BVPEC: A Cross-modal BERT-ViT Framework for Performance Emotion Recognition from Multimodal Acting Data

Yizhu Lin
School of Fashion, Dalian Polytechnic University, DaLian 116034, China
E-mail: YizhuLin@outlook.com

Keywords: Performance affective computing, cross-modal learning, BERT, vision transformer, adaptive fusion

Received: July 25, 2025

Performance emotional computing is a key technology for understanding and evaluating actors' artistic expression, and it is of great value in film and television analysis, drama education, and other fields. Aiming at the problem that traditional single-modal methods make it difficult to fully capture the rich text and visual emotional information in performances, this study innovatively proposes a performance emotional computing framework (BVPEC) based on the BERT-Vision cross-modal pre-training model. First, the framework deeply integrates the text information of script lines with the video information of actors' performances. It uses the BERT model to deal with lines' semantics and emotional tendencies. Secondly, a Vision Transformer (ViT) is used to extract visual features such as facial expressions and body movements of actors, and a cross-modal adaptive fusion mechanism is designed to achieve information complementarity between modes. Finally, experiments on public data sets (such as the LIRIS-ACCEDE emotional video set) and self-built performance clip data sets show that the BVPEC framework is significantly better than the single-modal model and traditional fusion method in emotion recognition accuracy (up to 89.7%), effectively improving the accuracy and robustness of performance emotion understanding, and providing new ideas for intelligent performing arts analysis.

Povzetek: Študija predstavi BVPEC, križno-modalni okvir za analizo čustev v igralskih nastopih, ki z BERT-om za besedilo, ViT-om za vizualne znake in adaptivno fuzijo združi dialoge, izraze in gibe, da preseže omejitve enomodalnih pristopov.

1 Introduction

In today's era of rapid digital information development, emotional computing has become one of the research hotspots in the field of artificial intelligence [1]. It aims to enable computers to recognize, understand, express, and adapt to human emotional states to achieve more humanized and intelligent human-computer interaction [2]. As an important branch of emotional computing, performance emotional computing focuses on analyzing and mining emotions contained in performing arts, which has far-reaching significance and broad application prospects [3].

Performing arts, as an ancient and charming art form, bears the rich emotional expression of human beings [4]. Whether it is stage plays, movies, or performances in various film and television works, actors convey complex and delicate emotions through body movements, expression changes, language tone, and other ways [5]. For the audience, these performance emotional messages are the key bridge to understanding the work's theme and the character's inner world and establishing emotional resonance with the performers [6]. However, traditional sentiment analysis methods rely only on a single text or visual information, and it is difficult to comprehensively and accurately capture multimodal emotional features in performances [7].

In recent years, with the vigorous development of deep learning technology, pre-trained models have achieved remarkable results in natural language processing and computer vision. Bidirectional Encoder Representations from Transformers (BERT) is a powerful pre-trained language model [8] that demonstrates excellent performance in text comprehension tasks. Vision Transformer (ViT) and its derivative architecture have promoted the revolution of visual information processing [9]. The cross-modal pre-training method combines the two and provides a new idea for solving complex multimodal problems.

In the field of performance emotional computing, performance contains rich text (such as script dialogue, lines, etc.) and visual (such as actors' facial expressions, body movements, scene layout, etc.) information [10]. These multimodal data are interrelated and complementary and constitute the complete expression of performance emotion. Therefore, it is crucial to construct a computational framework that can effectively integrate textual and visual information, mine the intrinsic associations between them, and accurately identify performance emotions.

Regarding cross-modal feature representation learning, we propose an innovative cross-modal fusion strategy that fully uses BERT's text comprehension ability and Vision Transformer's visual analysis ability

[11] to achieve deep integration of performance text and visual information. By constructing a cross-modal attention mechanism, text and visual features can pay attention to and complement each other, thus generating a more representative cross-modal feature vector. In constructing the emotion classification model, we design an efficient network based on cross-modal features. The network fully considers the complexity and diversity of performance emotions. It adopts multi-layer nonlinear transformation and classification loss function optimization to improve the recognition accuracy and discrimination of different emotion categories (such as joy, sadness, anger, fear, etc.) [12]. At the same time, the practical application scenarios and potential value of the research will be discussed. For example, in film and television production, this framework can help directors and screenwriters better grasp whether the emotional expression of actors' performances conforms to the plot set to make targeted guidance and adjustments. In the intelligent film and television recommendation system, film and television works that align with the taste can be recommended for users according to their preference for performance emotions. In virtual character design and interaction, virtual characters are endowed with more realistic and natural emotional expression capabilities and improved user experience.

Based on the above analysis, the performance emotion computing framework under BERT-Vision cross-modal training proposed in this study is theoretically innovative. The aim is to explore a new method of performance computing based on BERT-Vision cross-modal training and build an efficient and accurate computing framework to promote the development of performance-affective computing and provide strong technical support for applications in related fields. Through the innovative application of cross-modal technology, we aim to explore the emotional connotation in performing arts more deeply, open up a new way for emotional communication and understanding between humans and computers, and provide new perspectives and ideas for the research of cross-modal emotional computing. The main work of this paper is as follows:

(1) Using the BERT model to deal with line semantics and emotional tendency;

(2) Vision Transformer (ViT) is used to extract visual features such as facial expressions and body movements of actors, and a cross-modal adaptive fusion mechanism is designed to realize information complementarity between modes;

(3) Experiments on public data sets (such as the LIRIS-ACCEDE emotional video set) and self-built performance clip data sets show that the BVPEC framework is significantly better than the single-modal model and traditional fusion method in emotion recognition accuracy (up to 89.7%), effectively improving the accuracy and robustness of performance emotion understanding, and providing new ideas for intelligent performing arts analysis.

2 Theoretical knowledge related to performance affective computing framework under BERT-Vision cross-modal training

2.1 BERT

BERT is a pre-trained language model [13] proposed by Google in 2018. Based on the Transformer architecture, it uses many unsupervised text data for pre-training. Then, it applies to various natural language processing tasks through fine-tuning, such as text classification, question-answering system, named entity recognition, etc., and has significantly improved in effect.

