ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Research on Personalized Emotion Dynamic Deep Learning Modeling for Human-Centered Emotional Interaction

Yukai Liu [1] *and Hao Wang* [2,*]

[1] School of Computer Science and Engineering, University of New South Wales, Sydney 2052, New South Wales, Australia
[2] College of Music and Dance, Nanjing Normal University of Special Education, Nanjing 210038, Jiangsu, China

SUMMARY: *Aiming at the problems of continuous fluctuation of emotional expression, significant individual differences, and difficulty in unified modeling of multimodal heterogeneous data in human-centered emotional interaction, this paper proposes a personalized emotion dynamic deep learning model. Focusing on multi-source data such as text, speech, expression images and interactive behaviors, this paper constructs a unified data representation and preprocessing process, and combines multi-modal feature coding, emotional dynamic time series modeling and personalized adaptation mechanism to jointly depict the user's emotional state and its change trajectory. In the experimental part, MELD and CMU-MOSEI are selected as the main data sources, and the performance of the model is evaluated by Accuracy, Recall, F1-score and MSE. The results show that the Accuracy of the proposed model reaches 88.64%, the Recall is 87.91%, the F1-score is 88.23%, and the MSE is reduced to 0.098. Compared with Multimodal MLP, the Accuracy is increased by 10.23%, and the F1-score is increased by 10.41%, indicating that the method has good effect and application potential in emotion recognition, dynamic characterization and personalized adaptation.*
Povzetek: Aiming at the needs of human-centered emotional interaction, this paper constructs a personalized emotion dynamic deep learning model, which integrates multi-modal information such as text, speech, expression and interactive behavior, and realizes emotion recognition and change trajectory characterization through time series modeling and personalized adaptation. Experimental results show that the proposed model is superior to the comparison models in accuracy, F1 value and mean square error.

KEYWORDS: *Human-centered emotional interaction; Multi-modal emotion recognition; Emotion dynamic modeling; Personalized adaptation*

1 Introduction

With the continuous expansion of intelligent terminals, wearable devices, social media platforms and generative artificial intelligence applications, human-computer interaction is shifting from a task-oriented mode that focuses on functional response to an emotion-oriented mode that has the capabilities of understanding, perception and feedback. Affective computing has become an important research field in the intersection of artificial intelligence, computer technology, human-computer interaction, pattern recognition and signal processing. Emotion recognition based on multi-source data such as text, speech, facial expression and physiological

*hw090022@163.com
https://doi.org/10.65102/is2026077
Published: 30 April, 2026

signals has become an important way to improve the natural interaction ability and service adaptation ability of intelligent systems. In recent years, deep learning has promoted emotion recognition from shallow feature engineering to deep semantic representation and cross-modal joint modeling. Methods such as convolutional neural network, recurrent neural network, long short-term memory network, gated recurrent unit, Transformer and pre-training model have been widely used in emotion feature extraction, context-dependent modeling and cross-modal fusion. However, the emotional expressions in real interaction scenarios often have the characteristics of continuous fluctuation, implicit transfer, cross-modal heterogeneity and strong individual differences. Traditional static recognition frameworks are difficult to meet the actual needs of human-centered emotional interaction.

Although the existing research has achieved many results in the accuracy of emotion recognition, multi-modal fusion and model training paradigm, there are still some shortcomings. Some methods simplify emotion recognition into discrete classification tasks at a single moment, and pay insufficient attention to emotional evolution trajectory, intensity change and context dependence. Some general models pay more attention to the average performance of the group, and lack of sufficient modeling of user personality characteristics, interaction habits and historical preferences, which leads to limited long-term interaction and personalized response ability. At the same time, multimodal emotion data still have great challenges in time synchronization, feature scaling, noise suppression and semantic mapping. It can be seen that the research of emotion computing for emotional interaction is not only a problem of recognition accuracy, but also a computer technology problem of multi-source heterogeneous data processing, dynamic deep modeling and individual adaptive learning.

Based on this, this paper focuses on the characterization and personalized modeling of emotion dynamics in human-centered emotional interaction scenarios. This paper reviews the research progress in affective computing, emotion dynamic analysis, personalized emotion modeling, multimodal deep learning and emotional interaction, constructs a unified emotional information representation and preprocessing process, introduces a temporal deep learning model to describe the change law of emotion in continuous interaction, and combines user portrait, individual embedding and transfer adaptive mechanism. The learning ability and adaptation ability of the model for different user emotion patterns were enhanced. The paper is divided into five chapters: the first chapter is the introduction, the second chapter is the literature review, the third chapter is the model design, the fourth chapter is the experimental design and result analysis, and the fifth chapter is the conclusion.

2 Literature Review

2.1 Research on Affective Computing and Emotion Recognition

Affective computing is an important research field that realizes emotion recognition, emotion understanding and emotion interaction by acquiring, representing and analyzing human emotion information. With the development of artificial intelligence, human-computer interaction, pattern recognition and signal processing, affective computing has gradually shifted from the early research paradigm relying on artificial rules and shallow features to the technical path with data-driven, deep learning and multi-modal fusion as the core. Garcia-Hernandez et al. (2024) proposed that current emotion recognition research has formed a development trend of evolution from single-modal analysis to multi-modal collaborative modeling, and the joint utilization of multi-source data such as text, speech, vision and physiological signals is becoming an important direction to improve the emotion perception ability and system robustness [1]. In terms of specific methods, Chutia and Baruah (2024) proposed that deep

learning promoted text emotion recognition from traditional classification models to deep semantic representation based on attention mechanism and pre-training models, which enhanced the model's ability to understand complex contexts and implicit emotions [2]. George and Ilyas (2024) proposed that speech emotion recognition relies on features such as pitch, energy, formant, Mel-frequency cepstral coefficient and spectrogram, in which convolutional networks, recurrent networks and transformers all play an important role [3]. Kaur and Kumar (2024) proposed that convolutional neural networks, transfer learning, and attention mechanism have become mainstream methods for expression recognition [4]. Kacimi and Adda (2025) proposed that emotion recognition from physiological signals has unique advantages in arousal and valence analysis, but the problems of acquisition cost and subject differences are still prominent [5]. In general, existing research has presented multimodal and deep features, but there are still shortcomings in cross-modal collaboration and continuous emotion characterization.

2.2 Research on Modeling Emotion dynamics

Dynamic emotion modeling focuses on the generation, transmission and evolution of emotional states in the process of continuous interaction. The core of dynamic emotion modeling is not the static discrimination of emotion labels at a single moment, but the temporal correlation structure formed by context dependence, speaker state change and emotional cues accumulation. Early time series emotion analysis mainly relies on deep learning models such as recurrent neural network (RNN), long short-term memory (LSTM) network and gated recurrent unit (GRU) to continuously encode historical information by recursively transmitting hidden states. These methods can retain the previous emotional cues at the sequence level, and have good modeling ability for short-term context dependence and local emotional fluctuations. However, there are still shortcomings in long-distance dependence capture, cross-round information transfer and multi-speaker emotional interaction modeling.

In response to the above problems, in recent years, the modeling of emotion dynamics has gradually shifted from a simple sequence recursive mechanism to a composite modeling path combining individual attributes, future cues and global attention. Wang et al. (2024) proposed a conversational emotion recognition method based on dynamic personality [6], which introduced the speaker's individual characteristics into the emotional state update process, and enhanced the model's ability to explain subject differences. Khule et al. (2024) proposed a pseudo-future enhanced dynamic conversational emotion recognition model [7] to improve current emotion judgment by adding future context cues. With the development of attention mechanisms and Transformer architectures, modeling of emotion dynamics begins to place more emphasis on global context modeling and key cue aggregation. Fu et al. (2025) proposed a large language model conversational emotion recognition method combined with speaker features [8], which improved the model's ability to capture cross-round semantic dependencies and speaker state changes. Zhang and Tan (2025) proposed the evidence-inducement attention network [9], which strengthened the collaborative modeling ability of temporal and causal cues. Shen et al. (2025) proposed the emotional cue-driven framework [10], which further improved the model's ability to represent the implicit emotional transfer path. In general, the modeling of emotion dynamics is shifting from single time series learning to comprehensive modeling that integrates subject features, context relationships and multi-source cues.

2.3 Research on Personalized Emotion Modeling

Personalized emotion modeling emphasizes the explicit introduction of individual differences into the emotion computing process. Its goal is to break through the modeling method of general

models centered on the population sample distribution, so that the system can adaptively recognize and predict different users according to their behavior patterns, emotional expression habits and physiological response characteristics. With the development of wearable devices, mobile perception and continuous interaction scenarios, the subject heterogeneity of emotional data has become increasingly prominent, which makes individual difference modeling, user portrait modeling and adaptive learning gradually become important research directions of emotion computing. Han et al. (2024) proposed a systematic evaluation of personalized deep learning emotion recognition models [11], and pointed out that personalized modeling needs to comprehensively consider the impact of individual differences on recognition results at the levels of data organization, model structure and training strategy.

