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

Research on the migration technology of animation character emotion expression based on multimodal fusion and attention mechanism

Xiawei Lu[1,*]

[1] College of Architecture and Art, Taiyuan University of Technology, Jinzhong, Shanxi, 030600, China

SUMMARY: *In this paper, we design a multimodal sequence feature extraction model based on self-attention mechanism and propose an improved RoBERTa-MEN to realize emotion classification. Combined with the influence of cognition on emotion, the animated character emotion modeling and behavior modeling methods are optimized. Select mainstream models for emotion recognition performance comparison to explore the superiority of the proposed method. Combine the results of correlation visualization analysis to verify the effectiveness of the improvement scheme in this paper. Conduct an animated character emotion transfer experiment to analyze the feasibility of the proposed method. In the performance test, the four indexes of this paper's model are improved by 1.5%~2.5% than the TBJE model. The research on the effectiveness of facial emotion shows that the classification accuracy of this paper's model in seven emotion labels is above 80%, and the average value of the emotion index evaluation reaches 82 points. It proves that the scheme of this paper can integrate multimodal data and realize the effective characterization and transfer of animated characters' emotions.*

KEYWORDS: *animated characters; self-attention mechanism; multimodal sequence characterization; emotion classification; emotion migration*

1 Introduction

Animated film and television, as a unique art form, is able to convey emotions in a profound and delicate way by virtue of its dual visual and narrative appeal [1]. Literature [2] emphasizes the importance of emotional elements for animation film and television, and through the use of literature analysis and semi-structured interviews, it is shown that comprehensibility and emotional elements in animation production are able to attract the audience's attention. Whether it is through the expression and movement of the characters, or the color and music of the scene, animated films have shown great skill in expressing emotion [3]. Literature [4] points out that expression is the soul of animation characters, and the expressions of joy, anger, sadness and happiness of excellent animation works will resonate with the audience. Literature [5] analyzes the relationship between color and emotion in animation scenes, and draws on the interaction between “scene” and ‘emotion’ discussed in “Night Talks in Bed” to clarify the inseparable and intertwined characteristics of animation scenes and emotion, emphasizing the key role of color. The key role of color is emphasized. Literature [6] examined the influence of music on the audience's perception of animation, by focusing on the popular form of sand painting, explored the audience's cognitive differences and preferences to reveal the core factors, the results show that the audience's evaluation of animation works by the creativity of the music, the cultural

*13303443946@163.com
https://doi.org/10.65102/is2026260
Published: 30 April, 2026

connotation, and the influence of personal preferences. The important premise for realizing this emotional expression is the animation character emotional expression migration, which uses migration technology to give human emotions to animation characters through visual symbols, behavioral logic and other methods, so as to realize emotional expression and obtain the audience's emotional identity [7, 8].

Common animation character emotion migration techniques include facial expression migration, body movement migration, etc., which realize the transmission of emotional expression through appearance and dynamic information, audio and gesture data, etc. [9]. However, when targeting character emotions with complex information and salient features, it is difficult to perform a better emotion matching, which leads to problems such as inconsistent style and color, and loss of information in some emotion migrations [10, 11]. In order to solve these problems, based on multimodal fusion and attention mechanism is gradually applied.

Multimodal fusion refers to the combination of multiple modal information from different sensors or data sources to understand and analyze emotional information more comprehensively [12]. Its advantage lies in the ability to utilize multiple modal perception channels, thus improving the accuracy and robustness of sentiment analysis [13]. Regarding the application of multimodal fusion, literature [14] describes the value of applying multimodal information in complex scenes of animation, and by describing the motion relationship of pixels between frames, introducing multimodality into the task of semantic segmentation of video can reduce the redundancy in the network based on independent image segmentation. Literature [15] launched a technical review of the models and learning methods available for multimodal intelligence, especially the combination of visual and natural language modalities, and conducted a systematic review and analysis of multimodal related applications from multiple perspectives, which provides a reference for related fields in the emerging research of multimodal intelligence mention.

The principle of the attention mechanism is to assign weights to different input elements according to their importance. This mechanism enables the model to focus on the parts that are relevant to the current task while processing sequence data, thus ignoring the parts that are not relevant to the task [16, 17]. Through the attention mechanism, the model is able to dynamically adjust the focus of its attention, simulating the way human beings behave in paying attention to things [18]. Regarding the related application of the attention mechanism, literature [19] proposed a deep learning model based on the channel attention mechanism to realize the animation generation process that combines painting and live footage, and verified the effectiveness of the model, providing a theoretical foundation and technical solution for the integration of virtual and reality in digital art creation. Literature [20] proposes an unsupervised image animation technique based on fusion attention mechanism, which is able to reduce the computational parameters and accurately extract the motion patterns of objects to generate realistic animation effects by introducing multiple attention mechanisms and null convolution, and reveals that it realizes excellent visual effects. As for the migration of emotional expression of animated characters, by applying multimodal fusion and attention mechanism to the migration technique, the efficient conversion of emotional style is realized by fusing information such as text, image, and audio, and focusing on the key emotional features of the characters using attention mechanism [21-23].

In this paper, we propose a multimodal sequence feature extraction network based on self-attention mechanism and neural network, which employs the self-attention mechanism to enhance the contextual information within a single modality and the contextual information across time between multiple modalities. Sequence reorganization and modal enhancement are performed on the multimodal sequence data, and the fused feature representation of the information-enhanced multimodal sequences is directly extracted. Change the processing of

sentence embedding vectors and word embedding vectors based on the RoBERTa-BiLSTM-Attention model to improve the polarity judgment of textual emotions. Design specific emotional situations, manage behaviors in the animation environment through the behavioral organization form of behavioral tree, and realize the animation character emotional behavior. Conduct performance tests based on the MELD dataset to verify the effectiveness of the proposed model. Combine subjective and objective evaluations to assess the practical effect of the model in this paper in the task of animated character emotion expression.

2 Animation character emotion and behavior modeling method design

With the rapid development of digital media technology, animation, as a comprehensive art form integrating visual, auditory and narrative, plays an increasingly important role in the entertainment industry. The design of traditional animation character emotion expression is highly dependent on manual experience, which is not only time-consuming and laborious, but also faces challenges in achieving delicate, coherent and personalized emotion transmission. In recent years, the development of deep learning technology has provided new possibilities for the intelligent generation of animation content. However, for the specific medium of animation, how to build a method that can accurately model the emotional state of animated characters still needs to be explored in depth. To this end, the aim of this paper is to study the emotion extraction and classification model based on multimodal information and attention mechanism, to open up new paths for simulating and migrating complex emotional expressions, with a view to realizing intelligent and automatic migration of emotional expressions of animated characters.