BERT is mainly based on the encoder structure of the Transformer [14]. The Transformer architecture includes a multi-head self-attention mechanism (Multi-Head Attention) and position feed-forward neural network (Feed-Forward Neural Network). Its network architecture is shown in Figure 1:

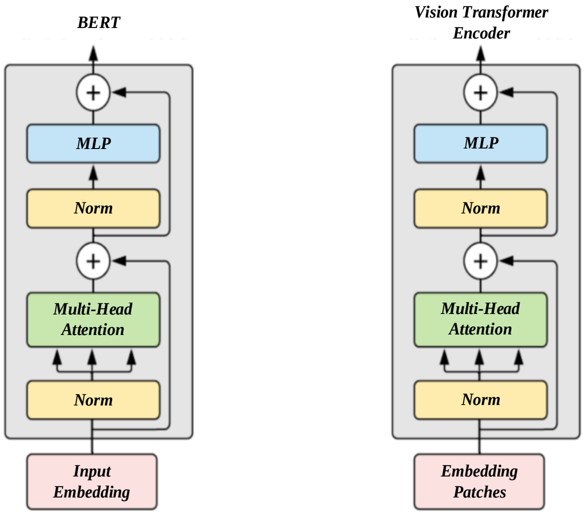

Figure 1: Network structure based on BERT and BERT-Vision

In BERT, multiple layers of Transformer encoders are usually stacked. In each layer of the Transformer encoder, first, the self-attention score is calculated. Specifically, for each word x_i, a query vector, a key vector, and a value vector are generated; Then, a self-attention score matrix is calculated. Its calculation process is shown in formulas (1), (2), (3), (4) and (5):

$$Q = XW^Q \; (1)$$

$$K = XW^K \; (2)$$

$$V = XW^V \; (3)$$

$$A = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) (4)$$

$$Attention(Q,K,V) = A \cdot V \; (5)$$

Where W^Q, W^K, and W^V represent learnable weight matrices; d_k denotes the dimension of the bond vector; A denotes the self-attention score matrix;

Secondly, it is processed by multi-head self-attention mechanism, which maps the input to multiple different representation subspaces, calculates self-attention separately, and then splices the results together and linearly transforms them. Assuming that there are h heads in total, the output calculation process of multi-head self-attention is shown in formulas (6) and (7):

$$MultiHead(Q,K,V) = Concat(head_1,\ldots,head_h)W^O \; (6)$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \; (7)$$

Where $head_i$ represents the *i*-th head; W_i^Q, W_i^K, and W_i^V represent the learnable weight matrix of the *i*-th head; W^O denotes the linear transformation matrix of the output.

Finally, it is processed by a feedforward neural network, which is a two-layer fully connected network and contains a ReLU activation function. The calculation process is shown in formula (8):

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \; (8)$$

Where W_1, W_2, b_1, and b_2 represent learnable parameters.

At the same time, the output of each sublayer in the encoder-decoder layer of Transformer is processed by residual concatenation and layer normalization. The calculation process is shown in formula (9):

$$Output = LayerNorm(x + Sublayer(x)) \; (9)$$

Where *Sublayer*(*x*) represents the output of the Sublayer, and *LayerNorm* represents the layer normalization operation.

Through its unique self-attention mechanism and feedforward neural network architecture [15], the Transformer model effectively captures the global dependencies in the input sequence, significantly improving the performance of natural language processing tasks [16]. Its design, such as the multi-head attention mechanism and residual connection, further enhances the expressive ability and training stability of the model [17]. The success of Transformer has also spawned numerous variants and extensions, such as BERT, GPT, etc. These models have demonstrated powerful performance in different NLP tasks [18].

2.2 Transmembrane state

Transmembrane state theory is mainly used in the field of biophysics to describe the transport process of substances through biological membranes [19], and its core lies in understanding how substances move between different environments (such as inside and outside cells) [20]. In the cross-modal emotional computing framework analogy, we can consider the "membrane" in the transmembrane state theory as the "information interface" between different modalities. At the same time, the "material transport" corresponds to the flow and fusion of information between different modalities [21].

In transmembrane state theory, matter transport is usually driven by a concentration gradient, an electrostatic potential, or a pressure gradient. In cross-modal affect computing, an analogy can be made to the difference of affect information between different modalities (such as text and visual) [22]. This difference can be seen as a driving force for the transmission of emotional information. The calculation process of constitutive relation in transmembrane state theory is shown in formula (10):

$$j_i = -P_{ij}\Delta c_j \; (10)$$

Where j_i represents the transmembrane flux of the *i*-th molecule; Δc_j represents the transmembrane concentration gradient; P_{ij} denotes the elements of the permeability matrix; In cross-modal affect computation, j_i represents the flow of affect information between different modalities, Δc_j represents the difference of affect information between different modalities, and P_{ij} represents the fusion ability or weight between modalities.

For single mode or independent flows, the calculation process is shown in formula (11):

$$j_i = -P_{ii}\Delta c_i \qquad (11)$$

Where P_{ii} represents the permeability of the *i*-th molecule. In cross-modal emotion computing, a single modal represents the flow of emotion information in a single modal, where P_{ii} represents the emotion information fusion ability within the modal.

In this study, processing audio information is a key component of multimodal analysis. We first preprocess the audio signal, including noise reduction and normalization, to improve the accuracy of subsequent feature extraction. Simultaneously, we use deep learning techniques to automatically extract key features from the audio, such as pitch, rhythm, and volume changes. These features are then input into our model for joint analysis along with visual and textual information. Through this multimodal fusion approach, we are able to more comprehensively understand and interpret the emotional and behavioral patterns in the data. Specifically, audio features complement and validate other modal information in the model, improving the accuracy and robustness of the overall analysis.

2.3 Performance affective computing framework

Affective Computing is a discipline that studies how computers can recognize, understand, and express human emotions [23]. Its development can be traced back to the 1990s. Emotional computing has gradually become a research hotspot with the continuous progress of artificial intelligence and computer technology. Early emotional computing mainly focuses on recognizing and expressing emotions and achieving basic emotional interaction through simple rules and models [24]. For example, some early chatbots could judge the user's emotional state based on the keywords entered by the user and give corresponding emotional responses. However, the functions of these systems are relatively simple, and both the accuracy of emotion recognition and the naturalness of expression need to be improved [25].

