World Journal of Information Technology
Print ISSN: 2959-9903
Online ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3104

SENTIMENT ANALYSIS MODEL BASED ON ATTENTION MECHANISM MULTIMODAL FUSION AND SPATIO-TEMPORAL GRAPH NEURAL NETWORK

AiQun Zhu[1], Jin Lu[2*]
[1]*Shenzhen City Polytechnic, Shenzhen 518000, Guangdong, China.*
[2]*Shenzhen Polytechnic University, Shenzhen 518000, Guangdong, China.*
[*]*Corresponding Author: Jin Lu*

Abstract: As a key task in the field of artificial intelligence, the application of sentiment analysis has expanded from a single text modality to cover multi-modal information such as vision and hearing. However, the existing multi-modal sentiment analysis methods often ignore the time sequence and structure of the dynamic interaction between modalities, and fail to fully model the evolution of emotional States in the time dimension. Therefore, this study proposes a sentiment analysis model based on multi-modal fusion of attention mechanism and spatio-temporal neural network. The model first extracts the high-dimensional features of text, visual and auditory modalities through a dedicated encoder, and then designs a hierarchical attention fusion mechanism to adaptively weight the contributions of different modalities at the feature level and decision level. In order to capture the dynamic evolution of emotional expression, the spatio-temporal graph neural network module is introduced into the model, and the fusion features of each time step are regarded as graph nodes, and the dynamic edges are constructed based on the correlation and temporal continuity between modalities, so as to model the spatio-temporal dependence of cross-modal interaction. Experiments are conducted on three open multimodal emotion datasets, including CMU-MOSEI, IEMOCAP and MOSI. The results show that the proposed model significantly outperforms the baseline method in both sentiment classification and sentiment regression tasks. On the CMU-MOSEI data set, the accuracy of the model reaches 88. 7% in the binary classification and 53. 2% in the seven-class classification, both of which reach the current advanced level. The ablation experiment further verifies the effectiveness of the hierarchical attention mechanism and the spatiotemporal neural network module. This study provides a new framework for multimodal sentiment analysis that can simultaneously model the complex interaction and temporal dynamics between modalities, which has theoretical significance for understanding the complex mechanism of human emotional expression, and provides technical support for the development of more accurate emotional intelligence applications.
Keywords: Multimodal sentiment analysis; Attention mechanism; Feature fusion; Space-time graph neural network; Dynamic modeling; Temporal dependence

1 INTRODUCTION

With the global popularity of social media platforms, the increasing naturalization of human-computer interaction interfaces and the continuous expansion of the scale of online education, artificial intelligence systems that can deeply understand and properly respond to human emotions have become one of the key technologies to promote the development of related fields [1-2]. Traditional sentiment analysis techniques mainly rely on text content to infer emotional tendencies by analyzing vocabulary, syntax and semantics [3-5]. However, human emotional expression is a natural multi-channel process, which is transmitted and revealed simultaneously and often synchronously through multiple channels such as language content, micro-expressions and macro-expressions composed of facial muscle activity, body posture and gestures, and prosodic features of speech such as pitch, loudness and speed [6-7]. Only relying on a single modality, especially text for analysis, will not only lose a large number of key emotional information contained in non-verbal channels, but also lead to serious misjudgments in the face of emotional expressions such as irony, sarcasm or context-dependent, resulting in information loss and understanding bias [8]. Therefore, multimodal sentiment analysis, an interdisciplinary research field, emerges as the times require and develops rapidly. Its core goal is to integrate signals from text, vision, hearing and even more physiological modalities through computational models, and to achieve more comprehensive, more robust and closer to human perception of emotion recognition and understanding by using the complementary and synergistic effects between modalities [9-10].

Although multimodal sentiment analysis has made remarkable progress in the past few years, it still faces several deep and interrelated core challenges in its mature application [11]. First of all, the contribution of different modalities to the final affective state is not constant, but highly dynamic and context-sensitive. This dynamic is manifested at multiple levels, for example, in ironic discourse, the literal meaning of the text may be completely opposite to the emotion carried by the phonetic intonation, and the visual blink or smile may further modulate its true emotional intention [12]. Second, cross-modal interactions themselves exhibit complex dynamic patterns. For example, the peak of facial expression may be precisely synchronized with the occurrence of speech stress, or there may be a slight advance or lag, and this temporal alignment and misalignment is itself an important emotional cue [13]. However, most of the existing

fusion strategies, whether early feature splicing, decision voting, or methods based on simple weighting or bilinear products, tend to adopt static or shallow interaction modeling methods, which are difficult to capture such nonlinear and context-dependent dynamic cross-modal associations [14-16]. Finally, emotion itself is a psychological state with internal time continuity and evolution. During a conversation, a lecture, or a lesson, emotions accumulate, transfer, mix, or dissipate as the topic progresses, the interaction object feeds back, and the internal thinking process [17]. At present, many research methods divide the input signal into independent short segments for processing, largely ignoring the continuity, dependence and state transition law of emotion in time series, thus losing important temporal context information [18-19].

In order to systematically address these challenges, this study focuses on two key scientific issues. First, how to design a fusion mechanism that can adaptively and finely balance the importance of different modalities and their mutual influence, which needs to be able to cope with the imbalance of modal signal quality and the variability of context. Second, how to go beyond simple sequence modeling to effectively capture the structural evolution of multimodal emotional signals in the time dimension, that is, to model the dynamic dependencies between emotional representations at different time points and between different modalities. In recent years, progress in the field of deep learning has provided ideas to solve these problems. Attention mechanism, especially self-attention and cross-attention, has achieved great success in natural language processing and computer vision tasks, which proves its powerful ability in dynamic weight allocation and information screening, and provides core inspiration for the design of adaptive multimodal fusion. At the same time, because of its powerful ability to model relationships between entities in non-Euclidean data structures, graph neural network technology shows unique advantages in spatio-temporal data modeling tasks such as social network analysis and traffic forecasting, which provides a novel paradigm for treating multi-modal time-series data as dynamic evolution graphs and reasoning about relationships.

In view of this, this study innovatively constructs a new affective computing model that deeply integrates hierarchical attention mechanism and spatio-temporal neural network. The core innovation of this work is embodied in two tightly coupled levels. In terms of fusion strategy, we design a hierarchical attention architecture, which carries out multiple rounds of adaptive interaction and weight calibration between modalities at the local feature level and the global semantic level, respectively, to replace the simple single-layer fusion operation. In terms of dynamic modeling, we innovatively construct a dynamic graph structure for multi-modal time series data, in which nodes represent multi-modal features at different time steps, while edges encode inter-modal consistency relations across time and continuity relations between adjacent times, and then explicitly learn the complex interactions between nodes and their evolving patterns over time through spatio-temporal graph neural networks. This comprehensive framework aims to improve the performance, robustness and interpretability of multimodal sentiment analysis tasks through more sophisticated fusion strategies and more powerful structured dynamic relationship modeling, thus taking a more solid step in the journey of understanding human complex emotional expressions.

2 RELATED WORK

The research history of multimodal sentiment analysis profoundly reflects the evolution path of artificial intelligence from perception to cognition. Early work focused on simple fusion strategies, such as feature-level stitching and decision-level voting [20]. Although these methods are intuitive, they fail to deeply explore the complex nonlinear interaction between modalities, and often regard multimodal information as a simple superposition of independent channels. With the rise of deep learning technology, the research focus has shifted to joint representation learning [21]. Tensor fusion networks explicitly model interactions by computing outer products of modal features, but their computational complexity grows exponentially with the number of modes [22]. Low-rank multimodal fusion reduces the computational burden through tensor decomposition techniques, however its interaction modeling is still relatively static [23]. Memory fusion networks introduce temporal memory units to integrate multimodal information in the time dimension, but their memory mechanisms have limited ability to capture long-range dependencies [24].

In recent years, new technologies such as attention mechanism and graph neural network have injected new vitality into multimodal sentiment analysis. Attention-based fusion methods, such as multi-modal attention network, realize the dynamic screening of key information by calculating cross-modal attention weights, and partially solve the problem of fixed modal contribution [25]. However, many attention models are only integrated at a single level, which makes it difficult to balance local feature alignment and global semantic coordination [26]. At the same time, graph neural networks, with their powerful relational induction bias, are beginning to be used to model complex relationships between speakers in conversations or structures within modalities [27]. For example, some studies construct discourse graphs or modal interaction graphs, and use graph attention networks for information dissemination [28]. These methods have shown potential in capturing structured dependencies, but most of the works regard graph structure as static or constructed only based on prior knowledge, which fails to fully reflect the dynamic evolution of multimodal interaction itself with the process of dialogue or emotional flow. In addition, existing graph-based methods tend to focus on modeling social relationships between speakers or co-occurrence relationships between modalities, while paying insufficient attention to how emotional States evolve structurally through multimodal signals on the time axis [29].

