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

Efficient Multimodal Sentiment Recognition with Dual Cross-Attention for Multi-Scale Features

Xinyu Ye[1], Tingsong Ma[2], Yi Feng[3], Yiming Zhai *

Xihua University, Chengdu, Sichuan 610039, China

yxy@stu.xhu.edu.xn

Abstract. Multimodal sentiment analysis has emerged as a key research field, particularly for decoding emotions conveyed through text and images on social media platforms. However, many approaches encounter difficulties when integrating textual and visual features across diverse dimensions, often leading to suboptimal performance. To address this, we propose a novel approach for multimodal sentiment recognition that designs a simple yet efficient network, inspired by feature pyramids. In this model, feature vectors are split into high-dimensional and low-dimensional representations, which are then processed through distinct cross-attention mechanisms tailored to their scales, followed by a fusion step to capture comprehensive cross-modal interactions. This strategy enhances the network's ability to model relationships between modalities effectively. We evaluated our approach on the well-established MVSA-Single and MVSA-Multiple datasets, where it consistently surpasses existing techniques. Specifically, it achieves an accuracy of 78.27% and an F1 score of 77.95% on MVSA-Single, and an accuracy of 71.18% and an F1 score of 68.92% on MVSA-Multiple. These results demonstrate the potential of combining high- and low-dimensional features with dual cross-attention for social media sentiment analysis.

Keywords: Multimodal Sentiment Analysis, Cross-Attention Mechanism, Multi-Scale Features.

1 Introduction

Multimodal sentiment analysis[1][2][3] plays a crucial role in understanding human emotions expressed through text and images, particularly in the context of social media, where it supports applications like public opinion monitoring and mental health assessment. In real-world scenarios, challenges such as inconsistent feature representations, varying emotional intensities across modalities, and complex interactions between textual and visual cues often result in inaccurate sentiment predictions or overlooked emotional nuances. Conventional methods typically rely on unimodal feature extraction or simplistic fusion techniques, which struggle to scale effectively, fail to capture intricate cross-modal dependencies, and fall short of delivering robust performance in diverse

and dynamic datasets, thus limiting their applicability in modern sentiment analysis tasks.

Traditional machine learning techniques have long been applied to multimodal sentiment analysis tasks. These approaches typically emphasize independent feature extraction, such as statistical measures for text and local descriptor-based representations for images, aiming to identify sentiment indicators within each modality. While they provide some ability to handle basic noise and modality differences, their dependence on separately crafted features hinders effective integration of textual and visual information across varying dimensions, often leading to incomplete capture of emotional nuances and diminished performance on complex social media data. This highlights the pressing need for more advanced strategies that can overcome these integration challenges and enhance sentiment recognition[4][5][6] in real-world multimodal settings.

The rapid progress in deep learning, spanning natural language processing and computer vision, has transformed multimodal sentiment analysis, enabling a range of algorithms to decode emotions from text and images, particularly in the expansive domain of social media. Joint-fusion approaches, such as CLIP[7][8], VisualBERT[9][10], and LXMERT[11], seek to integrate modalities directly within a single architecture, leveraging large-scale pretraining to align text and visual cues. Despite their promise, these methods often stumble in real-world scenarios, where varying emotional intensities, informal language, and noisy or ambiguous images degrade their ability to consistently capture nuanced cross-modal interactions, limiting their effectiveness for fine-grained sentiment tasks. On the other hand, unimodal-pretrained approaches, including BERT[12][13], RoBERTa[14], and XLNet[15] for text, alongside ResNet[16][17], VGG[18][19], and Inception[20] for images, first extract features independently using well-established pretrained models before attempting fusion. While these techniques excel at producing strong modality-specific representations, they frequently encounter inefficiencies in the subsequent combination phase, relying on intricate fusion layers or attention mechanisms that struggle to align high- and low-dimensional features, resulting in scalability issues and suboptimal performance on diverse datasets. Currently, multimodal sentiment recognition remains challenged by the complexity of integrating heterogeneous modalities, with existing models often failing to balance accuracy and adaptability, underscoring the need for more effective solutions in practical applications.

