Available Online: 13 April 2026 DOI: 10.54254/2977-3903/2026.32737

Multimodal sentiment analysis based on dynamic mode selection and contrastive learning alignment

Bing He

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, China

hebing7778@163.com

Abstract. Multimodal Sentiment Analysis (MSA) seeks to predict a speaker's sentiment orientation by comprehensively utilizing modalities such as text, vision, and audio. As deep learning and cross-modal fusion technologies evolve, key challenges include alleviating heterogeneity across modality feature spaces, avoiding bias from fixed main-modal fusion strategies, and enhancing model adaptability to dynamic changes in modality contribution across different samples. To address these issues, this paper proposes a multimodal sentiment analysis framework based on adaptive modality selection and contrastive learning alignment, named Adaptive Modality Selection and Guided Fusion Network (AMSGFN). The framework first employs a cross-modal contrastive learning alignment mechanism to map text, vision, and audio features into a shared semantic space, mitigating semantic discrepancies among heterogeneous modalities. A lightweight modality scoring module then evaluates the discriminability and reliability of each modality for the current sample, adaptively identifying the dominant modality. Building on this, a dominant modality-guided fusion mechanism selectively integrates supplementary information from auxiliary modalities around the dominant modality, highlighting key emotional semantics while suppressing noise and redundant information. Experimental results demonstrate that the proposed method achieves superior performance compared to existing approaches across multiple public datasets, confirming the effectiveness and robustness of the framework in multimodal sentiment analysis.

Keywords: multimodal sentiment analysis, contrastive learning, modal alignment, modality importance evaluation, dominant modality guidance, adaptive fusion

1. Introduction

The core of multimodal learning lies in the collaborative understanding of heterogeneous information from different modalities, which typically include vision, hearing, and text [1-3]. With the massive emergence of multimodal content in social media, how to integrate heterogeneous modal data to accurately infer emotional states has become a key research issue in the field of Multimodal Sentiment Analysis (MSA). Current mainstream methods can be roughly classified into two categories: one is based on triplet symmetric structure for fusion modeling [4-7], defaulting to an equal importance of each modality; the other builds an analysis

framework centered on the text modality [8-11]. However, both methods implicitly assume a common hypothesis: the importance of modalities is either equal or the text always dominates.

However, the real-world situation is far more complex than this assumption. The actual contribution of each modality in different samples is not the same; it is neither a constant average nor always dominated by text. For example, the text content "Your finances are still in need of an assistant" has a clearly negative sentiment, the audio conveys a similar negative tendency in a teasing and mocking tone, while the visual modality shows a smile, which presents an positive emotion. In this situation, if the fusion is still based on text as the fixed center, the model is prone to be misled by superficial semantics, thereby generating misjudgments that deviate from the true emotions. Therefore, in the fusion process, whether it can dynamically identify the dominant modality of the current sample and adaptively adjust the fusion weights according to its relationship with the auxiliary modality, not only directly affects the accuracy of sentiment prediction, but also determines the robustness of the model in complex and variable real-world scenarios.

Figure 1. Modalities convey different emotions

Furthermore, multimodal sentiment analysis also faces another fundamental challenge: the semantic separation between heterogeneous feature spaces. Text, vision, and audio data are naturally distributed in different representation spaces [12], with their own unique structural attributes and noise distributions. For example Figure 1, text mainly conveys emotional tendencies through the semantic network of words, while vision and audio express emotions through facial action units, gesture trajectories, speech rate and rhythm, etc. If simple vector concatenation or averaging is used for fusion, it is likely to destroy the original structural information of each modality, causing semantic distortion, and thereby inducing the model to learn false associations unrelated to the true emotions, significantly reducing its performance. Therefore, before cross-modal fusion, the heterogeneous features must be mapped to a unified semantic space that is semantically comparable.

This paper proposes a multimodal sentiment analysis framework based on adaptive modality selection and contrastive learning alignment, namely Adaptive Modality Selection and Guided Fusion Network (Adaptive Modality Selection and Guided Fusion Network, AMSGFN). This framework first uses the cross-modal contrastive learning alignment mechanism to map text, vision, and audio features to a shared semantic space to alleviate the semantic offset problem between heterogeneous modalities; then, through a lightweight modality scoring module, the discriminative and reliable nature of different modalities in the current sample is evaluated, and the dominant modality is adaptively determined; on this basis, a dominant modality-guided fusion mechanism is further designed to selectively integrate supplementary information from auxiliary

modalities centered on the dominant modality, thereby highlighting key emotional semantics while suppressing the interference of noise and redundant information.

The main contributions of this paper are as follows:

1.A multi-modal sentiment analysis framework based on adaptive modal selection and contrastive learning alignment was proposed to address the issues of modality space heterogeneity, fixed dominant modality fusion bias, and dynamic changes in modality contribution in multi-modal sentiment analysis.

2. A cross-modal contrastive learning alignment module was designed. By mapping the text, visual, and audio modalities to a shared semantic space, the semantic consistency among heterogeneous modalities' representations was enhanced, and the problem of cross-modal semantic offset was alleviated.

