying communicative intention do you infer the tweeter aims to convey through this visual content?**
P3:	Affective Atmosphere	Analyze this image, including the potential mood, emotions of the people, and the overall atmosphere. **How do these affective cues contribute to an implicit communicative intent or ironic implication suggested by the image?**
P4:	objective entity	Identify and list only the most essential entities (people, objects, setting) in this image briefly and concisely. **Based on these key entities and their configuration, what intended inference or communicative purpose do you believe the creator wants the viewer to draw?**

Then, to measure the relations for sentimental cues of each modality, We use Graph Attention Networks (GAT) [15] to further process e_t and e_v into the token-level textual graphs features T_g and the region-level visual graphs features V_g for the text and image modalities, respectively. Specifically, following HKEmodel [5], we treat text tokens as graph nodes and employ relations between words extracted via spaCy2 as edges. For the visual graphs, we establish edges between object regions according to the cosine similarity scores of their representations.

In order to achieve explicit and hierarchical modeling of incongruity between image and text pairs, we have developed a three-stage approach that leverages cross-modal inconsistency to guide sarcasm prediction.

Early Incongruity Detection: To quantify cross-modal semantic inconsistency, we construct inconsistency features by integrating global relationships. Shallow inconsistency D_{vt} is formed from the vector differences between images and text, while the implicit meanings that are difficult for images to convey are represented with visual intent semantics, and the vector differences between these and the text form deep inconsistency D_{ct}:

$$D_{vt} = [h_t; h_v; |h_t - h_v|], \tag{2}$$

$$D_{ct} = [h_t; h_c; |h_t - h_c|], \tag{3}$$

where [;] denotes concatenation, the global average pooled features h_t, h_v, and h_c are generated from the text features e_t, image features e_v, and visual intent semantic features e_c, respectively.

Intermediate Incongruity-Guided Fusion: To mitigate noise in feature alignment, incongruity features are introduced into text and image graph-structured features to guide the enhancement of fine-grained graph-structured features:

$$V_g{}' = V_g + ReLU\left(W_g D_{vt} + b_g\right), \tag{4}$$

$$T_g{}' = T_g + ReLU\left(W_g D_{ct} + b_g\right), \tag{5}$$

where T_g and V_g are both graph-baesd features, and $W_g \in \mathbb{R}^{d\times 3d}$ and $b_g \in \mathbb{R}^d$ are learnable parameters.

Late Incongruity-Enhanced Classification: To ensure that sarcasm-relevant incongruities are effectively involved in the final classification stage, we employ incongruity features to dynamically adjust cross-modal differences to generate more targeted incongruity-enhanced features, as formalized below:

$$D'_{vt} = \sigma(MLP\left(D_{vt}\right)) \odot |h_v - h_t|, \tag{6}$$

$$D'_{ct} = \sigma(MLP\left(D_{ct}\right)) \odot |h_t - h_c|, \tag{7}$$

where $\sigma(\cdot)$ denotes the sigmoid function, and MLP represents a two-layer feed-forward network with ReLU activation.

To further improve the discriminability of the model, we incorporate both fused multimodal features and incongruity-enhanced features to form an enhanced representation:

$$x = [x_{fused}; D'_{vt}; D'_{ct}], \tag{8}$$

where x_{fused} is the feature fused via a cross-modal Transformer [16] and [;] denotes concatenation.

C. Adaptive Gaussian Weighting Module

This module extends the conventional nearest-neighbor classification paradigm by density-aware adaptive Gaussian kernel weighting. Given a test sample and a set of training samples $\{(x_i, y_i)\}_{i=1}^N$, where y_i is the label of training samples. We first select the top-k nearest neighbors based on Euclidean distance. To quantify the reliability of the neighborhood, the Gaussian kernel width is dynamically adapted to the local feature density, the adaptive kernel width is computed as:

$$\sigma = \sigma_0\Big(1 + \text{sigmoid}\big((\mu_d - 0.5)\rho\big)\Big), \tag{9}$$

where σ_0 is the base kernel width, μ_d is the mean distance of neighbors estimating local dispersion, and ρ is the reciprocal of the distance standard deviation representing an estimate of density.

Finally, the prediction score is obtained by a Gaussian-kernel weighted combination of neighbor labels:

$$\hat{y} = \sum_{i=1}^{k} \frac{\exp\left(-\frac{d_i^2}{2\sigma^2}\right) y_i}{\sum_{j=1}^{k} \exp\left(-\frac{d_j^2}{2\sigma^2}\right)}, \tag{10}$$

where d_i and d_j are the distance between the test and top-k nearest neighbors training samples.