In the specific technical path, user portrait modeling and individual embedding mechanism are important means to realize personalized emotion recognition. Li and Washington (2024) proposed to compare the differences between personalized methods and general methods in emotion recognition of wearable devices [12], and the results show that personalized models usually have stronger adaptability in user-level recognition tasks. Transfer learning and cross-agent adaptive learning provide technical support for alleviating the problems of insufficient individual samples and cold start of new users. Shi et al. (2024) proposed a multi-source manifold metric transfer learning method for cross-subject EEG emotion recognition [13], which enhanced the cross-subject generalization ability of the model. Kim et al. (2025) proposed an emotion recognition and prediction method for wearable data based on clustering-guided attention and cross-species pre-training [14], indicating that personalized modeling has gradually extended from static recognition to dynamic prediction. Kovacevic et al. (2024) proposed a research on multimodal emotion recognition for real human-computer dialogue scenes [15], indicating that it is difficult to meet the requirements of natural interaction by relying solely on general models in an open environment. In general, personalized emotion modeling has become an important direction to improve the quality of emotional interaction and enhance the ability of model scene adaptation.

2.4 Research on Multimodal Deep Learning and Affective Interaction

The core of multimodal deep learning and emotional interaction research is to comprehensively use multi-source data such as text, speech, expression, physiological signals and behavior trajectories to construct a unified computational framework to represent semantic information, temporal features and emotional states. Compared with single-modal emotion recognition, multi-modal modeling can better simulate the complexity of emotion expression in real interaction, and reflect the role of computer technology in perception, understanding, fusion and response. Geetha et al. (2024) pointed out that multimodal emotion recognition has shifted from feature concatenation to deep representation learning and cross-modal collaborative reasoning, and deep learning technology has promoted the transformation of emotion understanding from local signal analysis to global semantic fusion [16].

At present, around multi-modal feature fusion, previous studies have proposed early fusion, middle fusion and late decision fusion paths. Hazmoune and Bougamouza (2024) proposed that the Transformer architecture can effectively improve the long-distance association between different modalities and the extraction ability of key emotional cues in multimodal emotion recognition through the self-attention mechanism [17]. Zhang et al. (2024) pointed out that multimodal emotion recognition of audio, vision and text is evolving to hierarchical fusion and dynamic weight allocation [18]. In the collaborative modeling of multi-source heterogeneous data, the asynchrony of modal time, the difference of noise distribution and the semantic inconsistency are still the performance bottlenecks. Wu et al. (2025) proposed that the key in

the future is to solve the problems of unified representation of heterogeneous features and missing mode compensation [19]. Wang and Zhang (2025) improved the fine-grained ability of emotion detection by jointly modeling EEG and text [20]. Overall, multimodal deep learning pushes emotional interaction from surface response recognition to deep state understanding.

2.5 Lack of existing research

Based on the existing literature, it can be seen that the research on affective computing has formed a relatively clear technical path in terms of single-modal recognition, temporal modeling, personalized learning and multi-modal fusion, which provides an important foundation for the research on emotion understanding. However, from the perspective of the requirements of human-centered emotional interaction, the existing research still focuses on a certain type of features, a certain stage task or a certain local scene, and has not formed a unified modeling framework that takes into account "dynamics, individuality and interactivity". Some studies emphasize the recognition accuracy, but lack the description of the emotion evolution process. Some studies focus on individual adaptation, but lack of support for multimodal collaboration and interactive feedback. Although some studies have introduced deep learning and Transformer methods, there is still room for improvement in heterogeneous data alignment, long-term dependency modeling, and response loop closure for real interaction scenarios. Therefore, it is necessary to construct a personalized emotion dynamic deep learning model for human-centered emotional interaction, which integrates multimodal perception, time series evolution modeling, individual embedding expression and interactive response support.Table 1 shows the analysis of the shortcomings of the existing research and the improvement direction of this research.

Table 1: Analysis of existing research deficiencies and directions for improvement of this research.

Research Direction	Representative References	Strengths of Existing Studies	Main Limitations	Improvements in This Study
Affective Computing and Emotion Recognition	[1]–[5]	Established a foundation for multimodal emotion recognition based on text, speech, facial expressions, and physiological signals, with relatively strong deep feature extraction capability	Mainly focuses on static classification; insufficient cross-modal collaboration; difficult to characterize continuous emotional changes	Integrates multi-source emotional information within a unified framework and strengthens dynamic representation
Emotion Dynamic Modeling	[6]–[10]	Effectively exploits contextual information and improves temporal dependency modeling in conversational emotion recognition	Still insufficient in modeling long-term dependencies, implicit transitions, and individual differences	Introduces a joint mechanism of individual embedding and dynamic modeling to enhance long-range dependency characterization
Personalized Emotion Modeling	[11]–[15]	Pays attention to user differences, transfer learning, and adaptive learning, improving cross-subject adaptability	Mostly limited to single scenarios or single-type data; insufficient generalizability and weak support for interaction	Combines user profiling with adaptive learning to improve personalized interaction capability
Multimodal Deep Learning and Affective Interaction	[16]–[20]	Has made progress in heterogeneous data fusion, Transformer-based modeling, and interaction scenario applications	Data alignment, modality missingness, response closed-loop design, and real-time deployment remain prominent issues	Builds an integrated model of “perception–understanding–adaptation–response” to enhance practical applicability

3 Personalized Emotion Dynamics Deep Learning Model for Human-Centered Emotional Interaction

3.1 Task definition and overall framework

The deep learning modeling of personalized emotion dynamics for human-centered emotional interaction is essentially a joint optimization problem that integrates multi-modal representation learning, temporal dependence modeling and individual adaptive learning. The core goal is not to distinguish the emotion category at a certain moment in isolation, but to comprehensively use the user's multi-modal emotion signals and individual attribute information in the continuous interaction process to describe the generation, accumulation, fluctuation and transfer rules of emotional states, and output the emotion understanding results that are more in line with the user's differences. Compared with traditional static emotion recognition, this task emphasizes more on two dimensions: one is the dynamics of emotion changes over time, and the other is the individual heterogeneity of different subjects in expression style, reaction intensity and interaction habits. Therefore, the model needs to have the ability of cross-modal feature fusion, long-term and short-term temporal dependence modeling, and individual-oriented parameter adjustment and adaptation.

From the task form, let the multimodal input of the user at interaction time t be represented as follows:

$$\mathrm{X_t} = \{\mathrm{x_t^{text}}, \mathrm{x_t^{audio}}, \mathrm{x_t^{vision}}, \mathrm{x_t^{physio}}\} \tag{1}$$

They correspond to text semantic features, speech acoustic features, visual expression features and physiological signal features. The individual attributes of users are represented as Pu, including static attributes, historical interaction preferences, behavior profiles and individual embedding vectors. An interaction sequence of length T can be expressed as follows:

$$\mathcal{X} = \{\mathrm{X_1}, \mathrm{X_2}, \ldots, \mathrm{X_T}\} \tag{2}$$

The model needs to learn from $(\mathcal{X}, \mathrm{P_u})$ to a sequence of emotional states $\mathcal{Y} = \{\mathrm{y_1}, \mathrm{y_2}, \ldots, \mathrm{y_T}\}$, that is:

$$\mathcal{F}: (\mathcal{X}, \mathrm{P_u}) \rightarrow \mathcal{Y} \tag{3}$$

Among them,y_t can be represented as a discrete emotion category, a continuous emotion intensity, or an emotional state vector described by a two-dimensional space of valence-arousal. Furthermore, in order to reflect the trend of emotion change, the model also needs to output the dynamic transition result $\Delta y_t = y_t - y_{t-1}$ between adjacent moments, so as to jointly describe the direction, range and stage fluctuations of emotion evolution. Therefore, the task of this paper can be summarized as a sequential deep learning problem of "multimodal sequence input-individual attribute constraints-dynamic emotional state output".

On this basis, this paper constructs the overall framework of personalized emotion dynamic deep learning for human-centered emotional interaction (as shown in Figure 1). The framework is composed of data input layer, feature representation layer, dynamic modeling layer, personalized adaptation layer and output layer. The relationship between each layer is not linear stacking, but a progressive computing link is formed around "perception, representation, evolution, adaptation, output". The data input layer is responsible for receiving multi-modal emotion data and user individual attribute information, and completes preprocessing operations

such as time synchronization, missing completion, scale normalization and sequence slicing, which provides a unified input interface for subsequent modeling. The feature representation layer uses heterogeneous encoders to extract semantic, acoustic, visual and physiological features for different modalities, and completes modal alignment and fusion through a shared representation space or cross-modal attention mechanism to form a moment-level joint emotion representation. The dynamic modeling layer takes the fusion feature sequence as input, and uses recurrent neural network, long short-term memory network, gated recurrent unit or Transformer module to learn the context dependence and emotion transfer law, so that the model can capture local fluctuations and long-term dependence at the same time. The personalized adaptation layer introduces user portrait, individual embedding and historical emotion pattern into the state update process. Through conditional modulation, gated weight allocation or parameter bias correction, the same emotional cue is interpreted in different ways for different users. The output layer completes emotion category discrimination, intensity estimation and dynamic trend prediction, and provides a computable emotional state basis for subsequent interactive response strategy generation.

