Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.085839

ARTICLE

A Multi-Modal Approach to Emotion Recognition Fusing EEG and Eye Movement in Virtual Reality

Junjie Wu, Yang Liu, Danyi Sheng and Shiwei Cheng*

College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China

*Corresponding Author: Shiwei Cheng. Email: swc@zjut.edu.cn

Received: 19 May 2026; Accepted: 01 July 2026; Published: 15 September 2026

ABSTRACT: With the development of brain-computer interfaces (BCI), more and more studies are using electroencephalography (EEG) for emotion recognition. Traditional emotion recognition often uses 2D videos and pictures to stimulate emotions, which do not provide an immersive feeling. Virtual reality (VR) can provide a more immersive and realistic experience, and recent studies are beginning to utilize EEG for emotion recognition in VR. However, due to the limited information on single-modal features, it is not possible to fully recognize individual emotions. To address this problem, we proposed a multi-modal approach in VR, which utilized a VR scene featuring videos to stimulate participants' emotions and simultaneously extracted and complementarily fused the features of EEG and eye movement. Then, we introduced a cross-attention mechanism to recognize multiple emotional states. The experimental results showed that our approach achieved 90.04% classification accuracy (for four emotional states) and 97.17% (for three emotional states) on the public dataset SEED-IV and the self-collected dataset VR-EED, respectively. This results not only outperformed the single-modal emotion recognition approach (pure EEG or pure eye movement-based) but also surpassed the traditional multi-modal emotion recognition approach (feature layer fusion or decision layer fusion-based). This verifies the superiority of the proposed fusion strategy and is valuable for multi-modal emotion recognition, promoting the development of affective computing applications in VR. This work demonstrates how emotion-aware VR systems can enable adaptive and context-aware interactions in immersive environments.

KEYWORDS: Brain-computer interaction; affective computing; eye movement; multi-modal interaction

1 Introduction

Emotion is a subjective and conscious mental state or process that people experience in response to internal or external stimuli, reflecting their feelings and cognition of a particular thing. Emotional communication is a vital component of people's social lives and can have a profound impact on their daily lives and professional work. With the introduction of "affective computing," scholars are committed to the mathematization of affective concepts, enabling computers to recognize, process, identify, and classify affective states. With the continuous progress of human-computer interaction technology, machine learning, and deep learning, affective computing has gradually penetrated various fields, including healthcare, media and entertainment, information retrieval, education, and intelligent wearable devices [1], showcasing a wide range of application prospects [2]. Therefore, improving the accuracy of emotion recognition effectively has become an important issue.

In emotion recognition, researchers have employed various types of signals as the basis for analysis, which can be broadly categorized into two main types: non-physiological signals and physiological signals.

Among them, physiological signals are controlled by the human body's autonomic, nervous, and endocrine systems. They are not controlled by subjective consciousness, so they can objectively and realistically reflect an individual's physiological, psychological, and emotional changes [3,4]. The main physiological signals commonly used in the field of emotion recognition include electroencephalography (EEG), eye tracking, electromyography (EMG), electrocardiography (ECG), galvanic skin response (GSR), and respiratory signals [5,6]. Among them, EEG, which records brain activity in the central nervous system, has good temporal characteristics, can directly reflect brain activity, and has been shown to provide informative features in response to emotional states [7]. Therefore, EEG-based emotion recognition has been widely studied in recent years [8–10].

Virtual reality (VR) provides users with a more immersive experience compared to traditional 2D displays. As a result, an increasing number of researchers are utilizing VR environments to evoke emotions [11], which in turn yields higher-quality physiological signals. However, due to the blocking of VR headsets, few existing studies have considered eye movement signals when using VR to induce emotions, although recent work has shown the feasibility of synchronized EEG and eye-tracking recording in fully immersive VR [12]. At the same time, eye-tracking has gained popularity in many fields as a means of emotion assessment because it gives researchers a window into the user's visual, emotional, and cognitive processes [13,14]. Studies have shown that eye-tracking can be a reliable method for evaluating emotional responses when individuals are stimulated [15]. It is necessary to consider eye movement features as a basis for emotion recognition.

EEG and eye-tracking signals reflect internal neural patterns and external subliminal behavioral information, respectively, and recent EEG-based multimodal emotion-recognition studies emphasize the complementarity between EEG and other physiological or behavioral modalities [9,10]. This complementarity enables the effective fusion of EEG and eye movement signals, thereby improving emotion recognition accuracy. Therefore, utilizing EEG and eye movement data for emotion recognition in VR is a topic worthy of in-depth exploration. In recent years, researchers have attempted to integrate multi-modal data to enhance the effectiveness of emotion recognition. However, there are several limitations in the existing studies, including the insufficient extraction of single-modal EEG features and the lack of in-depth fusion of EEG and eye movement multi-modal complementarities. For this reason, this paper aims to investigate an emotion recognition method that fuses EEG and eye movement data. First, we allowed users to watch emotion-evoking videos in VR to enhance user immersion. Second, we proposed a 3D-CNN-Transformer-based multidimensional feature extraction method for EEG signals in the time, spatial, and frequency domains to improve the classification accuracy of emotion recognition. Finally, we conducted the fusion of EEG and eye movement in VR and designed an effective fusion strategy to improve the accuracy of multi-modal emotion recognition.