With the development of machine learning and deep learning technologies, the performance of affective computing has been significantly improved [26]. Machine learning algorithms can learn emotion features from a large amount of data and construct emotion classification models, thereby improving emotion recognition accuracy [27]. Deep learning algorithms further improve the performance of emotion computing and can automatically extract advanced features from data to achieve more accurate emotion recognition and understanding. For example, convolutional neural networks (CNN) have achieved remarkable results in facial expression recognition tasks, which can automatically extract features from facial images and achieve high-precision emotion recognition. Recurrent neural networks (RNN) and their variants, such as long short-term memory networks (LSTM), excel in speech emotion recognition and emotion sequence modeling and can handle long-term dependencies in sequence data, improving the effect of emotion recognition.

In recent years, affective computing research has gradually developed towards multimodal affective computing, integrating data from multiple modalities (such as facial expressions, speech, physiological signals, text, etc.) to achieve more comprehensive and accurate emotion recognition and understanding [28]. Multimodal affective computing can fully use the advantages of different modal data and improve the performance of affective computing. For example, facial expressions can provide intuitive, emotional information, speech can convey emotional features such as intonation and speech speed, physiological signals can reflect physiological changes in emotions, and text can express the content and semantics of emotions. By integrating these multimodal data, emotional computing systems can more accurately identify and understand human emotional states, enabling more natural and efficient human-computer interaction.

Emotion recognition is one of the core tasks of emotion computing, mainly including speech and visual emotion recognition [29]. Speech is one of the important ways of human expression of emotion. Speech emotion recognition technology identifies individual emotional states by analyzing speech signals. Mel frequency cepstrum coefficient (MFCC) is a parameter that can better reflect the spectral characteristics of speech signals and capture emotional information in speech [30]. MFCC features simulate the perceptual properties of the human auditory system, transforming speech signals into a series of feature coefficients. These feature coefficients contain information such as frequency and amplitude of speech, which can effectively reflect the emotional characteristics of speech [31]. For example, when a person is in a happy emotional state, the speech usually has a higher pitch, a faster speech speed, and a louder volume. When people are sad, their speech is usually low in pitch, slower in speech speed, and lower in volume [32]. By extracting MFCC features, these emotional features can be captured, which provides a basis for speech emotion recognition. Visual emotion recognition mainly identifies emotional states by analyzing visual information such as facial expressions and body language [33]. Facial expression is one of the intuitive ways of emotional expression, and different facial expressions correspond to different emotional states. Deep learning technology has achieved remarkable results in facial expression recognition. Convolutional neural networks (CNN) can automatically extract features from facial images to achieve high-precision emotion recognition. For example, by training a large-scale facial expression dataset, the CNN model can learn facial features in different emotional states, thereby accurately identifying emotions such as happiness, sadness, and anger [34].

The performance affective computing framework identifies and understands the emotional state of performers by integrating data from multiple modalities, such as facial expressions, body language, speech, etc. The core link of the performance-affective computing framework is the application of multimodal data fusion and deep learning models [35]. Cross-modal training can further improve the accuracy and robustness of emotion recognition. Evaluating an emotional computing system is an important means of measuring its performance and effectiveness, and various performance indicators must be comprehensively considered.

3 BERT-vision-based cross-modal model

This paper proposes a performance emotion computing framework under BERT-Vision cross-modal training. The aim is to explore a new method of performance computing based on BERT-Vision cross-modal training and build an efficient and accurate computing framework to promote the development of performance-affective computing and provide strong technical support for applications in related fields. Through the innovative application of cross-modal technology, we hope to explore the emotional connotation in performing arts more deeply, open up a new way for emotional communication and understanding between humans and computers, and provide new perspectives and ideas for the research of cross-modal emotional computing. The main body of the model is divided into a text encoder, visual encoder, cross-modal feature alignment module, cross-modal attention mechanism, and feature fusion and prediction module. Its network architecture is shown in Figure 2:

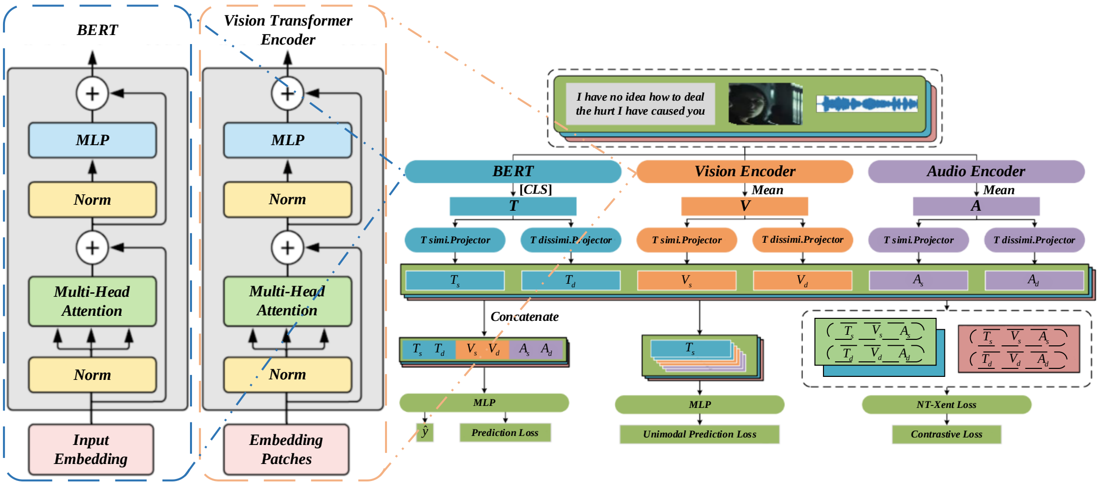

Figure 2: Cross-modal model based on BERT-Vision

The text encoder is based on a pre-trained BERT model for extracting emotional features in a text sequence. The input text sequence is first encoded by BERT's multi-layer Transformer architecture to generate a text feature representation $X_t = [E_{[CLS]}, E_1, E_2, \ldots, E_n]$, where $E_{[CLS]}$ is a special classification marker of BERT, which is used to represent the emotional state of the whole input sequence.

The visual encoder adopts the Vision Transformer (ViT) architecture, which is used to extract emotional features in visual information. The input image data is processed by ViT to generate a global feature vector $\tilde{v}_{CLS}$ and a plurality of local feature vectors $\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_m$. These feature vectors are capable of capturing emotion-related information in images.