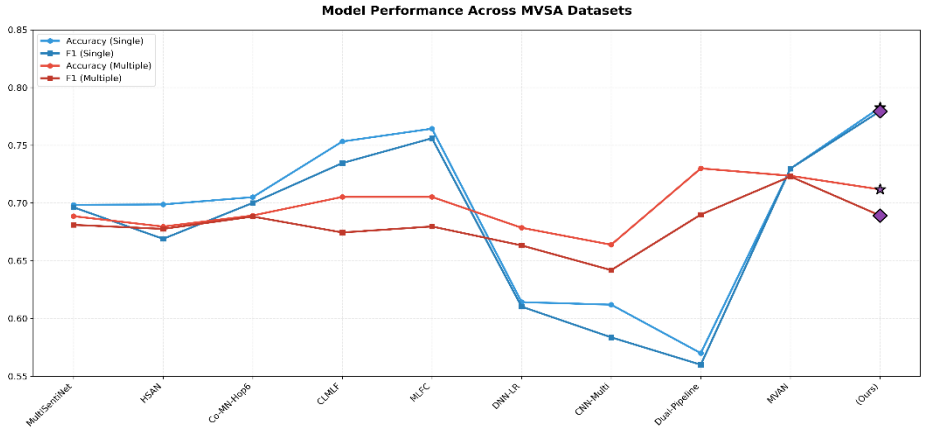

Fig. 1. This figure compares the performance of state-of-the-art methods on the MVSA-Single and MVSA-Multiple datasets, showing Accuracy and F1 scores for both datasets. Our approach (Ours)，achieves the highest scores of 78.27% Accuracy and 77.95% F1 on MVSA-Single, and 71.18% Accuracy and 68.92% F1 on MVSA-Multiple, demonstrating its superior performance in multimodal sentiment recognition.

To tackle these challenges, we propose a novel and efficient deep learning-based approach for multimodal sentiment recognition, designed to effectively integrate features across modalities by leveraging dual cross-attention mechanisms. Inspired by feature pyramids, our method first splits feature vectors into high-dimensional and low-dimensional representations, which are then processed through two tailored cross-attention mechanisms: Distance-Modulated Exponential Cross Attention (DECA) for low-dimensional features, which enhances interaction through exponential distance modulation, and Interaction-Gated Differential Cross Attention (IDCA) for high-dimensional features, which prioritizes salient cross-modal interactions via differential gating. These features are subsequently fused to capture comprehensive emotional cues. The effectiveness of our method is demonstrated in Fig. 1, which compares its performance against other leading methods on two benchmark datasets: MVSA-Single and MVSA-Multiple. The figure shows our approach achieving promising results with 78.27% Accuracy and 77.95% F1 on MVSA-Single, and 71.18% Accuracy and 68.92% F1 on MVSA-Multiple, highlighting its potential for advancing multimodal sentiment analysis.

2 Our Method

In our framework for multimodal sentiment recognition, we propose a novel architecture designed to facilitate effective cross-modal feature integration, as illustrated in Fig. 2. The framework consists of three main stages: an input stage, a feature extraction and interaction stage, and a prediction stage. Initially, paired text and image inputs are processed to extract visual and textual embeddings using CLIP, a pretrained vision-language model. These embeddings are then transformed into coarse-grained and fine-

grained feature spaces to capture different levels of abstraction. In the feature extraction and interaction stage, the features are processed by two cross-attention mechanisms, DECA and IDCA, to model cross-modal relationships, followed by a fusion step to combine the processed features. The prediction stage then uses these fused features for sentiment classification, enabling accurate emotion recognition across diverse social media datasets through a simple yet highly efficient design.

