MITG-CU: Multimodal Interaction Temporal Graphs Approach for Conversational Emotion Recognition

Qian Xing, Yaqin Qiu, Minglu Chi, Xuewei Li, Changyi Gao
School of Intelligent Engineering, Henan Institute of Technology, Xinxiang, Henan 453003, China

***Abstract*—In the emotion recognition of conversations, the complementary relationship between the context information and multimodal data cannot be fully exploited. This results in insufficient comprehensiveness and accuracy in emotion recognition. To address these challenges, this paper proposed a Multimodal Interactive Temporal Graph Conversation Understanding model (MITG-CU) composed of textual, audio and visual modalities. Firstly, the pre-extracted textual, audio, and visual features are adopted as the input of the Transformer, and the attention mechanism is utilized to capture the cross-modal context correlation information. Furthermore, structural relationships and temporal dependencies between utterances are captured through a local-level relational temporal graph module. Inter-modal interaction weights are dynamically adjusted by a global-level pairwise cross-modal interaction mechanism. By integrating two complementary hierarchical structures, a hierarchical multimodal information fusion was achieved, and at the same time, the model's adaptability to complex conversation scenarios was enhanced. Finally, feature fusion is carried out by using the gating mechanism and sentiment classification is conducted. Experimental results demonstrated that the proposed model outperforms six common baseline methods across metrics including accuracy, precision, recall, and F1-score. Especially in Weighted-F1 and Accuracy have improved by 0.28 % and 0.39 % respectively, which confirmed the effectiveness of the model.**

***Keywords*—Emotion recognition; multimodal interaction; relational temporal graph; cross-modal interaction; feature fusion**

I. Introduction

Emotional expression is multifaceted, extending beyond verbal language to encompass variations in vocal tone, facial expressions, and body movements. Most existing emotion recognition models are based on single modalities, such as textual [1], vision [2] and audio [3], but human emotions are often elusive. Therefore, a single text modality may not be able to correctly identify emotions in certain scenarios. For instance, the emotions directly expressed in text are neutral, but the corresponding facial expressions actually represent another emotion, such as anger.

As multimodal technologies are increasingly approaching real-world application scenarios, multimodal emotion recognition [4], [5], [6] has become a research hotspot. By analyzing and integrating information from different sensory channels, it aims to more comprehensively and accurately capture and interpret human emotions. The development of this technology can not only enhances the naturalness and efficiency of human-computer interaction, but also plays an important role in multiple fields such as mental health monitoring, education, and entertainment [7], [8]. Due to their limited representational capacity and insufficient information coverage, single modal methods exhibit application limitations. In contrast, multimodal methods significantly enhance the recognition accuracy and robustness of the system through a systematic feature complementation mechanism. Consequently, multimodal methods better align with the intrinsic characteristics of human multi-channel emotional expression. Unlike traditional emotion recognition tasks in single-modal and non-conversational scenarios, conversational emotion recognition involves complex multimodal relationships and contextual information, thereby posing more significant challenges.

In response to the limitations of existing methods in the task of conversational emotion recognition, where it is difficult to simultaneously utilize context information and the complementary relationship among various modalities. This paper proposes a Multimodal Interaction Temporal Graph Conversation Understanding Model (MITG-CU), which consists of two core modules: The Relational Temporal Graph Module (RTGM) and The Cross-modal Interaction Module (CIM). Contextual features at dual levels can be simultaneously extracted and effectively utilized by this model, thereby enabling the precise recognition of conversational emotions. At the local level, through the relationship temporal graph convolution network module, a time series graph structure is constructed to effectively capture the discourse temporal characteristics and local context information during multimodal interaction, further enhancing the understanding of Conversation evolution. At the global level, through the cross-modal feature interaction mechanism, the global context information is fully integrated to enhance the accuracy and robustness of discourse-level emotion recognition.

The rest of the paper is structured as follows: Section II is dedicated to related works, Section III presents our methodology, Section IV exposes the experimental results, The discussion is provided in Section V. Section VI introduced the conclusion for the overall work and talk about future directions.

II. Related Works

Multimodal emotion recognition primarily identifies emotional states by integrating multiple sensory modalities such as facial expressions, speech, and physiological signals, thereby enhancing the accuracy and robustness of emotion detection across diverse scenarios. The core issue of multimodal conversational emotion recognition primarily revolve around two aspects: achieving effective interaction and fusion of information across different modalities, and systematically modeling the global and local dynamics of session emotion.

Currently, the primary research approaches for multimodal emotion recognition in conversations include: those based on

contextual information [9], [10], [11] and those based on the complementary relationship among different modalities [12], [13], [14], [15]. Among context-based approaches, models such as LSTM, RNN, and GCN are widely adopted.

The study in [9] proposed a multimodal emotion recognition model based on fuzzy guidance and context self-attention network. By leveraging contextual self-attention mechanisms to capture dynamic interactions between different modalities and using fuzzy logic to handle uncertainty in multimodal data. However, High computational complexity, strong dependence on data quality, and limited generalization ability and other issues increase the training cost and affect the overall performance of the model. The study in [10] developed a contextual relationship modeling method that combines high-order relationships among multiple nodes. Complex dynamic interactions between different modalities in conversations can be effectively captured by this method. Nevertheless, constructing high-order relationships among multiple nodes requires complex prior knowledge or heuristic rules. If the structure is improperly designed, it will affect the performance of the model. The emotional features in conversations are highlighted through sequential analysis of conversational context and employment of a memory mechanism in [16], However, limitations have been identified in retaining contextual details from long-range dialogue sequences. While the study in [17] employed graph convolutional networks to model speaker interactions and conversation context for emotion recognition. It simply concatenated the multimodal information ,without deeply considering the potential interaction among different modalities.