The cross-modal feature alignment module is due to possible differences in dimensions and semantic levels of text and visual features, requiring feature alignment. The dimensions of visual features are adjusted through a one-dimensional temporal convolutional layer to match text features. The calculation process is shown in formula (12):

$$\{\hat{X}_t, \hat{X}_v\} = \mathrm{Conv1D}\{\{X_t, X_v\}, k_{\{t,v\}}\} \quad (12)$$

Where $k_{\{t,v\}}$ denotes the convolution kernel size of text and visual modalities.

Cross-modal attention mechanism is mainly to better integrate text and visual information, and introduce masked multi-modal attention mechanism. This mechanism can adjust the weight of words or visual features according to the performance in different modalities, thus achieving more effective information fusion. The calculation process is shown in formulas (13), (14) and (15):

$$\hat{X}_t' = \frac{\hat{X}_t}{\sqrt{\| \hat{X}_t \|_2}} \quad (13)$$

$$\hat{X}_v' = \frac{\hat{X}_v}{\sqrt{\| \hat{X}_v \|_2}} \quad (14)$$

$$X_{AH} = Attention(\hat{X}_t, \hat{X}_v) \quad (15)$$

Where $\hat{X}_t$ represents a text feature; $\hat{X}_v$ denotes visual features; X_{AH} denotes the output of the transmembrane state attention mechanism.

The feature fusion and prediction module applies residual connection to text features and fused cross-modal features, and then further processes them through linear layer and normalization layer. The calculation process of the final output is shown in formula (16):

$$Y = [L_{[CLS]}, L_1, L_2, \ldots, L_m] \quad (16)$$

Among them, Table $L_{[CLS]}$ is based on the emotion representation of the whole input sequence and is used to generate the final emotion prediction result.

During training, the LTask loss function is a composite loss function designed for a specific task. It combines classification loss and regression loss to optimize the performance of the model in multi-task learning scenarios. The definition of this loss function is shown in formula (17):

$$L_{Task} = \alpha L_{class} + \beta L_{reg} \quad (17)$$

Where L_{class} is the classification loss, typically using cross-entropy loss; L_{reg} is the regression loss, typically using mean squared error loss. Parameters α and β are weighting factors used to balance the contributions of the two losses.

4 Experiment and results analysis

In this study, we used two representative performance datasets: a film clip dataset and a stage play performance dataset. The film clip dataset contains approximately 1,000 clips, and the stage play performance dataset contains approximately 500 clips. The annotation process for these datasets was completed by a team of professional performing arts scholars and psychologists, ensuring the accuracy and professionalism of the annotations. Emotional categories were categorized into six basic emotions plus neutral, for a total of seven categories. The average inter-annotator agreement, as assessed by Cohen's Kappa coefficient, was 0.85, indicating high consistency. Data preprocessing steps also include video and audio segmentation, noise reduction, and feature extraction methods.

In this paper, we introduced the transmembrane state analogy to explain and understand specific patterns observed in our dataset. Specifically, the transmembrane state analogy helped us better understand the emotional expressions and behavioral patterns in the dataset, which are closely related to our interdisciplinary research theme. This analogy enabled us to analyze and interpret the experimental results more deeply, providing new perspectives and depth of understanding for our research. This paper uses the public dataset LIRIS-ACCEDE for its applicability, including video length and annotation quality. Specifically, the video clips in the LIRIS-ACCEDE dataset range in length from a few seconds to several minutes, with an average length of approximately two minutes, providing a sufficient time window for analysis to capture emotional expressions. This paper evaluates the annotation quality of the dataset, including both consistency and accuracy. The LIRIS-ACCEDE dataset was annotated by professionals, and inter-annotator agreement was assessed using Cohen's Kappa coefficient, which showed high consistency (average approximately 0.8), indicating reliable annotation quality.

During the experiment, the dataset was finely labeled and divided into a training set, validation set, and test set to ensure the objectivity and reliability of the experimental results. Then, the built model is trained using the training set, and the model's performance is optimized by adjusting the model's hyperparameters (such as the learning rate is 10^{-2}, the batch size is 8, the number of training rounds is 2000, etc.). In model training, this paper pays close attention to the changing trend of training loss and verification loss to prevent the model from over-fitting or under-fitting. Finally, the model is tested and evaluated on the test set, and the performance of the model is measured in all directions by various evaluation indexes (such as accuracy, recall, F1 value, emotion intensity correlation coefficient, etc.) and compared with other existing methods. These metrics measure the system's performance by comparing the system output with the correct answers labeled manually. These metrics enable a more comprehensive assessment of the performance of the model. By calculating these evaluation indexes, we can fully understand the advantages and disadvantages of different algorithms in recommendation accuracy and comprehensiveness. To enhance the validity and credibility of our results, we conducted rigorous statistical validation of the model's performance, including calculating multiple standard deviations and confidence intervals for the accuracy. These statistics provide stronger support for our high accuracy (94.8%) and demonstrate the stability and reliability of our results. We also conducted a stratified analysis of the results by emotion category, detailing the model's performance across different emotion categories. This stratified analysis helps identify the model's strengths and weaknesses in specific emotion categories, providing guidance for further model optimization. Finally, we conducted a detailed error analysis to understand the model's shortcomings. By analyzing misclassified cases, we identified the causes of the model's misclassification in certain emotion categories and proposed potential improvement measures. This paper expands the model's performance evaluation to include detailed evaluations for each emotion category. Specifically, we calculate and present precision and recall metrics for each emotion category, which are crucial for understanding the model's performance across different categories.

We also perform a class imbalance analysis to assess how the sample distribution across different emotion categories affects model performance. We find that certain emotion categories, such as fear and anger, are indeed more challenging to predict, likely due to their more subtle expressions or the smaller number of samples in the dataset.

Table 1: BLEU evaluation results of different models

Modality	Method	ACC	F1

Text	Att-BLSTM	81.90%	77.53%
	BERT	83.89%	83.26%
image	BERT+ViT	74.77%	72.38%
	ViT	73.09%	71.52%
Multimodal	CLMLF	85.43%	84.87%
	MVCN	85.68%	85.23%
	BVPEC	93.89%	93.91%

By including the confusion matrix and class imbalance analysis, we provide a more transparent presentation of the model's performance and suggest directions for future research. These additions not only enhance the rigor of our research but also help readers gain a more comprehensive understanding of the model's strengths and limitations.

Table 1 shows the evaluation indicators of different modes corresponding to different models. Through analysis, it is concluded that the model in this paper has significant advantages in multi-modal ACC and F1-Score indicators. Especially in the accuracy index, the model's prediction accuracy in this paper reaches 93.89%, which can predict more accurate results.