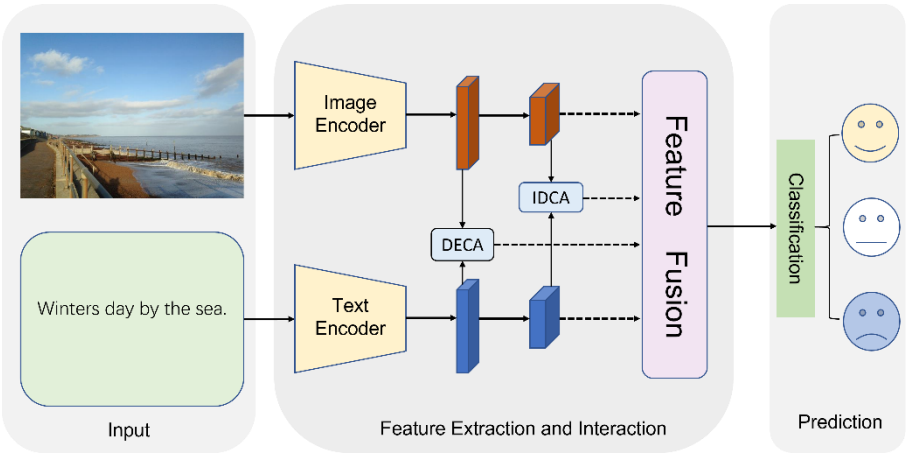

Fig.2. This figure depicts the overall architecture of our framework for multimodal sentiment recognition, featuring an input stage for paired text and image data, a feature extraction and interaction stage with dual cross-attention mechanisms (DECA and IDCA), and a prediction stage for sentiment classification, optimized for efficient and accurate performance on diverse social media datasets.

Our framework processes the multimodal inputs through a series of carefully designed steps to achieve effective sentiment recognition. Starting with the visual and textual embeddings extracted by CLIP, the features are first projected into two distinct spaces to represent different levels of abstraction, followed by normalization, non-linear activation, and regularization to ensure robust representations. The features capturing broader patterns are then passed through DECA, which applies a distance-modulated attention mechanism to compute interactions between text and image modalities, while the features focusing on detailed interactions are processed by IDCA, which leverages a differential gating mechanism to prioritize emotionally significant relationships. The outputs from DECA and IDCA are weighted and combined to balance their contributions, followed by a fusion layer that integrates all features into a unified representation. This fused representation is further refined before being fed into a classifier, which outputs the final sentiment predictions. By streamlining the cross-modal interaction process, our architecture achieves a balance between computational efficiency and predictive accuracy, making it well-suited for practical multimodal sentiment analysis tasks.

2.1 Distance-Modulated Exponential Cross Attention (DECA)

In multimodal sentiment recognition, fusing low-dimensional features poses a challenge in capturing global cross-modal interactions, as these coarse-grained features often lack fine-grained alignment between text and image modalities, leading to suboptimal interaction modeling. To address this, we propose Distance-Modulated Exponential Cross Attention (DECA), a novel attention mechanism that enhances the fusion of low-dimensional features by dynamically scaling attention scores based on feature distances. As depicted in Fig. 3, DECA takes the image low features as query (Q) and value (V) alongside the text low features as key (K) for one direction of cross-attention, and symmetrically, the text low features as Q and V with the image low features as K for the other direction, ensuring comprehensive interaction modeling through a bidirectional attention process, followed by normalization to stabilize the output.

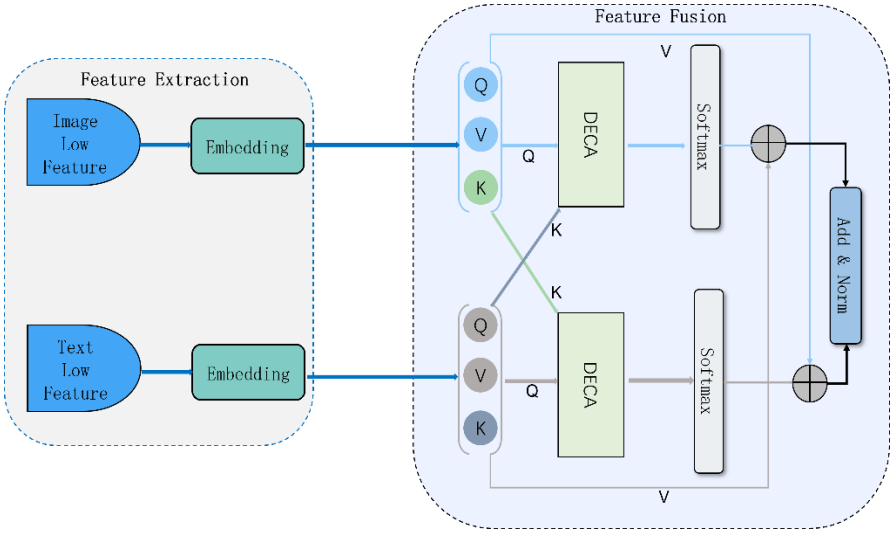

Fig. 3 This figure illustrates the architecture of Distance-Modulated Exponential Cross Attention (DECA), where image and text low features (Q, K, V) undergo distance and scale computation, followed by softmax normalization and an add & norm step to model cross-modal interactions effectively.

