MPID: A Modality-Preserving and Interaction-Driven Fusion Network for Multimodal Sentiment Analysis

Tianyi Li
Shanghai University of Electric Power
tianyi@mail.shiep.edu.cn

Daming Liu
Shanghai University of Electric Power
ldm@shiep.edu.cn

Abstract

The advancement of social media has intensified interest in the research direction of Multimodal Sentiment Analysis (MSA). However, current methodologies exhibit relative limitations, particularly in their fusion mechanisms that overlook nuanced differences and similarities across modalities, leading to potential biases in MSA. In addition, indiscriminate fusion across modalities can introduce unnecessary complexity and noise, undermining the effectiveness of the analysis. In this essay, a Modal-Preserving and Interaction-Driven Fusion Network is introduced to address the aforementioned challenges. The compressed representations of each modality are initially obtained through a Token Refinement Module. Subsequently, we employ a Dual Perception Fusion Module to integrate text with audio and a separate Adaptive Graded Fusion Module for text and visual data. The final step leverages text representation to enhance composite representation. Our experiments on CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets demonstrate that our model achieves state-of-the-art performance.

1 Introduction

With the development of online social media platforms such as TIKTOK, more and more people use various social media to express their emotions and views, such as sharing pictures and videos. Consequencely, the task of Multimodal Sentiment Analysis (MSA) has been extended to multimodel data, and is no longer limited to unimodel data. With the increase of multimodel data, MSA has become a popular research direction (Baltrušaitis et al., 2018). Extracting user emotions from multimodel data can help decision makers understand historical situations, predict future trends, and make more wise decisions (Chatterjee et al., 2019). These emotions are usually classified as positive, negative or neutral.

Previous works mostly focused on integrating three modalities with equal importance, while ignoring the individual information of each modality itself. Some works simply utilize feature cascading as a multimodal fusion mechanism for fusion (Majumder et al., 2019; Joshi et al., 2022). Although some works focus on text as the main modality for fusion (Ma et al., 2023), in some cases, analyzing the speaker's emotions solely through text modality is not enough, such as the following sentence."The song is sick." The meaning conveyed by this sentence is ambiguous, but if the speaker is smiling while saying it, it will be considered positive. Conversely, if the speaker is frowning while saying this sentence, it will be considered negative. And different tones will also affect the result, as shown in Figure 1.

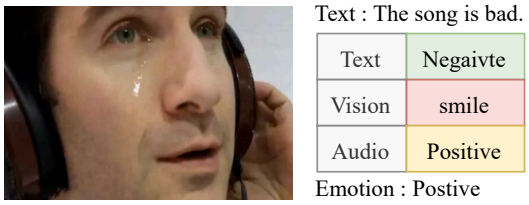

Figure 1: Illustration of the significance of different modalities for accurate prediction.

To address the above challenges, we need to consider the individual information of each modality in order to better understand and utilize their unique characteristics. To the best of our knowledge, when considering semantic alignment and contextual coherence, the similarity between text and audio, as well as text and visual content, is higher. The exploitation of this higher similarity is crucial to our research, as it directly impacts the effectiveness of feature fusion in MSA. Text and audio both convey linguistic and semantic information, with text providing clear language and audio adding elements like tone and intonation. Visual content, such as

Proceedings of the 31st International Conference on Computational Linguistics, pages 4313–4322
January 19–24, 2025.

text or symbols in images, can complement textual information and enhance contextual understanding. While text and audio share semantic similarities, visual and audio modalities differ significantly. Visual content relies on images, colors, and shapes, whereas audio uses sound waves, intonation, and speech rate.

The perceptual mechanisms and expression forms of these modalities are fundamentally different, leading to significant differences in information representation and understanding. Although cross-modal learning can integrate these pieces of information, it still does not fully and effectively utilize the unique information of each modality. In this paper, we propose a **M**odality-**P**reserving and **I**nteraction-**D**riven Fusion Network (MPID), which includes two modules: the **D**ual **P**erception **F**usion (DPF) Module and the **A**daptive **G**radual **F**usion (AGF) Module, designed to ensure that the unique information of each modality is thoroughly utilized.

Furthermore, the information density of multimodal data varies, and noise and irrelevant information may be introduced during the fusion process. We propose a **T**oken **R**efinement **M**odule (TRM) to condense the information of each modality, thereby reducing redundant data, improving the processing efficiency of the model, and reducing computational and storage overhead.