3. An adaptive dominant mode selection and dominant mode-guided fusion strategy has been introduced. This enables the model to dynamically identify the current dominant mode based on the sample characteristics, and selectively integrate supplementary information from the auxiliary modes around it, thereby enhancing the model's ability to model complex contexts and sample variations.

2. Proposed methodology

2.1. Overview of AMSGFN architecture

The AMSGFN proposed in this paper is mainly composed of three core modules: the Cross-modal Contrastive Alignment Module (CCAM), the Dominant Modality Selection Module (DMSM), and the Dominant-modality Guided Fusion and Prediction Module (DGFPM). The overall framework is shown in Figure 2.

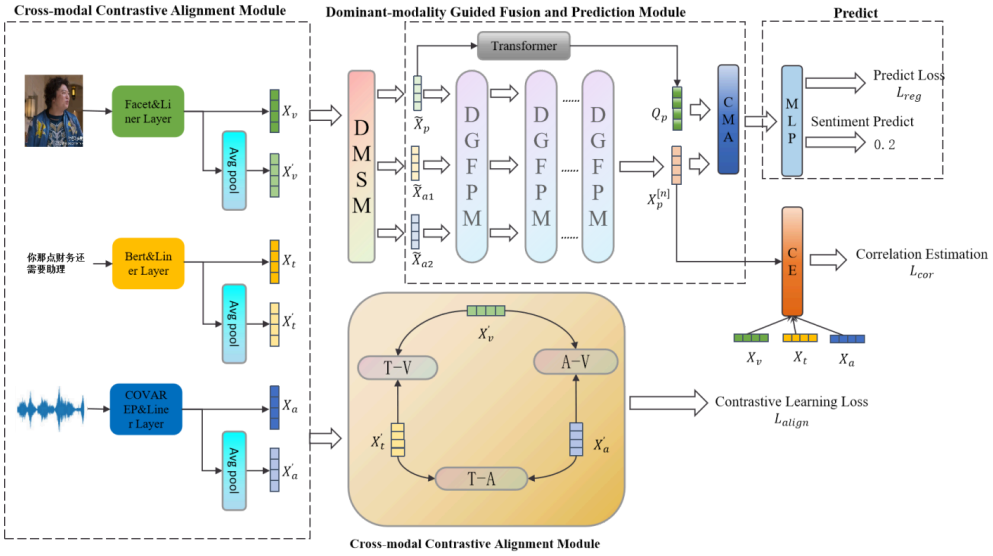

Figure 2. AMSGFN architecture

Specifically, the model first extracts corresponding single-modal feature representations from text, vision, and audio modalities; then, through CCAM, different modalities are mapped to a shared semantic space to alleviate the semantic offset problem caused by the heterogeneity of the modal spaces; on this basis, DMSM dynamically scores each modality and adaptively determines the dominant modality p of the current sample as well as the two auxiliary modalities a_1 and a_2 ; finally, under the guidance of the dominant modality, the fusion module selectively integrates the information of the auxiliary modalities to obtain the enhanced

representation $X_p^{[n]}$. Further, the model constructs the prediction query Q_p using the dominant modality and interacts with the final fusion representation through the attention readout mechanism to obtain the discriminative representation X_{pred}, and then inputs the Multi-Layer Perceptron (MLP) to obtain the final sentiment prediction result y. Next, each module will be introduced in detail.

2.2. Single-modal feature extraction

Multimodal Sentiment Analysis (MSA) aims to infer the speaker's sentiment. Typically, the input consists of three modalities: text t, vision v, and audio a. The original data of each modality is denoted as F_m, where $m \in \{t, v, a\}$.

For the text modality, we use the BERT language model [12] to extract features, and then project the dimensions through a linear layer to d, obtaining sequential representations $X_t \in R^{L_t \times d}$.

For non-text modalities F_m, where $m \in \{v, a\}$, we respectively use FACET [13] and COVAREP [14] to extract features from the original inputs F_v and F_a, and project them through linear layers to the same dimension as the text features, obtaining $X_v \in R^{L_t \times d}$, $X_a \in R^{L_t \times d}$.

2.3. Cross-modal contrastive alignment module

The cross-modal contrastive alignment module is used to map text, visual, and acoustic features to a shared semantic space, thereby enhancing the consistency and comparability between heterogeneous modal representations. Firstly, the sequence features $X_m \in R^{L_m \times d}, m \in \{t, v, a\}$ are globally averaged pooled to obtain the global representation vectors X'_m. Subsequently, the representations of each modal are projected onto the shared semantic space and cross-modal contrastive constraints are imposed in this space. Specifically, this paper aligns the three pairs of modalities: text-audio, text-visual, and audio-visual, constructing a triangular closed-loop alignment structure. By minimizing the formula L_{align}, it simultaneously constrains the three modal pairs: text-audio, text-visual, and audio-visual. This symmetric loss function forces the three encoders to generate representations with semantic consistency by pulling similar source samples closer and pushing distant negative samples apart within the batch.