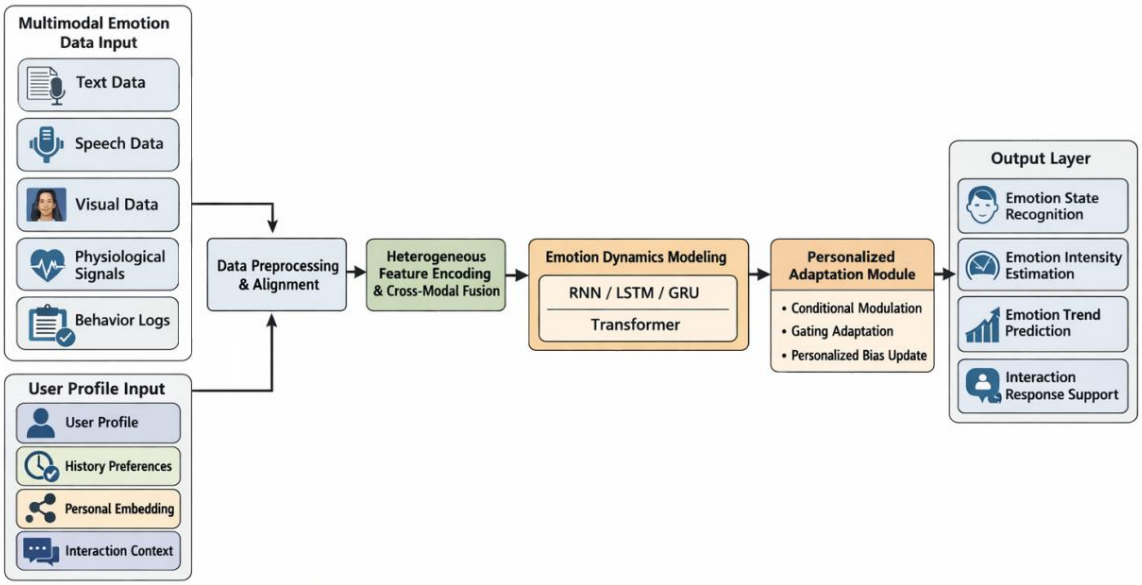

Figure 1: Overall framework diagram of personalized emotion dynamic deep learning for human-centered emotional interaction.

3.2 Multimodal Emotion Data Representation and Preprocessing

Personalized emotion dynamic modeling for human-centered emotional interaction does not simply concatenate multi-source data such as text, speech, expression and interactive behavior, but aims to establish a unified, computable, and alignable multimodal emotion representation space around the emotional expression mechanism of users in the continuous interaction process. Due to the significant differences in sampling frequency, data structure, noise distribution and semantic bearing way between different modalities, the original data cannot be directly entered into the deep learning model. It must go through a series of preprocessing steps such as collection, cleaning, time alignment, feature encoding and sample construction to form the standard input suitable for personalized emotion dynamic modeling. Let the original multimodal data set of user u during the interaction period be as follows.

$$\mathcal{D}^{(\mathrm{u})} = \left\{\mathcal{D}^{\mathrm{txt}}, \mathcal{D}^{\mathrm{aud}}, \mathcal{D}^{\mathrm{vis}}, \mathcal{D}^{\mathrm{beh}}\right\} \tag{4}$$

where $\mathcal{D}^{\mathrm{txt}}$ represents the text sequence, $\mathcal{D}^{\mathrm{aud}}$ represents the speech signal, $\mathcal{D}^{\mathrm{vis}}$ represents the expression image sequence, and $\mathcal{D}^{\mathrm{beh}}$ represents the interaction behavior log. For any mode m, the original observation can be written as follows.

$$\mathcal{D}^{\mathrm{m}} = \{(\mathrm{o}_{\mathrm{i}}^{\mathrm{m}}, \tau_{\mathrm{i}}^{\mathrm{m}})\}_{\mathrm{i}=1}^{\mathrm{N_m}} \tag{5}$$

where $\mathrm{o}_{\mathrm{i}}^{\mathrm{m}}$ is the *i* th mode observation, $\tau_{\mathrm{i}}^{\mathrm{m}}$ is the corresponding timestamp, and $\mathrm{N_m}$ is the number of samples of this mode. This representation preserves the asynchrony of the original sampling of each modality and provides the basis for subsequent temporal alignment.

In the data acquisition phase, the text modality is mainly derived from chat records, voice transcribed text and Q&A input, and its original form can be expressed as a sequence of lemma:

$$\mathrm{s}_{\mathrm{i}}^{\mathrm{txt}} = [\mathrm{w}_{\mathrm{i},1}, \mathrm{w}_{\mathrm{i},2}, \ldots, \mathrm{w}_{\mathrm{i},\mathrm{L_i}}] \tag{6}$$

where $\mathrm{L_i}$ is the length of the *i* th text. Speech modes are collected through microphones to form discrete time signals:

$$\mathrm{x}^{\mathrm{aud}}(\mathrm{n}), \quad \mathrm{n} = 0,1, \ldots, \mathrm{N}-1 \tag{7}$$

The visual modality comes from the sequence of facial expression frames captured by the camera, denoted as:

$$\mathcal{V} = \{\mathrm{I}_1, \mathrm{I}_2, \ldots, \mathrm{I}_{\mathrm{T}}\}, \quad \mathrm{I}_{\mathrm{t}} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{C}} \tag{8}$$

The interactive behavior mode is composed of events such as click, pause, scroll, response delay, input frequency and interface switch, which can be expressed as follows:

$$\mathcal{D}^{\mathrm{beh}} = \{(\mathrm{e_j}, \tau_{\mathrm{j}}, \rho_{\mathrm{j}})\}_{\mathrm{j}=1}^{\mathrm{N_b}} \tag{9}$$

where $\mathrm{e_j}$ is the behavior event type, τ_{j} is the event time, and ρ_{j} is the attribute value associated with the event. Different from static emotion recognition, the behavior log is retained here to incorporate the user's operational rhythm and feedback way in the interaction process into the emotion modeling, so that the model is closer to the "human-centered" emotional interaction scenario.

Since raw data usually have missing, noise, outliers, and duplicate records, data cleaning must be done before unified modeling. For continuous numerical features, the anomaly detection strategy with standard deviation normalization is adopted:

$$\mathrm{z_i} = \frac{\mathrm{x_i} - \mu_{\mathrm{x}}}{\sigma_{\mathrm{x}}} \tag{10}$$

where μ_{x} and σ_{x} denote the sample mean and standard deviation, respectively. If $|\mathrm{zi}|>\delta$, the observation is considered as an outlier and corrected or eliminated. For missing values in the time series, linear imputation based on adjacent moments is used:

$$\hat{\mathrm{x}}_{\mathrm{t}} = \mathrm{x}_{\mathrm{t_1}} + \frac{\mathrm{t}-\mathrm{t_1}}{\mathrm{t_2}-\mathrm{t_1}}(\mathrm{x}_{\mathrm{t_2}} - \mathrm{x}_{\mathrm{t_1}}), \quad \mathrm{t_1} < \mathrm{t} < \mathrm{t_2} \tag{11}$$

For speech signals and behavioral statistics with strong high-frequency noise, moving average is introduced to suppress random fluctuations:

$$\tilde{x}_t = \frac{1}{2k+1}\sum_{j=-k}^{k} x_{t+j} \tag{12}$$

In order to eliminate the influence of different dimensions on model training, the continuous features are further normalized and transformed:

$$x_t^* = \frac{x_t - \mu}{\sigma + \varepsilon} \tag{13}$$

where ε is the smoothing factor. The above processing is not a simple data collation, but to ensure the stable convergence of the subsequent deep model in the multi-modal joint training, and reduce the interference of individual modal noise on the emotional state estimation.