The main contributions of our paper can be roughly summarized as follows:

- We propose a Dual Perception Fusion Module that addresses inconsistencies between text and audio while preserving contextual integrity. It employs semantic and distance-based fusion to enhance text and audio utilization.
- We present a Adaptive Gradual Fusion Module dynamically adjusts attention weights to better combine local and global features, minimizing information loss and improving fusion, especially for complex visual and textual data.
- We conducted experiments on the representative CMU-MOSI, CMU-MOSEI and CH-SIMS datasets, demonstrating that our model achieves state-of-the-art performance.

2 Related Work

MSA seeks to harness data from diverse modalities to achieve a comprehensive understanding of sentiment, thereby mitigating ambiguity and enhancing accuracy in sentiment classification. Previous works has mainly focused on unimodal representation learning and multimodal fusion.

For unimodal representations, data from different modalities are integrated into a single feature representation before being fed into the model. (Zadeh et al., 2017; Pham et al., 2019) propose respective feature vectors concatenated to create an extended feature vector. This combined representation is then input into the neural network, allowing the model to simultaneously process and consider information from all modalities during its learning and inference stages. (Hazarika et al., 2020) attempts to decompose modal features in the joint space to represent modal invariance and specific representations.

For multimodal fusion, a good method should effectively integrate information from various modalities and address the heterogeneity between them. (Rahman et al., 2020; Tsai et al., 2018) uses word boundary alignment to learn associations within and between modalities. Due to the increasing popularity of (Vaswani, 2017; Tsai et al., 2019) proposed a cross-modal fusion method for unaligned sequences. Recently, (Yu et al., 2023) introduces knowledge injection based on Adapter architecture for comparative learning with general knowledge representation in order to ignore the influence of domain specific knowledge.

Unlike the previous work, our work proposes a novel fusion network that maximizes interaction between modalities while retaining specific personality information for each modality.

3 Methodology

3.1 Overview

The overall architecture of our model (MPID) is shown in Figure 2. The model mainly consists of four modules, Token Refinement Module, Dual Perception Fusion Module, Adaptive Graded Fusion Module, and **T**ext **A**ugmented **T**ransformer (TAT).

In TRM, unimodal features are first processed to uniformly condense the features of each modality. Subsequently, DPF is used to fuse the condensed text and audio features, while AGF is used to fuse the condensed text and visual features. Considering the importance of textual information, we further enhance the fusion representation obtained using TAT to improve the efficiency and accuracy of the model's utilization of textual information.

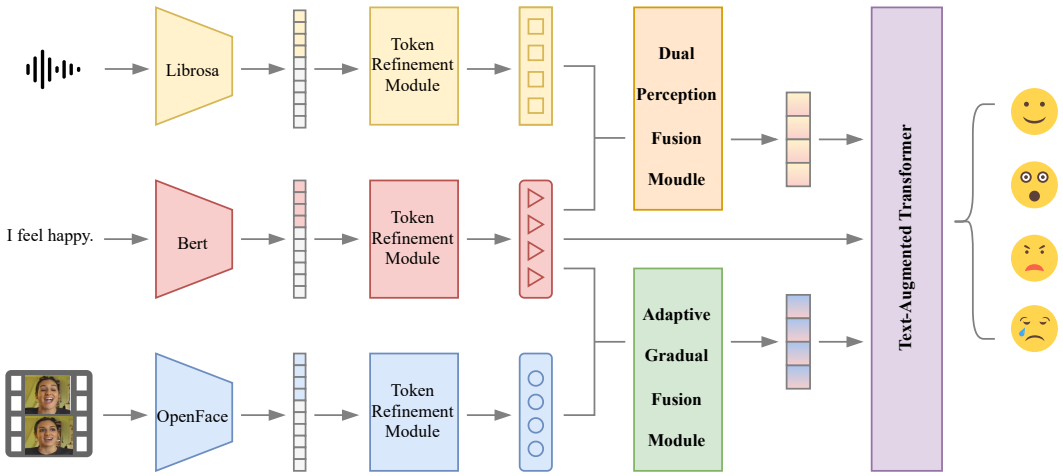

Figure 2: The overall architecture of MPID. This network utilizes Dual Perception Fusion Module and Dual Perception Fusion Module to achieve efficient complementary learning while retaining specific personality information for each modality.

3.2 Problem Definition

In this work, the MSA task encompasses three primary modalities, text (t), visual (v), and audio (a). Each modality contributes unique information to the analysis of sentiment within a video clip.

The input data for our model is derived from video clips, each consisting of a sequence of frames. We leverage precomputed representations for each modality, which are obtained through established methods: BERT (Kenton and Toutanova, 2019) for text, Librosa (McFee et al., 2015) for audio, and OpenFace (Baltrušaitis et al., 2016) for visual features.

For each modality $X_m \in \mathrm{R}^{l_m \times d_m}$, we denote l_m as the sequence length and d_m as the vector dimension, which $m \in \{t, v, a\}$.