$$Q = W_q Q, \quad K = W_k K, \quad V = W_v V \tag{1}$$

Firstly, in Eq.(1) , Q, $K$and V represent the query, key, and value features, respectively, where Q and V are typically the image low features, and K is the text low feature (or vice versa in the bidirectional process). The matrices W_q, W_k, and W_v are learned weights that project the features into a common space, ensuring alignment for attention computation.

$$\text{distance} = \left| Q_{\text{expanded}} - K_{\text{expanded}} \right| \tag{2}$$

Next, in Eq.(2), Q_{expanded} and K_{expanded} are the expanded forms of the query and key features, adjusted to enable pairwise comparison across the batch. The distance term captures the absolute difference between these expanded features, quantifying the dissimilarity between query and key along each dimension for attention weighting.

$$\text{scale\_factor} = \sigma\left(Q_{\text{expanded}} + K_{\text{expanded}}\right) \tag{3}$$

After that, in Eq. (3), scale_factor has the same shape as the expanded features. The Q_{expanded} and K_{expanded} compute the element-wise sum of the expanded query and key features, and σ denotes the sigmoid function, mapping the sum into the range (0,1). This scale factor modulates the influence of the distance, emphasizing contextually relevant features.

$$\text{attn\_scores} = \exp(-\sum(\text{distance} \cdot \text{scale\_factor})) \tag{4}$$

Subsequently, Eq.(4) combines the feature distance and the scale_factor by multiplying them element-wise, then summing the result across the feature dimension to produce a matrix of raw attention scores. By applying an exponential function with a negative sign, the computation ensures that smaller distances, indicating more similar features, result in higher attention scores, while the scale factor fine-tunes the sensitivity of this distance-based weighting to enhance relevant cross-modal interactions.

$$\text{attn} = \text{softmax}(\text{attn\_scores}) \quad \text{output} = \text{attn} \cdot V \tag{5}$$

Finally, as shown in Eq.(5), attn_scores contain the raw attention scores, and the softmax function is applied across the keys for each query, producing attn, where each row sums to 1, ensuring the attention weights are properly distributed. The attn weights are then multiplied with V, the value feature, to compute the output, which captures the enhanced cross-modal representation, where each query feature is informed by the most relevant key features. This process effectively aligns the query and key features, enabling the model to focus on the most pertinent information. Ultimately, the output provides a refined representation that enhances the model's ability to capture cross-modal relationships.

2.2 Interaction-Gated Differential Cross Attention (IDCA)

During the integration of high-dimensional features in multimodal sentiment analysis, a major difficulty arises in pinpointing emotionally significant cross-modal relationships, as these fine-grained features often involve intricate interactions that are challenging to prioritize effectively. To overcome this, we introduce Interaction-Gated Differential Cross Attention (IDCA), a novel attention mechanism tailored for high-dimensional features to selectively focus on the most relevant cross-modal connections. As shown in Fig. 4, IDCA takes the image high features as query (Q) and value (V) alongside the text high features as key (K) for one direction of cross-attention, and symmetrically, the text high features as Q and V with the image high features as K for the other

direction, employing an interaction and difference computation followed by softmax and normalization to enhance the precision of fine-grained feature fusion for improved sentiment recognition.