Another important technical route is the introduction of pre-training large models. The multi-modal large model based on Transformer architecture learns a powerful cross-modal alignment representation through self-supervised pre-training on massive data. These models have achieved remarkable performance on downstream tasks such as sentiment classification [30]. However, large models usually have huge parameters and high computational cost, and

their "black box" characteristics make the model decision-making process difficult to explain. More importantly, the general large model is not designed for sentiment analysis tasks, and may not accurately capture the specific cross-modal interaction patterns and temporal dynamics on which subtle changes in sentiment depend [31].

Despite the continuous progress of technology, there are still some gaps and significant deficiencies in current research that need to be filled urgently. First of all, at the level of fusion mechanism, most of the existing methods fail to achieve dynamic adaptive fusion in a real sense. The importance of modality and its interaction mode should be adjusted in real time according to the specific context, emotional content and signal quality, but most of the current static or shallow dynamic fusion strategies are difficult to meet this requirement. Secondly, at the level of temporal dynamic modeling, the mainstream methods either treat the input as an independent fragment, which separates the emotional continuity, or use sequential models such as recurrent neural networks, which are vulnerable to the problem of gradient disappearance or explosion, have limited ability to model long-distance emotional dependence, and are difficult to capture irregular emotional transitions. Thirdly, at the level of fine-grained and structured representation of interaction, the existing research on the modeling of cross-modal interaction is still not fine enough. Many approaches only focus on the modal alignment at the global level, ignoring the semantic correspondence between local feature points, such as the instantaneous association of specific facial action units with specific prosodic features in speech. At the same time, there is a lack of research that treats multimodal time series data as a dynamically evolving graph system and explicitly models the time-varying association between nodes. Fourthly, at the task and data level, most of the existing studies focus on coarse-grained emotion classification, and lack the ability to identify fine-grained emotion dimensions or complex mixed emotions. In addition, the scarcity of high-quality, large-scale, multi-lingual and fine-labeled multimodal sentiment data sets seriously restricts the generalization ability and performance ceiling of the model. The subjective inconsistency of data annotation and the difference of cross-cultural emotional expression have not been properly solved.

To sum up, although the field of multimodal sentiment analysis has achieved fruitful results, it still faces fundamental challenges in achieving fine, dynamic, interpretable and robust emotion understanding. Specifically, how to design a hierarchical fusion mechanism that can adaptively adjust the modal weights according to the context, and how to construct a dynamic graph neural network framework that can explicitly model the complex spatiotemporal dependencies among multimodal signals to capture the continuous evolution and structured interaction of emotions at the same time, are the key problems that have not been fully solved by current research. Aiming at these core gaps, this study proposes a new model that integrates hierarchical attention and spatio-temporal map neural network, aiming to promote multimodal sentiment analysis to a deeper dynamic understanding.

3 METHODS

3.1 Model Architecture

In order to meet the aforementioned research challenges and achieve dynamic and accurate multimodal sentiment analysis, this study proposes a novel model architecture, as shown in Figure 1. The core design idea of this architecture is to realize the adaptive deep fusion of multi-modal signals through a hierarchical attention mechanism, and explicitly model the dynamic evolution and structural interaction of features after fusion with the help of spatio-temporal map neural network [32]. The whole model is trained and optimized in an end-to-end manner, aiming to directly learn the discriminative representation of emotion from multimodal temporal input. The workflow of the model can be systematically divided into four main components, namely, multi-modal feature extraction, hierarchical attention fusion module, spatio-temporal neural network module and the final prediction output layer. These components are connected in turn and work together to complete the complete calculation process from the original multimodal data to the emotion prediction. The following will first show the overall model design through the architecture diagram, and then detail the technical details of each component one by one.

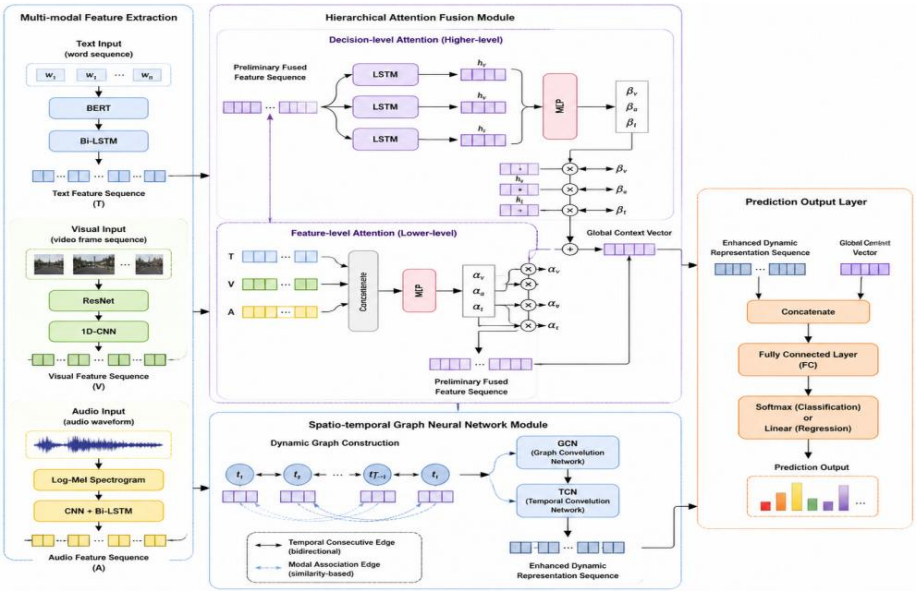

Figure 1 Multimodal Sentiment Analysis Model Architecture Based on Hierarchical Attention and Spatiotemporal Neural Network

As can be seen from the figure, the model processing flow is from left to right. Firstly, the original text, visual and auditory input are extracted by the pre-training model and the encoder network respectively, and three aligned feature sequences are obtained. These sequences are then fed into the hierarchical attention fusion module. The module first calculates the adaptive modal weights for each time step at the feature level to generate a preliminary fused feature sequence, and calculates another set of modal weights based on the context of the entire sequence at the decision level to generate a global context vector. At the same time, each time-step feature of the preliminary fusion feature sequence is defined as a node of the dynamic graph, and edges are constructed between the nodes according to the temporal adjacency relationship and the cross-temporal feature similarity to form the spatio-temporal graph. The graph is processed by a graph convolutional network and a time sequence convolutional network to capture the spatio-temporal dependency and output an enhanced dynamic representation sequence. Finally, the prediction output layer combines the enhanced dynamic representation sequence with the global context vector to generate the sentiment classification or regression results through the fully connected layer and the corresponding activation function. The whole architecture realizes a complete closed loop from feature extraction, adaptive fusion, dynamic modeling to final prediction.

3.2 Multi-Modal Feature Extraction

For the text modality, the pre-trained BERT model is used to obtain the context-dependent embedding of each term, and then further encoded by a bidirectional LSTM network to obtain the text feature sequence [33-34]. For the visual modality, the ResNet model pre-trained on the facial expression dataset is used to extract facial features from the video frames, and then a one-dimensional convolutional network is used to smooth the time dimension to obtain the visual feature sequence [35]. For the auditory modality, acoustic features such as log-Mel spectrogram are first extracted, and then input into an encoder composed of a convolutional neural network and a bidirectional LSTM to obtain an auditory feature sequence [36]. All feature sequences are unified to the same time step and feature dimension. Feature extraction is the basic step of multi-modal sentiment analysis, and its goal is to transform the original heterogeneous input data into a high-quality, aligned feature representation suitable for subsequent fusion and modeling. For the three core modalities of text, visual and auditory, this model uses the pre-trained model after domain adaptation combined with the specific network structure to construct feature extractors respectively, so as to fully capture the important information related to emotion in each modality. Figure 2 shows the detailed architecture and data flow of multimodal feature extraction.