2.1 Multimodal sequence feature extraction network

2.1.1 Contextual-unimodal information extraction module

For multimodal data obtained from videos, each modality X_n is a sequence of data containing a timestamp. In this module, we use three independent GRU networks to obtain the internal modal information of each modal sequence. Similar to the traditional Long Short-Term Memory (LSTM) network, the GRU network with reset and update gates is specialized to deal with sequence data with long-term dependencies, and the GRU formulas are shown in the following equations (1), (2), (3), and (4):

$$z^t = \sigma\left(W_z x^t + U_z h^{t-1} + b_z\right) \tag{1}$$

$$r^t = \sigma\left(W_r x^t + U_r h^{t-1} + b_r\right) \tag{2}$$

$$\tilde{h}^t = \tanh\left(W x^t + U\left(r^t \circ h^{t-1}\right) + b\right) \tag{3}$$

$$h^t = \left(1 - z^t\right) \circ h^{t-1} + z^t \circ \tilde{h}^t \tag{4}$$

In the above equation, $W_i \in \mathbb{R}^{h \times d}$, $U_i \in \mathbb{R}^{h \times h}$ and $b \in \mathbb{R}^h$ are the training parameters of the GRU, and z^t, r^t are the updating gate and reset gate, respectively. $\tilde{h}^t$ is a candidate activation that accepts the input vector x^t at the current t moment and the hidden state

vector h^{t-1} at the $t-1$ moment. $h^t \in \mathbb{R}^h$ is the hidden state of the GRU, where h denotes the dimension of the GRU hidden state vector. The $\circ$ is denoted as the Hadamard product and σ is a function of sigmoid.

We assign a separate GRU network to each modal input sequence X_l, X_a, X_v, which helps to obtain the modal internal feature representation $H_n = \left[h_n^t : t \le T, h_n^t \in \mathbb{R}^h \right]$, where h denotes the first dimension of the nth GRU hidden state vector, which can be expressed as Eq. (5):

$$H_i = I - GRU(X_i) \left(i \in \{l, v, a\}, I \in \{L, V, A\} \right) \tag{5}$$

Finally, the acquired sequence information for the three modalities passes through the fully connected layer (D) of dimension d at each time step t, which is intended to maintain the consistency of the feature dimensional orientation when reorganizing the sequence data. The final output of the module $U_l, U_a, U_v \in \mathbb{R}^{T \times d}$ is obtained through equation (6).

$$U_i = D_i(H_i) \left(i \in \{l, v, a\} \right) \tag{6}$$

2.1.2 Information enhancement and data reorganization module

In our proposed model, we aim to perform multimodal sentiment analysis using contextual information within modalities and public contextual information between modalities. It is well known that the information conveyed by a video is multimodal. These multimodal signals are intertwined with each other, and there are certain delays and lags between the multimodal signals. Moreover, multimodal data have different importance and should not potentially be given the same weight for fusion. In particular, as the sequence proceeds, the strength of the information conveyed by each modality in the multimodal data is different at each time step. Therefore, this paper introduces information enhancement and reorganization modules.

(1) Information enhancement module

First, we perform information enhancement on the internal modal feature representation $U_n (n \in \{l, a, v\})$ obtained above in the Information Enhancement (IE) module. We obtain a modal internal matching matrix $E_n \in \mathbb{R}^{T \times T}$ via Eq. (7):

$$E_n = U_n^T \cdot U_n \tag{7}$$

We then use the softmax function to compute the matrix of correlation coefficients $M_n \in \mathbb{R}^{T \times T}$ for the intra-modal matching matrix E_n. The element $M_n(i, j)$ in the matrix represents the degree of contextual information correlation between time i and time j in the model. Finally, the soft attention mechanism performs information augmentation at time t in each modality, and the information-augmented sequence data $S_l, S_a, S_v \in \mathbb{R}^{T \times d}$ are obtained by Eqs. (8) and (9).

$$M_n(i, j) = \frac{e^{E_n(i,j)}}{\sum_{t=1}^{T} e^{E_n(i,j)}} \ for\ i, j = 1, \cdots, T \tag{8}$$

$$S_i = M_i \cdot U_i \left(i \in \{l, v, a\} \right) \tag{9}$$

(2) Data reorganization module

Previously, the feature vector cascade operation was used when further processing the data at each time step. This method has the same weights for multimodal data by default when dimension joining is performed, especially at a single time step. However, when people express emotions, the strength of the modal signals is not the same and it changes over time. Perhaps words with strong emotional meaning dominate at moment $t-1$, while rich facial behavior suppresses other modal signals at moment t. Feature vector cascading cannot solve this problem.

Therefore, in this paper, we use the multimodal sequence reconstruction method in the data reconstruction module (DR), combined with the information enhancement module, to assign different weights to the modal information at each time step. The method of multimodal data reorganization is shown in Fig. 1. Considering that the verbal information at time t may have some connection with the acoustic (visual) at time $t-1$ and the acoustic (visual) at time $t+1$ or have multimodal contextual information, at each time step t, we arrange the information-enhanced sequence data in the order of text-visual-acoustic to get the new multimodal sequence data $F=\left[S_l^t;S_v^t;S_a^t:t\le T,S_l^t;S_v^t;S_a^t\in\mathbb{R}^d\right]$. The aligned new sequence data are then passed through the Information Enhancement Module (IE) block again to compute the cross-modal matching matrix $E_f\in\mathbb{R}^{3T\times 3T}$ and the correlation coefficient matrix $M_f\in\mathbb{R}^{3T\times 3T}$. The correlation between the multimodal data is further strengthened to obtain the enhanced multimodal recombination sequence $S_f\in\mathbb{R}^{3T\times d}$, which is given in Eq. (10), Eq. (11), and Eq. (12):

$$E_f = F\cdot F^T \tag{10}$$

$$M_f(i,j)=\frac{e^{E_f(i,j)}}{\sum_{t=1}^{3T}e^{E_f(i,j)}}\ \ for\ i,j=1,\cdots,3T \tag{11}$$