The main research methods based on the complementary relationship among various modalities are: The research [12] proposed a facial emotion recognition method based on deep convolutional neural networks. Yet, this method failed to take into account the dynamic changes of facial expressions. and temporal dependencies in the time dimension were also unaddressed. The method presented in [15] adaptively selects nodes and edges, thereby enhancing both intra-modality and inter-modality interactions. Due to the presence of complex dependencies and heterogeneity gaps, this approach struggles to accurately capture and fuse multimodal information. The research in [18] employed a multi-channel and multimodal sentiment analysis method based on knowledge graph. The limitation of this method lies in its inability to effectively process cross-modal affective information, particularly under conditions with significant inter-modal distribution discrepancies and noise contamination. The research in [19] developed a dynamic weighted gated mechanism, which enhances cross modal interaction for multimodal sentiment. The framework's effectiveness is constrained by the requirement for precise parameter tuning in its dynamic gating mechanisms.

While the studies discussed above have made significant contributions to multimodal emotion recognition in conversation, they also exhibit certain limitations. Failing to simultaneously account for global information, local information, and the complementary relationships among different modalities, this may lead to misunderstandings of the speaker's true emotions or intentions. Therefore, our proposed MITG-CU model is capable of extracting and leveraging context features at two levels, thereby achieving precise emotion recognition. By constructing RTGM and CIM, unique information and complementarity of modalities such as textual, audio and visual are integrated. This achieves a comprehensive understanding of the conversation evolution and the full integration of global context information. In order to verify the accuracy and robustness of emotion recognition in conversations, our model was compared with six commonly used baseline methods. The results showed that our model outperforms six common baseline methods. Weighted-F1 and Accuracy were respectively improved by 0.28 % and 0.39 %. This demonstrates the good stability and generalization ability of our model.

III. Proposed Model

The proposed MITG-CU model for this research consists of four components: feature extraction and encoding, RTGM, CIM, and multimodal fusion and prediction. The overall framework of the network is shown in Fig. 1.

Firstly, the pre-extracted textual, acoustic and visual features are employed as inputs of the Transformer model. Meanwhile, its self-attention mechanism is leveraged to capture the cross-modal contextual correlation information. Subsequently, the extracted features are fed into the relational temporal graph attention network layer and the cross-modal interaction layer, where context modeling and cross-modal interaction are respectively carried out. Finally, the gated mechanism is utilized to fuse the features at the feature level, and then classified based on the emotions.

A. Feature Extraction and Encoding

Before performing modal information enhancement, sBERT, OpenSmile and OpenFace are used to extract Text, Audio and Visual features respectively, and the extracted features are input into the Transformer encoding layer. After the processing of the encoding layer, three modal feature representations are finally generated, as shown in Eq. (1)-Eq. (3).

$$T \in R^{L_T \times D_T} \tag{1}$$

$$A \in R^{L_A \times D_A} \tag{2}$$

$$V \in R^{L_V \times D_V} \tag{3}$$

Where, *T* represents textual modal information, *A* denotes audio modal information, and *V* indicates visual modal information. and L_T, L_A and L_V respectively denote the lengths of the data sequences, while D_T, D_A, and D_V are the dimensions of the extracted features for each Modality.

B. Relational Graph Attention Network

In the graph attention network framework, each node is associated with a feature vector. X=$\{h_1,h_2,...,h_n\}$ is the set of feature vectors, where h_i is the feature vector of node *i*. These feature vectors are composed not only of the original attributes of the nodes, but also integrated with high-order structural information extracted from the graph's topological structure.

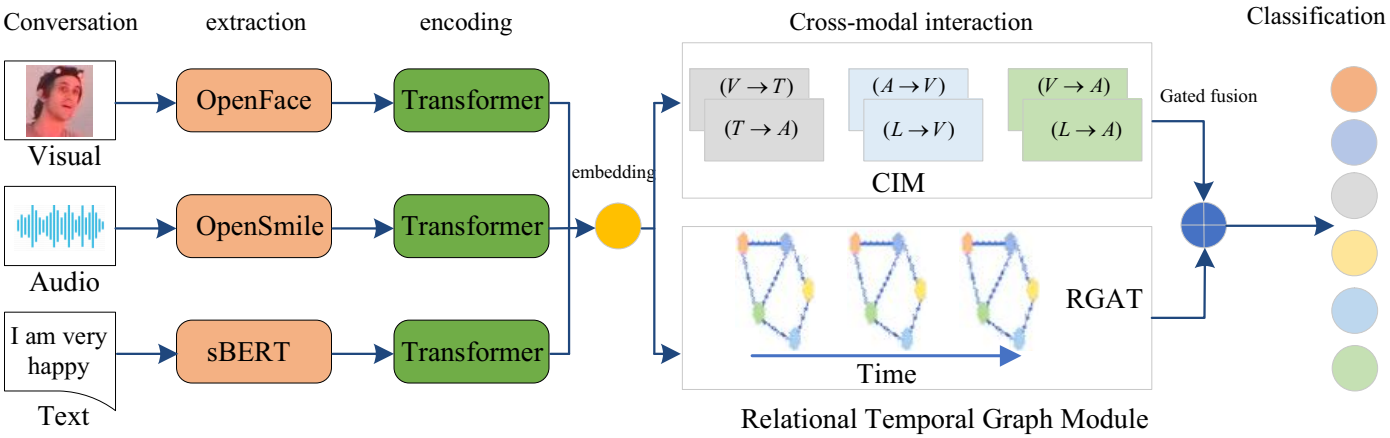


Fig. 1. Multimodal interaction temporal graph conversation understanding model.

The implementation steps of Relational Graph Attention Network (RGAT) are as follows:

Step 1: Relation Type Mapping Matrix

To distinguish different edge types or connection patterns, a relation-type mapping matrix $R=\{R_1,R_2,...,R_m\}$ is introduced. Each node i is associated with a feature vector $h_i \in R^m$, and all node feature vectors collectively form the input feature matrix $H \in R^{N\times m}$ of the graph, where N denotes the number of nodes.