Specifically, we use the NT-Xent loss [15] as the loss function for contrastive learning. For the i-th sample in a batch, the contrastive loss between modality m and modality n is defined as follows:

$$L_i^{(m,n)} = -log \frac{exp(\frac{sim(X_{m,i}, X_{n,i})}{\tau})}{\sum_{j=1}^{N} exp(\frac{sim(X_{m,i}, X_{n,j})}{\tau})} \tag{1}$$

Among them, $X_{m,i}$ and $X_{n,i}$ respectively represent the feature vectors of mode m and mode n of sample i in the projection space; $sim(\cdot, \cdot)$ represents the cosine similarity calculation function; τ is a temperature hyperparameter that controls the degree of distribution concentration [16]; N is the size of the current batch.

This module weights and sums the losses of the three modal pairs (text-audio, text-visual, audio-visual) along the bidirectional dimension to obtain the final alignment loss L_{align}, and the specific formula is as follows:

$$L_{align} = \sum_{(m,n) \in \{(t,a),(t,v),(a,v)\}} \frac{1}{N} \sum_{i=1}^{N} (L_i^{(m,n)} + L_i^{(n,m)}) \tag{2}$$

2.4. Dominant modality selection module

The dominant mode selection module (see Figure 3) is used to model the relative contributions of each mode in the current sample and adaptively determine the dominant mode and the auxiliary mode. Unlike the fixed

main mode setting, this module does not pre-specify a certain mode to always be dominant. Instead, it dynamically assesses the importance of each mode based on the multi-modal representation of the input sample.

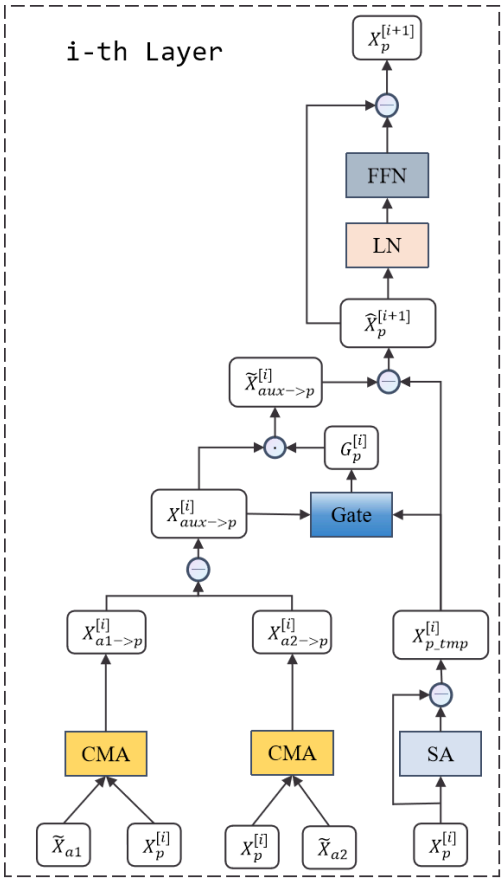

Figure 3. DGGFM module structure

Firstly, the previously obtained global representation vector X'_m is concatenated and input into a mapping network based on a Multi-Layer Perceptron (MLP) to learn the sample-level modal significance representation. The normalized contribution degree of each modality is obtained through the Softmax function:

$$s = Softmax(MLP([X'_t; X'_a; X'_v])) \tag{3}$$

Among them, $s = [s_t, s_a, s_v]$ correspond to the contribution degrees of text, audio and visual modalities respectively, and satisfy $\sum_{m\in\{t,a,v\}} s_m = 1$.

After obtaining the modal contribution degree, the model dynamically determines the dominant mode of the current sample based on the maximum contribution.

$$p = arg(s_k) \tag{4}$$

Among them, p represents the dominant mode of the current sample, while the other two modes are regarded as auxiliary modes.

To further enhance the role of the dominant mode and suppress the interference caused by the low-contribution modes, the corresponding contribution degrees are used to weight the features of each mode, resulting in the weighted mode representation:

$$\widetilde{X}_m = s_m X_m, \quad m \in \{t, a, v\} \tag{5}$$

Among them, $\widetilde{X}_m$ represents the sequence feature after being corrected by modal contribution.

2.5. Dominant-modality guided fusion and prediction module

2.5.1. Dominant mode guided fusion

After obtaining the dominant mode and the auxiliary mode, we proposed Dominant Modal Guided Fusion (DMGF). Specifically, we integrated the supplementary emotional cues in the auxiliary mode around the dominant mode, and through a gating mechanism, we adaptively controlled the injection intensity of the auxiliary information into the dominant mode, thereby reducing redundant information and noise interference and enhancing the discriminative ability of the fused representation.

Specifically, as shown in Figure 3, let the dominant mode representation of the input at the i -th layer be $X_p^{[i]}$, and the two auxiliary mode representations be $\widetilde{X_{a_1}}$ and $\widetilde{X_{a_2}}$. First, using the dominant mode as the query, perform cross-modal attention operations on the two auxiliary modes respectively, thereby modeling the information transmission process from the auxiliary modes to the dominant mode, and obtaining two cross-modal interaction results:

$$X_{a_1 \to p}^{[i]} = CMA(X_p^{[i]}, \widetilde{X_{a_1}}, \widetilde{X_{a_1}}) \tag{6}$$

$$X_{a_2 \to p}^{[i]} = CMA(X_p^{[i]}, \widetilde{X_{a_2}}, \widetilde{X_{a_2}}) \tag{7}$$

Among them, $CMA(\cdot)$ represents cross-modal attention operation, which is used to extract supplementary information that is most relevant to the current dominant modality from the auxiliary modality.