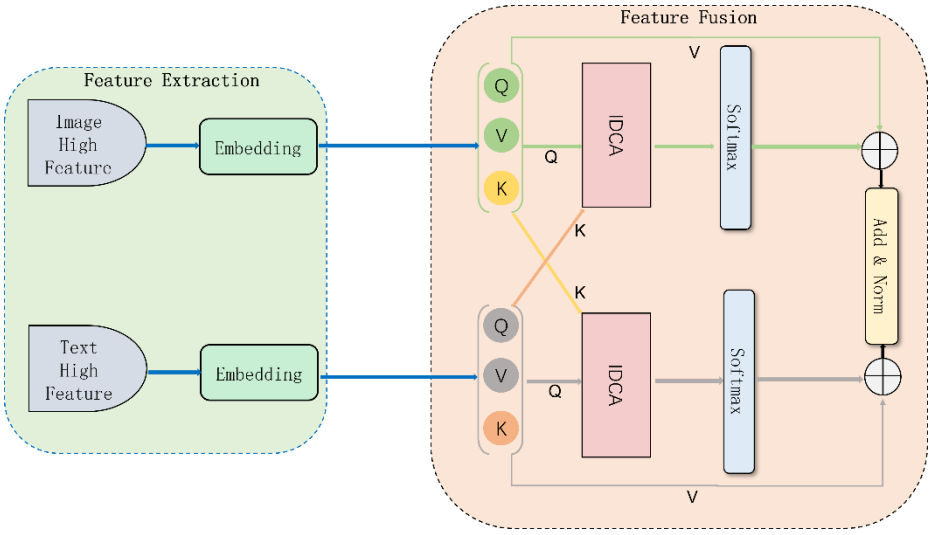

Fig. 4 This figure illustrates the architecture of Interaction-Gated Differential Cross Attention (IDCA), where image and text high features (Q, K, V) undergo interaction and difference computation, followed by softmax normalization and an add & norm step to model salient cross-modal interactions effectively.

$$\text{interaction} = Q_{\text{expanded}} \cdot K_{\text{expanded}} \tag{6}$$

Initially, as presented in Eq.(1), likewise， we first obtain the query (Q), key (K), and value (V) features through linear transformations. Secondly, in Eq.(6), Q_{expanded} and K_{expanded} are the expanded forms of the query and key features, adjusted to enable pairwise comparison across the batch. The interaction captures the element-wise multiplication of these expanded features, quantifying the interaction strength between query and key along each dimension to emphasize cross-modal relationships.

$$\text{difference} = \text{ReLU}\left(Q_{\text{expanded}} - K_{\text{expanded}}\right) \tag{7}$$

In Eq.(7), difference has the same shape as the expanded features. The Q_{expanded} and K_{expanded} compute the element-wise subtraction of the expanded query and key features, and the ReLU function ensures non-negative differences, highlighting the distinct aspects between query and key for attention weighting.

$$\text{attn\_scores} = \sum(\text{interaction} \cdot \text{difference}) \tag{8}$$

Eq.(8) combines the interaction strength and the difference by multiplying them element-wise, then summing the result across the feature dimension to produce a matrix of raw attention scores. This computation ensures that attention scores are higher for

features with both strong interactions and significant differences, effectively prioritizing emotionally salient cross-modal relationships.

$$\text{attn} = \text{softmax}(\text{attn\_scores}) \quad \text{output} = \text{attn} \cdot W \tag{9}$$

As shown in Eq.(9), attn_scores contain the raw attention scores, and the softmax function is applied across the keys for each query, producing attn, where each row sums to 1, ensuring the attention weights are properly distributed. The attn weights are then multiplied with W, the value feature, to compute the output, which captures the enhanced cross-modal representation, where each query feature is informed by the most relevant key features. This mechanism allows IDCA to effectively highlight emotionally significant interactions, significantly improving the accuracy of sentiment recognition in multimodal scenarios.

3 Experiments

The experiments are conducted on the MVSA datasets, specifically MVSA-Single and MVSA-Multiple, which are standard benchmarks for multi-modal sentiment analysis.As illustrated in Table 1, MVSA-Single comprises 4511 samples, while MVSA-Multiple includes 17024 samples, each with three label categories: positive, neutral, and negative. For both datasets, the samples from each label category are split into training, validation, and test sets in an 8:1:1 ratio, ensuring that all sets maintain a proportional distribution of labels and mitigating class imbalance. The implementation is developed using Python with PyTorch 2.3.1 and CUDA 12.1 to enable efficient deep learning computations. All experiments are performed on a single NVIDIA 4080 GPU, providing sufficient computational resources for training and evaluation. For the key hyper-parameter settings of our model which is presented in Table 2, we use a batch size of 32, an initial learning rate of 6e-5 for MVSA-Single and 8e-5 for MVSA-Multiple, and achieve the best results after training for 42 epochs on MVSA-Single and 20 epochs on MVSA-Multiple, leveraging the pretrained ViT-B/32 CLIP model for feature extraction.

Table 1. Statistics of the MVSA datasets

Dataset	Label	Train	Validation	Test	Total
MVSA-Single	Positive	2147	268	268	2683
	Neutral	376	47	47	470
	Negative	1088	135	135	1358
	All	3611	450	450	4511
MVSA-Multiple	Positive	9056	1131	1131	11318
	Neutral	3528	440	440	4408
	Negative	1040	129	129	1298
	All	13624	1700	1700	17024

Table 2. Hyper-parameter configuration for our model.