Before calculating the graph attention coefficients, the model first maps the original node features into an F' dimensional feature vector $Z_i \in R^{F'}$ through a linear transformation. The weight matrix $W \in R^{m\times F'}$ is shared among all nodes. The transformed node feature is expressed as $Z_i = Wh_i$.

Step 2: Relation-aware Attention

After establishing node features and relation-type representations, further computed using the self-attention mechanism. Attention weights $a^r_{i,j}$ between each pair of nodes *(i, j)* and relation type *r* as Eq. (4) and Eq. (5).

$$e^r_{i,j} = \text{Leaky Re}\,LU\left(a^T\left[Z_i \| Z_j\right]\right) \tag{4}$$

$$a^r_{i,j} = \frac{\exp\left(e_{ik}\right)}{\sum_{k\in N(i)}\exp\left(e_{ik}\right)} \tag{5}$$

Where, a^T represents the attention vector for learning, Z_i and Z_j respectively denote the transformed feature vectors of nodes i and j, and $N(i)$ represent the neighbor node sets of node i.

Step 3: Relation-specific Aggregation

The attention weights are obtained through the relationship-aware attention mechanism. These weights are weighted and aggregated with the neighborhood information. The node feature vector H_i is calculated as shown in Eq. (6).

$$H_i = \sigma\sum_{j\in N(i)} a^r_{i,j} W h_j \tag{6}$$

Where, h_j represents the feature vector of node j, and σ denotes a non-linear activation function such as *ReLU* or *LeakyReLU*. This process not only enhances the non-linear expression ability of the model, but also enables the model to more effectively capture the complex relationships between nodes.

Step 4: Multi-head Attention Mechanism

To further enhance the representational capacity of the model, a multi-head attention mechanism is introduced. This mechanism allows the model to learn information in parallel within different representation subspaces, and there are multiple independent attention heads for each relation type. By parallel computing the outputs of multiple attention heads and concatenating or averaging the results, the final representation of the node is obtained. The aggregation process of the multi-head attention mechanism is shown in Fig. 2.

Then, the final node representation h'_i is shown in Eq. (7).

$$h'_i = \text{concat}\left(\sigma\left(\sum_{j\in N(i)} a^{(k)}_{i,j} W^{(k)} h_j\right)\right) \tag{7}$$

Where, *concat* represents the concatenation operation, $a^k_{i,j}$ is the representation of the *k-th* head, and $W^{(k)}$ refers to the linear transformation weight matrix of the *k-th* head.

Step 5: Output Layer

After the feature extraction and update by the multi-head attention network, the updated node features are input into the output layer. Then, normalization is performed to generate a probability distribution of emotional categories, thereby achieving the classification and prediction of emotional states.

These steps collectively constitute the core computing process of RGAT module. The RGAT module can effectively incorporate relational type information, dynamically capture complex relationships, temporal dynamics, and multimodal information in conversations, thereby enhancing the performance of emotion recognition in conversations.

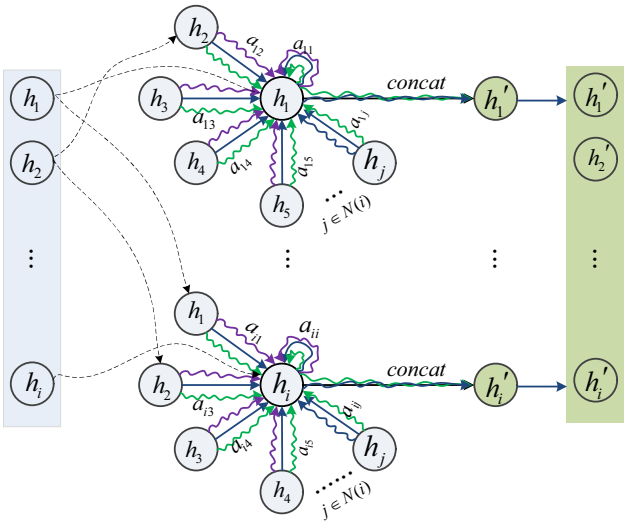


Fig. 2. The Aggregation process of the multi-head attention mechanism.

C. The Cross-Modal Interaction Module

In multimodal sentiment analysis tasks, the cross-modal heterogeneity inherent in human language often amplifies the complexity of linguistic interpretation. Utilizing cross-modal interaction is helpful in revealing the "unaligned" nature and long-term dependency of cross-modal phenomena. Hence, the pairwise cross-modal feature interaction module (P-CM) [20] is introduced into our proposed framework.

The P-CM module captures the subtle emotional correlations between textual and auditory modalities through the interaction of low-level features, thereby providing enriched semantic and affective information for multimodal sentiment analysis tasks. The cross-modal Transformer network based on P-CM is presented in Fig. 3.

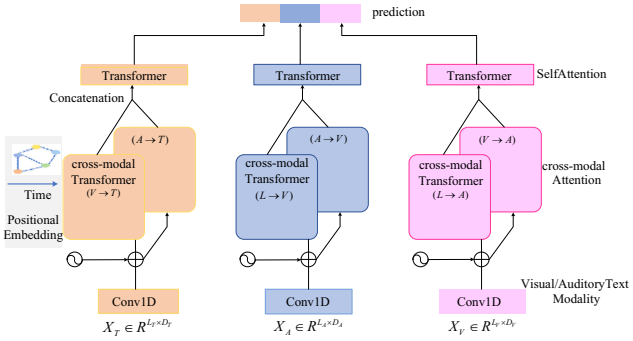


Fig. 3. The Cross-modal interaction module.