Meanwhile, in order to further enhance the context modeling ability of the dominant mode itself, a self-attention operation is applied to the dominant mode, resulting in an enhanced internal representation of the dominant mode:

$$X_{p_{tmp}}^{[i]} = SA(X_p^{[i]}) + X_p^{[i]} \tag{8}$$

Among them, $SA(\cdot)$ represents the self-attention operation, which is used to capture the temporal dependencies and semantic associations within the dominant mode.

After obtaining the interaction results of the two auxiliary modalities to the dominant modality, it is first aggregated into a unified representation of auxiliary information:

$$X_{aux \to p}^{[i]} = X_{a_1 \to p}^{[i]} + X_{a_2 \to p}^{[i]} \tag{9}$$

To prevent auxiliary information from being indiscriminately injected into the dominant mode, this paper further introduces a gating mechanism. Based on the matching relationship between the current dominant mode representation and the auxiliary information representation, an information control vector is adaptively generated:

$$G_p^{[i]} = \sigma(W_g[X_{p_{tmp}}^{[i]}; X_{aux \to p}^{[i]}] + b_g) \tag{10}$$

Among them, $[\cdot;\cdot]$ represents feature concatenation, W_g and b_g are learnable parameters, and $\sigma(\cdot)$ is the Sigmoid activation function. The gated vector $G_p^{[i]}$ is used to measure the effectiveness of the auxiliary information for the current dominant mode, and to perform element-wise filtering of the auxiliary information:

$$X_{aux \to p}^{[i]} = G_p^{[i]} \odot X_{aux \to p}^{[i]} \tag{11}$$

Among them, $\odot$ represents element-wise multiplication.

Subsequently, the dominant modal self-enhanced representation and the gated auxiliary information are fused to obtain the intermediate representation of the current layer:

$$\widehat{X}_p^{[i+1]} = X_{p_{tmp}}^{[i]} + X_{aux \to p}^{[i]} \tag{12}$$

Based on this, further updates are made to the fused representation through layer normalization and feedforward neural networks, resulting in the dominant mode output of the $(i+1)th$ layer:

$$X_p^{[i+1]} = FFN(LN(\widehat{X}_p^{[i+1]})) + \widehat{X}_p^{[i+1]} \tag{13}$$

Among them, $LN(\cdot)$ represents layer normalization, and $FFN(\cdot)$ represents feedforward neural network.

2.5.2. Model predict

After N layers of dominant mode guided fusion, the model obtains the final fused representation $X_p^{[N]}$. Meanwhile, the mode selection module outputs the weighted dominant mode representation X_p. To further enhance the guiding ability of the dominant mode in the final discrimination stage, this paper introduces a Transformer encoding block into X_p for deep modeling to obtain a more discriminative query representation. Specifically, this process can be expressed as:

$$Q_p = TB(\widetilde{X_p}) \tag{14}$$

Among them, $TB(\cdot)$ represents a Transformer encoding block, which consists of self-attention mechanism, residual connection, layer normalization, and feedforward neural network. Through this process, the model can further explore the context dependencies within the dominant mode, thereby constructing a more stable and having higher-level semantic information query representation Q_p.

Subsequently, the final fused representation of $X_p^{[N]}$ is used as the key and value, and the dominant mode query representation Q_p is used as the query. An attention readout is performed, extracting the discriminative information most relevant to the current dominant mode from the fused representation:

$$X_{pred} = CMA(Q_p, X_p^{[N]}, X_p^{[N]}) \tag{15}$$

Among them, $CMA(\cdot)$ represents cross-modal attention operation. Unlike the fusion stage in the previous part, the attention operation here is not for the feature interaction between different modalities, but for further screening and reconfiguration of the fused representation before prediction, so as to make the final predicted representation more focused on the key emotional cues represented by the dominant modality.

Finally, the output result X_{pred} is input into the multi-layer perceptron to obtain the final prediction result:

$$y = MLP(X_{pred}) \tag{16}$$

2.6. Model training objective

After the dominant mode-guided fusion and prediction as described in Section 2.5, the model obtains the final prediction result y. During the training phase, this paper comprehensively considers the task supervision objective, the consistency constraints between the fusion representation and the single-modal representation, as well as the cross-modal shared space alignment constraints, and jointly optimizes the entire model.

For regression tasks, when the true label $\hat{y}$ is given, the mean squared error is used as the task loss:

$$L_{reg} = MSE(y, \hat{y}) \tag{17}$$

To further enhance the semantic consistency between the final fused representation and each individual modality representation, this paper introduces a correlation constraint based on noise contrast estimation.