After the cleaning is completed, the temporal alignment problem caused by multimodal asynchronous sampling needs to be solved. Text is usually represented by sentences or turns, speech is sampled by continuous waveforms, expression is represented by video frame sequences, and behavior data is represented by discrete event streams. Without establishing a unified time reference, it is difficult to jointly model the emotional state at the same time. In this paper, the continuous interaction process is divided into time segments of length Δ by the combination of fixed time window and aggregation within window:

$$\mathcal{T} = \{[0, \Delta), [\Delta, 2\Delta), \dots, [(K-1)\Delta, K\Delta)\} \tag{14}$$

For any modality m, the aggregate representation in the KTH time window is defined as follows:

$$x_k^m = \mathrm{Agg}(\{o_i^m \mid \tau_i^m \in [(k-1)\Delta, k\Delta)\}) \tag{15}$$

where Agg(·) denotes the window aggregation operator. The operator has different implementations in different modalities: text uses context-encoded sentence vector aggregation, speech uses frame-level feature pooling, vision uses expression frame time series aggregation, and behavior log calculates event frequency, average length of stay and time interval statistics. For modalities missing in a window, a missing mask is introduced:

$$m_k^m = \begin{cases} 1, & \text{Modal existence,} \\ 0, & \text{Missing mode,} \end{cases} \tag{16}$$

To avoid interpreting modal miscues as low-intensity emotion signals in the fusion stage.In the feature encoding stage, the core task of text modality is to extract emotion cues at the semantic level. For the text sequence in window k:

$$s_k = [w_1, w_2, \dots, w_{L_k}] \tag{17}$$

Let's map it to an embedding matrix:

$$E_k = [e_1, e_2, \dots, e_{L_k}], \quad e_i \in \mathbb{R}^{d_t} \tag{18}$$

Then the hidden state sequence is obtained by the context encoder:

$$H_k^{txt} = f_{txt}(E_k) \tag{19}$$

Considering that emotional expressions usually focus on a few key components, attention pooling is introduced to generate window-level representations:

$$\alpha_i = \frac{\exp(q^\top h_i)}{\sum_{j=1}^{L_k} \exp(q^\top h_j)}, \quad z_k^{txt} = \sum_{i=1}^{L_k} \alpha_i h_i \tag{20}$$

For speech modality, more emphasis is placed on time-frequency structure and rhythm variation. After framing and windowing the original speech signal, the short-time Fourier transform is used to calculate the frequency domain representation:

$$X_r(v) = \sum_{n=0}^{L_f-1} x_r(n) e^{-j2\pi vn/L_f} \tag{21}$$

The log-Mel spectrum is further constructed as follows:

$$S_r^{mel}(m) = \log\left(\sum_v |X_r(v)|^2 H_m(v) + \varepsilon\right) \tag{22}$$

The speech encoder output is denoted as:

$$Z_k^{aud} = f_{aud}(S_k^{mel}), \quad z_k^{aud} = \frac{1}{T_k}\sum_{r=1}^{T_k} Z_{k,r}^{aud} \tag{23}$$

The visual modality is centered around the facial expression image. The standardized face sequence is obtained by face detection and region cropping, and then each frame is fed into the visual encoder:

$$u_t = f_{vis}(F_t), \quad u_t \in \mathbb{R}^{d_v} \tag{24}$$

Due to the significant temporal continuity of expression changes, the window-level visual representation uses weighted aggregation:

$$\beta_t = \frac{\exp(g^\top u_t)}{\sum_{j=1}^{T_k} \exp(g^\top u_j)}, \quad z_k^{vis} = \sum_{t=1}^{T_k} \beta_t u_t \tag{25}$$

The interactive behavior mode describes the user interaction rhythm through event embedding and time interval coding. Let the JTH event in window k be encoded as follows:

$$b_{k,j} = [\phi(e_{k,j}), \Delta\tau_{k,j}, \rho_{k,j}] \tag{26}$$

where $\phi(e_{k,j})$ is the action type embedding, $\Delta\tau_{k,j}$ is the time interval, and $\rho_{k,j}$ is the event attribute. After feeding the behavior sequence into the behavior encoder, we obtain:

$$B_k = f_{beh}([b_{k,1}, b_{k,2}, \ldots, b_{k,M_k}]) \tag{27}$$

And generate the behavior representation in a pooling fashion:

$$z_k^{beh} = \left[\max(B_k); \frac{1}{M_k} \sum_{j=1}^{M_k} B_{k,j} \right] \tag{28}$$

In the sample construction stage, it is necessary to map the window-level features of each modality into a unified time-series modeling input. The joint representation of the KTH time window is defined as follows:

$$x_k = \left[z_k^{txt}; z_k^{aud}; z_k^{vis}; z_k^{beh}; m_k^{txt}; m_k^{aud}; m_k^{vis}; m_k^{beh} \right] \tag{29}$$

Considering that the title of this paper emphasizes "Modeling Personalized emotion dynamics", the input should not only contain the modal feature sequence, but also the user individual attribute vector pu. Let the sliding window length be L, then the ith training sample can be expressed as follows.

$$X_i = \{x_{i-L+1}, x_{i-L+2}, \ldots, x_i\}, \quad \mathcal{S}_i = (X_i, p_u, Y_i) \tag{30}$$

where Yi={yi−L+1,yi−L+2,...,yi} is the corresponding emotional state label sequence. If sentiment trend prediction is further considered, dynamic labels can be constructed as follows:

$$\Delta y_t = y_t - y_{t-1}, \quad \widetilde{Y}_i = (Y_i, \Delta Y_i) \tag{31}$$

After the above processing, the original multi-modal heterogeneous data is converted into a serialized input with a unified time scale, a unified numerical space and a unified sample structure, which not only solves the basic problems of collection, cleaning, alignment and coding, but also provides a strict data representation basis for the subsequent emotional dynamic modeling layer and personalized adaptation layer. Different from the preprocessing methods that only serve static recognition, the design of this process is always centered on the continuous interaction process of users, so that the data representation itself serves the "human-centered" personalized emotion dynamic learning goal.

3.3 Multi-modal emotion feature encoding

After multi-modal data preprocessing and sample construction, it is necessary to further map heterogeneous inputs such as text, speech, expression images and interactive behaviors into a unified emotion representation space to support subsequent emotion dynamic modeling and personalized adaptation. Let the four types of modal inputs in the KTH time window be X_k^{txt},X_k^{aud},X_k^{vis} andX_k^{beh} respectively, then the basic objective of modal feature coding can be expressed as follows.

$$z_k^m = f_m(X_k^m; \theta_m), \quad m \in \{txt, aud, vis, beh\} \tag{32}$$

where $f_m(\cdot)$ is the corresponding modal encoder, and zkm is the high-level feature at the modal level. Text modality mainly carries semantic polarity, sentiment words and context dependence information. After embedding the text sequence, the following is obtained:

$$E_k = [e_1, e_2, \ldots, e_L] \tag{33}$$

Then we use Transformer for context encoding:

$$H_k^{txt} = \text{Transformer}(E_k + P_k) \tag{34}$$

In order to highlight key emotional words and emotional segments, attention pooling is introduced to obtain the text representation:

$$z_k^{txt} = \sum_{i=1}^{L} \alpha_i^{txt} h_i, \quad \alpha_i^{txt} = \frac{\exp(q^\top h_i)}{\sum_j \exp(q^\top h_j)} \tag{35}$$

Speech modality focuses on extracting acoustic cues such as pitch, energy, rhythm and prosody. Log Mel spectrogram is input into convolutional network to extract local time-frequency features:

$$C_k^{aud} = \sigma(\text{Conv}(X_k^{aud})) \tag{36}$$

Then we use bidirectional GRU to model the time dependence:

$$H_k^{aud} = \text{BiGRU}(C_k^{aud}) \tag{37}$$

Finally, temporal attention is used to obtain the window-level acoustic representation:

$$z_k^{aud} = \sum_{t=1}^{T} \alpha_t^{aud} h_t \tag{38}$$

This "convolution + loop" method can simultaneously retain local acoustic patterns and cross-moment emotional change information. Visual modality is mainly oriented to facial expression texture and short-term dynamic changes. Let the sequence of face frames in the window be {F1,F2,... ,FT}, the frame-level features are first extracted from the convolution backbone:

$$v_t = \phi_{vis}(F_t) \tag{39}$$

Then the temporal Transformer is used to model the expression evolution relationship:

$$H_k^{vis} = \text{Transformer}_{vis}([v_1, v_2, \dots, v_T]) \tag{40}$$

The visual representation is obtained by frame-level weighted aggregation:

$$z_k^{vis} = \sum_{t=1}^{T} \alpha_t^{vis} s_t \tag{41}$$

Interactive behavior modality is used to describe the implicit emotional cues such as click frequency, dwell time, input delay and interface switching. Construct the encoding for the JTH event:

$$r_j = [\psi(e_j); \eta(\Delta\tau_j); \rho_j] \tag{42}$$

And feed it into the behavior encoder to obtain the sequence representation:

$$H_k^{beh} = f_{beh}([r_1, r_2, \dots, r_M]) \tag{43}$$

To enhance the expression of key behavioral events, gated pooling is used to generate behavioral features:

$$z_k^{beh} = \sum_{j=1}^{M} \gamma_j \odot q_j \tag{44}$$

In order to realize cross-modal joint modeling, this paper projects the modal features into a shared space:

$$h_k^m = W_m z_k^m + b_m \tag{45}$$

On this basis, a unified emotion representation is constructed by gated fusion:

$$h_k^{fusion} = \sum_m g_k^m \odot h_k^m, \quad g_k^m = \sigma(W_m^g h_k^m + b_m^g) \tag{46}$$

Therefore, different modalities not only maintain their respective feature advantages, but also complete collaborative expression in a unified semantic space, which provides stable input for subsequent emotional dynamic modeling. The encoding of multimodal emotion features and the construction of unified representation space are shown in Figure 2.