As illustrated in Fig. 3, $X_T = R^{L_T \times D_T}$, $X_A = R^{L_A \times D_A}$ and $X_V = R^{L_V \times D_V}$ respectively denote the modality-sensitive representations of the textual, audio and visual modes when using a single modality encoder.

Based on the Transformer architecture, taking audio and text as examples. The defined Query , Keyword and Value are respectively $Q_A = X_A W_{Q^A}$, $K_T = X_T W_{K^T}$ and $V_T = X_T W_{V^T}$. When modality *T* performs cross-modal attention on modality *A*, the enriched representation $CM^{T \to A}$ of modal *A* is computed as in Eq. (8).

$$CM^{T \to A} = \text{softmax}\left(\frac{X_A W_{Q^A} (W_{k^T}) X_T}{\sqrt{d_k}} X_T W_{v^T}\right) \quad (8)$$

Where, $W_{Q^A} \in R^{d_A \times d_Q}$, $W_{K^T} \in R^{d_T \times d_K}$ and $W_{V^T} \in R^{d_T \times d_V}$ are learning parameters, while d_Q , d_K and d_V represent the dimensions of Query, Key and Value respectively, $\sqrt{d_K}$ is the scaling factor.

Assume $Z_A^{[i]}$ is the modality-sensitive global context representation of modality *A* across the entire conversation at the *i-th* layer, and $Z_A^{[0]} = X_A$. When modality *T* performs cross-modal attention on modality *A*, information from modality *T* is transferred to modality *A*, and the enriched representation $Z_{T \to A}^{[i]}$ of modality *A* is computed as the following procedure Eq. (9), and $Z_{T \to A}^{[0]} = Z_A^{[0]}$.

$$\begin{aligned} \bar{Z}_{T \to A}^{[i]} &= CM_{T \to A}^{[i]}\left(LN\left(Z_{T \to A}^{[i-1]}\right), LN\left(Z_{T \to A}^{0}\right)\right) \\ &+ LN\left(Z_{T \to A}^{[i-1]}\right) \\ Z_{T \to A}^{[i]} &= \\ &\max\left(0, LN\left(\bar{Z}_{T \to A}^{[i]}\right) \times w_1 + b_1\right) w_2 + b_2 + LN\left(\bar{Z}_{T \to A}^{[i]}\right) \end{aligned} \quad (9)$$

Where, w_1 and w_2 represent the weight matrices, b_1 and b_2 denote the biases, *LN* indicates the layer normalization calculation, *max(0,0)* is the *ReLU* activation function. Through layer normalization calculation, the activation value distribution of each layer is adjusted to make the model training more stable and improve performance.

According to Eq. (9), we can easily compute the cross-modal representation $Z_{T \to A}^{[i]}$. Meanwhile, the outputs of the cross-modal Transformer (D layers) that share the same target modality are concatenated to form the final cross-modal global context representation $Z_{T \to A}^{[D]}$. Similarly, for other modal pairs $Z_{V \to T}^{[D]}$ and $Z_{V \to A}^{[D]}$, the same process can be applied to obtain them, and ultimately, the representations of the three modalities in the global cross-modal context can be obtained.

P-CM effectively reduces the complexity of joint multimodal modeling by simultaneously modeling the relationship between the two modalities. The framework precisely extracts local paired modal information. Additionally, it constructs a global cross-modal dependency relationship. This integration significantly enhances the model's adaptability and robustness across diverse scenarios.

D. Multimodal Fusion and Prediction

1) Gated multimodal fusion: The Gated Multimodal Fusion (GMF) layer is a key component of the multimodal interactive temporal graph conversation understanding model, responsible for integrating features from different modalities. This layer aims to selectively emphasize the relevant information in each modality and suppress the irrelevant or redundant information, thereby enhancing the model's ability to accurately identify emotions. The GMF layer consists of two main parts: the transformation mechanism and the gating mechanism.

The transformation mechanism is to convert the concatenated features through another linear layer, thereby extracting higher-level feature representations. The transformed features are denoted as H', as shown in Eq. (10).

$$H' = W_t \cdot [h'_1, h'_2, \ldots, h'_n] + b_t \tag{10}$$

Where, W_t is the weight matrix and b_t is the bias term.

The gating mechanism is mainly implemented through linear layer and *sigmoid* activation function to calculate the importance of each modality, thereby weighting the features of different modalities. For each modality *i*, its weight value g_i is calculated by the sigmoid activation function Eq. (11).

$$g_i = \text{sigmoid}\left(W_t \cdot [h'_1, h'_2, \ldots, h'_n] + b_t\right) \tag{11}$$

2) Emotion classification layer: The emotion classification layer is mainly responsible for precisely mapping the fused feature vectors to the corresponding emotion categories. It consists of two main fully connected layers: *lin1* and *lin2*. The *lin1* layer maps the input features to the hidden layer space, while the *lin2* layer further maps the features obtained from *lin1* to the prediction space of emotion categories. To prevent overfitting of the model, a Dropout layer is introduced between *lin1* and *lin2*. Subsequently, the argmax function is used to determine the emotion category with the highest probability, thereby serving as the final emotion prediction result for the discourse. The predicted emotion category of the discourse is shown in Eq. (12).

$$\hat{c}_j = \arg\left(\text{softmax}\left(g_k z_j + b_k\right)\right) \tag{12}$$

Where, C_j represents the predicted emotion category of the *j-th* utterance. g_k and b_k respectively denote the weight and bias parameters of the fully connected layer in the second layer. After training, we are able to extract the feature representation of the *j-th* utterance, which is denoted as Z_j.

Finally, the performance of the model is evaluated by calculating the difference between the probability distribution predicted by the model and the true labels. The negative log-likelihood loss (*NLLLoss*) is used as the loss function. The input of *NLLLoss* is the log-probability vector and the target label. Parameter optimization is achieved by minimizing the difference between the predicted probability and the true distribution. For each sample *i*, the calculation of *NLLLoss* is expressed as Eq. (13).

$$\text{NLLLoss}(x, y) = -\sum_{i=1}^{N} x_i \left[y_i\right] \tag{13}$$

Where, *N* is the number of samples, y_i is the true label of the sample, and x_i is the logarithmic probability vector of the *i-th* sample.