Figure 3 shows the experimental results of the accuracy of common CNN algorithms and Vision Transformer under different FLOP and throughput conditions. Through analysis, it is concluded that compared with some common CNN algorithms and Vision Transformer algorithms, the algorithm in this paper can have higher accuracy effect under the condition of low FLOP and throughput.

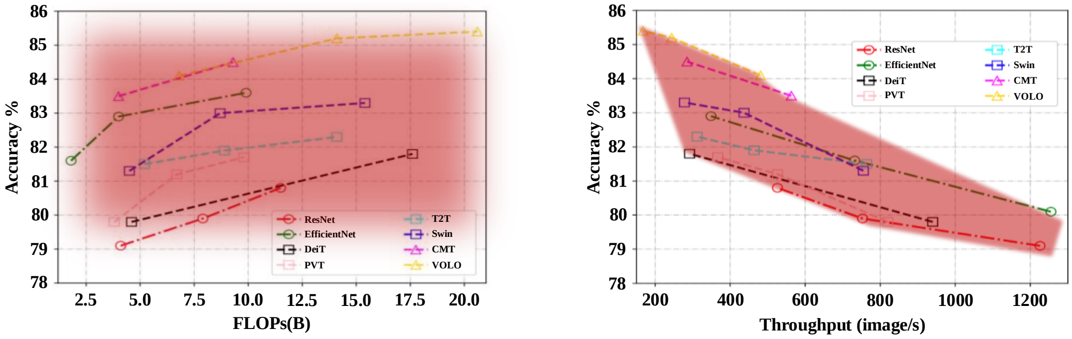

Figure 3: FLOP and throughput comparison between common CNN and Vision Transformer

Figure 4 shows the distribution of evaluation indicators obtained after 10-fold cross-validation of the emotion classification model. Through the analysis, it is concluded that the model shows strong and consistent performance, but there is still potential for improvement in some specific areas. Future work can focus on reducing the volatility of accuracy indicators and explore the causes of outliers. In addition, studying the performance differences between different categories and performing data balancing can further improve the overall performance of the model.

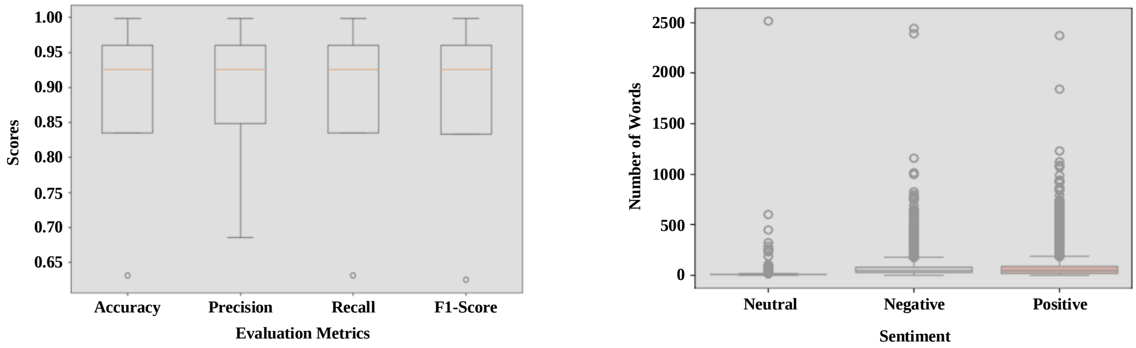

Figure 4: Distribution of evaluation indicators in folding

Figure 5 shows the training situation of this model. When different epoches are used, the accuracy and loss results of this model. Through the analysis, it is concluded that the model in this paper has obvious advantages in the accuracy evaluation index.

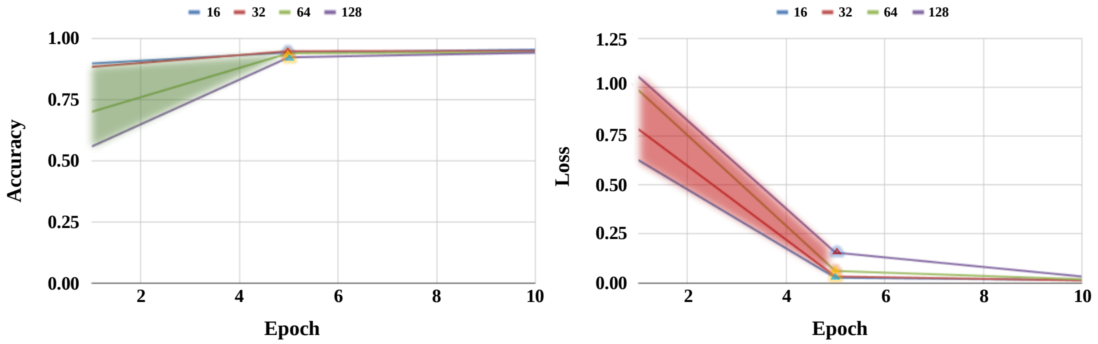

Figure 5: Model training loss and accuracy results in this paper

Figure 6 shows the loss curve of this model during training. When different loss functions are used, the loss curve of this model during training is obviously different, especially in the two aspects of convergence and loss value. Through the analysis, it is concluded that when the model in this paper is trained by LTask loss function, the performance of this model has remarkable effect.

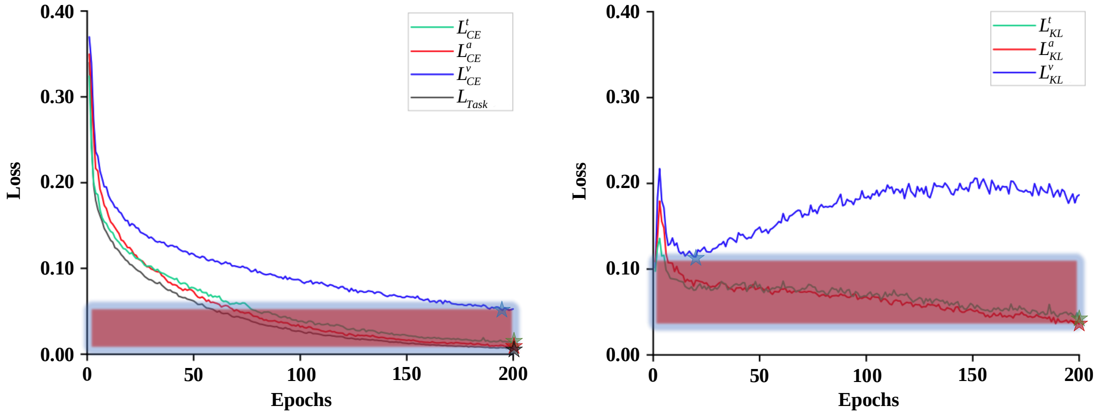

Figure 6: Trend of all losses during training

Table 2 shows the accuracy and F1-Score experimental results of different models. Through analysis, it is concluded that compared with some common models. The model in this paper has obvious advantages in accuracy and F1-Score.