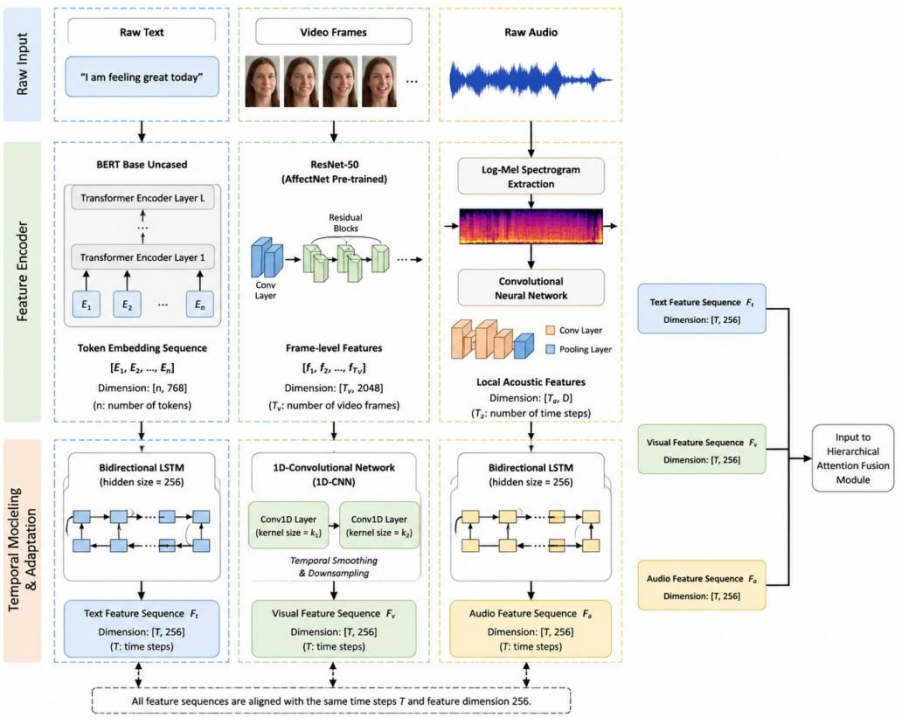

Figure 2 Multi-Modal Feature Extraction Architecture Diagram

For the input text sentence, we first use the BERT model to obtain its deep context-dependent representation. Specifically, after adding special classifiers and separators to the beginning and end of the input sequence, the word element sequence is input into the BERT base model. We take the last hidden state of BERT as the initial embedding vector of each word element. The embedding vector contains rich semantic and syntactic information. Considering the temporal dependence of emotional expression, we input the word vector sequence output by BERT into a two-way long and short-term memory network. Bi-LSTM processes the sequence from both forward and backward directions, and the result of the splicing of its final hidden state at each time step constitutes the final feature sequence of the text modality [37]. The sequence preserves both local context information and long distance dependencies.

For the input video clips, we first extract the video frames at a fixed sampling rate, and use the face detection alignment algorithm to crop out the face region in each frame [38]. Subsequently, the face images are fed into a ResNet model [39] pre-trained on a large facial emotion dataset. We remove its top classification layer and extract the feature vectors before the last fully connected layer as frame-level features. This feature captures key visual information related to facial expressions. As the high frame rate of video usually leads to high redundancy of adjacent frame features, we introduce a one-dimensional convolutional network to smooth and compress the frame-level feature sequence in the time dimension. The one-dimensional CNN is composed of two convolutional layers, and down-sampling is achieved by adjusting the step, so that the visual feature sequence matching the time step of the text feature is obtained.

For audio signals, we first perform pre-emphasis, framing, and windowing, and then calculate the log-Mel spectrum of each frame of audio [40]. Mel spectrum simulates the nonlinear auditory perception characteristics of the human ear, which is a common feature of speech emotion analysis. We treat the obtained spectrogram as a two-dimensional image and input it into a lightweight convolutional neural network. This CNN consists of multiple convolutional and pooling layers stacked to extract high-level, emotion-related acoustic local pattern features. The output of CNN is a feature sequence with a long time dimension [41]. To further model the long-term dependencies in speech, we input the feature sequences extracted by CNN into a bidirectional LSTM network. Bi-LSTM models the acoustic feature sequence in time series, and its output constitutes the final feature sequence of the auditory modality.

In order to ensure that the subsequent fusion module can work effectively, the feature sequences finally output by the above three processing flows need to be unified to the same number of time steps through time interpolation or down-sampling technology. In this model, the feature sequences of all modes are unified to the same time step, and the feature dimensions are unified to the same value through linear projection. Finally, we get three dense feature sequences aligned in time and feature dimensions, which represent the continuous and high-dimensional representations of text, visual and auditory modalities in emotional expression, respectively, and lay the foundation for subsequent deep integration and dynamic modeling.

3.3 Hierarchical Attention Fusion Module

The hierarchical attention fusion module is the core component of the model to realize the deep integration of multimodal information. Its design goal is to solve the problems of static modal weights and single interaction mode in traditional fusion methods, and to achieve dynamic and refined fusion of multi-modal information according to context

by applying attention mechanism at different abstract levels. The module consists of two core stages, feature-level attention fusion and decision-level attention fusion, which work cooperatively from the local time-step level and the global sequence level, respectively, to capture multi-granularity information from micro-feature complementarity to macro-semantic coordination. Figure 3 details the architecture and data flow of the module.

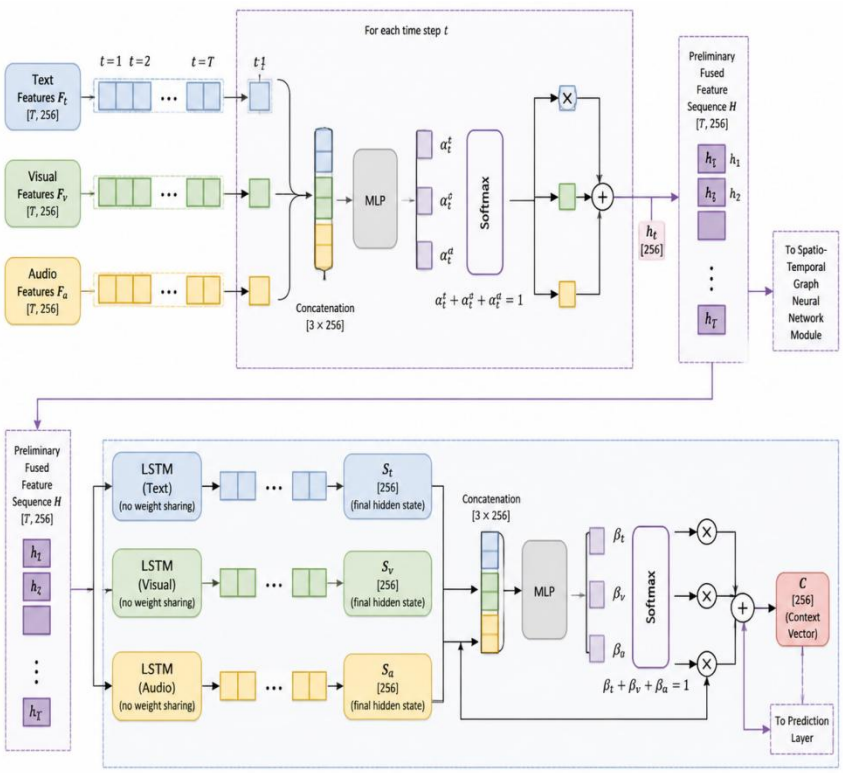

Figure 3 Hierarchical Attention Fusion Module Architecture Diagram

The first stage is feature-level attention fusion. The purpose of this phase is to dynamically evaluate and fuse the instantaneous features from different modalities at each time step. Assume that at the time step t, we have textual features $f\_t \wedge t$ Visual characteristics $f\_t \wedge v$. Auditory characteristics $f\_t \wedge a$. First, the three eigenvectors are concatenated to form a joint representation u_t. Then, through a learnable weight matrix $W1$ and offset $b1$, will u_t project to a shared space and calculate the unnormalized importance score for each modality. This process can be considered as a small neural network. Finally, the attention weight of each modality at the current time is obtained by normalizing the modality dimension with the sfotmax function [42]. These weights are used to perform a weighted summation of the original modal features, i.e., the preliminary fused features at the current time step are obtained h_t. This operation is repeated for all time steps in the sequence to obtain the entire preliminary fusion feature sequence H. This sequence preserves the temporal dynamics and achieves the adaptive fusion of modal information at each local time.

The second stage is the integration of strategic attention. The purpose of this phase is to re-fuse the modality-specific high-level representations that have been deeply encoded by LSTM from a global perspective of the entire sequence. We will initially fuse feature sequences H, input three independent LSTM networks respectively. Each LSTM independently processes the sequence and takes the hidden state of its final time step as the global characterization S of the path output. Here, each LSTM is expected to learn to reconstruct or emphasize the global context information of a particular modality from the mixed features. Subsequently, we concatenate these three global characterization vectors and pass through another learnable weight matrix $W2$ and offset $b2$ is the transformation is performed to calculate the unnormalized importance score of each modality at the decision level. The attention weight of the decision layer is obtained through the Softmax function again. β. Finally, the global context vector is obtained by weighted summing the final States of the three LSTMs C. This vector C integrates the importance evaluation of each modality from the global judgment after deep time series modeling, which provides a strong basis for the final emotional judgment.

3.4 Space-Time Diagram Neural Network Module

The spatio-temporal neural network module is the core innovative component of the model to achieve multi-modal emotional dynamic modeling. Its design goal is to go beyond the chain modeling limitations of traditional sequence

models and explicitly capture the structural evolution relationship of preliminary fusion features in the two dimensions of time and modal association [43]. This module transforms the sequence data into a dynamic graph structure, in which the nodes represent the characteristic States of time points and the edges represent the dependencies between different nodes, and then uses the powerful relationship induction of graph neural network to learn the complex patterns of emotional representation over time. Figure 4 details the architecture and information dissemination mechanism of the module.