3.3 Token Refinement Module

To compress the high-dimensional representations of each modality, we have designed a generic module termed the TRM. The initial representations of all three modalities will be subjected to this module for dimensionality reduction.

Taking text input as an example, we introduce Bottleneck Tokens, denoted as H_B. To achieve a condensed representation of multimodal information, the most straightforward approach is to concatenate the text input X_t with H_B to form a unified sequence. This sequence is then processed by the original transformer model without altering its architecture.

In the experiment, we first performed preliminary modal feature extraction through a mapping

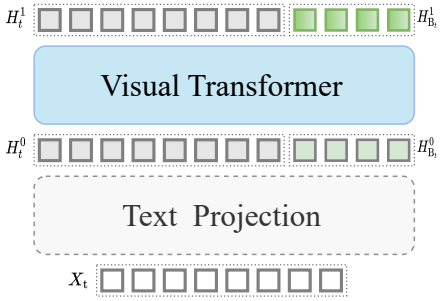

Figure 3: The framework of Token Refinement Module, realize compression of high-dimensional data.

layer,

$$\mathbf{X}_{\mathrm{t}}^{'} = g(\mathbf{X}_{\mathrm{t}}; \mathbf{E}_{\mathrm{t}}), \tag{1}$$

where X_t is the original text input and E_t is the projection matrix corresponding to the text.

And the structure of the Transformer layer was designed to be consistent with the Visual Transformer (ViT) with a depth of 1 (Yuan et al., 2021), as shown in Figure 3.

For layer i, the calculation procedure is as follows,

$$[\mathbf{H}_{\mathrm{t}}^{i+1} || \mathbf{H}_{\mathrm{B}_t}^{i+1}] = \mathrm{ViT}([\mathbf{H}_{\mathrm{t}}^{i} || \mathbf{H}_{\mathrm{B}_t}^{i}]; \theta_{\mathrm{i}}), \tag{2}$$

where $\mathbf{H}_{\mathrm{t}}^{1}$ denotes the final Bottleneck Tokens, $||$ representing the cascading operation, and θ_{i} refers to the associated parameter.

Through the TRM, basic modal information is mapped to randomly initialize Bottleneck Tokens. This process effectively reduces redundant information and achieves greater efficiency while maintaining a compact parameter space.

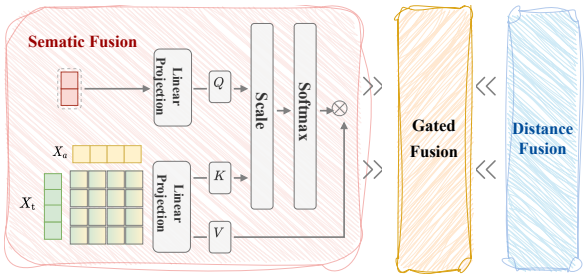

Figure 4: The structure of the Dual Perception Fusion Module adaptively adjusts the semantic and distance based fusion representations to obtain more representative fusion results.

3.4 Dual Perception Fusion Module

In cascading operations involving text and audio, crucial information is often concentrated in specific segments or phrases. For instance, in conversations, certain words are closely related to specific tones. Traditional global interaction models may not effectively capture these nuanced details. To address this, we proposed DPF that combines semantic and distance based dual perception fusion, as shown in Figure 4.

Based on semantic level, by processing and focusing on the specific pairs of each local interaction separately for audio tokens and text tokens, we have evaluated various interaction representation techniques, such as bi-linear and kernel-based approaches. Our findings suggest that linear aggregation is sufficient for capturing the semantic interactions between text and audio in practical tasks. Formally,

$$X_{ij}^{ta} = W^t X_i^t + W^a X_j^a, \tag{3}$$

$$X^{ta} = \left[X_{ij}^{ta} \mid i = 1, ..., n_t; j = 1, ...n_a \right], \tag{4}$$

where W^t and W^a are the projection matrices corresponding to X^t and X^a, respectively, and X^{ta} is the calculated density matrix.

