DDD: Image Description-Guided Dual-Chain Enhancement Model Based on Dynamic Fusion

Minghua Luan[1,2,3], Jun Lu[1,2,3*]
[1]College of Computer Science and Technology, Heilongjiang University, Harbin 150080, China
[2]Jiaxiang Industrial Technology Research Institute of Heilongjiang University, Jining 272400, Shandong Province, China
[3]Key Laboratory of Database and Parallel Computing of Heilongjiang Province, Heilongjiang University, Harbin 150080, China
Email:lujun@hlju.edu.cn, lmh@s.hlju.edu.cn

***Abstract*—Multimodal Aspect-Based Sentiment Analysis (MABSA) aims to extract aspect terms and perform sentiment classification from text-image pairs. Existing methods suffer from two limitations: Firstly, the differences in information density of intra-modal features are overlook. Secondly, the semantic correlation and collaboration mechanisms between modalities are insufficient. This paper proposes an Image Description Guided Dual Chain Enhancement Model Based on Dynamic Fusion (DDD), which synergistically tackles the above problems from two dimensions: intra-modal enhancement and inter-modal collaboration. Specifically, the Image Description Guided module (IDG) leverages image descriptions generated by GPT-3.5 to weight and guide text features, making up for the weak semantic correlation between text and images. The Dual Chained Enhancement (DCE) Module adopts a dual-chain architecture to filter and enhance features within each modality. The Dynamic Fusion module (DFM) dynamically optimize modal collaboration based on inter-modal correlations. Experiments on two benchmark datasets demonstrate that DDD achieves performance improvements compared with existing methods.**

***Index Terms*—Multimodal aspect-based sentiment analysis, Image Description Guided, Dual Chained Enhancement, Dynamic Fusion**

I. Introduction

Multimodal aspect-based sentiment analysis (MABSA) [1] has received great attention due to its important application in analyzing social media emotions. MABSA consists of three subtasks: Multimodal Aspect Term Extraction (MATE), Multimodal Aspect-oriented Sentiment Classification (MASC), Joint Multimodal Aspect-Sentiment Analysis (JMASA).

The existing MABSA methods can be divided into two categories: the first category, such as JML [2], can be classified as a "multimodal joint learning framework". The second category, such as TCMT [3], can be assisted in discrimination by introducing pretrained model and external tool. But existing methods have overlooked an important fact: multimodal information has inherent problems of different information densities within modalities and information imbalance between modalities. It is worth noting that this core issue belongs to the category of model architecture design, and its solution lies in optimizing feature extraction and interaction mechanisms, which is fundamentally different from the current mainstream large-scale pretraining technology route.

To address the aforementioned issues, this paper proposes a novel solution, with the following main contributions:

1) This paper proposes the DDD model (Image Description Guided Dual Chain Enhancement Model Based on Dynamic Fusion), which is synergistically optimized via three core technologies to address intra- and inter-modal issues.
2) To tackle the problem of weak semantic correlation between text and images, the Image Description Guided module (IDG) is proposed to capture the potential semantic associations between text and images, bridging the expression gap between abstract text and images.
3) Regarding the issue of different information densities within modalities, the Dual Chained Enhancement Module (DCE) is proposed. This module innovatively adopts the IFE-TFE dual chain architecture, achieving information specific noise filtering and fine-grained feature extraction within the modality.
4) To address the issue of information imbalance between modalities, a dynamic fusion module (DFM) is introduced. It optimizes modal collaboration, and thereby solves the problems of insufficient information utilization and poor modal synergy.

II. Related Work

MABSA integrates cross modal information from text and images to achieve fine-grained sentiment analysis. Although the JMASA task integrates image-text relationship detection, it lacks modeling of fine-grained aspect-emotion associations. VLP-MABSA [4] aligns modalities through image pre-training tasks, which relies on massive annotated data. CMMT [5] introduces a unified label system to optimize multi-angle detection but still faces bottlenecks in computational efficiency. Recent studies have attempted to break through modal limitations. For instance, the aspect-oriented method (AoM) [6] detects semantic and sentiment associations in images; the aesthetics-driven Atlantis [7] model is constructed to explore image sentiment, CORSA [8] locates relevant visual regions via conditional relationship detection, AETS [9] adopts an aspect enhancement module to improve aspect term sensitivity, and AESAL [10] designs a syntax-adaptive mechanism to capture aspect associations. However, these methods share

common limitations. (1) They fail to distinguish differences in information density within modalities. (2) Inter-modal fusion relies on fixed strategies, making it difficult to dynamically adjust according to input characteristics.