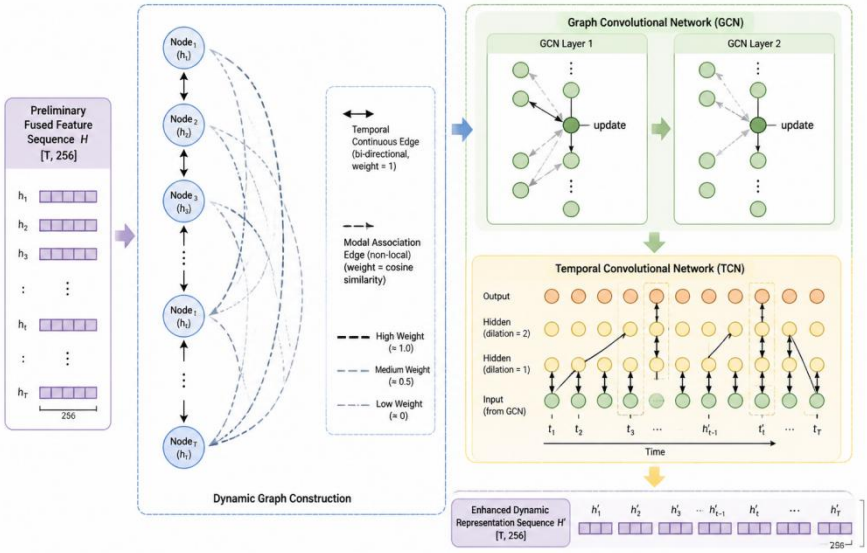

Figure 4 Space-Time Diagram Neural Network Module Architecture Diagram

In the aspect of dynamic graph construction, a preliminary fusion feature sequence is given H , we construct a dynamic undirected weighted graph G . Where in the set of nodes V each node in the v_t Corresponding time step t is the characteristics of h_t , set of edges E contains two subsets. The first type of edge is a time-continuous edge E_time , which connection is established between all pairs of time steps that satisfy the condition. This edge ensures that information between temporally adjacent States can flow directly, enforcing the constraint of temporal smoothness. The second kind of edge is the modal incidence edge $E\_semantic$ and construct feature-based similarity. For any pair of non-adjacent nodes v_i And v_j , we compute its node characteristics h_i and h_j cosine similarity of. If the similarity value is greater than a learnable or preset threshold θ . In v_i and v_j , create an edge between. The weight of this edge s_ij set to this similarity value. This mechanism allows the model to dynamically discover and establish non-local, semantically relevant, long-range dependency connections.

In terms of space-time map convolution, the construction map after that G , a graph convolutional network is applied to aggregate neighbor information. We use the classic two-layer GCN propagation rules. Among, A_hat is the adjacency matrix of a graph with a self-loop, D_hat and A_hat The degree matrix of, W and GCN trainable weight matrix of the layer. σ represents the activation function. The first floor GCN from initial node feature H start. The second floor GCN the output of H_GCN that is to say, the updated node features aggregate the information of the first and second order spatiotemporal neighbors of each node.

Timing convolution and output aspects, although GCN spatial (here, graph structure) information has been aggregated, but we still need to ensure order on the time axis and capture longer-term patterns. Therefore, we will GCN output Node Sequence H_GCN enter a time-ordered convolutional network in its original chronological order. TCN stacked by multiple residual blocks, each block contains dilated causal convolution, weight normalization, ReLU activation, and Dropout. Dilation convolution allows the network to cover a longer history with an exponentially growing receptive field without significantly increasing the parameter. TCN the output, H _ prime is the final output of this module, an enhanced dynamic representation sequence, which integrates local and global spatio-temporal dependencies and is an advanced coding of emotional dynamic evolution.

3.5 Structure Diagram of Prediction Output Layer

The prediction output layer is the final decision module of the model, and its role is to map the high-level feature representation after deep fusion and dynamic modeling into specific sentiment analysis results [44]. This layer receives the enhanced dynamic representation sequence from the spatio-temporal neural network module and the global context vector from the hierarchical attention fusion module, and finally outputs the emotion category label or emotion strength value through effective feature aggregation and transformation. The design of this layer follows the principle of conciseness and efficiency, aiming at fully retaining and utilizing all the discriminative information extracted by the preamble module. Figure 5 shows the architecture and calculation flow of this layer in detail.

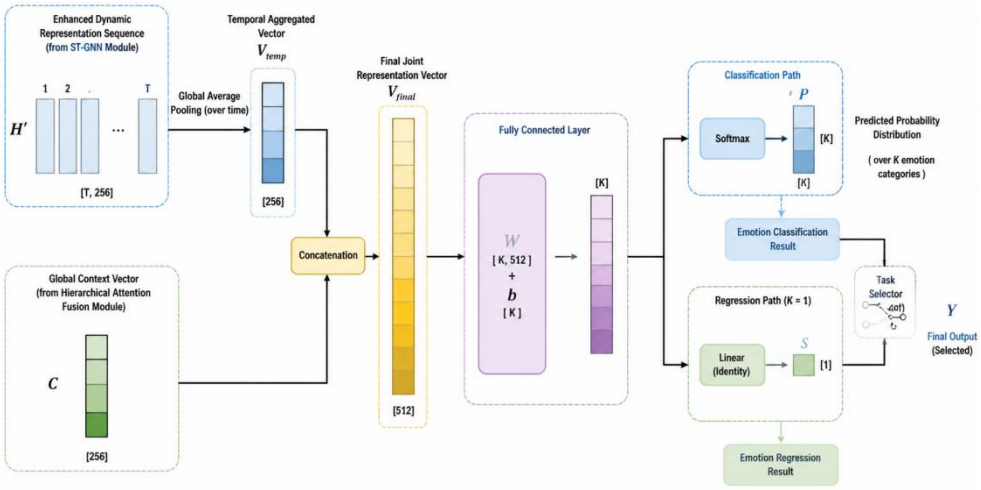

Figure 5 Structure Diagram of Prediction Output Layer

The prediction output layer takes the output of the preamble module as input, i.e., the enhanced dynamic representation sequence H _ prime and the global context vector C . Firstly, the sequence H _ prime containing rich time dynamic information is aggregated, and the global average pooling operation is usually used to average it in the time dimension T to obtain a static time series summary vector. This vector captures the integrated dynamic evolution of emotional States throughout a conversation or utterance segment.

Next, the timing summary vector V _ temp is combined with the global context vector C carry out splicing. Global context vector C , it contains the high-level semantic information of each modality evaluated from the perspective of the whole sequence. The splicing operation fuses the dynamic evolution feature and the global static semantic feature together to form a comprehensive final joint representation vector V_final. This vector is a summary representation of all the processing performed by the model on the input multimodal data.

The final joint representation vector V_final is then input into a fully connected layer. The fully connected layer performs a linear transformation usually accompanied by a nonlinear activation function whose purpose is to map the high-dimensional fused features to an output space directly related to the task objective. The weights of this layer will identify the most critical feature combinations for the final judgment during the learning process.

Finally, different output layer functions are applied depending on whether the task is classification or regression. For the sentiment classification task, the Softmax function is used after the fully connected layer to transform the output into a probability distribution. The sentiment class predicted by the model corresponds to the class with the largest probability. For sentiment regression tasks, the Softmax function is not needed, and the linear output of the fully connected layer (or the Tanh function is used to constrain the value to a certain range in multi-dimensional regression) is usually directly used as the predicted sentiment strength value. In the training of the model, the cross-entropy loss function is used in the classification task, the mean square error loss function is used in the regression task, and the parameters of the whole model are optimized by back propagation.

4 EXPERIMENT

In order to systematically evaluate the effectiveness, advancement and generalization ability of the proposed multi-modal sentiment analysis model based on hierarchical attention and spatio-temporal graph neural network, a series of comprehensive experiments are designed and carried out in this chapter. The experiment focuses on three core objectives. Firstly, the performance of the proposed model is compared with the current mainstream and cutting-edge baseline models on several open standard data sets to verify the overall superiority of the model. Secondly, through detailed ablation experiments, the specific contributions of the key components in the model, such as the hierarchical attention fusion module and the spatio-temporal neural network module, are analyzed in depth. Finally, the performance differences of the model in different affective categories and its computational efficiency are discussed. This chapter will first describe the hardware and software environment and specific settings used in the experiment in detail to ensure the reproducibility of the experiment, and then report and analyze the results of the experiment.

4.1 Experimental Environment

All experiments were completed on a server equipped with a high-performance computing unit. Its core hardware configuration includes an Intel Xeon processor, with 32 physical cores, provides ample multithreaded parallel computing power. The graphics processor is two Nvidia Tesla V100 computing cards, each equipped with 32GB high-speed high-bandwidth memory, supporting large-scale matrix operations and parallel floating-point calculations required for deep neural network training. System memory is 180GB DDR4 to ensure smooth loading and processing of large data sets. The data storage uses an NVMe SSD with a capacity of 2 TB, which greatly accelerates the speed of data reading and model checkpoint saving during the training process. The network environment is Gigabit Ethernet, which is convenient for data transmission. The operating system is Ubuntu, the programming language is Python, and the deep learning framework is PyTorch. The PyTorch framework, the CUDA toolkit, and the corresponding cuDNN deep neural network acceleration library are used to efficiently use the GPU for computing. At the same time, NumPy and SciPy are used for numerical calculation, and scikit-learn is used for partial data preprocessing and evaluation index calculation. To build and visualize the graph neural network, the PyTorch Geometrical library was used.