IV. Model Training and Evaluation

A. Data Det

This paper utilizes the Interactive Emotional Dyadic Motion Capture (IEMOCAP) sentiment analysis dataset. IEMOCAP is a dataset commonly employed for multimodal sentiment recognition, collected by the Speech Analysis and Interpretation Laboratory (SAIL) of the University of Southern California (USC). This dataset encompasses approximately 12 hours of multimodal data, including visual, audio, facial motion capture, and text transcription. IEMOCAP is divided into two parts. Participants engage in either impromptu performances or performances based on specific scripts in these parts. These scripts are specially designed to elicit emotional expressions.

B. Experimental Configuration and Evaluation Indicators

All experimental code was developed using Python 3.11 and the PyTorch 2.0 deep learning framework, and the models are trained and tested on a single NVIDIA RTX 3060 8G GPU. After extensive experimentation, the optimal performance was achieved with a batch size of 10 and the learning rate of 0.00025. The Adam algorithm is selected as the optimizer. To reduce overfitting, the dropout rate is set to 0.5.

Use Accuracy *(Acc.)*, Macro-Average Precision *(P)*, Macro-Average Recall *(R)*, and Weighted F1 score (W_F1) as the evaluation metrics for model performance, as computed in Eq. (14) and (15).

$$p = \frac{1}{k} \sum_{i=1}^{k} \frac{TP_{C_i}}{TP_{C_i} + FP_{C_i}} \tag{14}$$

$$F_1 = 2 \times \frac{P \times R}{P + R} \tag{15}$$

$$\text{Acc.} = \sum_{i=1}^{T} I\left(y_p = y_t\right) \tag{16}$$

$$W\_F1 = -\sum_{i=1}^{N} f_{reqk} \times F1_i \tag{17}$$

Where, *TP*, *FP*, and *FN* respectively stand for true positive, false positive, and false negative. *k* is the number of categories, and C_i denotes the *i-th* category, Furthermore, y_p, y_t and *T* represent the predicted label, the true label, and the total number of samples in the test set. f_{reqk} is the relative frequency of class *k*, and $F1_i$ is the F1 score of the *i-th* category.

C. Controlled Experiment

1) Baseline: In order to verify the performance of MITG-CU model proposed in this paper, the following models are selected for experimental comparison: 1) DialogueRNN [16]: It is an attention-based recurrent neural network for emotion detection in conversations, which enhances emotion classification accuracy by tracking participants' states and contextual information. 2) COGMEN [21]: The multimodal sentiment

recognition method based on graph neural networks, which predicts the emotional state of the speaker in a conversation by integrating local and global contextual information. 3) CoMPM [22]: It is a model that combines the pre-trained memory of the speaker and context modeling. By tracking the importance of the speaker's historical utterances and their distance from the current utterance, it enhances the focus on the current utterance, thereby captures the emotional information in the conversation. 4) MM-DFN [23]: The dynamic fusion module of the graph structure is utilized to fuse multimodal context features, thereby reducing redundancy and enhancing the complementarity among modalities. 5) GraphCFC [24]: The method of cross-modal feature complementation based on directed graphs is employed in this paper. By utilizing multiple subspace extractors and paired cross-modal complementary strategy, the heterogeneity gap in multimodal fusion can be alleviated. 6) SACL-LSTM [25]: Based on the supervised adversarial contrastive learning framework, robust emotional features are generated by combining adversarial training and contrastive learning.

2) Experiment comparing with the baseline method: The comparative results of different models are presented in Fig. 4. The results indicate that the proposed model exhibits performance advantages over existing approaches, particularly in the key metrics of Weighted-F1 and Accuracy.

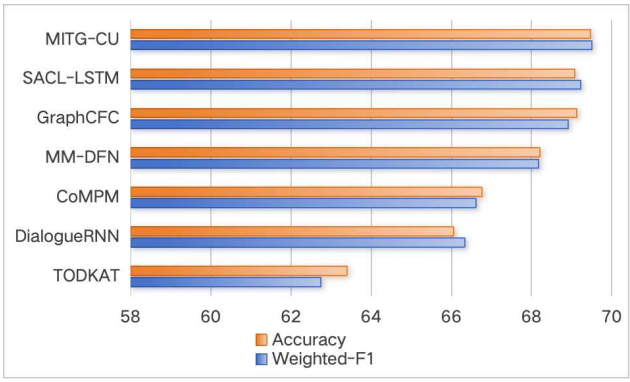


Fig. 4. Comparative analysis of models.

The comparative experimental results of different methods on the IEMOCAP dataset as presented in Table I. As shown, the proposed model achieves a Weighted-F1 score of 69.50, surpassing the second-best SACL-LSTM model by 0.28%. Similarly, it demonstrates superior Accuracy with a value of 69.47, outperforming SACL-LSTM by 0.39%.

TABLE I. COMPARATIVE EXPERIMENTAL RESULTS

Model	Weighted-F1	Accuracy
DialogueRNN	62.75	63.40
CoMPM	66.33	66.05
COGMEN	67.27	67.04
MM-DFN	68.18	68.21
GraphCFC	68.91	69.13
SACL-LSTM	69.22	69.08
MITG-CU	69.50	69.47

Fig. 5 provides comparison chart of F1-score and Weighted Average performance for different emotions. As can be seen in Fig. 5, the model achieves higher precision and recall in recognizing Emotion sad, indicating that the emotion of sad is the best among all emotions, while Emotion frustration exhibits relatively lower performance. Moreover, the F1-Score values for most emotions are clustered between 0.60 and 0.75, suggesting that the proposed model performs relatively balanced in most emotions and does not show extreme high or low values.