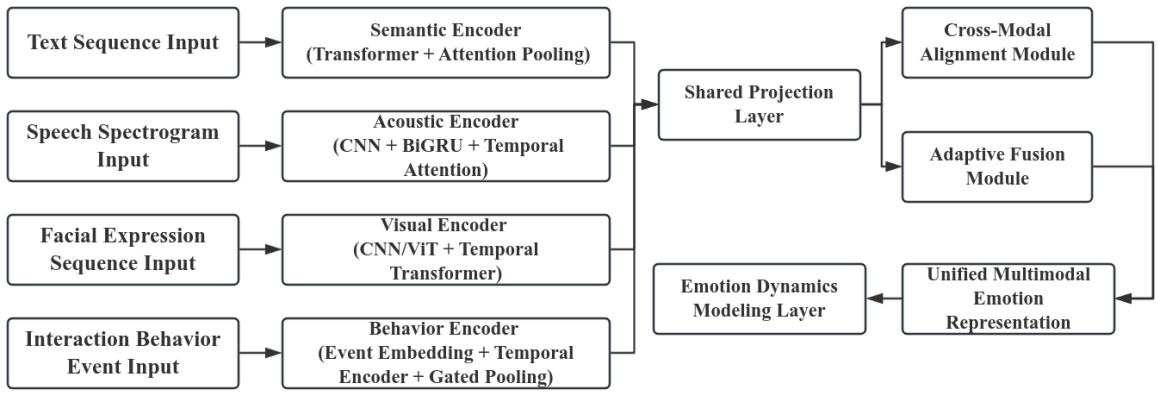

Figure 2: Construction diagram of multi-modal emotion feature encoding and unified representation space.

3.4 Temporal Modeling of emotional dynamics

After the construction of multimodal unified representation space, the key of emotion modeling has shifted from "feature extraction" to "state evolution characterization". In the process of real emotional interaction, the user's emotion is not generated discrete and isolated, but is jointly affected by preorder semantic stimuli, interactive feedback, behavioral inertia and stage events, showing obvious characteristics of volatility, persistence and stage conversion. If only the fusion features of a single time window are used for discrimination, the model is easy to ignore the emotional cumulative effect and lag effect, resulting in insufficient recognition of implicit turning points, short-term fluctuations and continuous states. To this end, this paper formulates

emotion dynamics modeling as a sequential state update problem. Let the unified multimodal representation of the KTH time window be $\mathbf{h}_k^{fusion} \in \mathbb{R}^d$ and an interaction sequence of length T be expressed as follows:

$$\mathrm{H} = [\mathrm{h}_1^{\mathrm{fusion}}, \mathrm{h}_2^{\mathrm{fusion}}, \ldots, \mathrm{h}_\mathrm{T}^{\mathrm{fusion}}] \tag{47}$$

The model goal is to learn a mapping function:

$$\Phi: \mathrm{H} \rightarrow \mathrm{Y} = [\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_\mathrm{T}] \tag{48}$$

where yt represents the emotional state vector at time t, which can correspond to discrete class probabilities or continuous emotion intensities. Considering that emotional changes contain both short-range dependence and cross-stage continuation relationships, this paper adopts a dynamic modeling method of "gated loop modeling + temporal attention enhancement". For the input sequence H, the long short-term memory network is used to complete the recursive state update:

$$\begin{aligned} \mathrm{f}_\mathrm{t} &= \sigma(\mathrm{W}_\mathrm{f}[\mathrm{h}_\mathrm{t}^{\mathrm{fusion}}, \mathrm{s}_{\mathrm{t}-1}] + \mathrm{b}_\mathrm{f}), \\ \mathrm{i}_\mathrm{t} &= \sigma(\mathrm{W}_\mathrm{i}[\mathrm{h}_\mathrm{t}^{\mathrm{fusion}}, \mathrm{s}_{\mathrm{t}-1}] + \mathrm{b}_\mathrm{i}) \end{aligned} \tag{49}$$

$$\begin{aligned} \tilde{\mathrm{c}}_\mathrm{t} &= \tanh(\mathrm{W}_\mathrm{c}[\mathrm{h}_\mathrm{t}^{\mathrm{fusion}}, \mathrm{s}_{\mathrm{t}-1}] + \mathrm{b}_\mathrm{c}), \\ \mathrm{c}_\mathrm{t} &= \mathrm{f}_\mathrm{t} \odot \mathrm{c}_{\mathrm{t}-1} + \mathrm{i}_\mathrm{t} \odot \tilde{\mathrm{c}}_\mathrm{t} \end{aligned} \tag{50}$$

$$\begin{aligned} \mathrm{o}_\mathrm{t} &= \sigma(\mathrm{W}_\mathrm{o}[\mathrm{h}_\mathrm{t}^{\mathrm{fusion}}, \mathrm{s}_{\mathrm{t}-1}] + \mathrm{b}_\mathrm{o}), \\ \mathrm{s}_\mathrm{t} &= \mathrm{o}_\mathrm{t} \odot \tanh(\mathrm{c}_\mathrm{t}) \end{aligned} \tag{51}$$

where ft, it and ot denote forgetting gate, input gate and output gate respectively, ct is memory cell and st is hidden state. The structure can retain the previous emotional traces in the sequence recursion, and suppress the interference of irrelevant disturbances on the current state, so as to enhance the expression ability of the emotion persistence and slow migration phenomenon.

It is still difficult to fully capture long-distance dependencies and key turning segments by only relying on the recursive structure, so the temporal self-attention mechanism is further introduced on the hidden state sequence S=[s1,s2,...,sT]. For any time tt, construct the query, key, and value vectors:

$$\mathrm{q}_\mathrm{t} = \mathrm{W}_\mathrm{q}\mathrm{s}_\mathrm{t}, \quad \mathrm{k}_\mathrm{j} = \mathrm{W}_\mathrm{k}\mathrm{s}_\mathrm{j}, \quad \mathrm{v}_\mathrm{j} = \mathrm{W}_\mathrm{v}\mathrm{s}_\mathrm{j} \tag{52}$$

The dependency weight of time t on historical time j is defined as follows:

$$\mathrm{a}_{\mathrm{t,j}} = \frac{\exp\left(\mathrm{q}_\mathrm{t}^\mathrm{T}\mathrm{k}_\mathrm{j}/\sqrt{\mathrm{d}_\mathrm{a}}\right)}{\sum_{\mathrm{r}=1}^{\mathrm{T}} \exp\left(\mathrm{q}_\mathrm{t}^\mathrm{T}\mathrm{k}_\mathrm{r}/\sqrt{\mathrm{d}_\mathrm{a}}\right)} \tag{53}$$

This gives us the enhanced context state:

$$.\hat{\mathrm{s}}_\mathrm{t} = \sum_{\mathrm{j}=1}^{\mathrm{T}} \mathrm{a}_{\mathrm{t,j}}\, \mathrm{v}_\mathrm{j} \tag{54}$$

This mechanism can directly establish the emotional association between distant Windows, so that the model can identify cross-stage emotional trajectories such as "early depression - middle volatility - late relaxation", without being obviously limited by the length of recursive transmission.

In order to further characterize the direction and magnitude of emotional changes, this paper adds emotional incremental modeling in addition to state prediction. Let the current emotion output be:

$$y_t = \mathrm{softmax}(W_y \hat{s}_t + b_y) \tag{55}$$

Then the emotion change of adjacent moments is expressed as follows:

$$\Delta y_t = y_t - y_{t-1} \tag{56}$$

When $\|\Delta y_t\|_2$ is large, it indicates that there is an obvious transition of emotion at this stage. When its value is small and remains stable continuously, it indicates that the emotional state is persistent. In order to make the model learn state recognition and dynamic trend at the same time, a joint loss function is constructed:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\mathrm{cls}} + \lambda_2 \mathcal{L}_{\mathrm{dyn}} \tag{57}$$

Among them:

$$\mathcal{L}_{\mathrm{cls}} = -\sum_{t=1}^{T}\sum_{c=1}^{C} y_{t,c}^{*} \log y_{t,c},$$
$$\mathcal{L}_{\mathrm{dyn}} = \sum_{t=2}^{T} \|\Delta y_t - \Delta y_t^{*}\|_2^2 \tag{58}$$

Here, Lcls is used to constrain emotion category discrimination, Ldyn is used to constrain emotion change trend fitting, and λ1 and λ2 are weight coefficients. Through this modeling method, the model no longer only pays attention to "what emotion is" at a certain moment, but can learn "how emotion changes, whether the change lasts, and where the turning point occurs" at the same time, so as to more accurately describe the user's emotional dynamic trajectory during continuous interaction. The dynamic temporal modeling framework for emotion change trajectories is shown in Figure 3.