Additionally, we have introduced Semantic Tokens as an intermediate representation post-fusion, which enhances flexibility in adapting to diverse data and scenarios. For the semantic fusion representation of layer X_l^s, formally,

$$X_l^s = CrossAttention\left(X_{l-1}^{ta}, X_{l-1}^s\right). \tag{5}$$

Based on distance perception, we calculated the Manhattan distance between input sequences. Manhattan distance is a measure that sums the absolute differences between point pairs in a feature space. It can capture linear changes between features, which is particularly effective for handling high-dimensional sparse data, formally,

$$D_{ij} = \sum_{k=1}^{d} |x_{t,ik} - x_{v,jk}|, \tag{6}$$

$$AttD_{ij} = \frac{\exp(-\varphi D_{ij})}{\sum_j \exp(-\varphi D_{ij})}, \tag{7}$$

$$\tilde{x}_{v,i} = x_{v,i} \cdot \frac{1}{|S_t|} \sum_j D_{ij}, \tag{8}$$

$$\tilde{x}_{t,j} = x_{t,j} \cdot \frac{1}{|S_v|} \sum_i D_{ij}, \tag{9}$$

where D_{ij} represents the distance between the i-th position in the input sequence and the j-th position in the target sequence. This distance can be Manhattan distance, Euclidean distance, etc. Through experiments, we have chosen Manhattan distance here. And φ is used to adjust the degree of influence of distance on attention weights. $\tilde{x}_{v,i}$ and $\tilde{x}_{t,j}$ are weighted feature representations, and $|S_t|$ and $|S_v|$ are the sequence lengths of x_a and x_v.

The output based on distance fusion O_d can be represented as,

$$O_d = \mathrm{FC}_{out}(\mathrm{concat}(\tilde{x}_v, \tilde{x}_t)). \tag{10}$$

where FC_{out} is a fully connected layer.

Since direct averaging or concatenation may weaken the representational capacity, we propose a gated fusion method inspired by (Mai et al., 2019) to adaptively combine semantic and distance features. This method integrates these features from both local and global perspectives, and is formulated as follows,

$$\mathrm{gate_c} = \mathrm{FusionFC}(\mathrm{Concat}(O_s, O_d)), \tag{11}$$

$$\mathrm{Semg} = f(\mathrm{gate_c}[:, :, 0]), \tag{12}$$

$$\mathrm{Distg} = f(\mathrm{gate_c}[:, :, 1]), \tag{13}$$

$$\mathrm{Output} = \mathrm{Semg} \odot O_s + \mathrm{Distg} \odot O_d, \tag{14}$$

where f is the activation function, $\odot$ is the dot product.

Through the above operations, we can automatically adjust the contribution of different levels of fusion results based on the characteristics of input data, thereby capturing more complex interactions and improving the representation ability of the model.

3.5 Adaptive Gradual Fusion Module

Text and visual features inherently possess different representations and contextual dependencies, and it is important to consider the role of local visual information. Local details, such as facial expressions, postures, and other subtle visual cues, are critical for accurately understanding emotions. Local features aid in identifying and interpreting nuanced emotional changes, while global information provides the overall context and background. Therefore, integrating local visual information with textual data can offer a more comprehensive analysis and recognition of emotional states. To address these challenges, we propose AGF.

Initial fusion is achieved by calculating cross-attention maps. The text features X_t and visual features X_v are mapped into queries, keys, and values. Attention scores are computed by taking the dot product of queries and keys, and local attention maps M are generated using softmax. This stage of local information fusion focuses on aligning each text position with the specific visual features, ensuring precise local feature alignment,

$$M = \text{softmax}\left(\frac{X_t \cdot X_a^T}{\sqrt{d_k}}\right) \cdot X_a. \quad (15)$$

Building on local information fusion, we introduce global average attention map G and compute the average of the attention maps, which represents the overall average attention of the features. Formally,

$$G = \frac{1}{n}\sum_{i=1}^{n} M_i. \quad (16)$$

By combining global and local information through a dynamic weighting coefficient G,

$$F = \alpha \cdot M + (1-\alpha)(\text{softmax}(G \cdot M)). \quad (17)$$

This approach facilitates a gradual transition to global information fusion, enabling the feature fusion process to integrate both detailed local information and global context, thereby enhancing the overall representational capacity of the features.

3.6 Text-Augmented Transformer and Output

Cross-modal attention operations leverage information from one modality to enhance another modality by learning directed attention between paired modalities. Given that advanced semantic features from text, speech, and visual fusion have been acquired, we apply cross-modal attention operations to further enhance the text modality. By weighting these representations and then utilizing a fully connected layer for classification, the text modality serves as a consistent reference. It not only guides the learning process of visual and audio representations but also improves sentiment analysis accuracy by integrating the fused representations.

3.7 Overall Learning Objectives

In summary, this method has only one learning objective, which is the loss function (L) of sentiment analysis, which is used to measure the gap between predicted emotions and true emotions.

$$\mathcal{L} = \frac{1}{N}\sum_{n=0}^{N} \|y^n - \hat{y}^n\|_2^2, \quad (18)$$

where N represents the number of samples in the training set, y^n denotes the sentiment label of the n-th sample, and $\hat{y}^n$ is the prediction made by MPID.