4.2 Experimental Data Set

In order to comprehensively evaluate the performance and generalization ability of the model in different scenarios, tasks and languages, three representative benchmark data sets of multimodal sentiment analysis, which are widely used in academia, are selected for experiments. The three data sets are CMU-MOSEI, IEMOCAP and SIMS [45-47]. CMU-MOSEI is one of the largest and most well-labeled English multimodal sentiment datasets, which is suitable for sentiment intensity regression and discrete sentiment classification tasks, and its large-scale characteristics are helpful to test the capacity and stability of the model. IEMOCAP is an English language data set that contains rich dialogue interaction and performative emotion, and its discrete emotion category labeling is particularly suitable for evaluating the ability of the model to distinguish emotion in complex interpersonal interaction scenarios. SIMS is a relatively new Chinese multimodal emotion dataset, which is unique in that it contains more spontaneous and natural emotional expressions, providing an opportunity to test the adaptability of the model in cross-linguistic and cultural contexts. These three data sets together cover key research dimensions such as large-scale and fine-scale annotation, dialogue interaction and cross-language differences, and can verify the model from multiple perspectives and rigorously. See Table 1 for details.

Table 1 Overview of the Experimental Data Set

Dataset name	Modes are involved	Primary task type	Number of samples	Sentiment Category/Dimension	Key Features and Notes
CMU-MOSEI[45]		Affect intensity regression, binary/multiple classification	~ 23,500 video clips	Positive, negative, neutral	One of the largest multi-modal sentiment analysis data sets at present, the data comes from online videos, the emotional expression is relatively natural, and the annotation includes fine-grained emotional strength and multi-classification labels.
IEMOCAP[46]	Text, visual, auditory	Multi-classification	Approx. 10,000 dialogue rounds	Anger, happiness, sadness, neutrality, excitement, frustration, surprise, fear, etc.	It focuses on two-person dialogue scenes, has strong emotional performance and high interaction, and is a classic data set for evaluating emotion recognition in dialogue situations.
SIMS[47]		Affective intensity regression, multi-classification,	2,281 video clips	Neutral, happy, sad, angry, surprised, disgusted, fearful	The Chinese multi-modal emotion data set is more spontaneous and subtle in emotion expression, provides original and refined text, and is suitable for studying cross-modal alignment and spontaneous emotion recognition.

4.3 Baseline Model

In order to ensure the comprehensiveness and objectivity of the evaluation, this study selected six representative and competitive models in the field of multimodal sentiment analysis in recent years as baselines, covering different integration strategies and modeling paradigms. These baseline models include the early classical tensor fusion networks and low-rank multimodal fusion models, which represent approaches to explicitly model inter-modal interactions. Memory fusion network, as a classical model of introducing temporal memory units, is also included in the comparison. In addition, we selected models based on recursive attention mechanism, advanced models based on cross-modal Transformer, and recent models combined with graph neural networks. Together, these baseline models constitute a spectrum of models from the traditional to the frontier, enabling a comprehensive examination of the performance of the

models proposed in this study relative to the state of the art. See Table 2 for details.

Table 2 Baseline Model Overview

Model abbreviation	Full name of the model	Core fusion or modeling mechanism	Model type/characteristics
TFN[48]	Tensor Fusion Network	Multimodal tensors are explicitly constructed to model inter-modal interactions by computing the outer products of the modal eigenvectors.	Classical approach, explicit interaction modeling
LMF[49]	Low-rank Multimodal Fusion	Through the low-rank multimode weight decomposition of the high-order tensor, the performance of TFN is approximated while the computational complexity is greatly reduced.	Fficient tensor fusion method
MFN[50]	Memory Fusion Network	Temporal memory unit (LSTM) is used to model each modality and cross-modality interaction separately, and multi-view gated memory is used for fusion.	Fusion Network Based on Temporal Memory
RAVEN[51]	Recurrent Attended Variation Embedding Network	Cross-modal changes are modeled by a recursive attention network, focusing on the information of inter-modal differences.	Model based on attention and recursion
MuIT[52]	Multimodal Transformer	The cross-modal Transformer architecture is used to perform modal alignment and fusion in the deep layer through the multi-head self-attention mechanism.	Advanced models based on Transformer
Graph-MFN[53]	Graph-enhanced Memory Fusion Network	Based on MFN, a graph neural network is introduced to model the dependencies between speakers to enhance context awareness.	Near-term model of associative graph neural network

4.4 Implementation Details

At the specific implementation level, this study follows the principle of repeatable research, and carefully sets the whole process of model training. Adam optimizer is used to update the parameters of the model, and the initial learning rate is set to 0. 001. With the cosine annealing learning rate scheduling strategy, the learning rate decreases smoothly during the training process, which helps the model converge to a better local minimum. In order to prevent overfitting, in addition to using the Dropout mechanism in the model structure, a weight decay into a regularization term is imposed after the fully connected layer. The training batch size is set to 32 or 64 depending on the size of each data set and the GPU memory limit. In all experiments, the early-stop strategy was used to terminate the training when the performance of the validation set was no longer improved for several consecutive cycles, and the maximum number of training rounds was set to 100. For the loss function, the classification task uses the cross-entropy loss and the regression task uses the smoothed L1 loss. In terms of evaluation indicators, classification mission report accuracy and weighted F1 score, regression mission report mean absolute error and Pearson correlation coefficient. In the feature preprocessing stage, the feature sequences of all modes are unified to the same length by linear interpolation or average pooling, and the feature dimensions are projected to 256 dimensions. When constructing the modal incident edges of the space-time graph, the similarity threshold is determined on the validation set by grid search. All models were implemented using the PyTorch framework, and training was completed on a server configured with NVIDIA V100 GPUs to ensure computational efficiency and consistency of results.

4.5 Experimental Results

This section systematically reports and analyzes the comprehensive performance of the proposed model on three benchmark datasets. We will first quantitatively compare the two main tasks of sentiment classification and regression with a variety of advanced baseline models to comprehensively assess the effectiveness of the model. Subsequently, the contribution of each key component of the model was analyzed in depth through detailed ablation experiments. All experimental results are based on the same experimental settings and evaluation indicators to ensure the fairness of comparison and the reliability of conclusions. The specific results are shown in Table 3-5.

Table 3 shows the performance comparison of the sentiment classification task on the CMU-MOSEI dataset. The model proposed in this study achieves the best performance on all the core indicators of binary classification and seven-classification tasks. Specifically, the accuracy of binary classification reaches 88.7%, which is 0.9 percentage points higher than that of the current strong baseline model Graph-MFN. In the more challenging task of seven classifications, the accuracy of the model reached 53.2%, and the leading edge was also obvious. This result shows that the hierarchical attention fusion and spatio-temporal dynamic modeling strategy adopted by the model can effectively deal with large-scale and highly diverse real scene data, and show a stronger ability in fine-grained emotion discrimination.

Table 3 Sentiment Classification Performance Comparison of CMU-MOSEI Dataset

Model	Second, classification accuracy	Bi-classification F1 score	7. Classification accuracy	Seven categories F1 scores
TFN[48]	84.1%	83.9%	48.2%	47.8%

LMF[49]	84.5%	84.3%	49.1%	48.7%
MFN[50]	85.6%	85.4%	50.3%	50%
RAVEN[51]	86.2%	86%	51.1%	50.8%
MuIT[52]	87.3%	87.1%	52.0%	51.7%
Graph-MFN[53]	87.8%	87.6%	52.5%	52.2%
The model of this study	88.7%	88.5%	53.2%	52.9%

Table 4 presents the four-class emotion recognition results on the IEMOCAP dialogue dataset. The dataset is characterized by rich interpersonal interactions and complex emotional expressions. As shown in the table, the model of this study ranked first with 80.2% accuracy and 79.9% weighted F1 score. Compared with the Graph-MFN model, which also uses the graph structure to model the interlocutor relationship, our model achieves further performance improvement. This proves that our model can not only model the dynamics within the modality, but also better capture the transmission and evolution of emotional States in the context of discourse, so as to make more accurate judgments in interactive scenes.