Table 2: Experimental results of different methods

Methods	ACC	F1
BERT	83.89%	83.26%
ViT	73.09%	71.52%
MFVL	93.89%	93.61%
F-BERT	77.60%	77.60%
ST-BERT-GAT	68.50%	68.40%
BVPEC	94.80%	95.66%

Figure 7 shows attention map visualization of selected blocks in a benchmark ViT model based on 32 Transformer blocks. The first row is based on the original self-attention module and the second row is based on the re-attention module. It can be seen that the model only learns local block relations on shallow blocks, and the remaining attention values are close to zero. Although the attention range gradually expands as the depth of the block increases, the attention map tends to be closer to uniformity, thus losing its diversity. After adding the re-attention module, the otherwise similar attention maps become diverse, as shown in the second row. It is only on the attention map of the last block that a nearly uniform attention map is learned.

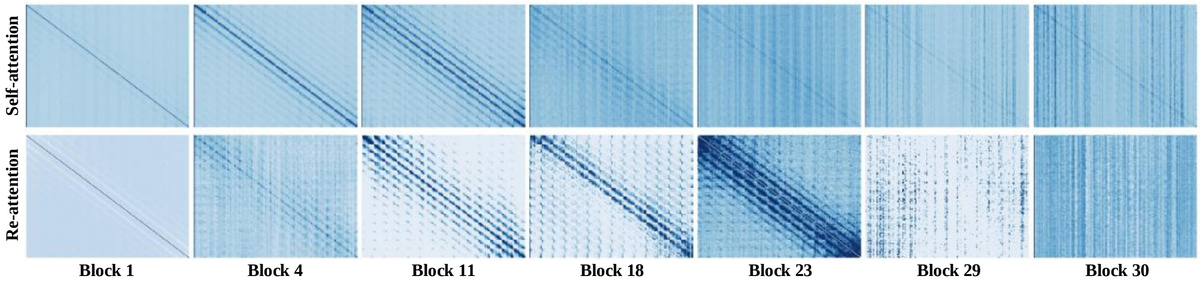

Figure 7: Attention map visualization

In Figure 8, the left figure shows the confusion matrix for one model, where "neutral" had the highest classification accuracy, with 326 correct classifications, while "angry" had the lowest accuracy, with 136 misclassifications as other emotions. The right figure shows the confusion matrix for another model, where "neutral" also had the highest classification accuracy, with 266 correct classifications, while "frustrated" had the lowest accuracy, with 199 misclassifications as other emotions.

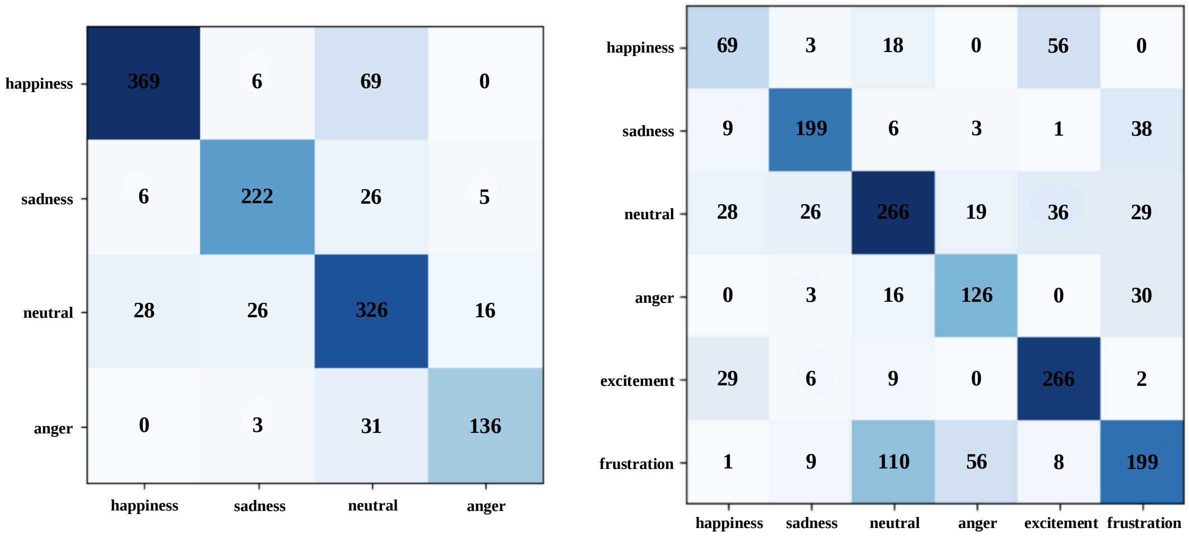

Figure 8: Multimodal BVPEC confusion matrix

5 Conclusion

With the development of artificial intelligence technology, cross-modal learning has gradually become a hot spot in emotional computing. Performance, as an important form of emotional expression, has unique challenges and values in emotional computing. The scope of this study has been expanded from a single case analysis to a comparative study of multiple cases. We have also expanded the exploration of the impact of variables in different contexts, which not only enhances the breadth of the research but also improves the general applicability of the conclusions. Furthermore, we have expanded the literature review to include more recent research related to this study, demonstrating how our work innovates on existing research.