Additionally, due to the simplicity of its optimization objective, MPID is easier to train compared to advanced methods with multiple optimization objectives (e.g.Hazarika et al., 2020), as it does not require the tuning of additional hyperparameters.

4 Evaluation

4.1 Datasets and Metrics

In this work, we employ the CMU-MOSI,CMU-MOSEI and CH-SIMS dataset to evaluate our proposed method.

CMU-MOSI (Zadeh et al., 2016) includes 93 opinion videos sourced from YouTube, which have been carefully processed into 2,199 distinct utterance clips. Each clip is required to feature both the speaker's voice and face clearly, without interference from other individuals. Additionally, each clip must have a sufficient duration and is assigned a sentiment score within the range of [-3, 3]. The sentiment scores are not fixed but encompass all values within this interval. For dataset division, 1,284 utterances are used for training, 229 for validation, and 686 for testing.

CMU-MOSEI (Zadeh et al., 2018b), created by the same authors as CMU-MOSI, offers a larger and more advanced dataset. CMU-MOSEI includes 22,852 annotated video clips (utterances) from 1,000 different speakers and 250 topics from online video platforms. Each utterance is labeled with sentiment intensity on a scale from [-3, 3].

CH-SIMS (Yu et al., 2020) comprises 60 original videos sourced from movies, TV shows, and variety programs, segmented into 2,281 individual clips. Each clip is annotated with sentiment intensity ranging from [-1, 1]. The dataset is partitioned into training, validation, and test subsets in a 60% to 20% to 20% split.

Consistent with recent research (Han et al., 2021), sentiment in the CMU-MOSI and CMU-MOSEI datasets is classified into seven levels based on intensity: [-3, -2) for highly negative, [-2, -1) for negative, [-1, 0) for weakly negative, [0] for neutral, (0, 1] for weakly positive, (1, 2] for positive, and (2, 3] for highly positive.

Our evaluation metrics include 7-class accuracy (Acc-7), 5-class accuracy (Acc-5), 2-class accuracy (Acc-2), F1 score (F1), Mean Absolute Error (MAE), and Pearson Correlation Coefficient (Corr). The 5-class classification includes negative, weakly negative, neutral, weakly positive, and positive sentiments. Acc-2 and F1 are assessed in two contexts: one excluding neutral (negative/non-negative) and one including neutral (negative/positive). The Pearson Correlation Coefficient (Corr) measures the covariance normalized by the standard deviations of the variables, while MAE indicates the average magnitude of errors between predicted and actual values.

We train our model on a single RTX 3080 GPU. The codes for our implementation are available at `https://github.com/Caroline-L11/MPID`.

4.2 Baselines

To verify the performance and effectiveness of the proposed model, we compared its performance with the following models.

EF-LSTM, the EF-LSTM model (Williams et al., 2018) integrates various feature vectors and inputs them into a Long Short-Term Memory (LSTM) network to forecast sentiment polarity.

MFN, the MFN model (Zadeh et al., 2018a) merges LSTM with a memory mechanism to capture the temporal dependencies.

MulT, the MulT model (Baltrušaitis et al., 2018) employs a cross-modal attention module based on the Transformer architecture to capture interactions between sequences.

Self-MM, the Self-MM model (Yu et al., 2021) is a multi-task framework based on self-supervised learning, where unimodal sentiment labels are automatically generated by the model.

MMIM, the MMIM model (Han et al., 2021) maximizes mutual information (MI) at a hierarchical level between unimodal input pairs and between multimodal fusion results and unimodal inputs.

ALMT, the ALMT model (Zhang et al., 2023) proposed an adaptive language guidance network through the AHL module.

MISA, the MISA model (Hazarika et al., 2020) learns modal invariants and representations of specific modalities.

4.3 Main Results

The experimental results on the MOSI and MOSEI datasets are shown in Table 1.

For the MOSEI dataset, the MPID model outperforms baseline models across all evaluation metrics. In the binary classification task, the MPID model achieves an accuracy of 85.1%, representing a 4% improvement over the ALMT and other models, highlighting its enhanced ability to predict sentiment polarity. In the multi-classification task, the MPID model scores 53.64% on Acc-7, exceeding all baseline models and indicating superior performance in capturing fine-grained emotional information. Additionally, the model achieves a MAE of 0.533, the lowest among all models, demonstrating its accuracy in predicting emotional intensity.

For the MOSI dataset, MPID exceeded the performance of both ALMT and Self MM, achieving the highest F1 score of 84.52%, reflecting excellent stability and accuracy in sentiment classification.

And the experimental results on the CH-SIMS datasets are shown in Table 2.