Hyper-parameter	MVSA-Single	MVSA-Multiple
Image size	224	224
Text length	77	77
Batch size	32	32
Optimizer	Adam	Adam
Initial learning rate	6e−5	8e−5
Learning rate decay	0.4	0.4
Epoch	42	20
Pretrained CLIP model	ViT-B/32	ViT-B/32

3.1 Compared with Sota

Table 3. Performance Comparisons of Different methos on MVSA-Single

Model	Accuracy	F1
MultiSentiNet	0.6984	0.6963
HSAN	0.6988	0.6690
Co-MN-Hop6	0.7051	0.7001
CLMLF	0.7533	0.7346
MLFC	0.7644	0.7561
DNN-LR	0.6142	0.6103
CNN-Multi	0.6120	0.5837
Dual-Pipeline	0.5700	0.5600
MVAN	0.7298	0.7298
(Ours)	**0.7827**	**0.7795**

Table 4. Performance Comparisons of Different methos on MVSA-Multiple

Model	Acc(%)	F1(%)
MultiSentiNet	0.6886	0.6811
HSAN	0.6796	0.6776
Co-MN-Hop6	0.6892	0.6883
CLMLF	0.7053	0.6745
MLFC	0.7053	0.6797
DNN-LR	0.6786	0.6633
CNN-Multi	0.6639	0.6419
Dual-Pipeline	0.7300	0.6900
MVAN	0.7236	0.7230
(Ours)	0.7118	0.6892

The performance of our model on the MVSA datasets is presented in Table 3 and Table 4. On MVSA-Single, our method achieves an accuracy of 78.27% and an F1-score of 77.95%, while on MVSA-Multiple, it records an accuracy of 71.18% and an F1-score of 68.92%. These results demonstrate the exceptional effectiveness of our approach, showcasing its robust capability to handle multi-modal sentiment analysis tasks with high accuracy and balanced performance across both datasets. The superior performance of our model can be attributed to its innovative design, which integrates

multi-scale feature representations to capture a wide range of cross-modal interactions. By processing both coarse- and fine-grained features, our method ensures a comprehensive understanding of the relationships between visual and textual modalities, leading to enhanced sentiment prediction accuracy. Additionally, the model's ability to adaptively focus on relevant cross-modal cues further contributes to its strong performance, making it highly effective for multi-modal learning tasks.

3.2 Ablation Study

Table 5. Module Comparisons of MVSA

model	MVSA-Single		MVSA-Multiple	
	Accuracy	F1	Accuracy	F1
None	0.6874	0.6727	0.6343	0.6312
DECA	0.7340	0.7261	0.6652	0.6611
IDCA	0.7528	0.7327	0.6842	0.6792
Ours	**0.7827**	**0.7795**	**0.7118**	**0.6892**

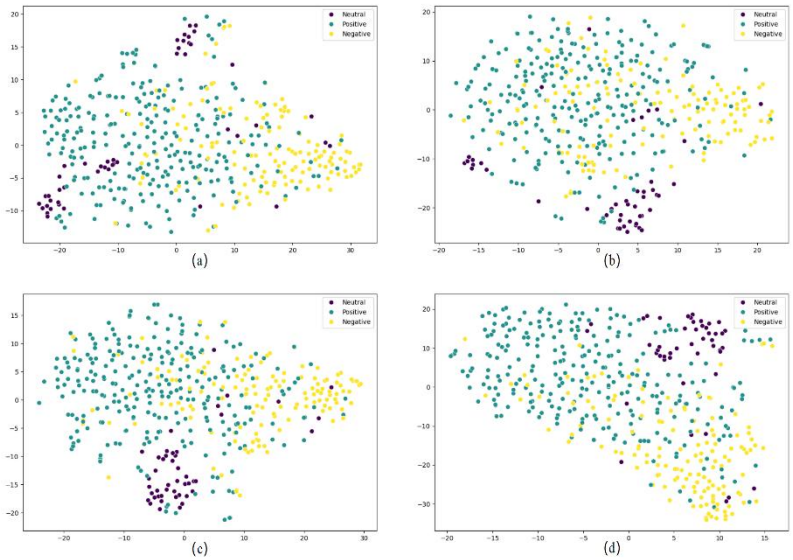

Fig. 5 Clustering visualizations using t-SNE for MVSA-Single and MVSA-Multiple datasets, where more concentrated colors indicate better clustering of positive (cyan), neutral (purple), and negative (yellow) classes. (a) Clustering on MVSA-Single without any mechanism, showing dispersed distributions with significant overlap. (b) Clustering on MVSA-Single using DECA, displaying moderate concentration but still with noticeable overlap. (c) Clustering on MVSA-Multiple using IDCA, exhibiting improved but limited class concentration. (d) Clustering on MVSA-Multiple with our dual cross-attention mechanism, demonstrating highly concentrated and well-separated class clusters.