Specifically, the final fused representation $X_p^{[N]}$ is taken as the reference representation, and contrast estimations are conducted with the text, visual, and audio modality representations respectively, thereby constructing a correlation estimation loss:

$$L_{cor} = \sum_{m\in\{t,v,a\}} L_{NCE}(X_p^{[N]}, X_m) \tag{18}$$

Among them, $L_{NCE}(\cdot)$ represents the contrastive loss function based on noise contrast estimation, and X_m represents the individual modality feature representation. Through this constraint, it can encourage the fused representation to retain the cross-modal supplementary information while being consistent with the key semantic information in each modality, thereby enhancing the robustness of the model representation.

Furthermore, in the cross-modal contrastive alignment module of this paper, a shared space alignment loss L_{align} is introduced to alleviate the spatial heterogeneity issues between different modalities. Ultimately, the training objective of the model is composed of the task loss, the correlation estimation loss, and the cross-modal alignment loss:

$$L_{task} = L_{reg} + \alpha L_{align} + \beta L_{cor} \tag{19}$$

Among them, α and β are the weight coefficients of the cross-modal alignment loss and the correlation estimation loss.

3. Results

3.1. DataSets

The model was evaluated on commonly used multimodal sentiment analysis datasets, including CMU-MOSI [1], CMU-MOSEI [17], and CH-SIMS [18]. Each sample in these datasets contains visual, audio, and text modalities. The sentiment intensity annotations for MOSI and MOSEI range from -3 to +3, while for CH-SIMS it ranges from -1 to 1. Table 1 presents the statistical information of these datasets.

CMU-MOSI: This MSA dataset contains 2,199 multimodal samples, including visual, audio, and textual modalities. Specifically, the training set consists of 1284 samples, the validation set contains 229 samples, and the test set contains 686 samples. Each sample is labeled with its emotional intensity, ranging from -3 (indicating a strong negative emotion) to +3 (indicating a strong positive emotion).

CMU-MOSEI: This dataset contains 22,856 audio clips collected from YouTube, covering 250 different topics and 1,000 different speakers. The dataset is divided into 16,316 training samples, 1,871 validation samples, and 4,659 test samples. Each sample has been precisely labeled, with its sentiment score ranging from -3 to +3, representing the degree of sentiment from the most negative to the most positive.

CH-SIMS: This Chinese multimodal sentiment analysis dataset contains several multimodal samples, each of which includes text, visual, and audio modalities. The sentiment intensity annotation for each sample ranges from -1 to +1, where -1 represents negative sentiment and +1 represents positive sentiment. This dataset is suitable for multimodal sentiment analysis research in Chinese social media scenarios and can support models in making sentiment predictions based on the combined effects of different modal features.

Table 1. Introduction to datasets

DataSets	Train	Valid	Test	Total
CMU-MOSI	1,284	229	686	2,199
CMU-MOSEI	16,326	1,871	4,659	22,856
CH-SIMS	1,368	456	457	2,281

3.2. BaseLine

The AMSGFN model was compared with several baseline methods in the field of multimodal sentiment analysis (MSA), including TFN4], LMF [5], MulT [21], MISA [22], MMIM [23], Self-MM [2], DMD [24], PriSA [20], DTN [26], MIM [25], DLF [27], DIB [28], and DashFusion [19].

3.3. Quantitative analysis

From Table 2 to Table 4, the experimental results demonstrate that the proposed AMSGFN achieves overall state-of-the-art performance on the CMU-MOSI, CMU-MOSEI, and CH-SIMS datasets, validating the effectiveness and robustness of the framework in multimodal sentiment analysis. On the CMU-MOSI dataset, AMSGFN attains binary classification accuracy (Acc-2) and F1 scores of 84.35 and 84.21, respectively, outperforming the best baseline by 0.79 and 0.39 percentage points; the seven-class accuracy (Acc-7) reaches 48.02, a gain of 0.52 percentage points; for regression metrics, MAE and Corr are 0.699 and 0.803, with MAE reduced by 0.013 and Corr increased by 0.003. On CMU-MOSEI, AMSGFN also achieves the best results, with Acc-2 and F1 improving by 0.76 and 0.43 percentage points, Acc-7 reaching 54.32 (a gain of 0.37 percentage points), MAE decreasing to 0.528, and Corr increasing to 0.801, representing improvements of 0.001 and 0.011 over the strongest baselines. On the CH-SIMS Chinese dataset, AMSGFN shows even more significant gains: Acc-2 and F1 improve by 1.93 and 2.17 percentage points, Acc-5 reaches 43.15 (a gain of 1.03 percentage points), and regression metrics MAE and Corr are 0.409 and 0.605, outperforming the best baselines by 0.006 and 0.012, respectively. Overall, AMSGFN achieves consistent improvements in binary classification, fine-grained classification, and continuous sentiment regression tasks, indicating that its cross-modal contrastive alignment and dominant modality-guided fusion mechanism effectively mitigate semantic discrepancies among heterogeneous modalities, enhance fine-grained sentiment modeling capabilities, and maintain strong generalization and robustness across different languages and complex scenarios.