For the CH-SIMS dataset, the MPID model significantly outperforms all baseline models on the CH-SIMS dataset. In binary classification tasks, MPID shows an improvement of approximately 1.5% to 2% compared to other models. In the multi-class classification task, MPID achieved an accuracy of 44.21% on Acc-5. In the regression task, our model achieved MAE of 0.421, outperforming all baseline models.

In summary, MPID performs better than existing SOTA models.

Model	CH-SIMS					
	Year	Acc-2 (%)↑	Acc-5(%)↑	MAE↓	F1 (%)↑	Corr↑
MulT	2019	77.55	36.84	0.443	78.99	0.554
MISA	2020	75.54	-	0.447	75.59	0.559
Self-MM	2021	79.92	42.53	0.435	79.44	0.596
MPID(ours)		**81.56**	**44.21**	**0.421**	**81.91**	**0.618**

Table 2: Performances of MPID on the CH-SIMS datasets.

Model	MOSI						MOSEI					
	Year	Acc-2 (%)↑	Acc-5(%)↑	Acc-7 (%)↑	MAE↓	F1 (%)↑	Corr↑	Acc-2 (%)↑	Acc-5(%)↑	Acc-7 (%)↑	MAE	F1 (%)↑	Corr↑
EF-LSTM	2018	77.39	40.12	35.6	0.949	77.36	0.668	77.82	50.19	50.03	0.603	78.33	0.683
MFN	2018	77.66	40.46	35.83	0.926	77.62	0.671	78.91	51.74	51.33	0.574	79.99	0.716
MulT	2019	78.6	-	33.59	1.147	78.29	0.662	80.15	-	46.61	0.651	79.8	0.662
Self-MM	2021	79.79	-	43.77	0.923	79.68	0.753	80.02	-	51.47	0.501	79.95	0.745
MMIM	2021	79.19	-	44.35	0.682	81.01	0.785	82.11	-	51.11	0.541	85.91	0.757
ALMT	2023	81.7	47.81	42.13	0.761	81.8	0.783	-	52.99	51.1	0.586	83.69	0.759
MPID(Ours)		**85.4**	**55.7**	**48.4**	**0.706**	**84.52**	**0.792**	**85.1**	**55.3**	**53.64**	**0.533**	**85.3**	**0.776**

Table 1: Performances of MPID on the CMU-MOSI and CMU-MOSEI datasets.

4.4 Ablation Experiment

The following ablation experiments will be conducted using the CMU-MOSI dataset. To assess the effectiveness of these experiments, we will focus on challenging multi-class indicators, such as ACC-5, as the primary evaluation metrics.

4.4.1 Effects of Different Components

To evaluate the effectiveness of our proposed module, we performed extensive experiments, the results of which are summarized in the Table 3.

Configs	Acc-5(%)	Acc-7(%)	MAE	F1 (%)	Corr
MPID	55.7	48.4	0.706	84.52	0.792
w/o TRM	48.98	44.02	0.7507	82.08	0.784
w/o TAT	46.5	41.2	0.779	81.65	0.777
w/o DPF	50.01	45.77	0.747	81.68	0.784
w/o AGF	50.87	45.34	0.749	80.94	0.78
w/o DPF,AGF	18.8	18.22	1.48	53.79	0.208

Table 3: Research on ablation of MPID under different module settings.

The absence of the TRM led to a notable decline in model performance, with Acc-5 and Acc-7 dropping to 48.98% and 44.02%, respectively, while MAE increased to 0.7507. This suggests that the TRM is crucial for enhancing feature representation and reducing errors. Removing the TAT resulted in a decrease of 9.2% in Acc-5 and 7.2% in Acc-7, highlighting the importance of the TAT in capturing fine-grained textual information. The overall performance significantly declined after its removal.

Excluding the DPF and AGF individually led to a decrease of approximately 5% in both Acc-5 and Acc-7, accompanied by an increase in MAE. This underscores the importance of incorporating multimodal information for optimal performance.

When both DPF and AGF were removed simultaneously, the model's performance deteriorated sharply, with Acc-5 and Acc-7 significantly dropping and MAE rising to 1.48. This indicates that cross-modal fusion is essential for effective MSA. The absence of these modules severely hampers the model's ability to integrate multimodal information, resulting in substantial performance degradation.

In summary, each module plays an indispensable role in MSA. Their effective integration significantly enhances the model's capability to understand and analyze multimodal data, thereby contributing to the outstanding performance of MPID in this task.

4.4.2 Effects of Different Settings in DPF

In order to substantiate the efficacy of the DPF, an extensive series of experiments was conducted. The outcomes of these experiments are delineated in Table 4.