$$S_f = M_f\cdot F \tag{12}$$

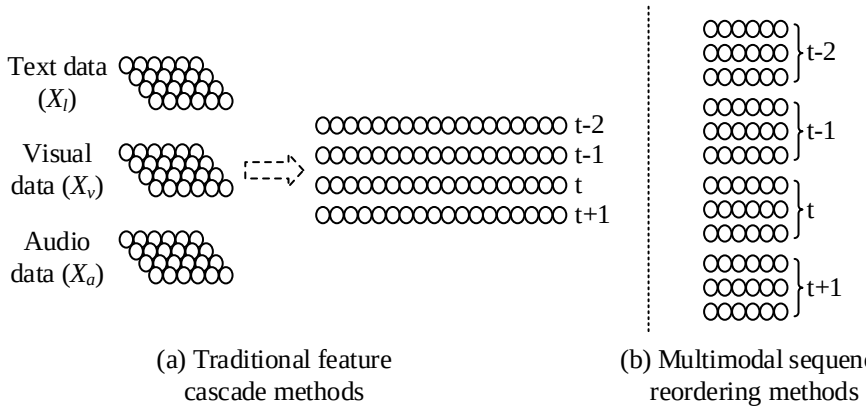


Figure 1: Methods for Multimodal Data Recon struction

The method prioritizes the enhancement of information within a single modality before reorganizing the data and enhancing the degree of common correlation between multimodal data. The reason for this is that the correlation of data within a single modality should be greater than the correlation of data between multiple modalities. The new augmented sequential data

S_f obtained after the data reorganization module (DR) and information enhancement module (IE) is then passed through the fusion GRU network to obtain the multimodal fused embedding representation. Unlike the unimodal GRU network, we obtain the contextual feature representation $h_f^T \in \mathbb{R}^{d_f}$ at the last moment T to represent the final representation output, i.e., equation (13).

$$h_f^T = Fusion - GRU\left(S_f\right) \tag{13}$$

2.2 Improved Unimodal Text Sentiment Classification Model

2.2.1 BiLSTM layer

A single-layer BiLSTM can be viewed as a combination of two LSTMs in opposite directions, one processing the input sequence in the forward direction and the other processing the sequence in the reverse direction, and splicing the two outputs together after the processing is completed, so that the output of the BiLSTM at the moment t contains the information of both the forward sequence and the reverse sequence, and the structure of the BiLSTM is shown in Fig. 2.

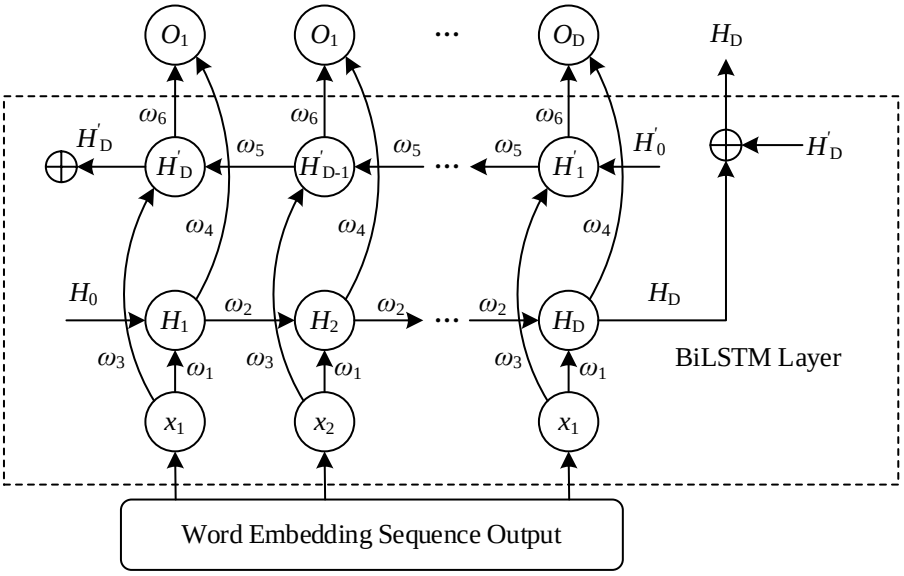


Figure 2: BiLSTM layer structure

BiLSTM has two outputs, the neural network output O at all time steps and the hidden state H_D at the last time step D. Their representations are shown in Eq. (14) and Eq. (15), respectively:

$$O = \left[O_1, O_2, \cdots, O_D\right] \in R^{B \times D \times 2H_d} \tag{14}$$

$$H_D \in R^{B \times (2nH_D)} \tag{15}$$

where B denotes the number of batch samples, H_d denotes the number of hidden states, n denotes the number of hidden layers of the neural network, and H_D denotes the hidden state of the last time step of the LSTM.

O considers all time steps up to the current moment and contains the outputs of all words, which can be regarded as a new text containing all historical information. And H_D is the hidden state of the last time step, which preserves all the history information, so it can be regarded as a feature of the text.

This model sets the number of hidden layers of LSTM network as 1. According to the experiments of related scholars, the LSTM network can achieve excellent results when the number of hidden layers is 1. Increasing the number of hidden layers will significantly increase the time complexity, but will not bring significant performance improvement.

2.2.2 Multiple Self-Attention Layers

The structure of the multi-head self-attention mechanism is shown in Fig. 3. Through the LSTM layer, the hidden state H_D of all time-step outputs $O=[O_1,O_2,\cdots,O_D]$ and the last time-step D is obtained. In order to recognize the importance of the word for the text, it is necessary to establish the self-attention relationship between H_D, which represents the features of the text, and O, which represents the features of the word, i.e., to establish the weights of the outputs O_t at each time step with respect to H_D using the dot product attention method. At time step t, K_t,V_t,Q, the score e_t and weight a_t are computed as shown in Equation (16):

$$\begin{cases} Q=\omega_Q\cdot H_D \\ V_t=\omega_V\cdot O_t \\ K_t=\omega_K\cdot O_t \\ e_t=Q\cdot K_t^T/\sqrt{d_k} \\ a_t=\dfrac{\exp(e_t)}{\sum_{t=0}^{D}\exp(e_t)} \end{cases} \tag{16}$$

where $\omega_Q,\omega_V,\omega_K$ are the parameters of the neural network, and all three parameters are modified during backpropagation. The output O_t at time step t is linearly transformed into K and V. Q is obtained by multiplying the parameter matrix ω_Q by the hidden state H_D at the last time step, which does not change with the time step. The text vector with self-attention is obtained by weighted summation of the weights a_t and V_t at each time step, and the text vector computation procedure is shown in Eq. (17):

$$z(Q,K,V)=\sum_t a_t Value_t \tag{17}$$

Multihead self-attention is to map Q, K and V linearly h times, and then carry out h self-attention operations to get h self-attention results for splicing. The specific process of its realization is to perform Eqs. (16) and (17) h times to get the text $z_1,\cdots,z_h$, and then this

h multi-head self-attention text is spliced and do a linear transformation once as the final output:

$$MultiHead\left(Q,K,V\right)=Concat\left(z_1,\cdots,z_h\right)\omega_z \tag{18}$$

where h denotes the number of self-attentive heads.