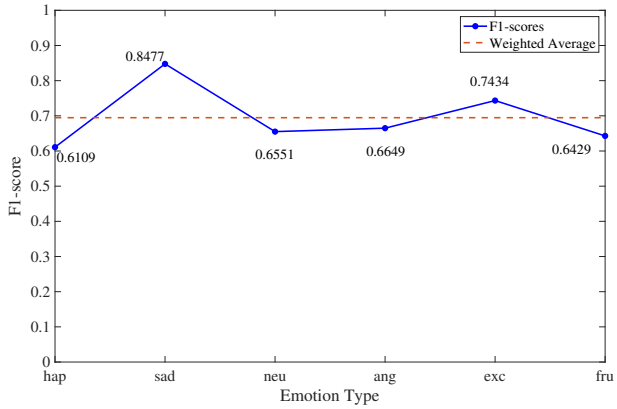


Fig. 5. Comparative chart of F1-scores and Weighted Average performance across different emotions.

The recognition results of different emotions on the IEMOCAP dataset are reported in Table II. As shown in Table II, the Macro-average F1-score is 0.6941, and the Weighted-average F1-score is 0.6947. The two scores are relatively close, indicating that the model has good stability and generalization ability across different emotion categories.

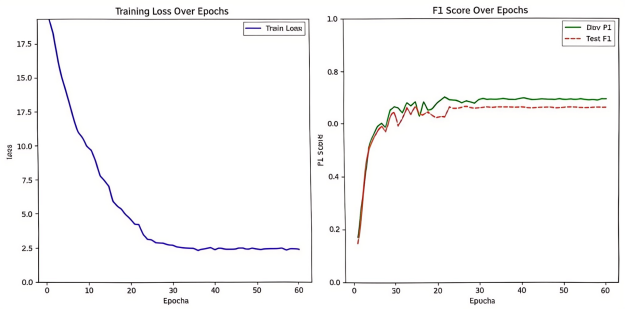
TABLE II. THE RECOGNITION RESULTS OF DIFFERENT EMOTIONS ON THE IEMOCAP DATASET

Model	Precision	Recall	F1-score
hap	0.5689	0.6597	0.6109
sad	0.8127	0.8857	0.8477
neu	0.6703	0.6406	0.6551
ang	0.6029	0.7412	0.6649
exc	0.7895	0.7023	0.7434
fru	0.6744	0.6142	0.6429
Macro-average	0.6864	0.7073	0.6941
Weighted-average	0.6986	0.6950	0.6947

As can be seen in Fig. 6, it can be observed that with the increase in the number of training epochs, the training loss decreases rapidly. This demonstrates that the model is constantly optimized during the learning process, and the fitting effect on the training data gradually enhances. In the early stage of training, the F1-score exhibits rapid growth, demonstrating the model maintains high accuracy while simultaneously improving recall, with its overall performance progressively enhancing throughout the learning process. After approximately 20 training epochs, the loss value plateaus, suggesting that the model has reached a good convergence state.

To comprehensively evaluate the performance of classification models, we analyzed the correspondence between the model's predicted outcomes and the true labels through comparison. As a result, the confusion matrix, which reflects the model's classification capability, is presented in Fig. 7.

As illustrated in Fig. 7, a strong positive correlation is observed between the F1 score and accuracy rate (0.97), which means that when F1 score increases, accuracy rate also tends



(a) Training loss (b) F1 score

Fig. 6. Training loss and F1 score variation curves.

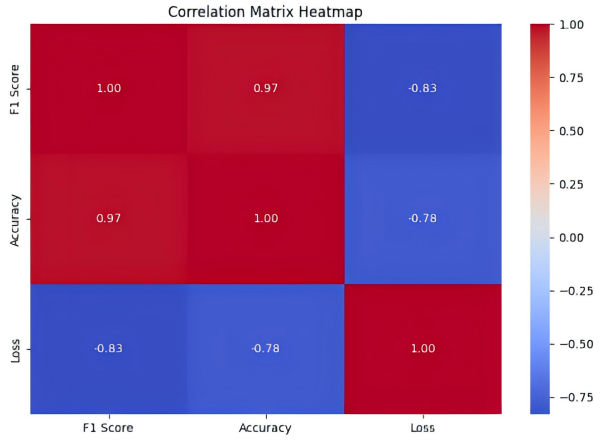


Fig. 7. Confusion matrix of evaluation indicators.

to increase. There is a strong negative correlation between F1 score and loss value (-0.83), indicating that as F1 score increases, loss value tends to decrease, which is what this model expects. This result shows that the model can improve the classification precision and recall while maintaining high overall prediction accuracy, demonstrating its balance and robustness in classification tasks.

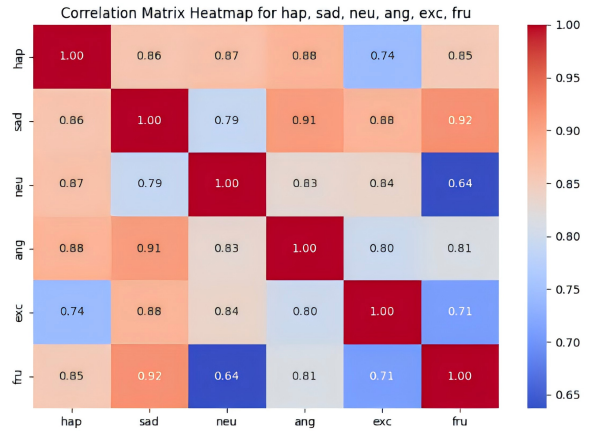


Fig. 8. Confusion matrix of various emotions.