Table 5 presents the performance of our with different modules on MVSA-Single and MVSA-Multiple, where our dual cross-attention mechanism achieves an accuracy of 78.27% and an F1-score of 77.95% on MVSA-Single, and 71.18% accuracy with a 68.92% F1-score on MVSA-Multiple, outperforming the variants without attention (68.74% accuracy on MVSA-Single, 63.43% on MVSA-Multiple), with DECA (73.40% on MVSA-Single, 66.52% on MVSA-Multiple), and with IDCA (75.28% on MVSA-Single, 68.42% on MVSA-Multiple). Fig 5 further illustrates this superiority through t-SNE clustering visualizations, where subfigure (a) (MVSA-Single, no mechanism) shows dispersed class distributions, subfigure (b) (MVSA-Single, DECA) displays moderate concentration, subfigure (c) (MVSA-Multiple, IDCA) exhibits limited concentration, and subfigure (d) (MVSA-Multiple, our dual cross-attention) reveals highly concentrated and well-separated class clusters. These results highlight the exceptional capability of our dual cross-attention mechanism in enhancing multi-modal sentiment analysis by effectively capturing inter-modal relationships. The strong performance stems from the mechanism's ability to integrate coarse- and fine-grained features, allowing the model to focus on the most relevant cross-modal interactions and produce more discriminative representations, as evidenced by the improved clustering in Fig. **5**. Overall, our approach demonstrates significant advancements in multi-modal learning through the strategic use of dual cross-attention.

$$\text{fused} = \text{Linear}\left(\text{C}\left(x \cdot \text{f}_{\text{t2v}}^{\text{low}}, x \cdot \text{f}_{\text{v2t}}^{\text{low}}, y \cdot \text{v}_{\text{high}} + z \cdot \text{f}_{\text{t2v}}^{\text{high}}, y \cdot \text{t}_{\text{high}} + z \cdot \text{f}_{\text{v2t}}^{\text{high}}\right)\right) \quad (10)$$

Table 6. Performance Comparison of Different Parameters On MVSA-Single

Parameters	Accuracy	F1
self-adaptation	0.7588	0.7498
x=0.5, y=0.5, z=0.5	0.7612	0.7527
x=1.2, y=0.7, z=0.3	**0.7827**	**0.7795**

Table 7. Performance Comparison of Different Extractions On MVSA-Single

Extration	Accuracy	F1
BERT+ResNet	0.6265	0.6178
BERT+VGG	0.6397	0.6236
CLIP+CLIP	**0.6874**	**0.6727**

We conduct ablation studies on the MVSA-Single dataset to evaluate the effectiveness of our proposed framework, focusing on the weighting parameters and feature extraction methods. In Eq.(10), the parameter x determines the contribution of low-dimensional cross-attention features, capturing global interactions between modalities, y weights the high-dimensional raw features, preserving fine-grained details, and z emphasizes the high-dimensional cross-attention features, highlighting emotionally significant relationships, all of which are concatenated and linearly transformed to produce the final fused representation. Table 6 examines the impact of these parameters, where

the self-adaptation setting dynamically adjusts x, y and z by computing the cosine similarity between the low- and high-dimensional features of each modality, averaged across the batch to reflect alignment strength; specifically, we normalize the similarity scores to the range [0, 1], then scale x proportionally to the low-dimensional similarity, y to the high-dimensional raw feature similarity, and z to the high-dimensional cross-attention similarity, ensuring adaptive weighting based on feature alignment, though this results in moderate performance due to over-balancing. In contrast, equal weights (x=0.5, y=0.5, z=0.5) slightly improve performance by treating all features uniformly, but through extensive experiments, we identify the optimal configuration (x=1.2, y=0.7, z=0.3) as the best, demonstrating the importance of emphasizing low-dimensional cross-attention features for effective multimodal fusion. Table 7 compares different feature extraction methods under the condition of excluding any fusion modules to isolate their impact, where BERT+ResNet and BERT+VGG show lower performance due to their limited ability to capture multimodal interactions, while our CLIP+CLIP configuration achieves the highest accuracy and F1 score, highlighting CLIP's superior capability in extracting and aligning textual and visual features for sentiment recognition. In summary, our ablation studies validate the effectiveness of the proposed weighting strategy and feature extraction approach, with the optimal parameter settings significantly enhancing the model's ability to capture cross-modal interactions.