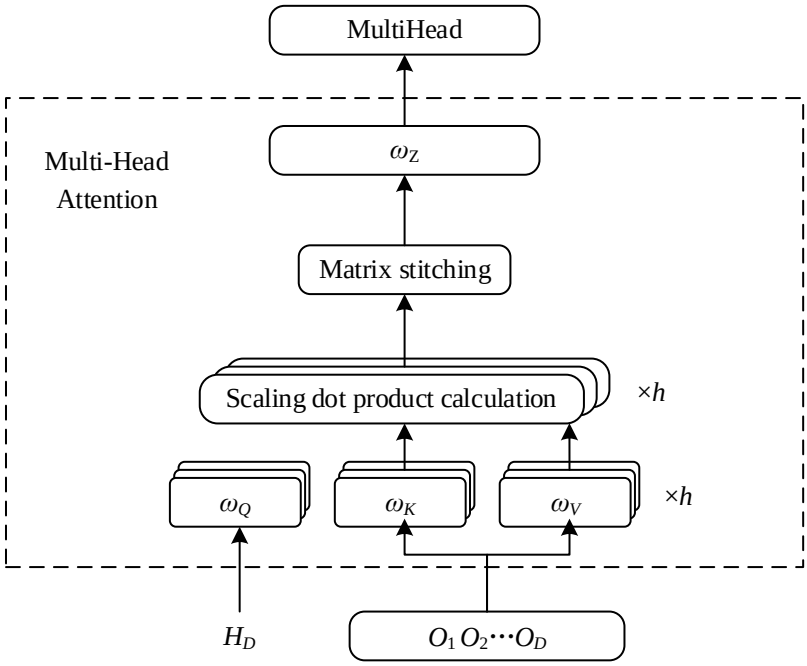


Figure 3: Structure of multi-head self-attention mechanism

2.3 Modeling Animated Character Emotion and Behavior

2.3.1 Affective Modeling

Moods and emotions have a similar structure to personality. Unlike the relative fixity of personality, both emotions and feelings can change over time. Emotion is a collection of emotional states with a certain intensity.In this paper, we define a one-dimensional vector e_t, which is used to represent the emotion of an intelligent body at the moment of time t, and the value of each component ranges from $(0,1)$, which is expressed by the following equation:

$$e_t^T=\langle\beta_1,\beta_2,\cdots,\beta_m\rangle,\forall i\in\{1,2,\cdots,m\},\beta_i\in(0,1),t>0 \tag{19}$$

2.3.2 Behavioral Modeling

Behavior is one of the most important forms of expressing feelings and emotions. The definition of behavior in this paper is: the sum of all the socially meaningful actions, behaviors, expressions, and gestures made by the animated characters in the animation environment with their own personalities as the basic reference, combined with their own emotions and feelings.

This paper focuses on two types of behavior: autonomous behavior $action$ and interactive behavior $action'$. The generator of autonomous behavior is generally active, often with a subjective purpose when generating autonomous behavior; the generator of interactive behavior

is often passive, without any subjective ideas when generating interactive behavior, and can only mechanically generate repetitive behaviors until the time or context changes so that the behavior is no longer in line with the conditions of generation.

(1) Autonomous Behavior $action$

Autonomous behavior is mainly discussed for animation characters. The so-called autonomous behavior of animation characters refers to the sum of intelligent behaviors produced by animation characters autonomously in order to express their own feelings and emotions under the premise of having certain personality, emotions and moods.

Autonomous behavior is produced by the animation character and can be imposed on all objects in the animation environment with autonomy and self-governance. When generating autonomous behavior, the animation character has clear emotional orientation, emotional expression and subjective purpose, and has absolute control over the autonomous behavior itself.

(2) Interactive behavior $action'$

Interaction behavior refers to the sum of intelligent behaviors produced by intelligent objects under the condition of meeting the conditions, which can have an effect on the emotion of animation characters within the effective influence range.

Interaction behavior is generated by intelligent objects, generally can only be imposed on animation characters in the animation environment, and its generation process is not autonomous and self-governing, but mechanical and relatively fixed and unchanging. As long as a smart object meets the conditions for generating an interaction behavior, it keeps generating the behavior uninterruptedly and acts on all animated characters within the effective influence of the behavior. The smart object does not consider the meaning of the behavior itself, nor does it have any expectation or purpose for the behavior.

(3) Behavior Rules

A behavior rule is a tuple, a formal representation of a behavior that contains the elements of the behavior. In this paper, we design behavior rules for autonomous and interactive behaviors in the following form:

$$action = \langle \inf, suf, tri(\ mood), aim(\ mood), state \rangle \tag{20}$$

$$action' = \langle \inf, suf, tri(\ con), state \rangle \tag{21}$$

where $\inf$ denotes the behavior inflicter, suf denotes the behavior bearer, and tri denotes the behavior triggering condition, where the autonomous behavior is triggered by the animated character's emotion $mood$ and the interactive behavior is triggered by the smart object's own condition con, and $state$ indicates the state of behavior execution. The aim is a component of the autonomous behavior and records the purpose of the autonomous behavior, which is reflected as what kind of change will happen to the animated character's mood $mood$.

(4) Behavior library

Behavior library is a database of all atomic behaviors, which consists of two parts: autonomous behavior library and interactive behavior library, where atomic behaviors refer to the smallest independent behavior units that can no longer be divided. The Autonomous Behavior Library stores the behavioral rules of all autonomous behaviors that may be generated by the animated characters, and the Interactive Behavior Library records the behavioral rules of all interactive behaviors that may be generated by the smart objects. Each animated object has a sub-list of autonomous behaviors that may be generated; each smart object also has a sub-list of interactive behaviors that may be generated.

(5) Behavior Tree

For each animated character in the animation environment, all of its possible completed behaviors should be predefined, and the animation environment scene management module cannot realize undefined animated character behaviors. This paper defines a data structure called behavior tree to realize the management of all possible autonomous behaviors of animated characters in the animation environment.