The confusion matrix for various emotions as shown in Fig. 8. It can be observed from the confusion matrix that the correlation between emotional categories significantly impacts model performance. The high correlation (0.92) between sadness and frustration indicates significant confusion between them in the feature space, this suggests that their features are less distinguishable during emotion recognition. In contrast, the neutral emotions have higher feature distinguish ability and are less influenced by other emotions. Especially, its correlation with frustration is the lowest (0.64). The confusion matrices of various emotions not only provide a basis for the accuracy of emotion recognition, but also offer an important basis for the subsequent optimization of emotion models.

3) Ablation experiment: In order to further investigate the contributions of different components in the model, four ablation studies were designed on the IEMOCAP dataset in this paper. The base, no-CM, no-RGAT, and no-GAT respectively represent the ablation of the basic module, removing the cross-modal module from the model, removing the graph-time convolution module from the model, and removing the graph attention module from the model. The experimental results are shown in Table III.

TABLE III. THE ABLATION EXPERIMENTS CONDUCTED ON IEMOCAP

Module	**IEMOCAP**	
	*F*1	*Acc.*
base	69.50	69.47
no-CM	65.67	66.61
no-RGAT	66.26	66.67
no-GAT	65.89	66.42

As shown in Table III, our proposed MITG-CU method achieves the best performance on the two key metrics of F1 score and accuracy. Specifically, no-CM has the most significant impact on the model performance. Meanwhile, no-GAT also had a significant impact on the model performance. This indicates that the cross-modal module and the graph attention module are crucial for the model in capturing complex emotional features and improving classification performance.

The reproduced comparison results are reported in Table IV. The modalities include text (T), audio (A), and visual (V).

TABLE IV. THE PERFORMANCE OF MITG-CU IN DIFFERENT MODALITIES

Module	**IEMOCAP**	
	Acc.	*F*1
T	65.74	65.35
A	50.65	49.92
V	32.35	31.80
T+V	66.24	66.23
T+A	65.06	64.94
A+V	46.09	46.00
A+V+T	69.47	69.50

As illustrated in Table IV, when solely utilizing the textual modality (T), the model achieves the highest values in both accuracy (65.74%) and F1 score (65.35%). This result demonstrates that textual information plays a crucial role in the emotion recognition task. By comparison, the performance of the audio modality (A) is moderately effective, whereas the performance of the visual modality (V) is comparatively limited.

As demonstrated by the experimental results of bimodal combinations, the combination of textual modality and visual modality (T+V) achieved accuracy and F1-scores of 66.24%

and 66.23% respectively on the IEMOCAP dataset. Its performance is slightly better than that of the combination of textual modality and audio modality (T+A). The main reason for this is that some audio categories contain noise, which will affect the prediction results. The combination of audio modality and vision modality (A+V) exhibited the lowest performance.

In practice, the trimodal combination (T+A+V) achieves the best performance, with accuracy of 69.47% and F1-scores of 69.50%. This result indicates that multimodal fusion can significantly enhance the accuracy of emotion recognition.

V. Discussion

A. Comparison with Existing Results

The author in [26] proposed a multimodal fusion method based on deep graph convolutional networks. This method can effectively capture the context-dependent relationships in conversations through the construction of a graph structure. However, the fixed graph structure cannot adjust the weights of each modality in real time as the conversation progresses dynamically. according to the dynamic progress of the dialogue. This situation significantly reduced the accuracy of emotion change recognition, with an accuracy rate of only 65.56%. In [27] models the dynamic dependencies between speakers in conversations, thereby improving the accuracy of emotion prediction. This method simultaneously captures the internal dependencies within speakers, and the inter-speaker dependencies, thereby resolving the limitation of focusing solely on a single type of dependency. Nevertheless, this model is primarily designed for a unimodal setting (text), and lacks a mechanism for cross-modal fusion. As a result, its accuracy rate is limited to 66.29%.

In response to this situation, the proposed method can comprehensively integrate global information, local information, and the complementary relationships among different modalities. According to the contribution of each modality, the interaction weights between modalities are dynamically adjusted, thereby achieving information interaction and deep integration among modalities. Through ablation experiments, it can be found that either no CIM or no GAT will lead to a decrease in the F1 score and accuracy of the system to varying degrees. This indicates that the cross-modal module and the graph attention module are crucial for the model to capture complex emotional features and improve classification performance. Meanwhile, the impact of different modality combinations on model performance varies. As can be seen from Table IV, It can be concluded that textual modality plays a pivotal role in emotion recognition, while the integration of visual modality further enhances performance. Although the audio modality underperforms compared to textual and visual modalities when used in isolation, its combination with other modalities improves overall system performance. Moreover, the combination of textual, audio, and visual modalities (T+A+V) achieves optimal performance on the IEMOCAP dataset.

B. Limitations and Challenges of the Proposed Method

The cross-modal interaction mechanism incorporated in our model has proven effective. However, it cannot fully capture the optimal interaction mechanisms across all scenarios, due to the dependencies among complex sentences. Meanwhile, the multi-level interaction mechanisms increase the computational complexity, which may limit its application ability in real-time scenarios. Therefore, for future work, we will focus on further optimizing the cross-modal interaction mechanism and exploring lightweight methods to enhance its practical application value.

VI. Conclusion

The MITG-CU module proposed in this paper can precisely control the recognition of conversational emotions at both local and global levels. By effectively integrating the distinctive characteristics and complementary relationships among textual, audio, and visual modalities, and combining the relationship timeline module and cross-modal interaction mechanism, the proposed framework significantly enhances both the accuracy and robustness of emotion recognition.

The experimental results show that the MITG-CU model proposed in this paper outperforms six common baseline methods in terms of accuracy, precision, recall rate and F1 score. Especially, it has improved 0.28% and 0.39% in Weighted-F1 and Accuracy respectively. Especially when dealing with long-term conversations and complex emotional scenarios, our model demonstrates excellent stability and generalization capabilities. Future will focus on optimization of cross-modal interaction mechanisms to better meet emotion recognition needs in scenarios, including mental health monitoring and intelligent customer service. Furthermore, additional modalities will be introduced to further enhance the performance of the model.