Table 2. The performance comparison of AMSGFN with the baseline model on the CMU-MOSIdataset

Model	Acc-2	F1	Acc-7	MAE	Cor
TFN	-/80.80	-/80.70	34.90	0.901	0.698
LMF	-/82.50	-/82.40	33.20	0.917	0.695
Mult	-/83.00	-/80.00	40.00	0.871	0.698
MISA	81.80/83.40	81.70/83.60	42.30	0.783	0.761
MMIM	83.32/85.79	83.23/85.75	46.65	0.712	0.796
Self-MM	82.54/84.77	82.68/84.91	45.79	0.712	0.795
PriSa	83.23/85.42	83.10/85.35	47.10	0.718	0.784
DMD	-/84.50	-/84.40	41.40	-	-
DTN	-/85.10	-/85.10	47.50	0.716	0.790

Table 2. Continued

MIM	-/84.80	-/84.80	46.40	0.718	0.792
DashFusion	83.56/85.26	83.82/85.22	45.25	0.712	0.786
DLF	-/85.06	-/85.04	47.08	0.731	0.781
DIB	-/85.60	-/85.60	47.40	0.715	0.800
Our	84.35/86.27	84.21/86.21	48.02	0.699	0.803

Table 3. The performance comparison of AMSGFN with the baseline model on the CMU-MOSEIdataset

Model	Acc-2	F1	Acc-7	MAE	Cor
TFN	78.50/81.89	78.96/81.74	51.60	0.573	0.714
LMF	80.54/83.48	80.94/83.36	51.59	0.576	0.717
Mult	81.15/84.63	81.56/84.52	52.84	0.559	0.733
MISA	83.60/85.50	83.80/85.30	52.20	0.555	0.756
MMIM	82.24/85.63	82.41/85.61	53.64	0.537	0.764
Self-MM	82.68/84.96	82.95/84.93	53.46	0.529	0.767
PriSa	82.10/85.20	83.30/85.20	53.95	0.536	0.761
DMD	-/85.20	-/85.20	53.10	-	-
DTN	-/85.50	-/85.50	52.30	0.572	0.765
MIM	-/85.70	-/85.60	51.80	0.779	0.579
DashFusion	82.15/85.93	82.17/85.91	52.46	0.532	0.782
DLF	-/85.42	-/85.27	53.90	0.536	0.764
DIB	-/86.00	-/86.00	53.50	0.588	0.790
Our	84.36/86.52	84.63/86.34	54.32	0.528	0.801

Table 4. The performance comparison of AMSGFN with the baseline model on the CH-SIMSdataset

Model	Acc-2	Acc-5	F1	MAE	Cor
TFN	78.38	39.30	78.62	0.432	0.591
LMF	77.77	40.53	77.88	0.441	0.575
Mult	76.54	-	76.59	0.447	0.563
MISA	74.44	-	71.75	0.492	0.399
MMIM	77.64	41.76	77.85	0.428	0.59
Self-MM	78.53	-	78.37	0.415	0.583
PriSa	78.38	39.30	78.62	0.432	0.591
DashFusion	78.03	42.12	78.25	0.421	0.593
Our	80.46	43.15	80.54	0.409	0.605

3.4. Ablation study

3.4.1. Effects of different modalities

From Table 5, it can be seen that different modal combinations have a significant impact on the model performance. Overall, the results of single-modal approach are significantly inferior to those of multi-modal approach, indicating that relying solely on a single information source is insufficient to fully capture the complex cues in emotional expression. Among them, the performance of the text modal is significantly better than that of the visual and audio modal, suggesting that in the emotion recognition task, text remains the most discriminative core modal; in contrast, when the visual and audio modal are used alone, their performance is lower, indicating that the emotional information contained in these two modalities is more susceptible to noise interference, and their discriminative ability is relatively limited. However, this does not mean they are ineffective; rather, it indicates that they are more suitable to be used as supplementary information and to be jointly modeled with text.

Table 5. The influence of different modalities in the CMU-MOSEI dataset

Model	Acc-2	F1	MAE	Cor
T	80.21/82.54	80.74/82.53	0.653	0.753
V	57.36/57.62	57.32/57.62	0.992	0.145
A	60.25/60.72	60.35/60.62	0.965	0.293
V+A	60.45/61.63	60.59/61.51	0.924	0.235
T+V	81.55/83.56	81.68/83.42	0.599	0.775
T+A	81.68/83.83	81.83/83.69	0.587	0.781
Our	84.36/86.52	84.63/86.34	0.528	0.801

Further examination of the results of the dual-modal combination reveals that when visual or audio information is introduced on top of the textual modal, the model performance improves. This indicates that the auxiliary modal can provide additional emotional cues for the text, thereby enhancing the model's ability to judge emotional states. Compared to using only the textual modal, when the visual modal is added, Acc-2 and F1 increase by 1.47 and 0.94 percentage points respectively, suggesting that visual information can supplement the speaker's facial expressions and movement features to some extent. When the audio modal is added, Acc-2 and F1 increase by 1.47 and 1.09 percentage points respectively, and the correlation coefficient also increases by 0.028, indicating that the intonation, rhythm, and intensity changes in the audio have a more direct help in modeling the emotional trend. In other words, although the visual and audio modalities have limited discriminative power on their own, when combined with the text, they can demonstrate significant complementary value.