Simply put, a behavior tree is a tree whose constituent nodes are autonomous behaviors. A global behavior tree is designed as a common ancestor of all the child behavior trees. Under the global behavior tree, any number, any type, and any depth of subbehavior trees can be defined. Sub-behavior trees have various forms of organization, each of which manages nodes according to its own characteristics. The designed behavior tree will obey the law of top-to-bottom and left-to-right, executing according to the rules of the organization form until all executions return successfully or fail somewhere in the middle. This paper defines six organizational forms for behavior trees: sequential execution; selective execution; filtered execution; parallel execution; repetitive execution; and modified execution. All behavior trees must be defined at startup, and once in the execution state, no modification to the behavior tree will be allowed. In the execution process, the behavior tree will manage all animated characters in the animation environment with reference to the emotion and mood state of the animated characters and based on the behavioral rules of autonomous behavior, so that the animated characters can make real and palpable emotional behaviors with the guidance of their own personalities, emotions and moods.

The performance and influence of the behavior tree on the animation character's emotion and mood is mainly reflected in five aspects:

1) All the judgment bases in the behavior tree come from the animation character's emotion and mood. Whether an autonomous behavior will be performed or not depends entirely on the state of the animated character that may perform the autonomous behavior.

2) The execution conditions and execution results of all behaviors in the behavior tree are completely based on the judgment criteria of behavior rules. The triggering conditions in the behavior rules are composed of the animated character's emotion and emotional state, and the results of the behaviors are all reflected as the effects on the animated character's emotion and mood.

3) The highest guiding idea of the organizational form in the behavior tree is the orientation of the behavior itself, and the orientation of the behavior is determined by the emotion and emotional state embodied in the behavior. For example, a behavior that embodies an animated character's happiness, whose orientation is happy, should not be organized as a parallel execution with a behavior whose orientation is sad.

4) The execution process of all behaviors in the behavior tree must follow the rules of behavior and the emotions and moods of the animated characters. When a behavior meets the execution conditions, it may be executed, and as soon as a behavior no longer meets the execution conditions, it will immediately abort execution.

5) Each behavior in the behavior tree itself is a reflection of the animated character's feelings and emotions.

3 Deep Learning-based Migration Analysis of Animated Character Emotional Expression

3.1 Model performance testing

The experiments in this section use a dataset containing multimodal features called MELD. This dataset extends the data volume of the Emotionlines dataset and adapts it further to multimodal scenarios, i.e., it adds multimodal features instead of just textual dialogues.

3.1.1 Evaluation of emotional categorization

(1) Unimodal Emotion Recognition Results in Multi-Person Conversations

In order to better validate the effectiveness of the methods proposed in this paper, methods that only consider unimodal and multimodal interactions (MFN), methods that consider coarse-grained information about the speaker (CMN, ICON, DiaRNN), methods that only consider contextual information (BC-LSTM), multitask learning methods that consider contextual information modeling (MTMM-ES), and joint unimodal Information Encoding approach (TBJE) as benchmark models for performance comparison.

The experimental results for the emotion recognition task in text modality are shown in Table 1. On textual modality, among all the benchmark methods, this paper's model can achieve a weighted average F1 value of 50.22%, which is 3.34% higher than the next best performing MTMM-ES.

Table 1: Experimental results of emotion recognition task in text modality(%)

Emotion	Anger	Disgust	Fear	Joy	Neutral	Sad	Surprise	Weighted average
BC-LSTM	41.32	0.71	1.55	48.22	75.37	1.24	50.42	43.30
CMN	30.18	0.12	0.00	50.45	72.18	18.96	44.34	41.33
ICON	28.11	0.00	0.25	50.13	72.47	12.55	54.68	42.35
DiaRNN	33.57	0.45	4.98	51.37	73.88	26.38	43.27	43.33
MTMM-ES	35.45	6.81	7.37	53.78	74.45	22.56	55.89	46.88
TBJE	35.19	6.53	7.02	53.25	74.12	20.43	50.12	45.26
The proposed	41.37	8.44	9.22	58.12	75.06	23.77	60.26	50.22

The experimental results for the emotion recognition task in speech modality are shown in Table 2. In speech modality, the TBJE model joint unimodal information encoding considers the global information and performs better than the MTMM-ES model. However, the weighted average F1 value of this paper's model is still optimal (40.37%) because the model can effectively enhance and attenuate the multimodal signals through both the data enhancement module and the data reorganization module, thus making better use of the multimodal information. According to the significance test, the model in this paper is significantly higher ($p<0.05$) than the currently available benchmark methods.

Table 2: Experimental results of emotion recognition task in speech modality(%)

Emotion	Anger	Disgust	Fear	Joy	Neutral	Sad	Surprise	Weighted average
BC-LSTM	29.83	1.25	0.33	0.00	70.37	0.11	20.32	24.19
CMN	40.28	0.00	0.18	22.78	68.55	0.00	4.65	27.26
ICON	43.94	0.24	0.00	20.23	69.42	0.15	4.27	27.60
DiaRNN	45.06	6.88	0.00	25.17	56.24	19.17	30.86	33.73
MTMM-ES	45.48	5.36	0.00	26.32	67.27	20.36	37.44	37.61
TBJE	45.89	5.49	0.00	26.89	68.02	20.47	38.95	38.27
The proposed	48.24	5.26	6.78	30.17	69.15	22.33	40.12	40.37

(2) Multimodal Emotion Recognition Results in Multi-Person Conversations

This section reports the experimental results and cause analysis of the emotion recognition task in the MELD dataset for multi-person conversations in a multi-modal scenario. The performance comparison results of different benchmark methods on multimodality and the methods in this paper are shown in Table 3. The performance of the multimodal fusion method MFN is significantly lower than that of other methods modeling conversation-related information, especially when compared with MTMM-ES, the average F1 value is 11.79% lower, which indicates that contextual information can indeed help to recognize emotions.The TBJE model is, on average, more advantageous than the other methods, with a weighted average F1 value of 58.75%. This is because TBJE mainly uses the Transformer network to encode information directionally from one modality to another, focusing on mutual information and long-term dependencies between the two modalities, which makes the model more inclined to deal with bimodal information. In contrast, the model in this paper focuses on the importance of the difference of multimodal information in the time dimension, while reorganizing and coding the multimodal data, and the weighted average F1 value is improved by 1.63% compared to the TBJE model.