In response to the above challenges, designing an image description-guided mechanism to enhance cross-modal semantic associations using generative descriptions, developing a chained enhancement module to systematically tackle the issue of intra-modal information density differences for the first time, and constructing a dynamic fusion module to realize an input-adaptive modal collaboration strategy.

III. Methods

A. Model Overview

This paper proposes model DDD, as shown in Fig. 1. In the feature extraction stage, text encoding generates n × d text features $H_T \in R^{n\times d}$ based on BART, visual encoding extracts 36 regional features through Faster R-CNN and projects them into a d × 36 visual feature $V \in R^{d\times 36}$, where n is the sequence length and d is the dimension. The IDG module balances inter-modal information weights to strengthen semantic correlation. The DCE is used to optimize intra-modal representations, and the DFM module adaptively assigns weights to multimodal features to enable adaptive fusion. Finally, the decoder outputs aspect-based sentiment results.

B. Image Description Guided Module

In cross-modal learning, text-image representation exists heterogeneity. To address this, this paper proposes an IDG module, as show in Fig.2, It uses the GPT-3.5 model to generate image descriptions, as an anchor to measure the semantic correlation between each token in the text and the corresponding image content. Based on this, the text features are weighted to enhance high correlation tokens and suppress low correlation tokens. This preserves core text semantics, strengthens cross-modal alignment, and balances cross-modal information weights.

The initial feature matrix obtained after the input text is processed by the encoder is $H_T \in R^{n\times d}$. The feature matrix obtained by encoding the image description is $E_D \in R^{n\times d}$, the process of IDG module is as follows.

Firstly, normalize the text features and image description features to obtain the $\hat{H}_T$ and $\hat{E}_D$.

Calculate the similarity between each text token feature and the text description feature, and construct a semantic association vector, as shown in Equation 1.

$$\mathbf{s} \in R^n : \ s_i = \hat{H}_{T,i} \cdot \hat{E}_D^{\top} \tag{1}$$

Where $\hat{H}_{T,i} \in R^{n\times d}$ is the i-th row of the normalized text feature matrix, $s_i \in [-1, 1]$ represents the semantic correlation between the i-th text token and the image description.

Map the semantic correlation vector S to the interval of $[0, 1]$, generate the final weight vector w_i, and based on the weight vector w_i, the original text features are weighted token by token to generate text features $\tilde{H}_{T,D}$ that fuse semantic clues in the image description.

C. Dual Chained Enhancement Module

1) TFE Module: The Text Feature Enhancement Chain (TFE) is shown in Fig. 3.

In order to capture the dependency relationships self-attention mechanism is used to preliminarily process the text features to obtain features $H_{\text{attn}} \in R^{n\times d}$.

To avoid gradient instability caused by excessive dot product values, the enhanced features $H_{\text{attn}} \in \mathbb{R}^{n\times d}$ are split into h heads, where each head generates a Query using independent heterogeneous projection matrices $W_Q^{(i)} \in \mathbb{R}^{d\times d/h}$.

Different from conventional multi-head attention, heterogeneous projection matrices $W_Q^{(i)}$ are utilized. A shared key-value space is applied to enforce global consistency of key and value information across all heads, as shown in Equation 2.

$$\overline{K} = \frac{1}{h}\sum_{j=1}^{h} H_{\text{attn}} W_K^{(j)}, \quad \overline{V} = \frac{1}{h}\sum_{j=1}^{h} H_{\text{attn}} W_V^{(j)} \tag{2}$$

where $W_K^{(j)}, W_V^{(j)} \in \mathbb{R}^{n\times d/h}$ denote the key and value projection matrices.

Attention outputs for each head are computed based on the shared keys and values, where each query is processed via scaled dot-product attention with shared keys and values to obtain $\text{CrossAttn}_i \in \mathbb{R}^{n\times d/h}$.