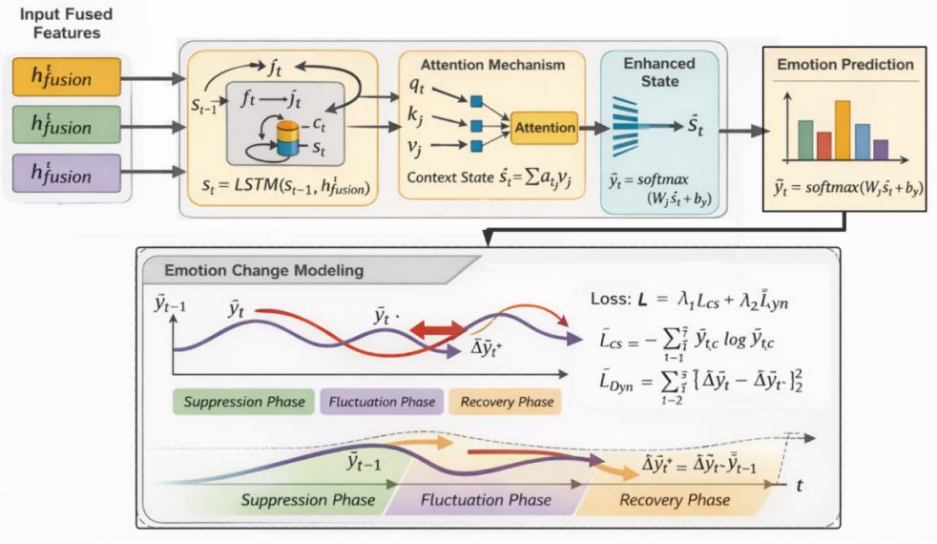

Figure 3: Framework diagram of dynamic temporal modeling oriented to emotional change trajectories.

3.5 Modeling Individual Differences and Personalized Adaptation

Based on the dynamic time series modeling of emotion, we further introduce the user's historical behavior, interaction preference and individual attribute information to construct the expression mechanism for individual differences. The reason is that the expression of the same emotional state on different users is not consistent. Some users are more inclined to release emotions through semantic expression, while some users are more likely to reflect state fluctuations in speech prosody, expression changes or interaction rhythm. If the unified parameter space is still used to model the emotions of all users, the model can learn the common rules at the group level, but its ability to describe the subject heterogeneity, long-term usage habits and individual response thresholds is limited, which affects the recognition accuracy and adaptation ability of the emotional interaction system.

Let the individual attribute vector of user u be represented as follows:

$$p_u = [a_u; b_u; c_u] \in \mathbb{R}^{d_p} \tag{59}$$

where au represents static attribute features, bu represents historical behavior statistical features, and cu represents interaction preference embeddings. The static attributes can include stable information such as age level, usage time preference, and device environment. The historical behavior characteristics can be composed of average response time, click frequency, stay time and emotional transfer statistics. To enable these heterogeneous individual information to enter the unified modeling process, the user embedding representation is first obtained through the mapping layer:

$$e_u = \tanh(W_p p_u + b_p) \tag{60}$$

where $e_u \in$ Rde is the compressed individual representation vector. On the one hand, this vector retains the difference information of users, and on the other hand, it avoids the disturbance of the model training caused by the high dimension of the original attribute.

In the phase of temporal dynamic modeling, this paper injects user embeddings as conditional information into the emotional state update process. For the enhanced emotional state s_t at time t, the conditional modulation method is used to realize personalized correction:

$$\tilde{s}_t = \gamma_u \odot \hat{s}_t + \beta_u \tag{61}$$

Among them:

$$\gamma_u = \sigma(W_\gamma e_u + b_\gamma), \quad \beta_u = W_\beta e_u + b_\beta \tag{62}$$

Here, γ_u denotes the user-dependent scale adjustment vector and β_u denotes the bias correction vector. The meaning of this process is that different users can apply different scaling and offset to the same emotional dynamic state, so that the model can adjust the interpretation way of the state according to the subject differences while maintaining the general dynamic law. Compared with the simple concatenation of individual attributes, this conditional modulation method is more suitable for embedding in the middle layer of the deep network, and can directly act on the emotional state space.

In order to further improve the responsiveness to user differences, a personalized gating mechanism is added to the fusion feature layer. Let the multimodal fusion at the current time be denoted as htfusion, then the user-conditional gating vector is defined as follows:

$$g_{u,t} = \sigma\left(W_g[h_t^{fusion}; e_u] + b_g\right) \tag{63}$$

The corresponding personalized features are expressed as follows:

$$r_{u,t} = g_{u,t} \odot h_t^{fusion} \tag{64}$$

This mechanism can dynamically adjust the effective contribution of different modal features at the current time according to the user embedding, so that the model can form differentiated responses to different users when faced with similar inputs. For example, for users who rely more on speech to express their emotions, the gating mechanism can improve the weight of acoustic features. For users with more discriminative behavior patterns, the effect of behavior modes can be strengthened.

Finally, the personalized emotion output is expressed as follows:

$$y_{u,t} = \mathrm{softmax}\left(W_y[\tilde{s}_t; r_{u,t}] + b_y\right) \tag{65}$$

Therefore, the model receives both the general temporal emotional state information and the personalized modulated user feature information, realizing the transition from "group commonality modeling" to "user condition modeling". This design not only improves the model's ability to describe the differences in emotional expression of different users, but also provides a more fine-grained individual emotion basis for subsequent interaction response generation, thereby enhancing the adaptive learning and personalized recognition ability for human-centered emotional interaction scenarios.

4 Experimental design and result analysis

4.1 Dataset and Experimental environment

In order to ensure that the experimental design is consistent with the research goals of personalized emotion dynamic deep learning modeling for human-centered emotional interaction, this paper selects MELD and CMU-MOSEI as the main experimental data sources. MELD contains text, voice and video information in multi-party dialogue scenarios. It can

better reflect the context dependence and dynamic fluctuation of emotion in the process of continuous interaction. CMU-MOSEI covers text, audio and visual modalities, and has a large sample size, which is suitable for verifying the effect of multi-modal deep representation learning and cross-modal fusion. Based on these two types of data, we further organize the interactive statistical characteristics such as round length, speech interval, duration and response frequency, and combine the user's historical behavior and preference information to construct individual attribute vectors, so that the experiment can not only test the ability of emotional dynamic modeling, but also support personalized adaptation analysis. In terms of sample composition, each sample is composed of a multi-modal input sequence under a continuous time window, the corresponding emotion label sequence and the user individual attribute vector. The multi-modal input is used to describe the emotional evolution trajectory, the label sequence is used to supervise the emotion recognition and trend prediction, and the individual attribute is used to realize the user difference modeling. The data is divided into three parts: training set, validation set and test set, and the ratio is set to 7:1:2. During the division process, it tries to ensure that the same user or the same dialogue segment does not appear in different subsets at the same time, so as to reduce the risk of information leakage and improve the credibility of the experimental results. The experimental platform is implemented based on Python 3.10 environment, PyTorch 2.0 is used for model construction and training, Transformers are used for text processing, LibROSA is used for speech feature extraction, and OpenCV is used for image preprocessing. The training process is completed in a CUDA-based GPU environment, and the batch size, learning rate, optimizer and early stop strategy are uniformly set to ensure that different comparison experiments are carried out under the same implementation conditions, so as to enhance the repeatability of model verification and the reliability of result analysis.

4.2 Comparative experimental design

In order to systematically test the effectiveness of the proposed model in emotion recognition, emotion dynamic characterization and personalized adaptation, this paper sets up comparative experiments from three dimensions of static recognition, dynamic modeling and personalized modeling. The static recognition model is mainly used to investigate the basic support ability of multi-modal emotion features for emotion classification tasks without explicitly modeling time dependence and individual differences. The dynamic modeling class is mainly used to compare the performance differences of different temporal deep learning structures in the modeling of emotion fluctuation, persistence and stage transition. The personalized modeling class is used to verify the adaptation ability of the model to the differences in emotional expression of different subjects after introducing user attributes, historical behaviors and individual embeddings. Through this hierarchical design, the source of improvement of the proposed method compared with the traditional static model, the general time series model and the personalized extended model can be analyzed more clearly.