Table 3: Performance comparison of different emotion recognition methods(%)

Emotion	Anger	Disgust	Fear	Joy	Neutral	Sad	Surprise	Weighted average
MFN	50.12	0.00	0.15	42.17	68.37	16.33	42.15	42.20
BC-LSTM	52.86	0.12	0.24	48.33	70.33	17.37	44.27	44.91
CMN	54.28	0.24	0.33	50.12	71.75	20.18	47.12	46.70
ICON	56.88	0.00	0.45	55.46	75.68	22.43	50.36	49.94
DiaRNN	57.12	0.00	0.51	58.48	78.37	25.38	52.41	51.84
MTMM-ES	60.28	0.33	5.78	60.25	80.12	27.86	53.85	53.99
TBJE	61.65	5.56	7.22	71.37	83.24	37.17	55.72	58.75
The proposed	63.46	9.78	10.86	68.42	84.26	40.25	60.47	60.38

3.1.2 Objective evaluation of results

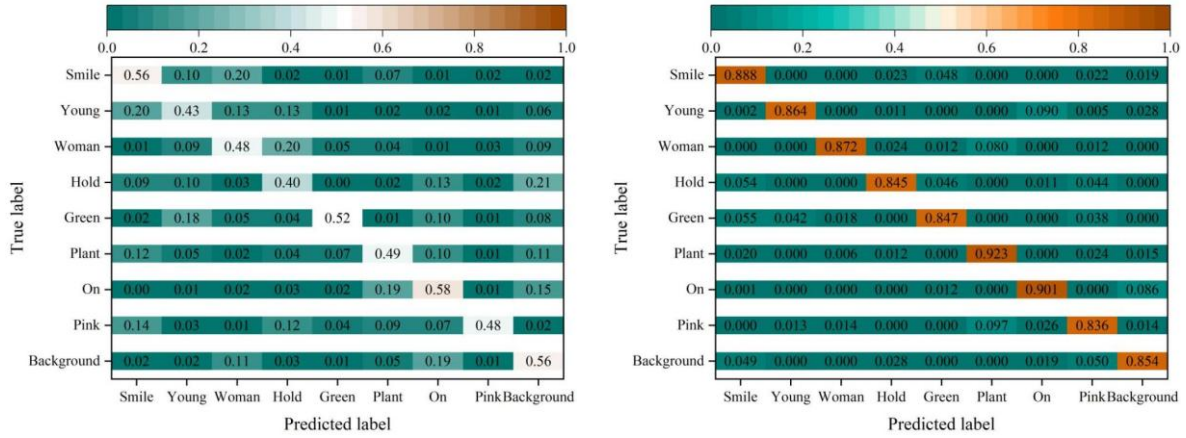
Four types of evaluation metrics, precision rate, recall rate, F1-measure and accuracy rate, are used to measure the performance of the model, and the results of the objective evaluation comparison are shown in Table 4. Compared to the baseline model, the model in this paper leads the existing algorithms in all four metrics by a huge margin, reaching 0.8158, 0.8314, 0.8235, and 0.8275, respectively.This is due to the powerful feature extraction capability of Transformer and the strong migration learning capability of the pre-trained model RoBERTa.RoBERTa uses a large amount of unsupervised data for pre-training, which enables it to learn a wider and more comprehensive knowledge of the language and semantic representation.

Table 4: Comparison results of objective evaluation

	Prec.	Rec.	F1	Acc.
MFN	0.6672	0.6945	0.6806	0.6778
BC-LSTM	0.6734	0.6983	0.6856	0.6894
CMN	0.6892	0.7026	0.6958	0.6947
ICON	0.7246	0.7373	0.7309	0.7365
DiaRNN	0.7420	0.7562	0.7490	0.7512
MTMM-ES	0.7731	0.7928	0.7828	0.7797
TBJE	0.7964	0.8085	0.8024	0.8042
The proposed	0.8158	0.8314	0.8235	0.8275

3.1.3 Analysis of the effectiveness of improvement programs

The comparison results of the attention weight visualization results of the original model and this paper's model with the addition of the multi-head self-attention mechanism are shown in Fig. 4(a~b). It can be seen that the model of this paper reasonably establishes the words related to the object, and the correlation all reaches more than 0.8, while the correlation of the original model's modeling results for the correlation of these words with the picture area is not more than 0.6, which verifies the effectiveness of the improvement scheme of this paper.



(a)Original model (b)The proposed

Figure 4: Comparison of attention weight visualization results

3.2 Animated character emotion expression migration

3.2.1 Lip-sound synchronization and quality of facial expression generation

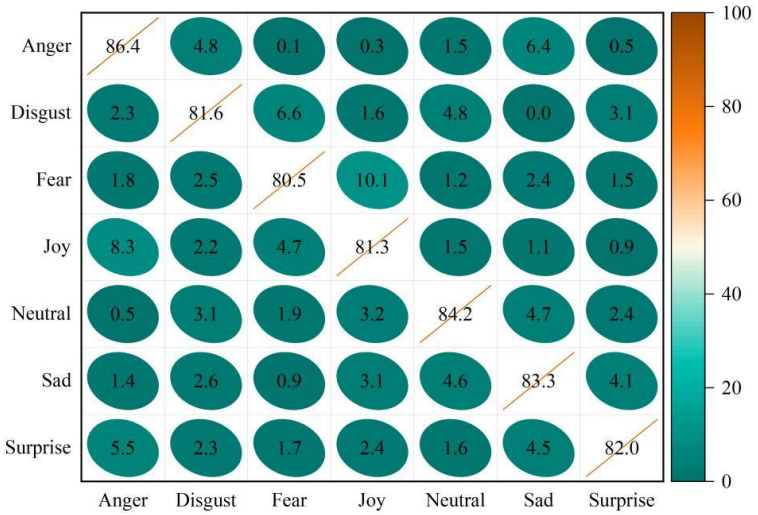
Further experiments on modeling emotion and behavior of animated characters are carried out. This paper will be compared and analyzed with four state-of-the-art methods in the field, which are AVTG, MakeItTalk, MEAD, and EVP.The quality of the animated characters' lip-synchronization and facial expression generation will be evaluated using the coordinate-distance discrepancy (LD) and coordinate-velocity-distance discrepancy (LVD), which are commonly used in the field. LD denotes the average Euclidean distance between the predicted coordinate position of this generated portrait and the reference position.LVD denotes the average Euclidean distance between the coordinate velocities of the two sequences, where the velocity is the difference in coordinate position between neighboring frames. Before evaluation,

the generated coordinate sequences need to be aligned with the real reference one and the facial coordinates are extracted from them. In this paper, two sets of LD and LVD metrics are set up, where the lip coordinate distance difference (L-LD) and lip coordinate velocity distance difference (L-LVD) are used to measure the level of lip-sound synchronization, and the facial coordinate distance difference (F-LD) and facial coordinate velocity distance difference (F-LVD) are used to measure the level of facial emotion. The results of lip-sound synchronization and facial expression generation quality are shown in Table 5. It can be seen that the four evaluation indexes of this paper's model are optimal values in all models, with L-LD, L-LVD, F-LD, and F-LVD being 2.33, 1.55, 2.43, and 1.26, respectively.