Configs	Acc-5(%)	Acc-7(%)	MAE	F1 (%)	Corr
Effect of Integrating Different Modalities					
DPF(a-t)	45.63	41.25	0.792	81.21	0.761
DPF(v-t)	43.73	39.8	0.784	81.98	0.774
DPF(a-v)	15.6	15.45	1.461	56.15	0.281
Impact of Activation Function					
MPID(sigmoid,tanh)	55.7	48.4	0.706	84.52	0.792
DPF(sigmiod,sigmoid)	51.9	46.06	0.734	82.52	0.791
DPF(tanh,sigmoid)	51.31	45.48	0.738	82.1	0.789
DPF(tanh,tanh)	49.27	44.02	0.747	81.19	0.785
DPF(relu,tanh)	47.81	43.88	0.776	80.05	0.771
DPF(relu,relu)	51.46	46.21	0.746	80.93	0.78
DPF(relu,sigmoid)	48.98	43.88	0.776	79.72	0.765
DPF(sigmoid,relu)	46.79	41.98	0.787	80.1	0.764
DPF(tanh,relu)	48.4	42.86	0.763	81.4	0.773

Table 4: Ablation studies for DPF, $m - n$ denotes the fusion of m and n modalities.

The outcomes presented in the upper section of Table 4 robustly substantiate the effectiveness of the DPF's design. The visual-audio fusion is less effective due to the weak correlation between visual and audio information, such as environmental sounds or speech that do not directly match image content. This results in less intuitive distance-based measurements and poorer fusion. In contrast, audio and text exhibit higher semantic alignment, allowing the DPF to effectively capture and align their semantic information. This enhances fusion accuracy by focusing on their complementary features.

Ablation studies on activation functions revealed that combining Sigmoid and Tanh optimizes our model by utilizing Sigmoid for precise weighting

of semantic information and Tanh for broader adjustment of distance information. As shown in Table 4, incorporating the ReLU function reduced performance, likely due to its range not being well-suited for the fine-tuning required in our module's fusion process.

4.4.3 Effects of Different Modalities in AGF

To evaluate the effectiveness of the AGF, we conducted experiments retaining only TRM, with results summarized in the Table 5.

Configs	Acc-5(%)	Acc-7(%)	MAE	F1 (%)	Corr
AGF(v-t)	50.44	44.75	0.739	82.14	0.782
AGF(v-a)	16.62	16.47	1.45	55.93	0.195
AGF(t-a)	50.29	44.46	0.742	81.41	0.78

Table 5: The effect of using AGF with different modalities for fusion,

The results indicate that visual-text fusion achieves the best performance, whereas visual-audio fusion performs poorly. This disparity may stem from the strong semantic correlation between visual and textual features, which facilitates more effective fusion. In contrast, audio features are more complex.

Additionally, visual features possess a clear spatial structure that supports effective local information preservation, while textual features have strong semantic associations and are easier to handle in terms of local information. Audio features, being time series data, have local information such as pitch and rhythm that is less intuitive compared to visual and textual features.

4.4.4 Visualization of Attention in AGF

To evaluate the effectiveness of AGF, we visualized attention maps at different training stages (Epoch 20, Epoch 60, and Epoch 100), as shown in Figure 5. These maps illustrate how the module dynamically attends to different visual features based on textual guidance during the fusion process.

At the early training stage (Epoch 20), the attention weights are concentrated on specific regions, suggesting that the module initially emphasizes localized visual features guided by the text. By Epoch 60, the attention maps show a broader distribution, indicating the incorporation of more contextual visual features influenced by the textual modality. At Epoch 100, the attention achieves a balanced distribution, effectively capturing both fine-grained local details and global contextual patterns in the visual information.

This progressive evolution highlights the capability of the module to adaptively fuse visual and textual modalities. By transitioning from local to global feature emphasis, the module effectively leverages complementary cues from both modalities.

4.4.5 Effects of Different Fusion Combinations

In addition, we evaluated the impact of various modal fusion combinations.

Configs	Acc-5(%)	Acc-7(%)	MAE	F1 (%)	Corr
MPID DPF(a-t),AGF(v-t)	55.7	48.4	0.706	84.52	0.792
DPF(a-t),AGF(a-t)	50.4	44.9	0.743	81.54	0.784
DPF(v-t),AGF(a-t)	50.2	45.3	0.735	82.45	0.789
DPF(v-t),AGF(v-t)	51.17	45.48	0.751	82.29	0.774

Table 6: The impact of various modal fusion combinations.

The Table 6 shows that omitting fusion for any modality leads to a significant drop in multi-classification accuracy. Specifically, using visual text fusion with DPF and audio text fusion with AGF results in a 5.5% decrease in ACC-5. This underscores the importance of designing DPF and AGF to leverage the unique characteristics of each modality, thereby optimizing the retention of modality-specific information.