D. Loss Function

We train the multimodal graph fusion model by minimizing the cross-entropy loss $\mathcal{L}_{\text{ce}}$ to optimize the sarcasm detection following prior work [1]. To further enhance representation learning, we also apply graph contrastive loss $\mathcal{L}_{\text{cl}}$ to encourage features of the same class to be closer:

$$\mathcal{L}_{\text{cl}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\mathbf{h}_i^\top \mathbf{h}_i^+ / \tau\right)}{\sum_{j\neq i} \exp\left(\mathbf{h}_i^\top \mathbf{h}_j / \tau\right)}, \tag{11}$$

where $\mathbf{h}_i$ denotes the normalized feature of sample i, $\mathbf{h}_i^+$ and $\mathbf{h}_j$ are its positive sample and negative sample, respectively, and τ is the temperature. Therefore, The total loss is:

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \lambda \mathcal{L}_{\text{cl}}, \tag{12}$$

where λ is a hyper-parameter to balance different losses.

III. EXPERIMENTAL SETUP

A. Dataset and Implementation Details

The primary experiments were carried out using the publicly accessible MMSD dataset [1]. Following the original split, the training set, validation set, and test set with the ratio of 80%:10%:10%. We also utilize the enhanced dataset

TABLE II: The Statistics of Multi-modal Sarcasm Detection Benchmark Datasets

MMSD/MMSD2.0	Training	Validation	Testing
Sarcastic	8642/9572	959/1042	959/1037
No-Sarcastic	11174/10240	1451/1368	1450/1372
Total	19816/19816	2410/2410	2409/2409

TABLE III: Performance Comparison on MMSD and MMSD2.0 Datasets

Modality	Method	MMSD				MMSD2.0			
		Acc(%)	Pre(%)	Rec(%)	F1(%)	Acc(%)	Pre(%)	Rec(%)	F1(%)
Text	TextCNN [18]	80.03	74.29	76.39	75.32	71.61	64.62	75.22	69.52
	SMSD [19]	80.90	76.46	75.18	75.82	73.56	68.45	71.55	69.97
	Bi-LSTM [20]	81.90	76.66	78.42	77.53	72.48	68.02	72.39	70.05
	BERT-Base [13]	83.85	78.72	82.27	80.22	78.35	74.24	72.13	73.17
Image	ResNet [21]	64.76	54.41	70.80	61.53	65.50	61.17	54.39	57.58
	ViT [14]	67.83	57.93	70.07	63.40	72.02	65.26	74.83	69.72
Multi-Modality	HFM [1]	83.44	76.57	84.15	80.18	70.57	64.84	69.05	66.88
	Att-BERT [3]	86.05	80.87	85.08	82.92	80.03	76.28	77.82	77.04
	InCrossMGs [4]	86.10	81.38	84.36	82.84	–	–	–	–
	HKE [5]	87.36	81.84	86.48	84.09	76.50	73.48	71.07	72.25
	Multi-View CLIP [17]	88.33	82.66	88.65	85.55	85.64	80.33	88.24	84.10
	DIP [7]	89.59	87.76	86.58	87.17	85.32	81.76	85.94	83.79
	DMSD-CL [22]	88.95	84.89	87.90	86.37	–	–	–	–
	MILNet [8]	89.50	85.16	89.16	87.11	–	–	–	–
	G^2SAM [6]	90.48	87.95	89.02	88.48	–	–	–	–
	DSP-GCN [23]	90.97	88.28	89.56	88.91	87.22	82.20	88.15	85.07
	MoBA [24]	88.40	82.04	88.31	84.85	85.22	79.82	88.29	84.11
	MMFN [11]	89.31	82.88	90.16	86.71	86.82	81.04	88.72	84.71
	IHIG(Ours)	**91.53**(±0.22)	**88.95**(±0.56)	**90.51**(±0.49)	**89.72**(±0.52)	**88.32**(±0.49)	**83.91**(±0.61)	**88.74**(±0.32)	**86.33**(±0.63)

MMSD2.0 [17], which addresses the shortcomings of the MMSD dataset by removing spurious cues and reassessing reassessing unreasonable samples. Detailed data are shown in Table II.

In our experiments, we extract 49 regions for visual graph modeling, connecting those with cosine similarity over 0.6. The fusion module uses 6 self-attention layers. We use the Adam optimizer during training, where the learning rate is set to 2e-5, weight decay to 5e-3, batch size to 64. To avoid overfitting, early stopping with a patience of 5 is implemented. In graph contrastive learning, τ is fixed at 0.07. The base kernel width σ_0 is set to 0.5 and the number of neighbors k used in the AGWM module is set to 20, which is determined through the analysis in Fig. 3.