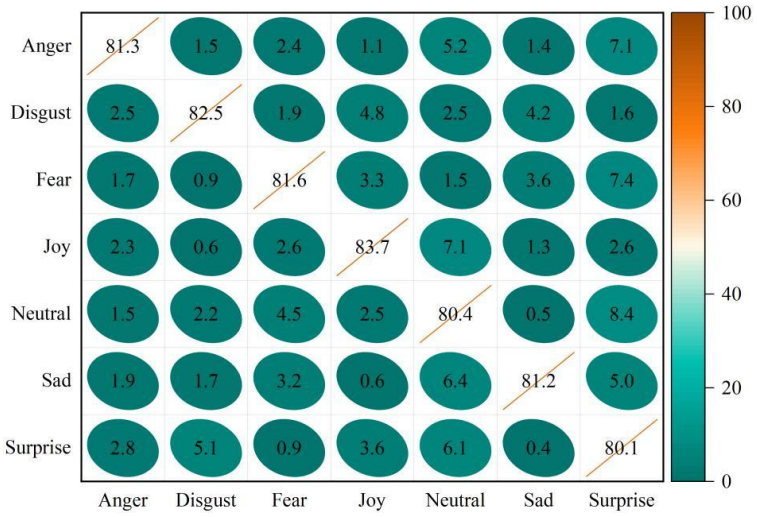
Table 5: Quality results for labial synchronization and facial expression generation

	L-LD	L-LVD	F-LD	F-LVD
AVTG	4.09	2.18	4.02	1.79
MakeItTalk	3.22	1.79	3.78	1.72
MEAD	3.01	2.06	3.42	1.64
EVP	2.56	2.48	2.69	1.43
The proposed	2.33	1.55	2.43	1.26

The research on facial emotion validity used a confusion matrix with the aim of comparing real data with the method proposed in this paper to investigate whether the proposed method is true and effective in rendering facial emotions of animated characters. A total of 20 subjects participated in the research where participants were asked to judge the emotions of the generated videos. A total of 128 animated character facial animations were generated, and the background sound was removed from these videos in order to allow participants to categorize them visually. The results of the facial emotion validity study are shown in Figure 5(a~b). The overall classification accuracy of the real data is 82.76%, and the overall classification accuracy of this paper's method is 81.54%, which verifies that the proposed method has a certain degree of effectiveness in emotion control, and demonstrates the ability of this paper's facial emotion coordinate animation generation network for emotion control.



(a)Real data



(b)The proposed

Figure 5: Survey results of facial affective validity

3.2.2 Affective Index Assessment Analysis

Fifteen videos were selected and participants were further invited to analyze the assessment of sentiment indices, which were assessed by two objects: (1) Participants selected the sentiment indices in the whole video that had the highest degree of conformity with the true sentiment and rated them. (2) Participants selected the sentiment index that most deviated from the true sentiment in the whole video and rated it. The evaluation score range was 0~100, with 0 representing that the sentiment index was extremely inconsistent with the real emotion, and 100 representing that the sentiment index was completely consistent with the real emotion. The results of the emotion index assessment are shown in Table 6.The emotion most consistent with the true emotion in the 15 groups was happiness (60%), and the emotion most deviated from the true emotion was disgust (46.67%).

Table 6: Evaluation results of affective index

Video number	The most consistent expression score	The most consistent expression	Least consistent expression rating	The least consistent expression	Overall rating
1	89.4	Joy	47.3	Neutral	82.35
2	78.6	Sad	38.4	Disgust	81.09
3	80.2	Fear	40.2	Disgust	82.44
4	85.1	Surprise	44.6	Neutral	83.92
5	90.5	Joy	49.1	Neutral	84.21
6	88.3	Joy	50.2	Sad	83.09
7	89.6	Joy	39.8	Disgust	80.47
8	82.4	Surprise	33.5	Disgust	82.55
9	83.9	Fear	40.6	Surprise	81.36
10	90.5	Joy	42.1	Fear	82.45
11	87.3	Sad	32.6	Disgust	80.23
12	89.1	Joy	33.7	Disgust	82.17
13	90.2	Joy	42.8	Surprise	83.42
14	83.4	Joy	33.5	Disgust	80.24
15	82.3	Angry	40.6	Neutral	80.12

A confidence interval gives a range within which the true value of a population parameter falls with a certain probability, and the probability that the population parameter falls within that range is the confidence level. For the choice of confidence level, statistically the results at the 95% confidence level are generally considered to be statistically significant. Based on the data in Table 6, the confidence level analysis of the scores of the emotional index and the true emotional match in the 15 videos was performed, and the results of the analysis are shown in Figure 6. The mean value of the video assessment scores was 82.02, and the confidence interval of the mean value was [80.23, 83.90] (95% confidence level). The interval [80.23, 83.90] contains the overall mean of the overall evaluation score of the sentiment index.