4 Conclusion

To address the challenges of capturing global cross-modal interactions in low-dimensional features and prioritizing emotionally significant relationships in high-dimensional features for multimodal sentiment recognition, this paper proposes a novel framework integrating Distance-Modulated Exponential Cross Attention (DECA) and Interaction-Gated Differential Cross Attention (IDCA). Firstly, DECA enhances the fusion of coarse-grained features by dynamically scaling attention scores based on feature distances. Secondly, IDCA selectively focuses on fine-grained cross-modal connections through an interaction-gated differential mechanism. Experimental results on diverse social media datasets demonstrate that our approach significantly improves sentiment recognition accuracy compared to existing methods, while maintaining computational efficiency. Future work will focus on incorporating additional modalities, such as audio, to further enhance the framework's robustness and applicability in real-world multimodal scenarios.

References

1. Das R, Singh T D. Multimodal sentiment analysis: a survey of methods, trends, and challenges[J]. ACM Computing Surveys, 2023, 55(13s): 1-38.
2. Gandhi A, Adhvaryu K, Poria S, et al. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions[J]. Information Fusion, 2023, 91: 424-444.

3. Zhu L, Zhu Z, Zhang C, et al. Multimodal sentiment analysis based on fusion methods: A survey[J]. Information Fusion, 2023, 95: 306-325.
4. Guo Z, Jin T, Zhao Z. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition[J]. arXiv preprint arXiv:2407.05374, 2024.
5. Singh U, Abhishek K, Azad H K. A survey of cutting-edge multimodal sentiment analysis[J]. ACM Computing Surveys, 2024, 56(9): 1-38.
6. Xu W, Chen J, Ding Z, et al. Text sentiment analysis and classification based on bidirectional gated recurrent units (grus) model[J]. arXiv preprint arXiv:2404.17123, 2024.
7. Hafner M, Katsantoni M, Köster T, et al. CLIP and complementary methods[J]. Nature Reviews Methods Primers, 2021, 1(1): 20.
8. Shen S, Li L H, Tan H, et al. How much can clip benefit vision-and-language tasks[J]. arXiv preprint arXiv:2107.06383, 2021.
9. Li L H, Yatskar M, Yin D, et al. Visualbert: A simple and performant baseline for vision and language[J]. arXiv preprint arXiv:1908.03557, 2019.
10. Li L H, Yatskar M, Yin D, et al. What does BERT with vision look at[C]//Proceedings of the 58th annual meeting of the association for computational linguistics. 2020: 5265-5275.
11. Tan H, Bansal M. Lxmert: Learning cross-modality encoder representations from transformers[J]. arXiv preprint arXiv:1908.07490, 2019.
12. Jawahar G, Sagot B, Seddah D. What does BERT learn about the structure of language[C]//ACL 2019-57th Annual Meeting of the Association for Computational Linguistics. 2019.
13. Zhou C, Li Q, Li C, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt[J]. International Journal of Machine Learning and Cybernetics, 2024: 1-65.\
14. Tan K L, Lee C P, Anbananthen K S M, et al. RoBERTa-LSTM: a hybrid model for sentiment analysis with transformer and recurrent neural network[J]. IEEE Access, 2022, 10: 21517-21525.
15. Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[J]. Advances in neural information processing systems, 2019, 32.
16. Koonce B. ResNet 50[M]//Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization. Berkeley, CA: Apress, 2021: 63-72.
17. Wu Z, Shen C, Van Den Hengel A. Wider or deeper: Revisiting the resnet model for visual recognition[J]. Pattern recognition, 2019, 90: 119-133.
18. Tammina S. Transfer learning using vgg-16 with deep convolutional neural network for classifying images[J]. International Journal of Scientific and Research Publications (IJSRP), 2019, 9(10): 143-150.
19. Rajab M A, Abdullatif F A, Sutikno T. Classification of grapevine leaves images using VGG-16 and VGG-19 deep learning nets[J]. TELKOMNIKA (Telecommunication Computing Electronics and Control), 2024, 22(2): 445-453.
20. Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818-2826.