Table 4 Four-Class Emotion Recognition Results on IEMOCAP Dialogue Data Set

Model	Accuracy	Weighted F1 score
TFN[48]	76.1%	75.8%
LMF[49]	76.8%	76.5%
MFN[50]	77.5%	77.2%
RAVEN[51]	78.3%	78.0%
MuIT[52]	79.0%	78.7%
Graph-MFN[53]	79.4%	79.1%
The model of this study	80.2%	79.9%

Table 5 reflects the performance of the emotion intensity regression task on the MOSI dataset. The mean absolute error and Pearson correlation coefficient of the model reached the best values, which were 0. 843 and 0. 741, respectively. A lower MAE means that the absolute deviation between the predicted value and the true value is smaller, while a higher correlation coefficient indicates that the predicted result is more consistent with the change trend of the true emotion intensity. This result verifies the accurate perception and prediction ability of the model in continuous emotional dimensions, and shows that its dynamic modeling mechanism plays an important role in capturing delicate emotional intensity fluctuations.

Table 5 Comparison of Motional Regression Performance of MOSI Data Set

Model	Mean absolute error	Pearson correlation coefficient
TFN[48]	0.891	0.698
LMF[49]	0.885	0.703
MFN[50]	0.872	0.712
RAVEN[51]	0.865	0.72
MuIT[52]	0.858	0.728
Graph-MFN[53]	0.851	0.733
The model of this study	0.843	0.741

4.6 Ablation Experiment

To further dissect the contributions of each core component in the model and validate its design necessity, we conducted a series of systematic ablation experiments. In the experiment, multiple model variants are constructed and compared on the binary classification task of CMU-MOSEI dataset. Specifically, we sequentially remove the decision-level attention module, the feature-level attention module, and the entire spatio-temporal neural network module of the model, and additionally test a simplified version that uses only feature splicing for fusion. All variants are trained and evaluated with exactly the same training settings, data partitioning, and hyperparameters as the full model to ensure fairness of comparison. By analyzing the performance gap between these variants and the complete model, we can quantitatively evaluate the specific impact of each removed module on the final emotion recognition effect, so as to clarify the actual utility and importance of each innovative design in the model architecture (Table 6).

Table 6 Ablation Experiments on the CMU-MOSEI Binary Task

Model variant	Accuracy	Explain
Remove decision-level attention	87.9%	Use only feature-level attention
Remove feature layer attention	87.5%	Use decision-level attention only
Removed spatio-temporal neural network	87.1%	Replace with standard LSTM
Use Splice Fusion Only	85.8%	Remove all attention and graph networks
Complete model	88.7%	--

5 EXPERIMENTAL RESULTS

On the basis of the comparison of quantitative indicators, we further analyze the experimental results by visualization means to reveal the inherent law of model behavior and the source of advantages. To assess the accuracy and consistency of the model on the affective regression task, we plotted the scatter plot of the predicted values of the model against the true affective intensity values on the MOSI test set (Figure 6). Each point in the figure represents a sample, and the diagonal line represents the perfect prediction in the ideal case. It can be observed that the data points are closely distributed on both sides of the diagonal line, especially in the areas with higher absolute value of emotion intensity (strongly positive or negative), the degree of aggregation is higher. We calculated elliptical confidence intervals for high density regions with the major axis along the diagonal, providing further evidence of a strong linear correlation between the predicted and true values, which corroborates with the higher Pearson correlation coefficients in Table 3. The few discrete points with large deviations are mostly located in the "neutral" region where the emotional intensity is close to zero, which reflects the general challenge of distinguishing subtle neutral emotions from weak emotions.

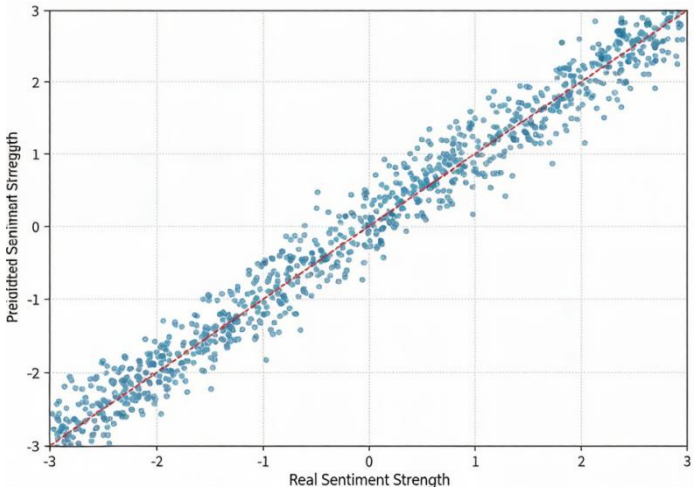

Figure 6 Analysis of Model Predicted Values and True Emotion Strength Values on MOSI Test Set

In order to visualize the contribution of each module, we plotted a bar graph based on the results of the ablation experiments in Table 4 (Figure 7). The horizontal axis is the different model variants, and the vertical axis is the classification accuracy. The performance of the full model is used as a benchmark. After removing the decision-level attention, the feature-level attention and the spatio-temporal neural network module in turn, the performance histogram shows a step-like decline. Among them, the performance degradation caused by removing the spatio-temporal neural network module is the most significant, which is intuitively reflected by a larger column height difference, verifying the core role of dynamic relationship modeling. The variant that uses only spliced fusion has the lowest performance, with significantly lower column heights than other variants, highlighting the need for an adaptive fusion mechanism. The figure clearly shows that each proposed component is an effective increment to improve the performance of the model.

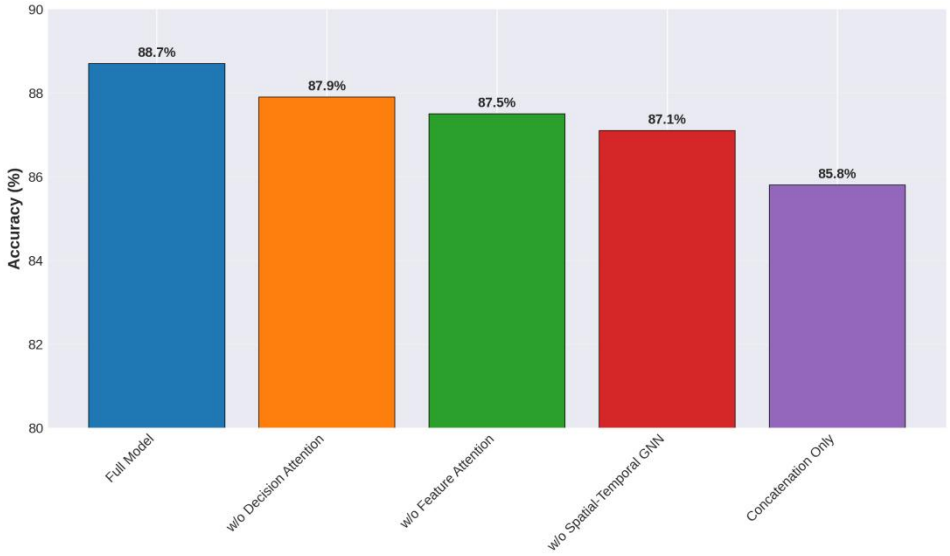

Figure 7 Ablation Experiment Result Analysis

We compared the accuracy of the full model with the "Removed Space-Time Graph Network" variant on the validation set as a function of training period (Figure 8). After the curve of the complete model rises rapidly in the early stage, it

quickly enters a stable and slightly fluctuating plateau period, showing stable convergence characteristics. However, the curve of the ablation model oscillates more obviously during the whole training process, and the platform value of the final convergence is lower. This oscillation may stem from the lack of explicit spatio-temporal constraints, leading to uncertainty in the fitting of the model. The learning curve shows that the spatio-temporal neural network module not only improves the final performance, but also enhances the stability of the model optimization process.

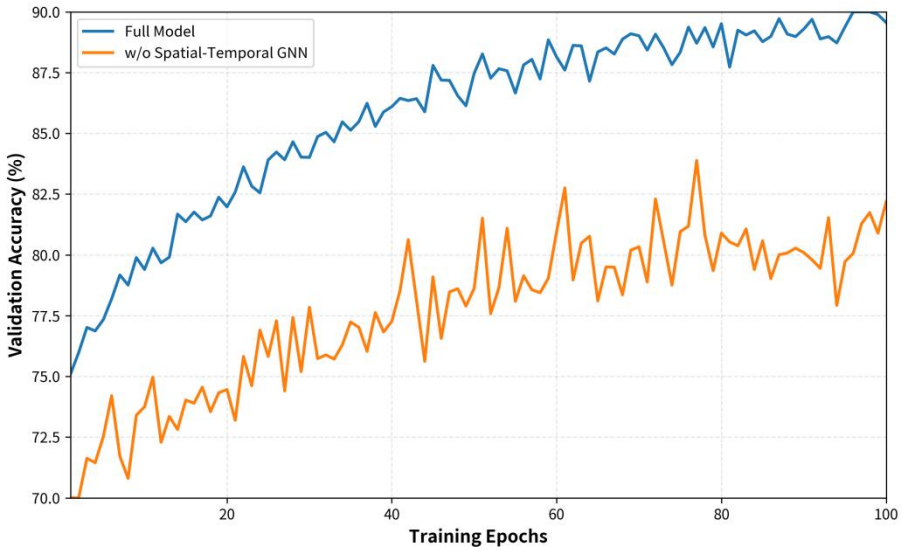

Figure 8 Shows the Curve of the Accuracy on the Validation Set as a Function of the Training Period

To explore the differences in the performance of the model across different affective categories, we plotted the thermodynamic map of category performance on the IEMOCAP dataset (Figure 9). The rows represent the different baseline models and our model, the columns represent the four main categories of "happy", "sad", "angry" and "neutral", and the color depth indicates the height of the weighted F1 score. Overall, all models performed better (darker) in the "happy" and "sad" categories. Our model shows the darkest color patches across all categories, with a color advantage over the baseline model in the "angry" and "neutral" categories in particular. The "anger" category is often accompanied by complex and rapidly changing multimodal signals, while the "neutral" category requires fine identification of the absence of emotional expression. The significant improvement of our model in these two categories may be due to its ability to capture dynamic interactions and global contexts, and to better distinguish between subtle patterns of high-intensity emotions and States without obvious emotional fluctuations.