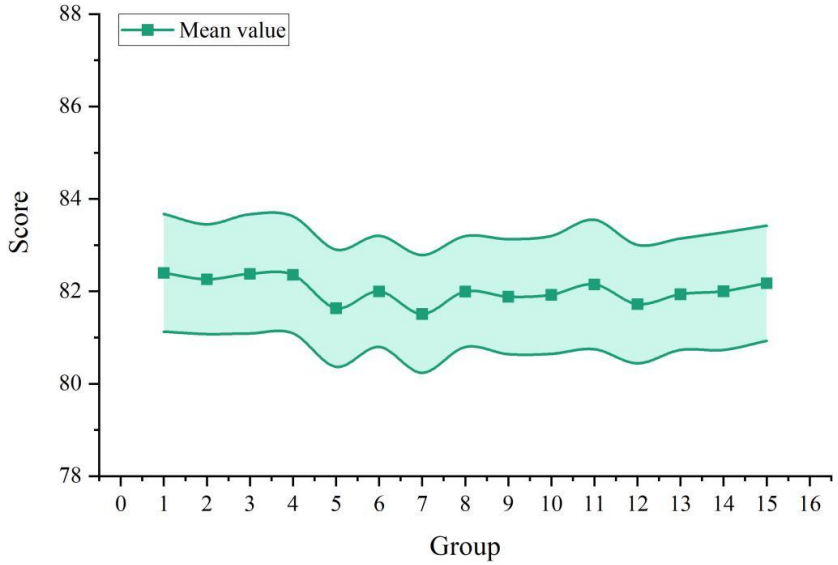


Figure 6: Results of confidence analysis

4 Conclusion

In this paper, an animated character emotion expression migration model is designed based on multimodal fusion and attention mechanism, and its practical effect is explored through experiments.

For the text and speech modal emotion recognition tasks in unimodal scenes, this paper's model achieves a weighted average F1 value of 50.22% and 40.37%, respectively. For the multimodal scene, the model in this paper still performs optimally (60.38%). Compared with the TBJE model with sub-optimal performance, the four indexes of precision, recall, F1-measure and accuracy of this paper's model are improved by 1.94%, 2.29%, 2.11% and 2.33%, respectively. In the animated character emotion expression migration task, the overall classification accuracy of this paper's method differs from the real data by only 1.22%, and the overall scores in the emotion index evaluation are all above 80.

The research results of this paper break through the limitation of single modal analysis, and through the deep integration of multimodal information and attention guidance, it significantly improves the recognition accuracy and migration effect of animation character emotion expression, and provides a feasible technical solution for the intelligent and automated emotion design in animation production.

Funding

This research was supported by the Shanxi Province Art Science planning project: "Research on Innovative Promotion of Cultural Heritage under Virtual Interactive Vision" (Project Number: 23BG092).

References

[1] Cheng, B. (2021). Evaluation and analysis of the spread effect of domestic animation film works based on behavioral psychology. Psychiatria Danubina, 33(suppl 6), 162-164.

[2] Jiang, L. (2022). Expression of emotion and art in film and television animation from the perspective of color psychology. Psychiatria Danubina, 34(suppl 5), 69-69.

[3] Karras, T., Aila, T., Laine, S., Herva, A., & Lehtinen, J. (2017). Audio-driven facial animation by joint end-to-end learning of pose and emotion. ACM Transactions on Graphics (ToG), 36(4), 1-12.

[4] Zong, M., Qi, Z., & Zong, Z. (2020, March). Research on character expression shaping in animation movies. In 4th International Conference on Culture, Education and Economic Development of Modern Society (ICCESE 2020) (pp. 151-155). Atlantis Press.

[5] Wang, Q., & Sharudin, S. A. (2025). The Impact of Color in Animation Scenes on Emotional Expression. International Journal of Literature and Arts Studies, 1(1), 64-71.

[6] Wu, J., Wu, J., Cheng, C. W., Shih, C. C., & Lin, P. H. (2021). A study of the influence of music on Audiences' cognition of animation. Animation, 16(3), 141-156.

[7] Garza, M., Akleman, E., Harris, S., & House, F. (2019, September). Emotional silence: Are emotive expressions of 3D animated female characters designed to fit stereotypes. In Women's Studies International Forum (Vol. 76, p. 102252). Pergamon.

[8] Wang, L. (2024). Artificial Intelligence in Animation Creation: Multi-Dimensional Innovation and Integration–Applications in Artistic Expression and Production Efficiency. Asia-pacific Journal of Convergent Research Interchange (APJCRI), 11-29.

[9] Pan, Y., Zhang, R., Wang, J., Ding, Y., & Mitchell, K. (2023, October). Real-time facial animation for 3d stylized character with emotion dynamics. In Proceedings of the 31st ACM international conference on multimedia (pp. 6851-6859).

[10] He, J. (2024). Exploring style transfer algorithms in Animation: Enhancing visual. Entertainment Computing, 49, 100625.

[11] Chen, Y., & Ma, F. (2024). Research on matching methods of detail features of animated characters in digital media virtual environments. Journal of Computational Methods in Sciences and Engineering, 14727978251360980.

[12] Xue, Z., & Marculescu, R. (2023). Dynamic multimodal fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2575-2584).

[13] Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., & Sun, C. (2021). Attention bottlenecks for multimodal fusion. Advances in neural information processing systems, 34, 14200-14213.

[14] Chu, K. (2024). Application of animation products via multimodal information and semantic analogy. Multimedia Tools and Applications, 83(9), 26031-26054.

[15] Zhang, C., Yang, Z., He, X., & Deng, L. (2020). Multimodal intelligence: Representation learning, information fusion, and applications. IEEE Journal of Selected Topics in Signal Processing, 14(3), 478-493.

[16] Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning. Neurocomputing, 452, 48-62.

[17] Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., ... & Hu, S. M. (2022). Attention mechanisms in computer vision: A survey. Computational visual media, 8(3), 331-368.

[18] Soydaner, D. (2022). Attention mechanism in neural networks: where it comes and where it goes. Neural Computing and Applications, 34(16), 13371-13385.

[19] Zhu, W., & Zhou, M. (2025). The application of channel attention mechanism for fusion of painting and live-action footage under deep learning and animation generation technology. Scientific Reports, 15(1), 41114.

[20] Liang, B., Li, Y., & Lv, Y. (2022, June). Image animation via joint attention mechanism. In 2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC) (Vol. 10, pp. 536-539). IEEE.

[21] Zhu, H., Wang, Z., Shi, Y., Hua, Y., Xu, G., & Deng, L. (2020). Multimodal Fusion Method Based on Self-Attention Mechanism. Wireless Communications and Mobile Computing, 2020(1), 8843186.

[22] Xie, J., Wang, J., Wang, Q., Yang, D., Gu, J., Tang, Y., & Varatnitski, Y. I. (2023). A multimodal fusion emotion recognition method based on multitask learning and attention mechanism. Neurocomputing, 556, 126649.

[23] Li, P., & Li, X. (2020, July). Multimodal fusion with co-attention mechanism. In 2020 IEEE 23rd International Conference on Information Fusion (FUSION) (pp. 1-8). IEEE.