On the baseline model setting, SVM, TextCNN and Multimodal MLP are selected as baselines in the static recognition part. Among them, SVM is used to reflect the basic performance of traditional machine learning methods in emotion recognition tasks, TextCNN is mainly used as the deep learning baseline for text emotion feature extraction, and Multimodal MLP is used to investigate the recognition effect under simple fusion methods by directly concatenating text, speech and visual features. In the dynamic modeling part, LSTM, BiGRU and Transformer are selected as comparison models. LSTM and BiGRU are used to compare the ability difference of gated loop structure in emotional temporal dependence modeling, and Transformer is used to verify the improvement of global dependence modeling for emotional dynamic analysis. In the part of personalized modeling, LSTM+User Embedding and

Transformer+Profile Fusion are selected as extended comparison models to test the influence of user-level condition information on the performance of emotion recognition. The model proposed in this paper is jointly modeled at three levels: multi-modal fusion, emotional dynamic modeling and personalized adaptation, as the final comparison object.

The experimental evaluation metrics are set according to the type of task. For the emotion classification task, Accuracy, Precision, Recall and F1-score were used to evaluate the overall recognition performance of the model. For the emotion dynamic modeling task, Macro-F1 is added to reduce the impact of class imbalance, and the ability of the model to describe the state transition is evaluated by combining the accuracy of the emotion change trend prediction. For the personalized modeling effect, the average F1 value on different user groups and the cross-user performance fluctuations are analyzed to test the stability and adaptability of the model. In order to ensure the fairness of the experiment, all models are run under the same training set, validation set and test set partition conditions, and the same optimizer, learning rate and training rounds are used.

Table 2: compares experimental model Settings with functional positioning.

Model Category	Model Name	Main Input	Modeling Characteristics	Comparison Purpose
Static Recognition	SVM	Handcrafted statistical features / concatenated features	No deep temporal modeling; serves as a traditional baseline	To verify the advantages of deep models over traditional methods
Static Recognition	TextCNN	Text features	Extracts local semantic patterns without considering temporal dependencies	To compare the effectiveness of deep textual feature extraction
Static Recognition	Multimodal-MLP	Text + speech + vision	Directly concatenates multimodal features for classification	To evaluate the performance limit of simple fusion methods
Dynamic Modeling	LSTM	Multimodal sequential features	Uses gated memory to model short- and medium-term temporal dependencies	To compare the effectiveness of basic temporal modeling
Dynamic Modeling	BiGRU	Multimodal sequential features	Models contextual dependencies bidirectionally with relatively fewer parameters	To compare differences among recurrent structures
Dynamic Modeling	Transformer	Multimodal sequential features	Models global dependencies based on self-attention	To evaluate the capability of long-range dependency modeling
Personalized Modeling	LSTM + User Embedding	Multimodal features + user embeddings	Introduces individual representations into the temporal model	To verify the contribution of user embeddings
Personalized Modeling	Transformer + Profile Fusion	Multimodal features + user profiles	Incorporates user attributes into global modeling	To compare the effect of profile fusion
Proposed Model	Proposed Model	Multimodal features + individual attributes + dynamic states	Jointly models multimodal fusion, emotion dynamics, and personalized adaptation	To verify the overall effectiveness of the proposed method

Through the above comparative experimental design, the performance improvement of the proposed model can be decomposed and analyzed from three levels of basic recognition ability, temporal dynamic modeling ability and individual user adaptation ability, which provides a clear basis for the discussion of subsequent experimental results.

4.3 Analysis of experimental results

In order to test the performance of the proposed model in emotion recognition, emotion dynamic characterization and personalized adaptation, this paper uses Accuracy, Recall, F1-score and MSE as evaluation indicators to compare and analyze different models. Among them, Accuracy is used to measure the overall classification accuracy, Recall is used to reflect the ability of the model to detect the real emotion categories, F1-score is used to comprehensively evaluate the balance performance of precision and recall, and MSE is used to describe the error level of the model in the estimation of emotion intensity and fitting of dynamic trends. The experimental results are shown in Figure 4.

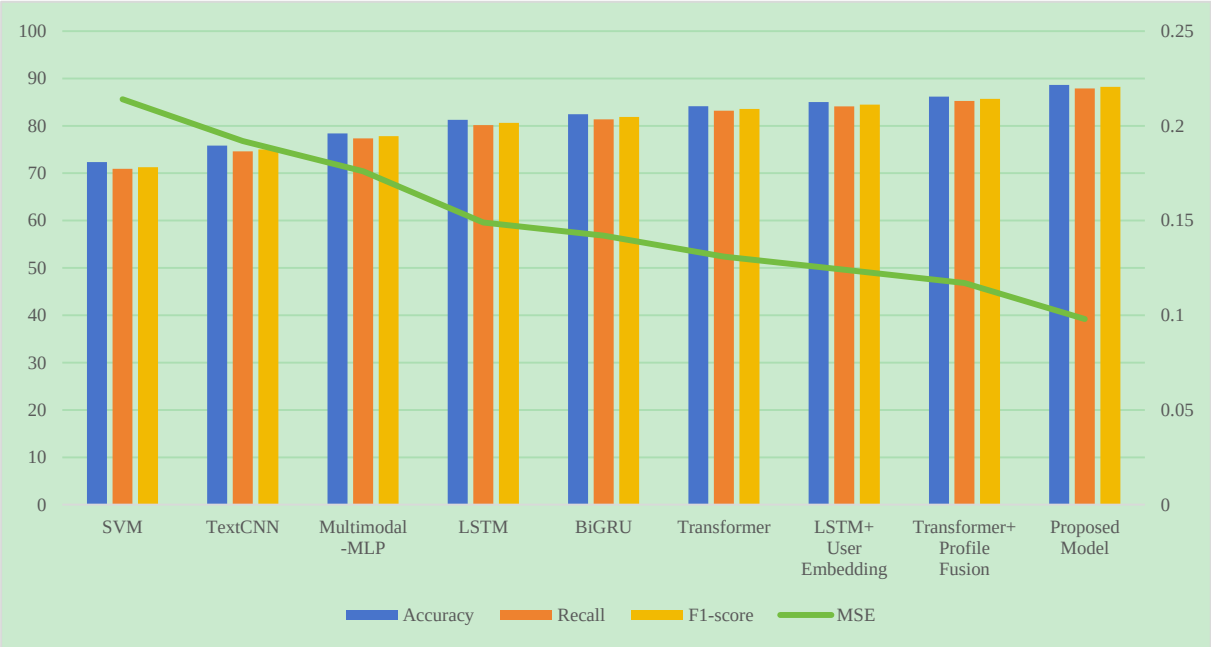

Figure 4: Comparison of experimental results of different models.

From the results distribution, the proposed model maintains optimal performance in four indicators, indicating that it has good performance in emotion recognition accuracy, true category coverage ability, comprehensive balance performance, and fitting error control of continuous emotion changes. The experimental results show that the joint introduction of multi-modal feature coding, emotion dynamic time series modeling and personalized adaptation mechanism makes the model achieve a higher level of performance in emotion recognition, emotion dynamic characterization and individual difference adaptation.

4.4 Analysis of ablation experiments

In order to further verify the contribution of multimodal fusion module, emotional dynamic modeling module and personalized adaptation module to the overall model performance, this paper designs an ablation experiment based on the complete model. Specifically, three pruning versions were set. First, the multi-modal fusion module was removed, and only a single main modal feature was retained for emotion recognition. Secondly, the dynamic modeling module

is removed, and the evolution process of emotional states in continuous time Windows is no longer explicitly modeled. Third, the personalized adaptation module is removed, and the user's historical behavior, interaction preference and individual attribute information are no longer introduced. By comparing with the full model, the actual role of each component in emotion recognition, dynamic characterization and individual adaptation can be more clearly identified. The experiment still uses Accuracy, Recall, F1-score and MSE as evaluation indexes, and the results are shown in Figure 5.

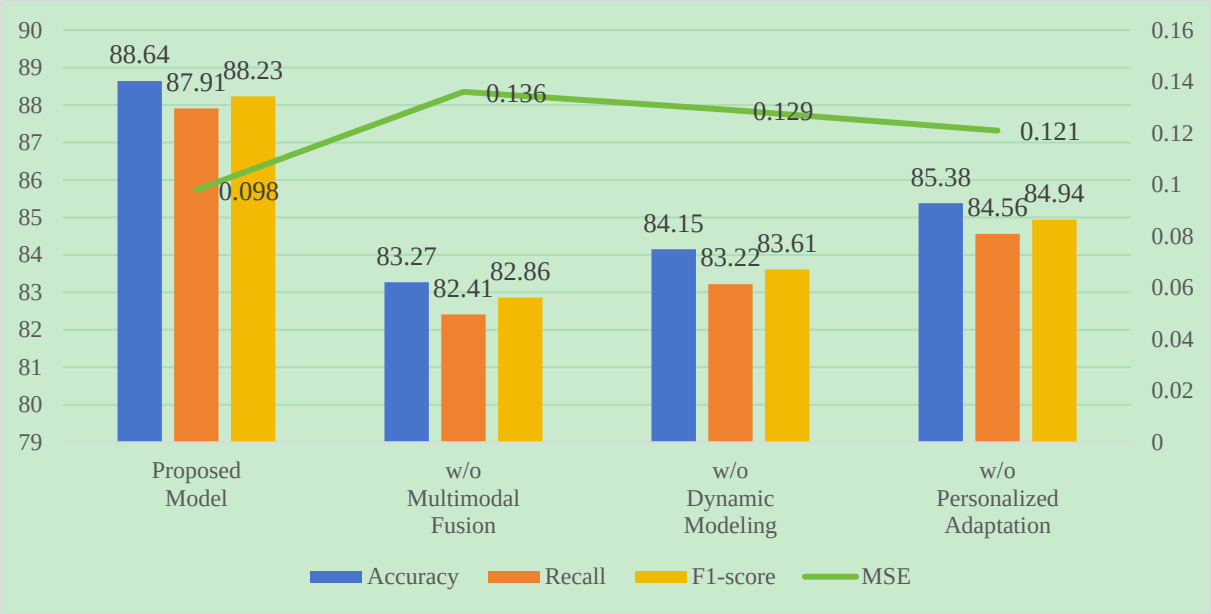

Figure 5: Comparison of the results of ablation experiments.