Based on this, the complete model achieved even better results. Compared with the best dual-modal combination, the complete model further increased by 2.68 percentage points on Acc-2, by 2.80 percentage points on F1, decreased by 0.059 on MAE, and increased by 0.020 on Corr. This indicates that the tri-modal joint modeling does not simply add more information, but after effectively coordinating the relationships between different modalities, it can significantly enhance the discriminative ability and robustness of the final representation. In particular, the significant decrease in MAE suggests that the deviation between the model's predicted values and the true sentiment labels is smaller; while the improvement in Corr indicates that the model can more accurately depict the continuous trend of sentiment changes.

3.4.2. Effects of different components

From Table 6, it can be seen that the complete AMSGFN achieved the best results on the CMU-MOSEI dataset, indicating that each core component is necessary and complementary to the improvement of the model performance. Overall, when any module is removed or the dominant modality is fixed to a single modality, the model performance will decline to varying degrees, which indicates that the advantage of the method proposed in this paper does not come from an isolated design, but from the collaborative effect of modules such as cross-modal alignment, dynamic dominant modality selection, guided fusion, and consistency constraints.

Table 6. The impact of different components on the performance of the CMU-MOSEI dataset isshown, with the best results displayed in bold

Model	Acc-2	MAE	Cor
w/o CCAM	83.19/84.91	0.584	0.776
t-oriented	83.51/85.79	0.552	0.776
v-oriented	82.37/84.43	0.563	0.753
a-oriented	82.51/84.67	0.567	0.741
w/o DMGF	80.12/80.50	0.595	0.722
w/o TB	83.46/85.92	0.541	0.791
Our	84.36/86.52	0.528	0.801

Firstly, after removing the cross-modal alignment module CCAM, the Acc-2 decreased by 1.17/1.61 percentage points, MAE increased by 0.056, and Corr decreased by 0.025. This indicates that if the alignment constraints in the shared semantic space are absent, the semantic offsets between different modalities will be more difficult to eliminate, thereby weakening the consistency of the fused representation. In other words, the role of CCAM is not only to "put the features together", but more importantly, to enhance the comparability between heterogeneous modalities, enabling text, vision, and audio to interact around a unified semantic, thus playing a fundamental role in improving the overall performance.

Secondly, based on the results of different dominant modalities settings, the performance is significantly better when text is the dominant modality compared to when vision or audio is the dominant modality. Moreover, the complete model outperforms the fixed text-dominant scheme even further. Specifically, compared to t-oriented, the complete model has increased Acc-2 by 0.85/0.73 percentage points, further reduced MAE by 0.024, and increased Corr by 0.025; while compared to v-oriented and a-oriented, the improvement of the complete model is more significant, with Corr increasing by 0.048 and 0.060 respectively. This indicates that the text modality is indeed the most stable and discriminative source of information for sentiment discrimination, but the dominant information in different samples is not completely fixed. The dynamic dominant modality selection mechanism can adapt to sample differences more flexibly and thus achieve better results than static setting of the dominant modality. This phenomenon also validates the design motivation of this paper that "dynamic selection is superior to fixed setting".

Again, the performance of the model deteriorated most significantly after removing DMGF. The accuracy (Acc-2) decreased by 4.24/6.02 percentage points, the mean absolute error (MAE) increased by 0.067, and the correlation (Corr) decreased by 0.079. This indicates that the dominant mode-guided gated fusion module is an extremely crucial part of the entire model. Its function lies not only in integrating multimodal information, but also in selectively absorbing and filtering the auxiliary modes based on the dominant mode, thereby suppressing the ineffective injection of noise and conflicting information. After removing this module, the model has difficulty effectively coordinating the relationship between the main and auxiliary modes, resulting

in significant impairments in overall discrimination ability and continuous emotion modeling ability. This also indicates from a certain perspective that AMSGFN can indeed control the information flow more precisely during the fusion stage.

Furthermore, after removing the Transformer encoding block TB, the model performance also declined. The accuracy (Acc-2) decreased by 0.90/0.60 percentage points, the Mean Absolute Error (MAE) increased by 0.013, and the Correlation decreased by 0.010. This indicates that in the final discrimination stage, it is necessary to further conduct deep contextual modeling on the dominant mode. TB can help the model extract more discriminative query representations from the dominant mode, thereby more accurately focusing on the key emotional cues during the prediction reading process. Therefore, it still has a stable boosting effect on the final performance.

Finally, after removing the correlation constraint Cor, the model performance also declined. The accuracy of Acc-2 decreased by 1.15/1.39 percentage points, and Corr decreased by 0.015. Although the improvement brought by this module is not as significant as that of DMGF, it can enhance the semantic consistency between the final fused representation and each single-modal representation, enabling the fusion result to retain the complementary information across modalities while not deviating too much from the key emotional semantics in the original modality. Therefore, it has a positive effect on the model's robustness and prediction stability.