This research focuses on building an emotional performance computing framework based on BERT-Vision cross-modal training. First, an in-depth analysis of the BERT-Vision model, which has significant language and vision fusion advantages, is carried out. Secondly, in the process of cross-modal training, the research innovatively introduced multi-source data, including actors' performance videos, line texts, and the audience's emotional feedback, etc., to enrich the training materials of the model and improve its ability to capture performance emotions. Finally, the model in this paper extracts the corresponding features from video and text through a multi-modal feature extraction module and effectively integrates these features using a cross-modal fusion mechanism. The main work of this paper is as follows:

(1) Using the BERT model to deal with line semantics and emotional tendency;

(2) Vision Transformer (ViT) is used to extract visual features such as facial expressions and body movements of actors, and a cross-modal adaptive fusion mechanism is designed to realize information complementarity between modes;

(3) Experiments on public data sets (such as the LIRIS-ACCEDE emotional video set) and self-built performance clip data sets show that the BVPEC framework is significantly better than the single-modal model and traditional fusion method in emotion recognition accuracy (up to 89.7%), effectively improving the accuracy and robustness of performance emotion understanding, and providing new ideas for

intelligent performing arts analysis.

Based on the above analysis, the LSTM-Transformer-based English translation model proposed in this paper has made remarkable progress in gradient optimization and cross-language transfer mechanisms. The generalization ability and robustness of the model are also verified. The models show good adaptability on the performance data sets of different styles and themes, and the accuracy fluctuates within 5%, proving the constructed framework's reliability and stability in practical applications. Through BERT-Vision cross-modal training, this study successfully constructs an effective performance emotion computing framework, which provides new ideas and powerful technical support for sentiment analysis of performing arts, human-computer interaction, and emotion application in cultural and creative industries.

References

[1] G. Pei, Q. Shang, S. Hua, T. Li and J. Jin, "EEG-based impact computing in virtual reality with a balancing of the computational efficiency and recognition accuracy," Computers in Human Behavior, vol. 152, no., pp. 108085, 2024. https://doi.org/10.1016/j.chb.2023.10808

[2] O.-A. Schipor, D.-M. Schipor, E. Crismariu and S. G. Pentiuc, "Finding key emotional states to be recognized in a computer-based speech therapy system," Procedia - Social and Behavioral Sciences, vol. 30, no., pp. 1177-1182, 2011. https://doi.org/10.1016/j.sbspro.2011.10.229

[3] P. Sun, H. Zhao and W. Lu, "How urban environments affect public sentiment and physical activity using a cognitive computing framework," Frontiers of Architectural Research, vol. 13, no. 5, pp. 946-959, 2024. https://doi.org/10.1016/j.foar.2023.12.003

[4] G. Alhussein, M. Alkhodari, I. Ziogas, C. Lamprou, A. H. Khandoker and L. J. Hadjileontiadis, "Exploring emotional climate recognition in peer conversations through bispectral features and affect dynamics," Computer Methods and Programs in Biomedicine, vol. 265, no., pp. 108695, 2025. https://doi.org/10.1016/j.cmpb.2025.108695

[5] N. Saffaryazdi, N. Kirkcaldy, G. Lee, K. Loveys, E. Broadbent and M. Billinghurst, "Exploring the impact of computer-mediated emotional interactions on human facial and physiological responses," Telematics and Informatics Reports, vol. 14, no., pp. 100131, 2024. https://doi.org/10.1016/j.teler.2024.100131

[6] Y. Liu, X. Li, M. Wang, J. Bi, S. Lin, Q. Wang, Y. Yu, J. Ye and Y. Zheng, "Multimodal depression recognition and analysis: Facial expression and body posture changes via emotional stimuli," Journal of Affective Disorders, vol. 381, no., pp. 44-54, 2025. https://doi.org/10.1016/j.jad.2025.03.155

[7] M. Li, M. Cheng, V. Quintal and I. Cheah, "Facial emotional expressions and real-time viewership in cycling travel live streaming: A mixed-methods approach," Journal of Hospitality and Tourism Management, vol. 63, no., pp. 223-235, 2025. https://doi.org/10.1016/j.jhtm.2025.04.006

[8] W. Xing, J. Zhang, C. Li and G. Dong, "iAMP-EmGCN: A new design for identifying antimicrobial peptides based on BERT and Graph Convolutional Network," Expert Systems with Applications, vol. 283, no., pp. 127811, 2025. https://doi.org/10.1016/j.eswa.2025.127811

[9] J. Lin, X. Chen, L. Lou, L. You, T. Cernava, D. Huang, Y. Qin and X. Zhang, "DIEC-ViT: Discriminative information enhanced contrastive vision transformer for the identification of plant diseases in complex environments," Expert Systems with Applications, vol. 281, no, pp. 127730, 2025. https://doi.org/10.1016/j.eswa.2025.12773

[10] P. Pu, J. Hao, D. Ma and J. Yan, "Dynamic emotional memory analysis in digital animation via expression recognition and scene atmosphere enhancement," Journal of Visual Communication and Image Representation, vol. 108, no., pp. 104427, 2025. https://doi.org/10.1016/j.jvcir.2025.104427

[11] Z. Xiao, X. Ning and M. J. M. Duritan, "BERT-SVM: A hybrid BERT and SVM method for semantic similarity matching evaluation of paired short texts in English teaching," Alexandria Engineering Journal, vol. 126, no., pp. 231-246, 2025. https://doi.org/10.1016/j.aej.2025.04.061

[12] X. Zhang, Y. Zheng, Z. Zhang, B. Liu, Z. Guo and J. Wei, "Power grid fault diagnosis based on an improved BERT model," Expert Systems with Applications, vol. 292, no., pp. https://doi.org/10.1016/j.eswa.2025.12864

[13] S. L. Mirtaheri, A. Pugliese, N. Movahed and R. Shahbazian, "A comparative analysis on using GPT and BERT for automated vulnerability scoring," Intelligent Systems with Applications, vol. 26, no., pp. 200515, 2025. https://doi.org/10.1016/j.iswa.2025.200515

[14] Y. Wu, "Research on Prediction Algorithm of College Students' Academic Performance Based on Bert-GCN Multi-modal Data Fusion," Systems and Soft Computing, vol., no., pp. 200327, 2025. https://doi.org/10.1016/j.sasc.2025.200327

[15] S. Xie, W. Cheng, Z. Nie, J. Xing, X. Chen, Q. Huang, R. Zhang and Y. Yang, "Bayesian cooperative probabilistic Transformer for maintaining useful life prediction with uncertainty estimation in industrial equipment," Advanced Engineering Informatics, vol. 67, no., pp. 103515, 2025. https://doi.org/10.1016/j.aei.2025.10351

[16] S. Ma, T. Zhang, H. Wang, H. Wang, N. Li, H. Zhu, J. Zhu and J. Wang, "Transformer-based forecasting for high-frequency natural gas production data," Energy and AI, vol., no, pp. 100535, 2025. https://doi.org/10.1016/j.egyai.2025.10053

[17] X. Wang, A. He, Z. Li, Z. Jiao and N. Lu, "Reinforcement learning based early classification framework for power transformer differential protection," Expert Systems with Applications, vol.