In the head merging stage, multi-semantic fusion and dynamic regulation are adopted for information integration. The outputs of each head is concatenated and mapped back to the original dimension, as shown in Equation 3.

$$\begin{aligned} \text{AttnConcat} &= \text{Concat}(\text{CrossAttn}_1, \ldots, \text{CrossAttn}_h) \\ H_{\text{cross}} &= \sigma(\Gamma) \odot (\text{AttnConcat}\mathbf{W}_o) \end{aligned} \tag{3}$$

where $W_O \in \mathbb{R}^{n\times d}$ is the output matrix, $\Gamma \in \mathbb{R}^{n\times d}$ denotes learnable gating parameters, σ is the sigmoid activation function, and $\odot$ denotes element-wise multiplication.

To balance attention output with original feature distribution and boost robustness, H_{cross} undergoes layer normalization and feedforward processing, then outputs a globally consistent $\tilde{H}$ via gate projection and double normalization. Adversarial perturbations (mimicking adversarial training) enforce robustness under minor noise, enhancing feature stability in chain reinforcement and yielding perturbation-enhanced H_{pert}, as shown in Equation 4.

$$H_{\text{pert}} = \tilde{H} + \alpha \cdot \Phi_{\text{adv}}(\tilde{H}) \tag{4}$$

where α is disturbance intensity coefficient, the fixed value is 0.1, $\Phi_{\text{adv}}()$ is adversarial disturbance generator, Composed of two linear layers with activation functions.

To avoid gradient instability caused by numerical discrepancies, the perturbation-enhanced feature H_{pert} and IDG-generated text feature $\tilde{H}_{T,D}$ are weighted and fused through a learnable gating vector α to produce intermediate features

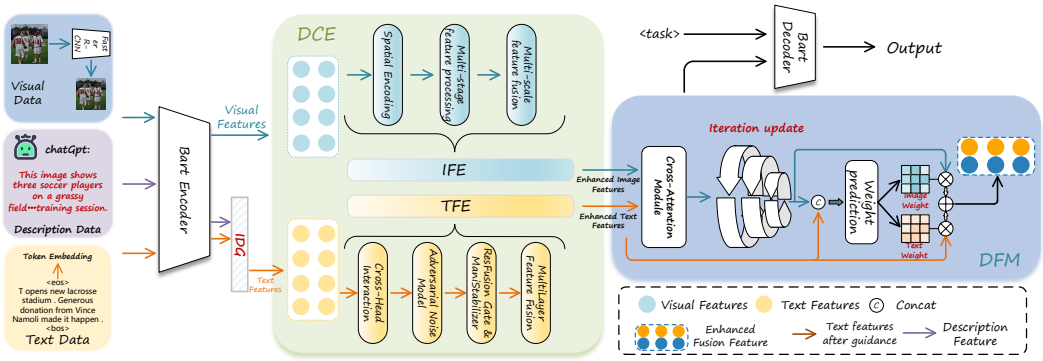

Fig. 1. The overall architecture of the proposed DDD.

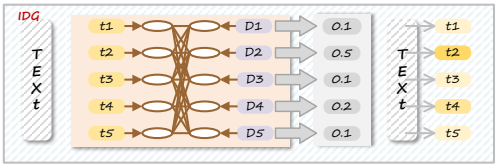

Fig. 2. IDG architecture.

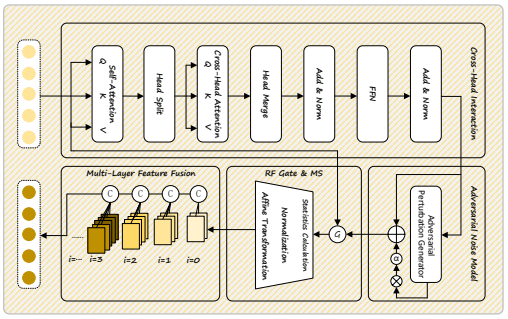

Fig. 3. TFE architecture.

H_{res}. LayerNorm is adopted to constrain the feature distribution range, as shown in Equation 5.

$$H_{\text{res}} = \text{LayerNorm}\Big(\alpha \odot H_{\text{pert}} + (1-\alpha) \odot \tilde{H}_{T,D}\Big) \tag{5}$$

where α is the learnable gating vector.