TABLE IV: Comparisons of Different Methods on MMSD With the Macro-Scores

Modality	Method	MMSD		
		Macro-Pre(%)	Macro-Rec(%)	Macro-F1 (%)
Text	TextCNN [18]	78.03	78.28	78.15
	SMSD [19]	80.87	78.20	79.51
	Bi-LSTM [20]	80.97	80.13	80.55
	BERT-Base [13]	81.31	80.87	81.09
Image	ResNet [21]	60.12	73.08	65.97
	ViT [14]	65.68	71.35	68.40
Multi-modality	HFM [1]	79.40	82.45	80.90
	Att-BERT [3]	80.87	85.08	82.92
	CMGCN [20]	87.02	86.97	87.00
	InCrossMGs [4]	85.39	85.80	85.60
	HKE [5]	81.84	86.48	84.09
	KnowleNet [25]	88.83	88.21	88.51
	DIP [7]	88.46	89.13	89.01
	DMSD-CL [22]	88.35	88.77	88.54
	MILNet [8]	88.88	89.44	89.12
	G^2SAM [6]	89.44	89.79	89.65
	MMFN [11]	88.07	89.93	88.99
	IHIG(Ours)	**90.73**(±0.35)	**91.07**(±0.19)	**90.88**(±0.22)

B. Comparison Models

To fully validate the performance of IHIG, We compare it with the following baselines.

Text-based methods: TextCNN [18] is a convolutional neural network designed for text classification. SMSD [19] employs an enhanced co-attention to capture text incongruity. BiLSTM [26] employs bidirectional long short-term memory networks for text classification and BERT-Base [13] is a pre-trained model for text classification.

Image-based methods: ResNet [21] employs pooling-layer image embeddings for sarcasm detection. ViT model [14] adopts a pre-trained vision transformer.

Multi-modal methods: HFM [1] applies hierarchical fusion of text, image and image attributes. Att-BERT [3] adds attention mechanisms for MSD. InCrossMGs [4] uses heterogeneous graphs to capture ironic expressions. HKE [5] assesses consistency at atomic and combinatorial levels. Multi-view CLIP [17] uses multi-grained cues by multiple perspectives. DIP [7] models incongruity from factual and affective perspectives. MILNet [8] introduces a tripartite graph structure to identify multimodal incongruities. G^2SAM [6] designs a novel paradigm by using graph-based global semantic awareness. DMSD-CL [22] proposes a debiased contrastive learning framework. MMFN [11] uses multi-task CLIP and multi-granularity fusion. KnowleNet [25] employs a knowledge fusion network for MSD. DSP-GCN [23] adopts dual synergetic perception graph convolutional networks. MoBA [24] leverages a mixture of bidirectional adapters for MSD.

TABLE V: Performance of Different Variants on MMSD and MMSD2.0 Datasets

Variants	MMSD				MMSD2.0			
	Acc	F1	△Acc	△F1	Acc	F1	△Acc	△F1
IHIG	**91.53**	**89.72**	-	-	**88.32**	**86.33**	-	-
w/o D_{ct}	88.63	86.42	-2.90	-3.30	85.61	83.66	-2.71	-2.67
w/o D_{vt}	89.46	87.14	-2.07	-2.58	86.20	84.07	-2.12	-2.26
w/o $HIGM$	90.60	88.38	-0.93	-1.34	86.16	83.96	-2.16	-2.37
w/o $AGWM$	89.72	87.71	-1.81	-2.01	86.84	84.41	-1.48	-1.92
w/o $\mathcal{L}_{\text{cl}}$	90.56	88.00	-0.97	-1.72	87.22	85.00	-1.10	-1.33

IV. EXPERIMENTS AND ANALYSIS

A. Main Results

To evaluate the performance of IHIG, we compare our model against following strong baselines on both the MMSD and MMSD2.0 datasets. As shown in Table III , text-based models generally outperform image-based models because the emotion-related information implicit in images is not fully utilized, while text contains comparatively more emotion-related information. In addition, multimodal models that integrate text and images further enhance the performance of sarcasm detection. However, relying solely on the superficial features of text and images also limits the improvement of the model. In contrast, our approach introduces an understanding and inference of the deeper emotional intentions implicit in images, and explicitly models inconsistencies with the text modality to capture more subtle sarcastic cues, significantly improving the effectiveness of sarcasm detection. Compared with the previous state-of-the-art work MMFN [11] on MMSD dataset, accuracy has increased by more than 2%, precision has improved by over 6%, and the F1 score has risen by more than 3%, demonstrating the effectiveness of our method.