4.4.6 Effects of the Guidance of Different Features in TAT

To assess the effectiveness of the TAT module, we used various modalities as reinforcement inputs, with results summarized in the Table 7.

Configs	Acc-5(%)	Acc-7(%)	MAE	F1 (%)	Corr
MPID(t)	55.7	48.4	0.706	84.52	0.792
TAT(a)	50.29	45.19	0.735	81.31	0.791
TAT(v)	50.15	45.34	0.743	80.62	0.79

Table 7: Performances of using TAT modules with different modalities.

Our proposed MPID approach, utilizing text as the reinforcement modality, outperforms the other modalities. Specifically, it achieves an ACC-5 score approximately 5% higher and an ACC-7 score about 3% higher. These results underscore the significant advantage and effectiveness of using text as the reinforcement modality.

5 Conclusion

Our findings confirm that the MPID network's approach to multimodal sentiment analysis effectively

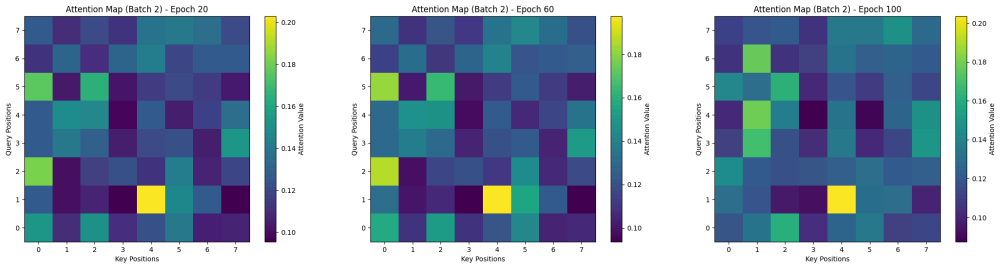

Figure 5: Attention maps of AGF across different training epochs (20, 60, and 100) for the same batch of data. The maps illustrate the module's evolving focus from localized visual features to progressively incorporating global contextual information, guided by textual inputs.

leverages and preserves the unique traits of each data modality. In this study, we found that the integration of DPF and AGF modules substantially enhances the fusion process, leading to improved sentiment analysis. Our results also indicate that the TRM and TAT module further refines the model's accuracy. The validation of each module's contribution reaffirms the network's robust and efficient design. Results from CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets indicate that our model achieves superior performance.

6 Limitation

Our MPID model uses multiple fusion modules, which may face efficiency challenges due to computing resource limitations and processing speed. Handling large feature data for effective fusion can result in high computational complexity and memory usage, creating potential bottlenecks.

Furthermore, integrating and managing information across the four modules may increase system complexity, potentially affecting overall model performance.

Meanwhile, we will conduct further research on the issues related to modal loss and noisy data in our future work.

Acknowledgements

We thank the anonymous reviewers for their constructive comments and suggestions.

References

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–10. IEEE.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. Semeval-2019 task 3: Emocontext contextual emotion detection in text. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 39–48.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412*.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1122–1131.

Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Singh, and Ashutosh Modi. 2022. Cogmen: Contextualized gnn based multimodal emotion recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4148–4164.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Feipeng Ma, Yueyi Zhang, and Xiaoyan Sun. 2023. Multimodal sentiment analysis with preferential fusion and distance-aware contrastive learning. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1367–1372. IEEE.

Sijie Mai, Songlong Xing, and Haifeng Hu. 2019. Locally confined modality fusion network with a global perspective for multimodal human affective computing. *IEEE Transactions on Multimedia*, 22(1):122–137.

N Majumder, S Poria, D Hazarika, R Mihalcea, A Gelbukh, and E Cambria DialogueRNN. 2019. An attentive rnn for emotion detection in conversations. *Association for the Advancement of Artificial Intelligence*, pages 6818–6825.

Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *SciPy*, pages 18–24.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6892–6899.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, page 2359. NIH Public Access.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Jennifer Williams, Steven Kleinegesse, Ramona Comanescu, and Oana Radu. 2018. Recognizing emotions in video using multimodal dnn feature fusion. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, pages 11–19.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3718–3727.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10790–10797.

Yakun Yu, Mingjun Zhao, Shi-ang Qi, Feiran Sun, Baoxun Wang, Weidong Guo, Xiaoli Wang, Lei Yang, and Di Niu. 2023. Conki: Contrastive knowledge injection for multimodal sentiment analysis. *arXiv preprint arXiv:2306.15796*.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 558–567.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018b. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. *arXiv preprint arXiv:2310.05804*.