Next, global statistics (mean μ, standard deviation σ)from fused features H_{res} eliminate distribution differences between samples or levels.Finally, normalized features are dynamically adjusted through learnable scaling parameters γ and translation parameters β, where the affine transformation restores the feature expression ability that normalization may lose, as shown in Equation 6.

$$\hat{H} = \gamma \odot \left(\frac{H_{\text{res}} - \mu}{\sigma}\right) + \beta \tag{6}$$

Finally, recursively fuse intermediate features and aggregate multi-scale ones to get enhanced text features H_{fused}.

2) IFE Module: The IFE uses spatial encoding as the initial stage, as shown in Fig. 4.

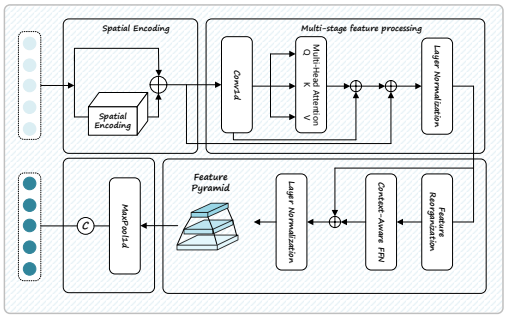

Fig. 4. IFE architecture.

Given the input image bounding box coordinates x_1, y_1, x_2, y_2 and area A, performs normalization to obtain S, as shown in Equation 7.

$$s = \left(\frac{x_1}{W}, \frac{y_1}{H}, \frac{x_2}{W}, \frac{y_2}{H}, \frac{A}{W \times H}\right) \in R^5 \tag{7}$$

where W and H are the width and height of the image, A represents the area of the object detection box.

High-dimensional spatial encoding is generated via nonlinear mapping and fused with the original visual feature $V \in \mathbb{R}^{d\times 36}$ by transposition, mapping low-dimensional geometric information to a aligned high-dimensional space to obtain $\tilde{\mathbf{V}}$. Convolution is applied to $\tilde{V}$ to capture local dependencies, and an attention mechanism is used to weight key regions; the enhanced features are fused via residual connection to form $\hat{V}$. Normalization is conducted on $\hat{V}$. Linear projection reduces multi-head attention redundancy, and a Sigmoid gating function generates weights $\sigma(W_g)$ to select critical features as shown in Equation 8.

$$G = \text{GELU}(V_{\text{reorg}} W_1) \odot \sigma(V_{\text{reorg}} W_g)$$
$$V_{\text{final}} = \text{LayerNorm}(G + V_{\text{norm1}}) \tag{8}$$

Where V_{norm1} undergoes normalization for the first time, using $V_{\text{reorg}} = V_{\text{norm1}} W_{\text{reorg}}$ for feature recombination,$W_{\text{reorg}}$is Learnable weight matrix.

Next enter the stage of constructing a multi-scale pyramid. Convolutional extraction of local patterns from V_{final} and generation of features at different levels through adaptive max pooling are performed, as shown in Equation 9.

$$V_{\text{conv\_relu}} = \text{ReLU}\big(\text{Conv1D}(V_{\text{final}}^{\top})\big)$$
$$V_{\text{pyramid}} = \text{AdaptiveMaxPool1d}\big(V_{\text{conv\_relu}}\big)^{\top} \tag{9}$$

Finally, all pyramid features are pooled to a fixed length of L=20 and concatenated along the feature dimension to form a multi-scale fused visual feature representation V_{fused}.

D. Dynamic Fusion Module

This paper proposes DFM to addresses the three core pain points of text-image modality heterogeneity, fusion dynamicity, and efficiency adaptation, as shown in Fig. 5.

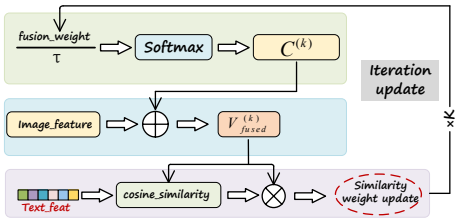

Fig. 5. Iteration update architecture.