Table IV further compares IHIG with competitive multimodal baselines and shows that it achieves improvements of 23.7% and 7.68% in accuracy compared with the state-of-the-art text-based and image-based models, respectively. Furthermore, compared with state-of-the-art multimodal models, IHIG surpasses the state-of-the-art method, achieving a substantial improvement of 1.89% macro-F1, which indicates that IHIG not only excels on dominant categories but also performs more reliably on minority or subtle sarcastic cases. Since there are fewer baselines for Macro-Score on MMSD2.0, only the results on MMSD are shown.

As shown by the results of our model IHIG in Tables III and Table IV, we further conducted statistical significance tests, which demonstrate that IHIG significantly surpasses the baseline model on most of the assessment metrics (p - value < 0.05). Notably, standard deviations remain below 1% across all evaluation metrics, indicating that the improvement is stable.

B. Ablation Study

In order to evaluate the effectiveness of different components, we conduct an ablation study on both MMSD and MMSD2.0 datasets, with the results presented in Table V.

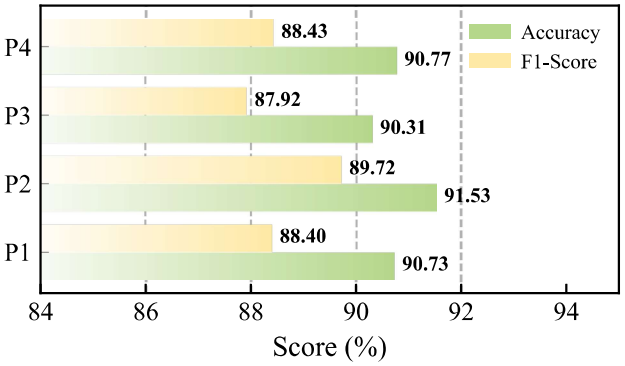

Fig. 2: Performance of Different Prompt Templates on MMSD Datasets

The results show that removing *(w/o D_{ct})* that models the inconsistency between the implicit visual intent semantic and text, as well as removing *(w/o D_{vt})* that models the inconsistency between images and text, both lead to a decrease of more than 2% in Accuracy and F1 scores, and the former decreases more than the latter. This observation indicates that implicit semantic intent derived from images provides richer emotional and sarcastic information than raw visual features, thereby enabling more efficient utilization of implicit visual intent information and text for inconsistency modeling. When the Hierarchical Incongruity Guidance Module*(w/o HIGM)* is removed and only the implicit visual intent semantics are concatenated with the text, a performance decline can be observed, indicating the important contribution of HIGM. In addition, when the Adaptive Gaussian Weighting Module *(w/o AGWM)* is removed, the performance exhibits a consistent decrease, demonstrating that AGWM also effectively improves the overall model performance. Finally, we observe that removing the contrastive learning loss module also leads to a certain decline in performance, as it influences the prediction

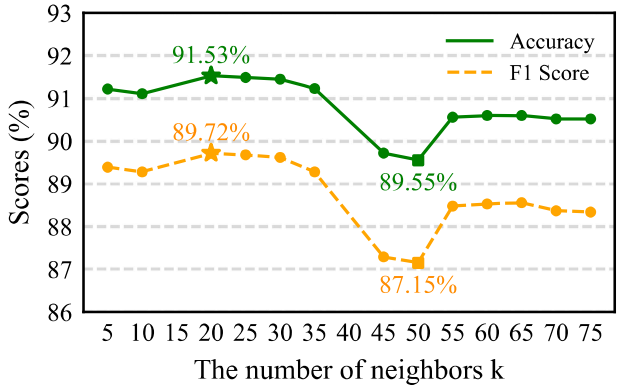

Fig. 3: Performance of Different k in AGWM on the MMSD Dataset

ability of the AGWM module by pulling positive samples closer together.

C. Impact of Different Prompts Templates

To investigate the effect of different prompt designs for MSD, we evaluate model performance using four candidate intent-oriented prompt templates. Each prompt guides the vision-language model LLAVA [12] to generate a semantic interpretation under a distinct perspective as introduced in Section II-A. For each prompt, we evaluate different prompt templates on the validation set using Accuracy and F1-score. As shown in Fig. 2, the results exhibit significant differences across different prompt designs, indicating that prompt formulation has a substantial impact for MSD. Notably, P2 achieves the best overall performance which is intended for subjective summary. Consequently, we choose P2 as the prompt template for our experiments.