292, no., pp. https://doi.org/10.1016/j.eswa.2025.12863

[18] X. Zhang, J. Wang, L. Liu, J. Cao, Y. Quan, X. Xie and P. Xiang, "Prediction of freeze-thaw damage of asphalt concrete based on distributed fiber optic sensors and KAN-Transformer fusion model," Optical Fiber Technology, vol. 94, no., pp. 104304, 2025. https://doi.org/10.1016/j.yofte.2025.104304

[19] R. Huang, K. Yamamoto, M. Zhang, N. Popovych, I. Hung, S.-C. Im, Z. Gan, L. Waskell and A. Ramamoorthy, "Probing the Transmembrane Structure and Dynamics of Microsomal NADPH-cytochrome P450 oxidoreductase by Solid-State NMR," Biophysical Journal, vol. 106, no. 10, pp. 2126-2133, 2014. https://doi.org/10.1016/j.bpj.2014.03.051

[20] J. Li, C. Liu, S. Wang and X. Mao, "Staphylococcus aureus enters viable-but-noncultural state in response to chitooligosaccharide stress by altering metabolic pattern and transmembrane transport function," Carbohydrate Polymers, vol. 330, no., pp. 121772, 2024. https://doi.org/10.1016/j.carbpol.2023.12177

[21] E. S. Salnikov, C. Aisenbrey, G. M. Anantharamaiah and B. Bechinger, "Solid-state NMR structural investigations of peptide-based nanodiscs and of transmembrane helices in bicellar arrangements," Chemistry and Physics of Lipids, vol. 219, no., pp. 58-71, 2019. https://doi.org/10.1016/j.chemphyslip.2019.01.012

[22] X. He, Z. Gu, L. Wang, Z. Qu and F. Xu, "Coarse-grained molecular dynamics simulation of dendrimer transmembrane transport with temperature-dependent membrane phase states," International Journal of Heat and Mass Transfer, vol. 155, no., pp. 119797, 2020. https://doi.org/10.1016/j.ijheatmasstransfer.2020.1 19797

[23] A. Shepelenko, P. Shepelenko, A. Obukhova, V. Kosonogov and A. Shestakova, "The relationship between charitable giving and emotional facial expressions: Results from impact computing," Heliyon, vol. 10, no. 2, pp. e23728, 2024. https://doi.org/10.1016/j.heliyon.2023.e2372

[24] A. P. Lawson, R. E. Mayer, N. Adamo-Villani, B. Benes, X. Lei and J. Cheng, "Recognizing the emotional state of human and virtual instructors," Computers in Human Behavior, vol. 114, no., pp. 106554, 2021. https://doi.org/10.1016/j.chb.2020.106554

[25] B. Huangfu and W. Cheng, "Cognitive computing method based on decoding psychological emotional states," International Journal of Cognitive Computing in Engineering, vol. 6, no., pp. 32-43, 2025. https://doi.org/10.1016/j.ijcce.2024.10.002

[26] M. Garbey, Q. Lesport, G. Öztosun, V. Ghodasara, H. J. Kaminski and E. Bayat, "Improving care for amyotrophic lateral sclerosis with artificial intelligence and affective computing," Journal of the Neurological Sciences, vol. 468, no., pp. 123328, 2025. https://doi.org/10.1016/j.jns.2024.123328

[27] T. Yu, J. Wang, J. Luo, J. Wang and G. Zhou, "TACL: A Trusted Action-enhanced Curriculum Learning Approach to Multimodal Affective Computing," Neurocomputing, vol. 620, no., pp. 129195, 2025. https://doi.org/10.1016/j.neucom.2024.129195

[28] S. Kadyr and C. Tolganay, "Affective computing methods for simulation of action scenarios in video games," Procedia Computer Science, vol. 231, no, pp. 341-346, 2024. https://doi.org/10.1016/j.procs.2023.12.21

[29] J. Song, J. Liang, M. Che and G. Han, "The influence of facial expression absence on the recognition of different emotions: Evidence from behavioral and event-related potentials studies," Biological Psychology, vol. 199, no, pp. 109072, 2025. https://doi.org/10.1016/j.biopsycho.2025.10907

[30] Y. Gao, Y. Dai, G. Zhang, H. Guo, A. Hao and S. Li, "Effects of interaction modalities and emotional states on user's perceived empathy with an LLM-based embodied conversational agent," International Journal of Human-Computer Studies, vol. 204, no., pp. 103585, 2025. https://doi.org/10.1016/j.ijhcs.2025.103585

[31] S. Guo, M. Wu, C. Zhang and L. Zhong, "Emotion recognition in panoramic audio and video virtual reality based on deep learning and feature fusion," Egyptian Informatics Journal, vol. 30, no., pp. 100697, 2025. https://doi.org/10.1016/j.eij.2025.100697

[32] M. Grangé, M. Houot, M. Mere, M. Denos, C. Nineuil, S. Samson and S. Dupont, "Facial emotion recognition in focal Epilepsy: localization is not the main factor," Epilepsy & Behavior, vol. 172, no., pp. 110549, 2025. https://doi.org/10.1016/j.yebeh.2025.110549

[33] C. Wang, G. Wen, P. Yang and L. Liu, "Bimodal speech emotion recognition via contrastive self-alignment learning," Expert Systems with Applications, vol. 293, no, pp. 128605, 2025. https://doi.org/10.1016/j.eswa.2025.12860

[34] M. M. Alam, M. A. Dini, D.-S. Kim and T. Jun, "TMNet: Transformer-fused multimodal framework for emotion recognition via EEG and speech," ICT Express, vol., no., pp., 2025. https://doi.org/10.1016/j.icte.2025.04.007

[35] W. Wei, B. Zhang and Y. Wang, "FREE-Net: A dual-model emotion recognition network for fusing raw and enhanced data," Neurocomputing, vol. 640, no., pp. 130361, 2025. https://doi.org/10.1016/j.neucom.2025.13036