Given enhanced text features $\mathbf{H}^{\text{fused}}$ and image features $\mathbf{V}^{\text{fused}}$, initial text-image alignment is constructed via cross-modal attention, where text features act as Query to guide selective attention on visual content, producing weighted visual features $\mathbf{A}^{\text{init}}$ and attention weight matrix $\boldsymbol{\alpha}^{\text{init}}$.

The fusion weights $\mathbf{W}^{(0)} = \boldsymbol{\alpha}^{\text{init}}$ are initialized with K iterative updates. The temperature-scaled softmax function converts weights into a probability distribution, as shown in Equation 10.

$$\mathbf{C}^{(k)} = \text{softmax}\left(\frac{\mathbf{W}^{(k-1)}}{\tau}\right) \in R^{B \times L_t \times L_v} \tag{10}$$

Based on the correlation matrix $\mathbf{C}^{(k)}$, image features are weighted and aggregated to obtain $\mathbf{V}^{(k)}_{\text{fused}}$.

The semantic correlation between fused text and visual features is computed to measure alignment, as shown in Equation 11.

$$\boldsymbol{s}_i^{(k)} = \frac{\mathbf{h}_i^{\text{fused}} \cdot \mathbf{v}_i^{(k)}}{\|\mathbf{h}_i^{\text{fused}}\|_2 \cdot \|\mathbf{v}_i^{(k)}\|_2} \tag{11}$$

where $\mathbf{h}_i^{\text{fused}}$ and $\mathbf{v}_i^{(k)}$ denote the i-th text token and corresponding visual feature.

The high-similarity correspondences are then updated and strengthened, as shown in Equation 12.

$$\mathbf{W}_{ij}^{(k)} = \mathbf{W}_{ij}^{(k-1)} \cdot \boldsymbol{s}_i^{(k)} \tag{12}$$

After K iterations, the refined visual representation $\mathbf{V}^{(K)}_{\text{fused}}$ is obtained. To overcome fixed fusion limitations, a lightweight network dynamically predicts the text-image fusion ratio for each token, as shown in Equation 13.

$$\mathbf{w}_{\text{text}}, \mathbf{w}_{\text{vis}} = \text{softmax}\left(\text{MLP}\left(\left[\mathbf{H}^{\text{fused}} \| \mathbf{V}^{(K)}_{\text{fused}}\right]\right)\right) \tag{13}$$

where $[\cdot\|\cdot]$ denotes feature concatenation, and MLP contains two fully connected layers (intermediate dimension $d/4$, NeLU activation). The weights $\mathbf{w}_{\text{text}}, \mathbf{w}_{\text{vis}} \in R^{B \times L_t \times 1}$ satisfy $\mathbf{w}_{\text{text}} + \mathbf{w}_{\text{vis}} = \mathbf{1}$. The final fused feature is given by Equation 14.

$$\mathbf{F}^{\text{out}} = \text{LayerNorm}(\mathbf{w}_{\text{text}} \odot \mathbf{H}^{\text{fused}} + \mathbf{w}_{\text{vis}} \odot \mathbf{V}^{(K)}_{\text{fused}}) \tag{14}$$

IV. Experiment

A. Experimental Settings

Using the Twitter-2015 and Twitter-2017 datasets [11] as benchmarks. Image descriptions are generated by GPT-3.5. Pretraining is conducted for 40 epochs with a batch size of 64 and learning rate of 5e-5, while downstream tasks are fine-tuned for 35 epochs with a batch size of 16 and learning rate of 5e-5, the multi-task balancing coefficients $\lambda_1 - \lambda_5$ are set to 1. For evaluation, MASC is assessed with accuracy (Acc) and F1-score, whereas JMASA and MATE use Micro-F1, precision (P), and recall (R).

B. Main Results

The experimental results of the proposed DDD model and current state-of-the-art baseline models in tasks such as JMASA, MASC, and MATE are shown in Table I, II, and III.

TABLE I
COMPARISON OF DIFFERENT MODELS IN JMASA.