4. Conclusion

This paper presents a new model named AMSGFN for multimodal sentiment analysis. By introducing a cross-modal contrastive alignment mechanism, this model maps text, visual, and audio features into a shared semantic space, effectively alleviating the semantic offset problem among heterogeneous modalities. Additionally, by combining a dynamic dominant-modal selection module, it can adaptively identify the dominant modalities based on the discriminative and reliable nature of each modality in different samples, thereby reducing the fusion bias caused by fixed dominant-modal settings. On this basis, this paper further designs a dominant-modal-guided gated fusion module, which selectively integrates auxiliary modal information centered around the dominant modalities, highlighting key emotional cues while suppressing redundant information and noise interference. Through the Transformer encoding block and cross-modal attention readout mechanism, a more robust final representation is obtained. Experimental results based on benchmark datasets such as CMU-MOSI and CMU-MOSEI demonstrate that AMSGFN achieves superior performance in multiple evaluation metrics, verifying the effectiveness of the proposed model in alleviating modality heterogeneity, enhancing sample adaptive modeling capabilities, and improving the robustness of multimodal sentiment prediction.

References

[1] Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2016). MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv preprint arXiv: 1606.06259.

[2] Yu, W., Xu, H., Meng, F., & Wu, P. (2021). Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(12), 10782–10790.

[3] Wen, Y., Li, Q., Zhou, Y., & Zhang, Y. (2025). DashFusion: Dual-stream alignment with hierarchical bottleneck fusion for multimodal sentiment analysis. IEEE Transactions on Neural Networks and Learning Systems. Advance online publication.

[4] Zadeh, A., Liang, P. P., Mazumder, N., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. arXiv preprint arXiv: 1707.07250.
[5] Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., & Morency, L. P. (2018). Efficient low-rank multimodal fusion with modality-specific factors. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 6137–6146.
[6] Tsai, Y. H. H., Liang, P. P., Zadeh, A., & Morency, L. P. (2018). Learning factorized multimodal representations. arXiv preprint arXiv: 1806.06176.
[7] Han, W., Chen, H., & Poria, S. (2021). Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. Proceedings of the 2021 International Conference on Multimodal Interaction, 221–229.
[8] Wang, D., Zhang, X., & Li, X. (2022). Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia, 25*, 4909–4921.
[9] Wang, D., Li, X., & Zhang, X. (2023). TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognition, 136*, 109259.
[10] Zhang, H., & Wang, H. (2023). Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. arXiv preprint arXiv: 2310.05804.
[11] Cao, Z., Xu, Q., Yang, Z., & Liu, Y. (2022). OTKGE: Multi-modal knowledge graph embeddings via optimal transport. *Advances in Neural Information Processing Systems, 35*, 39090–39102.
[12] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171–4186.
[13] Baltrušaitis, T., Robinson, P., & Morency, L. P. (2016). OpenFace: An open source facial behavior analysis toolkit. 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), 1–10.
[14] Degottex, G., Kane, J., Drugman, T., Raitio, T., & Scherer, S. (2014). COVAREP—A collaborative voice analysis repository for speech technologies. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 960–964.
[15] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. International Conference on Machine Learning, 1597–1607.
[16] Wu, Z., Xiong, Y., Yu, S. X., & Lin, D. (2018). Unsupervised feature learning via non-parametric instance discrimination. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3733–3742.
[17] Zadeh, A., Zellers, R., Pincus, E., & Morency, L. P. (2016). Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems, 31*(6), 82–88.
[18] Yu, W., Xu, H., Meng, F., & Wu, P. (2020). CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 3718–3727.
[19] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv: 1412.6980.
[20] Ma, F., Zhang, Y., & Sun, X. (2023). Multimodal sentiment analysis with preferential fusion and distance-aware contrastive learning. 2023 IEEE International Conference on Multimedia and Expo (ICME), 1367–1372.
[21] Tsai, Y. H. H., Bai, S., Liang, P. P., Zadeh, A., & Morency, L. P. (2019). Multimodal transformer for unaligned multimodal language sequences. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 6558–6569.
[22] Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: Modality-invariant and-specific representations for multimodal sentiment analysis. Proceedings of the 28th ACM International Conference on Multimedia, 1122–1131.
[23] Han, W., Chen, H., & Poria, S. (2021). Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 9180–9192.

[24] Li, Y., Wang, Y., & Cui, Z. (2023). Decoupled multimodal distilling for emotion recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6631–6640.
[25] Zeng, Y., Mai, S., Yan, W., & Li, X. (2023). Multimodal reaction: Information modulation for cross-modal representation learning. *IEEE Transactions on Multimedia, 26*, 2178–2191.
[26] Zeng, Y., Yan, W., Mai, S., & Li, X. (2024). Disentanglement translation network for multimodal sentiment analysis. *Information Fusion, 102*, 102031.
[27] Wang, P., Zhou, Q., Wu, Y., & Li, Z. (2025). DLF: Disentangled-language-focused multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence, 39*(20), 21180–21188.
[28] Huang, H., Gong, T., He, K., & Li, Y. (2025). Robust multimodal sentiment analysis via double information bottleneck. *Information Fusion*, 103964.