D. Impact of Neighbor Numbers k in AGWM

As illustrated in Fig. 3, although the performance is influenced by different neighbor numbers, the variation remains relatively small and consistently high performance over a broad range, confirming the robustness of AGWM. When k is too small, the local neighborhood information is insufficient, leading to unstable feature aggregation. Conversely, overly large k introduces redundant or noisy neighbors, which dilutes the discriminative signals for MSD. When $k = 20$, the model achieves the best performance, which indicates that AGWM achieves the optimal balance between information and noise. Therefore, we adopt $k = 20$ as the default AGWM setting in all subsequent experiments.

V. CONCLUSION

This paper presents an intent-oriented hierarchical incongruity guidance framework for MSD by leveraging visual intent semantics generated from a large vision–language model to derive shallow and deep inconsistency signals, which explicitly enhance cross-modal incongruity modeling and more effectively guide sarcasm prediction. Experimental results on benchmark datasets demonstrate consistent improvements over previous state-of-the-art methods, confirming the effectiveness of explicitly incorporating visual intent semantics into incongruity construction. Future work will consider extending this paradigm to broader multimodal understanding tasks.

REFERENCES

[1] 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*, 2019, pp. 2506–2515.

[2] Y. Tay, A. Luu, S. Hui, and J. Su, "Reasoning with sarcasm by reading in-between," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1010–1020.

[3] H. Pan, Z. Lin, P. Fu *et al.*, "Modeling intra and inter-modality incongruity for multi-modal sarcasm detection," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1383–1392.

[4] B. Liang, C. Lou, X. Li *et al.*, "Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4707–4715.

[5] 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*, 2022, pp. 4995–5006.

[6] Y. Wei, S. Yuan, H. Zhou, L. Wang, Z. Yan, R. Yang, and M. Chen, "G^{2}SAM: Graph-based global semantic awareness method for multi-modal sarcasm detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 9151–9159.

[7] C. Wen, G. Jia, and J. Yang, "Dip: Dual incongruity perceiving network for sarcasm detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2540–2550.

[8] Y. Qiao, L. Jing, X. Song, X. Chen, L. Zhu, and L. Nie, "Mutual-enhanced incongruity learning network for multi-modal sarcasm detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 8, 2023, pp. 9507–9515.

[9] D. Guo, C. Cao, F. Yuan *et al.*, "Multi-view incongruity learning for multimodal sarcasm detection," in *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 1754–1766.

[10] K. Li, Y. Chen, Q. Wu, W. Mai, F. Li, and Y. Xue, "Ambiguity-aware multi-level incongruity fusion network for multi-modal sarcasm detection," in *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 380–391.

[11] L. Ou and Z. Li, "Multi-modal sarcasm detection on social media via multi-granularity information fusion," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 21, no. 3, pp. 1–23, 2025.

[12] H. Liu, C. Li, Q. Wu *et al.*, "Visual instruction tuning," in *Advances in neural information processing systems*, vol. 36, 2023, pp. 34 892–34 916.

[13] J. Devlin, M. 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: human language technologies*, 2019, pp. 4171–4186.

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.

[15] P. Veličković, G. Cucurull, A. Casanova *et al.*, "Graph attention networks," *arXiv preprint arXiv:1710.10903*, 2017.

[16] L. Zhu and Y. Yang, "Actbert: Learning global-local video-text representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8746–8755.

[17] L. Qin, S. Huang, Q. Chen *et al.*, "Mmsd2.0: Towards a reliable multi-modal sarcasm detection system," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 10 834–10 845.

[18] B. Liang, C. Lou, X. Li, L. Gui, M. Yang, and R. Xu, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1746–1751.

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

[20] B. Liang, C. Lou, X. Li, M. Yang, L. Gui, Y. He, W. Pei, and R. Xu, "Multi-modal sarcasm detection via cross-modalgraph convolutional network," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, vol. 1, pp. 1767–1777, 2022.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[22] M. Jia, C. Xie, and L. Jing, "Debiasing multimodal sarcasm detection with contrastive learning," in *Proceedings of the 2024 AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 18 354–18 362.

[23] X. Zhuang and Z. Li, "Multi-modal sarcasm detection via dual synergetic perception graph convolutional networks," in *IEEE International Conference on Data Mining*, 2024, pp. 971–976.

[24] Y. Xie, Z. Zhu, X. Chen *et al.*, "Moba: Mixture of bi-directional adapter for multi-modal sarcasm detection," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 4264–4272.

[25] T. Yue, R. Mao, H. Wang, Z. Hu, and E. Cambria, "Knowlenet: Knowledge fusion network for multimodal sarcasm detection," *Information Fusion*, vol. 100, p. 101921, 2023.

[26] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005.