Methods	Twitter2015			Twitter2017		
	P	R	F1	P	R	F1
Multimodal						
OSCGA	63.1	63.7	63.2	63.5	63.5	63.5
JML	65.0	63.2	64.1	66.5	65.5	66.0
VLP-MABSA	65.1	68.3	66.6	66.9	69.2	68.0
M2DF [12]	67.0	68.3	67.6	67.9	68.8	68.3
Atlantis	65.6	69.2	67.3	68.6	70.3	69.4
MCPL-VLP [13]	67.2	69.2	68.2	69.0	69.4	69.2
RNG [14]	67.8	69.5	68.6	69.5	71.0	70.2
TCMT	69.3	70.4	69.8	70.2	71.5	70.8
CORSA	69.0	70.8	69.9	70.1	71.0	70.6
DDD (Ours)	**70.92**	**71.54**	**71.33**	**71.41**	**72.06**	**71.73**
LLMs						
ChatGPT-3.5	54.7	52.9	55.1	58.9	57.7	58.9
Llama 2 [15]	51.3	50.9	51.7	55.9	55.7	56.2
DDD (Ours)	**70.92**	**71.54**	**71.33**	**71.41**	**72.06**	**71.73**

TABLE II
COMPARISON OF DIFFERENT MODELS IN MATE.

Methods	Twitter2015			Twitter2017		
	P	R	F1	P	R	F1
RAN	80.5	81.5	81.0	90.7	90.7	90.0
OSCGA	81.7	82.1	81.9	90.2	90.7	90.4
JML	83.6	81.2	82.4	92.0	90.7	91.4
VLP-MABSA	83.6	87.9	85.7	90.8	92.6	91.7
M2DF	85.2	87.4	86.3	91.5	93.2	92.4
MCPL-VLP	84.8	87.4	86.1	91.9	92.4	92.2
CORSA	85.1	87.6	86.3	92.6	93.0	92.8
DDD (Ours)	**85.64**	**88.42**	**87.05**	**92.64**	**93.92**	**93.17**

TABLE III
COMPARISON OF DIFFERENT MODELS IN MASC.

Methods	Twitter2015		Twitter2017	
	ACC	F1	ACC	F1
JML	78.7	74.1	72.7	70.5
VLP-MABSA	78.6	73.8	73.8	71.8
M2DF	78.9	74.8	74.3	73.0
MGFN-SD [16]	79.36	74.81	72.77	72.07
MCPL-VLP	79.3	74.9	75.1	74.0
A2II [17]	79.46	75.16	74.39	72.35
DPFN [18]	79.9	76.0	73.9	72.6
CORSA	81.1	77.7	76.6	74.5
TCMT	81.4	76.7	77.3	75.8
DDD (Ours)	**81.47**	**77.91**	**77.57**	**75.90**

From Tables I, II, and III, DDD consistently outperforms most baselines, demonstrating its effectiveness. Conventional methods (e.g., VLP, JML) lack efficient modality-specific feature utilization, and even recent approaches (e.g., TCMT, MCPL, CORSA) remain limited by insufficient cross-modal collaboration and/or fine-grained intra-modal modeling. In contrast, DDD achieves a better balance between recall and misclassification, yielding notable gains (especially in F1 on MASC), which indicates its advantage in fine-grained classification and class-imbalance scenarios. Moreover, compared with ChatGPT-3.5 and Llama 2, DDD identifies both aspects and sentiments more accurately, further highlighting the benefit of its dynamic fusion mechanism.

C. Parameter Analysis

To investigate the impact of key parameters on model performance in the dynamic fusion module, analyzing two core parameters: the temperature coefficient and the number of iterations. The results are shown in Fig. 6.

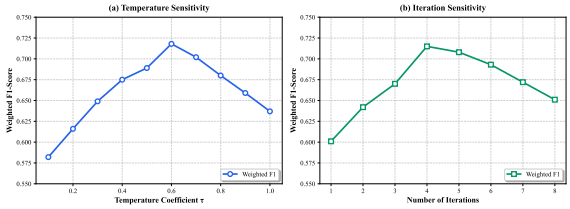

Fig. 6. Parameter sensitivity analysis.

Fig. 6(a) shows that the weighted score rises significantly as the temperature coefficient increases from 0.1 to 0.6 (F1 peaks at $\tau = 0.6$), since moderate values enhance feature allocation discriminability for efficient cross-modal fusion, while $\tau > 0.6$ causes over-sharpening and information loss. Fig. 6(b) illustrates the weighted score increases with iterations (peaking at 4), as multiple iterations optimize inter-modal semantic fusion; beyond 4, redundant computations and noise overfitting reduce stability.