It can be seen from Figure 5 that the multimodal fusion module, dynamic modeling module and personalized adaptation module respectively correspond to the key functions of the three levels of "multi-source emotional information integration", "emotional change trajectory description" and "adaptive learning of user differences". Together, these three factors constitute the main source of the performance improvement of the proposed model. Among them, the multimodal fusion module contributed the most, followed by the dynamic modeling module, and the personalized adaptation module further enhanced the adaptation ability of the model to human-centered emotional interaction scenarios.

4.5 Discussion of results

Combined with the above experimental results, it can be seen that the advantages of the proposed model in emotional interaction support are mainly reflected in three aspects. Firstly, multi-modal feature coding improves the model's ability to comprehensively utilize text semantics, speech prosody, expression changes and interactive behavior cues, so that the system no longer relies on a single signal to judge the emotional state, thereby enhancing the recognition stability in complex interactive environments. Secondly, the temporal modeling of emotion dynamics can capture the fluctuation, persistence and periodic changes of user emotions, so that the model can not only determine "what is the current emotion", but also analyze "how does the emotion change", which is particularly important for continuous dialogue and long-term interaction scenarios. Thirdly, the personalized adaptation mechanism introduces the user's historical behavior, interaction preference and individual attributes into the modeling process, so that the model can further approach the real expression of different

users on the basis of the group common law, so as to improve the pertinency and adaptation of emotional interaction. At the same time, the experimental results also show that the model still has some limitations. Due to the combined effect of multi-modal input, dynamic modeling and personalized modulation, the model parameter scale and training cost are relatively high, which are strongly dependent on data integrity and computing resources. The model performance may still be affected in the scenarios of modal loss, noise enhancement or user cold start. Although the individual difference modeling improves the adaptation ability, the dependence on long-term stable user profiles also means that the generalization ability of the model in new user scenarios still needs to be further strengthened. Nevertheless, the proposed model still has good application potential in intelligent companishment, educational assistance, psychological services and human-computer interaction systems. In the intelligent companion scenario, it can be used to identify the continuous emotional changes of users and adjust the feedback strategy. In the educational auxiliary scene, it can support the dynamic perception of learners 'emotional state and participation. In the psychological service scenario, it can provide auxiliary support for emotion monitoring and risk early warning. In human-computer interaction systems, it can be used to improve the system's ability to understand the user's emotional state and the naturalness of response. On the whole, the deep learning modeling of personalized emotion dynamics for human-centered emotional interaction has strong application value.

5 Conclusion

Focusing on the needs of human-centered emotional interaction, this paper constructs an emotional dynamic deep learning model that integrates multimodal feature encoding, dynamic emotional time series modeling and personalized adaptation mechanism. Starting from multi-source data such as text, speech, expression images and interactive behaviors, this study completes the overall design of unified representation, dynamic modeling and differential modulation of users, so that the model can not only recognize the user's current emotional state, but also depict the change trajectory of emotions in continuous interaction. On this basis, the adaptability of the model to the differences in emotional expression of different users is improved. The experimental results show that the proposed model is superior to the comparison models in multiple evaluation indicators, the Accuracy reaches 88.64%, the Recall reaches 87.91%, the F1-score reaches 88.23%, and the MSE is reduced to 0.098. Compared with Multimodal MLP, the Accuracy is increased by 10.23 percentage points, the F1-score is increased by 10.41 percentage points, and the MSE is reduced by 0.078. Compared with Transformer+Profile Fusion, the Accuracy is still improved by 2.46 percentage points, indicating that the proposed method has better comprehensive performance in emotion recognition, emotion dynamic characterization and personalized adaptation.

Ablation experiments further show that the multimodal fusion module, dynamic modeling module and personalized adaptation module have obvious contributions to the performance improvement. After removing the multimodal fusion module, the model Accuracy is reduced to 83.27%, and the F1-score is reduced to 83.61% after removing the dynamic modeling module. After removing the personalized adaptation module, the MSE rises to 0.121, indicating that the three types of modules together constitute the key basis for the effectiveness of the model. Future research can focus on the online emotion update mechanism in real-time interactive environment, the unified emotion understanding framework supported by multi-modal large model, and the lightweight compression and reasoning optimization method for terminal deployment, so as to improve the application efficiency and landing ability of the model in actual human-computer emotional interaction system.

References

[1] García-Hernández R A, Luna-García H, Celaya-Padilla J M, et al. A systematic literature review of modalities, trends, and limitations in emotion recognition, affective computing, and sentiment analysis[J]. Applied Sciences, 2024, 14(16): 7165.

[2] Chutia T, Baruah N. A review on emotion detection by using deep learning techniques[J]. Artificial Intelligence Review, 2024, 57(8): 203.

[3] George S M, Ilyas P M. A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise[J]. Neurocomputing, 2024, 568: 127015.

[4] Kaur M, Kumar M. Facial emotion recognition: A comprehensive review[J]. Expert Systems, 2024, 41(10): e13670.

[5] Kacimi Y, Adda M. Comprehensive review of physiological signal-based emotion recognition: methods, challenges, and insights on arousal and valence dimensions[J]. Procedia Computer Science, 2025, 257: 174-181.

[6] Wang Y, Wang B, Zhao Y, et al. Emotion recognition in conversation via dynamic personality[C]//Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). 2024: 5711-5722.

[7] Khule T, Agrawal R, Narayan A. Pfa-erc: Psuedo-future augmented dynamic emotion recognition in conversations[C]//Findings of the Association for Computational Linguistics: EMNLP 2024. 2024: 16196-16207.

[8] Fu Y, Wu J, Wang Z, et al. LaERC-S: Improving LLM-based emotion recognition in conversation with speaker characteristics[C]//Proceedings of the 31st International Conference on Computational Linguistics. 2025: 6748-6761.

[9] Zhang T, Tan Z. ECERC: evidence-cause attention network for multi-modal emotion recognition in conversation[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 2064-2077.

[10] Shen Z, Pang Y, Rao Y, et al. CoE: A clue of emotion framework for emotion recognition in conversations[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2025: 23548-23563.

[11] Han Y, Zhang P, Park M, et al. Systematic evaluation of personalized deep learning models for affect recognition[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2024, 8(4): 1-35.

[12] Li J, Washington P. A comparison of personalized and generalized approaches to emotion recognition using consumer wearable devices: Machine learning study[J]. JMIR AI, 2024, 3(1): e52171.

[13] Shi X S, She Q, Fang F, et al. Enhancing cross-subject EEG emotion recognition through multi-source manifold metric transfer learning[J]. Computers in biology and medicine, 2024, 174: 108445.

[14] Kim W, Kutsuzawa G, Maruyama M. Emotion recognition and forecasting from wearable data via cluster-guided attention with cross-species pretraining[J]. Intelligent Systems with Applications, 2025: 200560.

[15] Kovacevic N, Holz C, Gross M, et al. On multimodal emotion recognition for human-chatbot interaction in the wild[C]//Proceedings of the 26th International Conference on Multimodal Interaction. 2024: 12-21.

[16] Geetha A V, Mala T, Priyanka D, et al. Multimodal emotion recognition with deep learning: advancements, challenges, and future directions[J]. Information Fusion, 2024, 105: 102218.

[17] Hazmoune S, Bougamouza F. Using transformers for multimodal emotion recognition: Taxonomies and state of the art review[J]. Engineering Applications of Artificial Intelligence, 2024, 133: 108339.

[18] Zhang S, Yang Y, Chen C, et al. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects[J]. Expert Systems with Applications, 2024, 237: 121692.

[19] Wu Y, Mi Q, Gao T. A comprehensive review of multimodal emotion recognition: Techniques, challenges, and future directions[J]. Biomimetics, 2025, 10(7): 418.

[20] Wang J, Zhang C. Cross-modality fusion with EEG and text for enhanced emotion detection in English writing[J]. Frontiers in Neurorobotics, 2025, 18: 1529880.