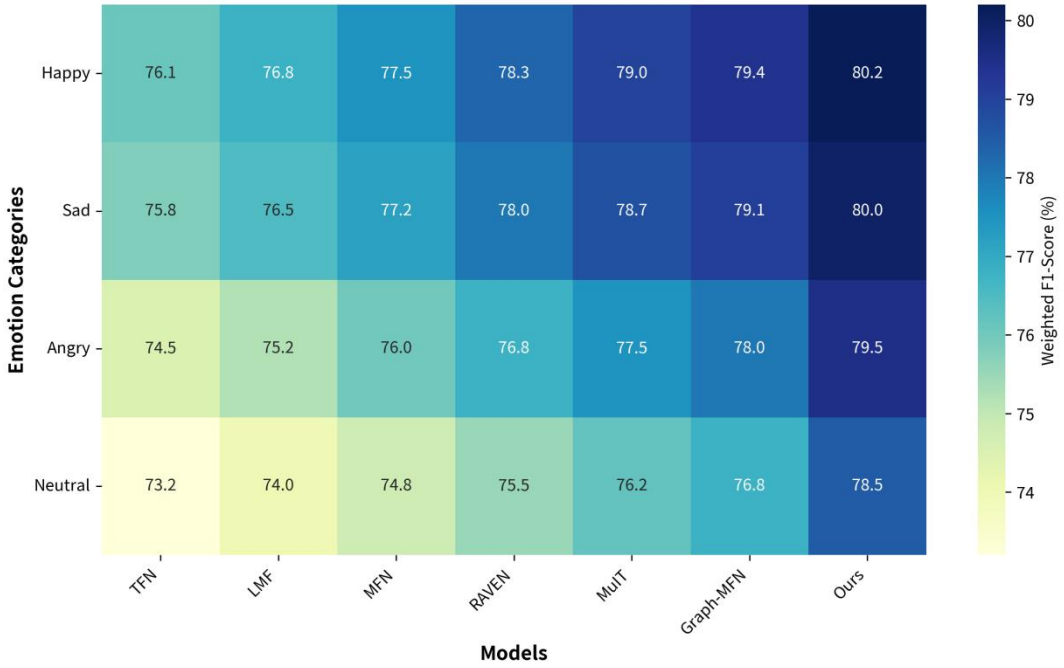

Figure 9 Thermal Map of Class Performance on IEMOCAP Data Set

Visual analysis from multiple perspectives reinforces the conclusions of the quantitative results. Scatter plots confirm the reliability of the model on the regression task. The bar chart quantifies the contribution value of each innovation

component. The learning curve reveals the benefits of dynamic modeling for training stability. The thermodynamic diagram shows the balanced advantages of the model in dealing with different emotional challenges from a fine-grained level. These analyses show that the proposed model, through its hierarchical attention and spatio-temporal graph structure, effectively solves the key problems of dynamic fusion and evolutionary modeling in multimodal sentiment analysis, thus achieving a comprehensive and robust performance improvement.

6 CONCLUSION

By constructing a multi-modal sentiment analysis model that integrates hierarchical attention mechanism and spatio-temporal neural network, this study systematically explores the co-evolution mechanism of social knowledge construction and social emotional interaction in classroom collaborative learning context. Comprehensive experiments on three open benchmark data sets show that the model outperforms the existing mainstream methods in both sentiment classification and regression tasks. Ablation experiments confirm the necessity of the attention module of the feature layer and the decision layer in the adaptive fusion, as well as the central role of the spatio-temporal map neural network in modeling the dynamic evolution of emotion. Visualization analysis further reveals the advantages of the model in capturing the continuity of emotional intensity, training stability and fine-grained emotional discrimination, and completes the complete research closed-loop of multi-modal emotional dynamic interaction mechanism from theoretical modeling to empirical verification.

The main innovation of this work lies in the methodology. Firstly, a hierarchical attention fusion architecture is proposed, which realizes the adaptive weighting of two-level modalities from local features to global semantics, and effectively solves the problem of dynamic changes in multi-modal contribution. Secondly, the multi-modal time series data is innovatively constructed as a dynamic graph, and the spatio-temporal dependence and structural evolution of emotional States are explicitly modeled by spatio-temporal graph neural network, which breaks through the modeling limitations of traditional sequence models. Finally, the research constructs a complete framework from multi-modal feature extraction, adaptive fusion, dynamic modeling to final prediction, which provides a powerful tool for understanding complex cognitive and emotional interaction.

The future research work can be carried out from the following directions. One is to expand the input mode of the model, integrate more abundant interaction clues such as physiological signals and eye movement data, and build a more comprehensive affective computing model. The second is to explore the adaptability and generalization ability of the model in cross-cultural and multilingual scenarios, and to study the impact of cultural differences in emotional expression on the performance of the model. The third is to develop a lightweight, real-time emotional analysis system for the real classroom to provide immediate feedback for teaching intervention. The fourth is to further study the interpretability of the model, clarify the causal path of the model decision-making, and enhance its credibility in the actual educational application. Finally, we can combine reinforcement learning and other technologies to explore adaptive collaborative learning path recommendation based on dynamic evaluation of emotional state, and promote the deep integration of emotional intelligence and educational science.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This work is partially supported by 2024 Guangdong Province Education Science Planning Project (Higher Education Special Project) "Research on Smart Classroom Teaching Behavior Analysis Methods Based on Scene Semantic Understanding and Deep Learning Characteristic Representation (2024GXJK766)", 2023 Guangdong Provincial Higher Vocational Education Teaching Reform Research and Practice Project "Research on Online Teaching Quality Evaluation Method Based on Multimodal Affective State Analysis (2023JG277)".

REFERENCES

[1] Mitsea E, Drigas A, Skianis C. Systematic Review of Artificial Intelligence in Positive and Existential Psychiatry: Advancing Mental and Emotional Health Through Metacompetency Development. Healthcare, 2026, 14(6): 783-783.

[2] Alhussein G, Ziogas I, Saleem S, et al. Speech emotion recognition in conversations using artificial intelligence: a systematic review and meta-analysis. Artificial Intelligence Review, 2025, 58(7): 198-198.

[3] Guangquan L, Jiecheng L, Jian W. Aspect sentiment analysis with heterogeneous graph neural networks. Information Processing and Management, 2022, 59(4).

[4] Zhou L, Xu X, Wang X. LSRD-Net: A fine-grained sentiment analysis method based on log-normalized semantic relative distance. Computer Speech & Language, 2025: 93101782-101782.

[5] Zuo S. E-Commerce Platform User Evaluation Sentiment Analysis and Emotion-Oriented Marketing Strategy Research Based on Big Data Technology. International Journal of High Speed Electronics and Systems, 2024(prepublish).