D. Ablation Study

To verify the effectiveness of each module in the DDD model, this paper focuses on the IDG, TFE, IFE, DCE and DFM in JMASA task (Table IV), MATE task (Table V), and MASC task (Table VI). By removing individual modules one by one, compare the performance of the model on metrics such as Precision (P), Recall (R), and F1.

TABLE IV
ABLATION STUDY ON JMASA TASK.

Methods	Twitter2015			Twitter2017		
	P	R	F1	P	R	F1
DDD (Full)	**70.92**	**71.54**	**71.33**	**71.41**	**72.06**	**71.73**
w/o IDG	69.83	70.32	70.07	69.79	71.89	70.82
w/o IFE	68.02	69.38	69.16	69.25	71.38	70.29
w/o TFE	68.85	69.92	69.11	69.82	70.97	70.30
w/o DCE	67.95	69.15	68.54	68.01	70.18	69.18
w/o DFM	67.25	68.65	67.91	67.31	68.18	67.74
w/o All Components	65.1	68.3	66.6	66.9	69.2	68.0

TABLE V
ABLATION STUDY ON MATE TASK.

Methods	Twitter2015			Twitter2017		
	P	R	F1	P	R	F1
DDD (Full)	**85.64**	**88.42**	**87.05**	**92.64**	**93.92**	**93.17**
w/o IDG	85.28	88.26	86.74	92.13	93.69	92.90
w/o IFE	84.83	88.30	86.53	91.78	93.40	92.58
w/o TFE	84.17	88.15	86.11	91.52	93.11	92.31
w/o DCE	84.02	88.01	85.96	91.26	92.98	92.11
w/o DFM	83.97	87.97	85.92	91.03	92.76	91.89
w/o All Components	83.6	87.9	85.7	90.8	92.6	91.7

TABLE VI
ABLATION STUDY ON MASC TASK.

Methods	Twitter2015		Twitter2017	
	ACC	F1	ACC	F1
DDD (Full)	**81.47**	**77.91**	**77.57**	**75.90**
w/o IDG	80.88	77.65	76.18	74.89
w/o IFE	80.71	76.58	75.21	73.37
w/o TFE	80.52	76.05	75.07	72.90
w/o DCE	79.58	75.37	74.52	72.62
w/o DFM	78.83	74.64	74.27	72.09
w/o All Components	78.6	73.8	73.8	71.8

Removing the IDG module leads to obvious performance drops, verifying its role as a semantic anchor for cross-modal alignment and noise filtering. Dropping IFE or TFE

separately degrades performance across tasks, indicating their ability to suppress modality-specific noise and refine feature representations. Removing the full DCE module causes greater performance degradation than removing IFE/TFE individually, revealing the strong synergy between the two enhancement chains. Removing DFM results in the most significant performance decline, as it deprives the model of dynamic cross-modal weight assignment for reliable decision-making. When all components are removed, performance drops to the lowest level, further confirming that the modules are not merely additive but complementary.

E. Case Study

To validate the performance of the DDD model, this paper conducts a case study experiment, selecting VLP, and M2DF as baseline models. The experimental results are shown in Fig. 7.

Fig. 7. Case study results. POS: positive, NEU: neutral, NEG: negative.

Case study results demonstrate that the DDD model accurately extracts aspect terms and identifies corresponding sentiment polarities from tweets. In Fig. 7(a), VLP and M2DF correctly extract aspect words but fail to predict accurate sentiment polarities and misextract aspects under multi-aspect scenarios. As shown in Fig. 7(c), VLP misclassifies the sentiment of "NW Rankin" as NEU, and M2DF produces similar errors, while DDD achieves correct aspect and sentiment prediction in all cases, showing stronger robustness.

V. Conclusion

This paper proposes the DDD model to address the issues of information density differences within modalities and insufficient semantic collaboration between modalities in fine-grained multimodal aspect level sentiment analysis tasks. This model compensates for the semantic gap between text and images through the IDG Module, utilizes the DCE Module to achieve inter-modality specific noise filtering and fine-grained feature extraction, and introduce DFM Module to optimize inter modal collaboration. The experimental results on two publicly available benchmark datasets show that DDD outperforms existing mainstream methods in terms of performance, validating its design effectiveness.