Acknowledgment

This work was supported by the Science and technology research project of Henan Province, under Grant 24210222012, Key Research Projects of Higher Education Institutions in Henan Province, under Grant 25A460020 and 25B413008, Educational and Teaching Reform Research and Practice Project of Henan Institute of Technology, under Grant 2024JG-YB039 and 2024JG-YB040.

References

[1] N. Alswaidan and M. E. B. Menai, "A survey of state-of-the-art approaches for emotion recognition in text," *Knowledge and Information Systems*, vol. 62, pp. 2937–2987, 2020.

[2] Z. Y. Huang *et al.*, "A study on computer vision for facial emotion recognition," *Scientific Report*, vol. 13, p. 8425, 2023.

[3] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE Access*, vol. 9, pp. 47 795–47 814, 2021.

[4] T. Meng, F. Zhang, Y. Shou, H. Shao, W. Ai, and K. Li, "Masked graph learning with recurrent alignment for multimodal emotion recognition in conversation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 4298–4312, 2024.

[5] F. Chen, J. Shao, A. Zhu, D. Ouyang, X. Liu, and H. T. Shen, "Modeling hierarchical uncertainty for multimodal emotion recognition in conversation," *IEEE Transactions on Cybernetics*, vol. 54, no. 1, pp. 187–198, 2024.

[6] A. Geetha, T. Mala, D. Priyanka, and E. Uma, "Multimodal emotion recognition with deep learning: Advancements challenges and future directions," *Information Fusion*, vol. 105, p. 102218, 2024.

[7] H. B. Xu, X. Wu, and X. Liu, "A measurement method for mental health based on dynamic multimodal feature recognition," *Frontiers in Public Health*, vol. 10, p. 990235, 2022.

[8] S. Chougule, S. Bhosale, V. Borle, V. Chaugule, and Y. Mali, "Emotion recognition based personal entertainment robot using ML & IP," *International Journal of Scientific Research in Science and Technology*, vol. 5, no. 8, pp. 73–75, 2020.

[9] D. Z. Jiang, H. Liu, R. G. Wei, and G. Tu, "CSAT-FTCN: A fuzzy-oriented model with contextual self-attention network for multimodal emotion recognition," *Cognitive Computation*, vol. 15, pp. 1082–1091, 2023.

[10] N. Lu, Z. Han, and Z. Tan, "A hypergraph based contextual relationship modeling method for multimodal emotion recognition in conversation," *IEEE Transactions on Multimedia*, pp. 1–13, 2024.

[11] A. Khalane, R. Makwana, T. Shaikh, and A. Ullah, "Evaluating significant features in context-aware multimodal emotion recognition with XAI methods," *Expert Systems*, vol. 42, no. 1, p. 13403, 2023.

[12] S. M. S. Abdullah and A. A. Mohsin, "Facial expression recognition based on deep learning convolution neural network: A review," *Journal of Soft Computing and Data Mining*, vol. 2, no. 1, pp. 53–65, 2021.

[13] A. A. Gladys and V. Vetriselvi, "Survey on multimodal approaches to emotion recognition," *Neurocomputing*, vol. 556, p. 126722, 2023.

[14] T. Zhang, S. Li, B. Chen, H. Yuan, and C. L. P. Chen, "AIA-Net: Adaptive interactive attention network for text–audio emotion recognition," *IEEE Transactions on Cybernetics*, vol. 53, no. 12, pp. 7659–7671, 2023.

[15] G. Tu, T. Xie, B. Liang, H. Wang, and R. Xu, "Adaptive graph learning for multimodal conversational emotion detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 19 089–19 097.

[16] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "DialogueRNN: An attentive RNN for emotion detection in conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, 2019, pp. 6818–6825.

[17] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019, pp. 154–164.

[18] S. Dong, X. Fan, and X. Ma, "Multichannel multimodal emotion analysis of cross-modal feedback interactions based on knowledge graph," *Neural Processing Letters*, vol. 56, pp. 1–17, 2024.

[19] N. Wang and Q. Wang, "Dynamic weighted gating for enhanced cross-modal interaction in multimodal sentiment analysis," *ACM Transactions on Multimedia Computing, Communications*, vol. 21, no. 38, pp. 1–19, 2024.

[20] Y. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 6558–6569.

[21] A. Joshi, A. Bhat, A. Jain, A. V. Singh, and A. Modi, "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*, 2022, pp. 4148–4164.

[22] J. Lee and L. Wooin, "CoMPM: Context modeling with speaker's pre-trained memory tracking for emotion recognition in conversation," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologie*, 2021, pp. 5669–5679.

[23] D. Hu, X. Hou, L. Wei, L. Jiang, and Y. Mo, "MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2022, Virtual and Singapore*, 2022, pp. 7037–7041.

[24] J. Li, X. Wang, G. Lv, and Z. Zeng, "GraphCFC: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition," *IEEE Transactions on Multimedia*, vol. 26, pp. 77–89, 2024.

[25] D. Hu, Y. Bao, L. Wei, W. Zhou, and S. Hu, "Supervised adversarial contrastive learning for emotion recognition in conversations," arXiv preprint arXiv:2306.01505, 2023.

[26] J. Hu, Y. Liu, J. Zhao, and Q. Jin, "MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021, pp. 5666–5675.

[27] Y. Bao, Q. Ma, L. Wei, W. Zhou, and S. Hu, "Speaker-guided encoder-decoder framework for emotion recognition in conversation," in *Proceedings of the Thirty-First Inter-national Joint Conference on Artificial Intelligence, IJCAL-22*, 2022, pp. 4051–4057.