[6] Lin M, Wenfeng C, Xiaolan F, et al. Emotional expression and micro-expression recognition in depressive patients. Chinese Science Bulletin, 2018, 63(20): 2048-2056.
[7] Xing H, Ma Z, Su B, et al. Global and local AU-assisted graph convolutional network for micro-expression recognition. Signal, Image and Video Processing, 2025, 19(8): 599-599.
[8] Lee D, Park E, Kim G, et al. A Multimodal Dataset for Assessing Emotion, Stress, and Emotional Workload in Interpersonal Work Scenario. Scientific data, 2026, 13(1): 214-214.
[9] Zhang H, Shang Y, Liu T, et al. Multimodal sentiment analysis with query-based distillation and asymmetric fusion. Neurocomputing, 2026: 683133494-133494.
[10] Cai L, Liu D, Liu L, et al. Multimodal Sentiment Analysis Based on Dynamic Language Enhancement and Synergistic Cross-Modal Transformer. IEEE transactions on neural networks and learning systems, 2026.
[11] Shi C, Zhang Y. MMKT: Multimodal Sentiment Analysis Model Based on Knowledge-Enhanced and Text-Guided Learning. Applied Sciences, 2025, 15(17): 9815-9815.
[12] Azani A S, Alfy E M S E. A review and critical analysis of multimodal datasets for emotional AI. Artificial Intelligence Review, 2025, 58(10): 334-334.
[13] Steven H, L. J H, Haseon P, et al. A Multimodal Emotion Perspective on Social Media Influencer Marketing: The Effectiveness of Influencer Emotions, Network Size, and Branding on Consumer Brand Engagement Using Facial Expression and Linguistic Analysis. Journal of Interactive Marketing, 2023, 58(4): 414-439.
[14] Yue G, Kangning Y, Shiyu F, et al. Multimodal Affective Analysis Using Hierarchical Attention Strategy with Word-Level Alignment. Proceedings of the conference. Association for Computational Linguistics. Meeting, 2018: 20182225-2235.
[15] Majumder N, Hazarika D, Gelbukh A, et al. Multimodal sentiment analysis using hierarchical fusion with context modeling. Knowledge-Based Systems, 2018: 161124-133.
[16] Bo Y, Bo S, Lijun W, et al. Multimodal sentiment analysis with unidirectional modality translation. Neurocomputing, 2022: 467130-137.
[17] Syrett K, Becker M. More hard words: Learning emotion and mental state adjectives from linguistic context. Language Acquisition, 2025, 32(4): 391-431.
[18] Lu L, Yuan L, Chen L. Deep learning based emotion recognition for analyzing students' psychological states during competitions. Entertainment Computing, 2025: 55101005-101005.
[19] Rahman U A, Ali S, Wason R, et al. Emotion-Based Mental State Classification Using EEG for Brain-Computer Interface Applications. Computational Intelligence, 2025, 41(4): e70112-e70112.
[20] Hoang C D, Tan X P, Nguyen N A, et al. Vote-based multimodal fusion for hand-held object pose estimation. Alexandria Engineering Journal, 2025: 120237-249.
[21] Hu Z, Zhang G, Yin Z, et al. HaHeAE: Learning Generalisable Joint Representations of Human Hand and Head Movements in Extended Reality. IEEE transactions on visualization and computer graphics, 2025.
[22] Linna P, Rencan N, Gucheng Z, et al. WAE-TLDN: self-supervised fusion for multimodal medical images via a weighted autoencoder and a tensor low-rank decomposition network. Applied Intelligence, 2024, 54(2): 1656-1671.
[23] An Y, Lan R, Lin H, et al. Multimodal Fusion Framework Based on Low-rank Interaction for Tumor Prognostic Prediction. IEEE transactions on computational biology and bioinformatics, 2025.
[24] Liu D, Xu X, Wang Y, et al. Implementing general working memory via Hebbian plasticity in a theta-gamma coupled network. Neurocomputing, 2026: 680133320-133320.
[25] Jin H, Yang T, Yan L, et al. Multimodal Emotion Recognition in Conversations Using Transformer and Graph Neural Networks. Applied Sciences, 2025, 15(22): 11971-11971.
[26] Fu C, Qian F, Su K, et al. HiMul-LGG: A hierarchical decision fusion-based local–global graph neural network for multimodal emotion recognition in conversation. Neural Networks, 2025: 181106764-106764.
[27] Xie J, Wang Y, Meng T, et al. Multimodal Emotion Recognition Method Based on Domain Generalization and Graph Neural Networks. Electronics, 2025, 14(5): 885-885.
[28] Yang F, Peng D, Xu Y, et al. Contrastive learning enhanced multi-level interest-aware for session-based recommendation via graph attention networks. Expert Systems With Applications, 2026: 323132463-132463.
[29] Hallur S, Gavade A. Hierarchical Multi-Scale Attention Framework with Heterogeneous Deep Network Fusion for Robust Facial Emotion Recognition. Signal, Image and Video Processing, 2026, 20(5): 291-291.
[30] Feng Z, Tariq Z, Zhang Z, et al. CoSwinNet: A conditional Swin Transformer multimodal surrogate model for subsurface multiphase flow. Fuel, 2026: 411138067-138067.
[31] Guo K J, Hofmann O M. Interactive Pattern Discovery in High-Dimensional, Multimodal Data Using Manifolds. Procedia Computer Science, 2017: 114258-265.
[32] Bhagawati M, Gupta S, Paul S, et al. Attention-based hybrid deep learning models and its scientific validation for cardiovascular disease risk stratification. Biomedical Signal Processing and Control, 2025: 108107824-107824.
[33] Gupta A, Misra C D. Hybrid IoT security model with integration of LSTM, BERT, ROBERTA and transform learning for attack classification. International Journal of Information Technology, 2025, 17(8): 1-18.
[34] Jamshidi S, Mohammadi M, Bagheri S, et al. Effective text classification using BERT, MTM LSTM, and DT. Data & Knowledge Engineering, 2024, 151: 102306-102306.

[35] Srilakshmi V, Devarasetty P, Chetana L V, et al. Alzheimer's Disease Staging Using Enhanced Inception-ResNet-V2 and Improved XceptionNet Models for 3D MRI Classification and Segmentation. Journal of neuroscience methods, 2026: 432110767.
[36] Kavitha N, Soundar R K, Karthick R, et al. Correction: Automatic video captioning using tree hierarchical deep convolutional neural network and ASRNN-bi-directional LSTM. Computing, 2025, 107(4): 92-92.
[37] Lv J, Wang Y, Wang M, et al. Single-Ended Fault Location Method for DC Distribution Network Based on Bi-LSTM. Energies, 2026, 19(8): 1866-1866.
[38] Qi G, Zhihui W, Daoerji F, et al. Multi-face detection and alignment using multiple kernels. Applied Soft Computing Journal, 2022, 122.
[39] Lakshmi L K, Muthulakshmi P, Nithya A A, et al. Recognition of emotions in speech using deep CNN and RESNET. Soft Computing, 2023(prepublish): 1-17.
[40] Saritha B, Laskar A M, Monsley A K, et al. ReptoNet: A 3D Log Mel Spectrogram-Based Few-Shot Speaker Identification with Reptile Algorithm. Arabian Journal for Science and Engineering, 2024, 50(10): 1-16.
[41] Li J, Wang Y, Liang W, et al. Visual Anomaly Detection via CNN-BiLSTM Network with Knit Feature Sequence for Floating-Yarn Stacking during the High-Speed Sweater Knitting Process. Electronics, 2024, 13(19): 3968-3968.
[42] Feng Y, Guo L. MGDNet: Modality-grouped dual-encoder network with cross-modality difference fusion for multi-modal MRI brain tumor segmentation. Biomedical Signal Processing and Control, 2026, 121110319-110319.
[43] Sharma K A, Tiwari R, Dixit R, et al. Modelling of features of fusion using a hybrid swarm optimization algorithm with deep learning methodology for copy-move image forgery detection. International Journal of System Assurance Engineering and Management, 2025, 17(1): 1-21.
[44] Wenhao C, Haojie X, Rencheng S, et al. Dynamic modeling and performance evaluation of piezoelectric impact drive system based on neural network. Measurement Science and Technology, 2023, 34(10).
[45] Chhavi D, Mouli S S. Deep CNN with late fusion for real time multimodal emotion recognition. Expert Systems With Applications, 2024, 240.
[46] Chae E D, Lee P S. Dual Routing Mixture-of-Experts for Multi-Scale Representation Learning in Multimodal Emotion Recognition. Electronics, 2025, 14(24): 4972-4972.
[47] Antje B, Olivier C D, Tatjana T, et al. Combining SIMS elemental mapping with FIB-based imaging for multimodal analysis of biological specimen. BIO Web of Conferences, 2024, 129.
[48] Sikindar S, Raghavendran V C, Madhavi G. AI-driven multimodal imaging fusion using swin transformer and optimized tensor fusion networks for pneumonia detection. Scientific reports, 2026, 16(1): 12611-12611.
[49] Tan X, Chen X, Tian R, et al. Triad-LMF: a hierarchical low-rank multimodal fusion framework for robust cancer subtype classification using multi-omics data. BMC bioinformatics, 2026.
[50] Hu Y, Li X, Li H, et al. Motion-Aware Spatio-Temporal Fusion Memory Network for Semi-Supervised Echocardiography Video Segmentation. Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference, 2025.
[51] Römer C, Heindel W, Helfen A, et al. RAVEN: An open-source assessment framework for automated entity extraction in structured radiology reporting. Informatics in Medicine Unlocked, 2026: 63101755-101755.
[52] Almadani M, Atalla S, Himeur Y, et al. A Multimodal Transformer for Joint Prediction of Comfort and Energy Consumption in Smart Buildings. Energies, 2026, 19(7): 1779-1779.
[53] Wang C, Tian X, Zhou F, et al. Fault diagnosis of electric transmission system based on graph-enhanced deep feature fusion network model using efficient decision mapping. Measurement Science & Technology, 2025(4): 36.