References

[1] N. Xu, W. Mao, and G. Chen, "Multi-interactive memory network for aspect based multimodal sentiment analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 371–378.

[2] X. Ju, D. Zhang, R. Xiao, J. Li, S. Li, M. Zhang, and G. Zhou, "Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 4395–4405.

[3] W. Zou, X. Sun, W. Wu, Q. Lu, X. Zhao, Q. Bo, and J. Yan, "TCMT: Target-oriented cross modal transformer for multimodal aspect-based sentiment analysis," *Expert Systems with Applications*, vol. 264, p. 125818, 2025.

[4] Y. Ling, J. Yu, and R. Xia, "Vision-language pre-training for multimodal aspect-based sentiment analysis," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 2149–2159.

[5] L. Yang, J.-C. Na, and J. Yu, "Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis," *Information Processing & Management*, vol. 59, no. 5, p. 103038, 2022.

[6] R. Zhou, W. Guo, X. Liu, S. Yu, Y. Zhang, and X. Yuan, "AoM: Detecting aspect-oriented information for multimodal aspect-based sentiment analysis," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 8184–8196.

[7] L. Xiao, X. Wu, J. Xu, W. Li, C. Jin, and L. He, "Atlantis: Aesthetic-oriented multiple granularities fusion network for joint multimodal aspect-based sentiment analysis," *Information Fusion*, vol. 106, p. 102304, 2024.

[8] X. Liu, R. Li, S. Ye, G. Zhang, and X. Wang, "Multimodal aspect-based sentiment analysis under conditional relation," in *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 313–323. [Online]. Available: https://aclanthology.org/2025.coling-main.22.

[9] L. Zhu, H. Sun, Q. Gao, Y. Liu, and L. He, "Aspect enhancement and text simplification in multimodal aspect-based sentiment analysis for multi-aspect and multi-sentiment scenarios," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 2, 2025, pp. 1683–1691.

[10] L. Zhu, H. Sun, Q. Gao, T. Yi, and L. He, "Joint multimodal aspect sentiment analysis with aspect enhancement and syntactic adaptive learning," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, ser. IJCAI Proceedings. ijcai.org, 2024, pp. 6678–6686.

[11] J. Yu and J. Jiang, "Adapting BERT for Target-Oriented Multimodal Sentiment Classification," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019, pp. 5408–5414.

[12] F. Zhao, C. Li, Z. Wu, Y. Ouyang, J. Zhang, and X. Dai, "M2DF: Multi-grained multi-curriculum denoising framework for multimodal aspect-based sentiment analysis," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 9057–9070.

[13] J. Zhang, J. Qu, J. Liu, and Z. Wang, "MCPL: Multi-model co-guided progressive learning for multimodal aspect-based sentiment analysis," *Knowledge-Based Systems*, vol. 301, p. 112331, 2024.

[14] Y. Liu, Y. Zhou, Z. Li, J. Zhang, Y. Shang, C. Zhang, and S. Hu, "RNG: Reducing multi-level noise and multi-grained semantic gap for joint multimodal aspect-sentiment analysis," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 2024, pp. 1–6.

[15] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," 2023. [Online]. Available: https://arxiv.org/pdf/2307.09288.pdf

[16] J. Yang, Y. Xiao, and X. Du, "Multi-grained fusion network with self-distillation for aspect-based multimodal sentiment analysis," *Knowledge-Based Systems*, vol. 293, p. 111724, 2024.

[17] J. Feng, M. Lin, L. Shang, and X. Gao, "Autonomous aspect-image instruction a2II: Q-former guided multimodal sentiment classification," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, ser. IJCAI Proceedings, 2024, pp. 1996–2005.

[18] D. Wang, C. Tian, X. Liang, L. Zhao, L. He, and Q. Wang, "Dual-perspective fusion network for aspect-based multimodal sentiment analysis," *IEEE Transactions on Multimedia*, vol. 26, pp. 4028–4038, 2024.