The main contributions of this work are summarized as follows. (1) We constructed a VR-based EEG–eye movement emotion dataset, VR-EED, in which immersive emotional video stimuli were used to induce emotional responses while EEG and eye-movement signals were simultaneously recorded. (2) We proposed a 3D-CNN-Transformer-based EEG feature extraction module to jointly model spatial, spectral, and temporal information from EEG signals. (3) We designed a cross-attention-based multimodal fusion strategy to capture the complementary relationship between EEG and eye-movement features. (4) We validated the proposed method on both the public SEED-IV dataset and the self-collected VR-EED dataset, and the experimental results demonstrate the effectiveness of the proposed multimodal fusion framework.

2 Related Work

2.1 Inducing Emotions in VR

Most studies usually use passive emotion stimulation mechanisms, such as viewing images, watching videos, and listening to music. Both static and dynamic visual stimuli are effective in inducing emotions [16]. For example, users can be asked to watch images or videos designed to elicit changes in emotion [11,17,18]. Alzeer Alhouseini et al. produced different emotions by having users watch pictures [19]. In contrast, Katsigiannis and Ramzan induced multiple emotions by having users watch movie clips corresponding to different emotions [20]. However, with the development of VR, which offers a high sense of immersion and can provide a more immersive experience, scholars have gradually been able to induce users' emotions in VR. For example, Felnhofer et al. induced different emotions by having users experience different scenes in a VR environment [21]. Compared to traditional 2D displays, VR can provide a more immersive environment for emotion induction [16,22]. To summarize, VR can make users feel as if they are in an immersive environment, allowing them to focus fully on the current scene and generate stronger emotions [23,24].

2.2 Multi-Modal Emotion Recognition

In recent years, an increasing number of researchers have begun to focus on feature correlation and information integration across multiple modalities. For example, Jin [25] proposed a new multi-modal fusion method, which utilizes a multi-spatial self-attention mechanism to extract higher-order features strongly correlated with the emotional state between the two modalities of EEG and eye movements. They used the joint attention for eye movement feature screening to take full advantage of the complementary nature of different modalities in terms of emotion representation ability, effectively improving the emotion recognition performance, achieving 73.03% recognition accuracy on the SEED-IV dataset, and 77.96% and 82.72% on the arousal and valence dimensions of the MAHNOB-HCI dataset, respectively. Qiu et al. [26] proposed a correlated attention network (CAN), a multi-modal emotion recognition model, which calculates different types of correlations between gated cyclic units. This incorporates feature correlations of EEG and eye movement signals into the attention mechanism, thereby realizing a coordinated representation of feature complementarity. Experimental results on the SEED-IV and DEAP datasets show that the CAN model significantly improves emotion recognition accuracy. In addition, existing methods often fail to fully utilize the joint dimensional information of EEG, i.e., the spatial and temporal relationships between different electrodes, which may limit the model's ability to extract and represent complex emotional features. Recent EEG-based emotion recognition studies have also explored temporal-spatial-graph modeling for more effective EEG representation learning. For example, Abdulwahhab et al. [27] proposed TSConv-GAT, a hybrid temporal-spatial encoder-decoder framework with mRMR graph topologies and GATv2Conv for EEG emotion recognition. This work provides a relevant benchmark for advanced EEG-based affective computing.

Although much progress has been made in emotion recognition by fusing EEG and eye-movement signals, some issues remain to be resolved. For example, the depth of information in different modal features may be insufficient, limiting the accuracy of emotion recognition. In addition, most fusion methods ignore the interaction information between modalities and restrict the performance of emotion recognition. Therefore, future research can further explore how to extract and utilize modal features, as well as how to fuse interaction information between different modalities more effectively, to enhance the accuracy and reliability of emotion recognition.

2.3 Multi-Modal Fusion Strategies

Current multi-modal fusion strategies based on EEG and eye movements focus on the fusion of feature or decision layers. Soleymani et al. [7] used SVM algorithms for emotion recognition research, conducted single-modal experiments with eye movements and EEG on the MAHNOB-HCI database, and then used the summation rule in the fusion of decision layers to perform multi-modal emotion recognition, and reached an accuracy rate of 76.4% and 68.5% on the two emotion dimensions of potency and arousal with an accuracy of 76.4% and 68.5%, respectively. Recent deep fusion approaches further indicate that direct splicing or late decision aggregation is insufficient for modeling the interaction between EEG and eye movement signals. For example, attention-based feature-fusion models can learn modality-specific representations and then selectively integrate complementary information across modalities [28], while recent reviews of EEG-based multimodal emotion recognition also emphasize the need to model inter-modal dependency rather than simply concatenating heterogeneous features [10]. Although decision-layer fusion integrates the emotion recognition results of individual modalities, it ignores the complementarity of features and potential redundant information between different modalities. Meanwhile, traditional feature-layer fusion methods often directly splice multi-modal features, failing to utilize the feature complementary relationship between modalities effectively.

2.4 Transformer-Based Multi-Modal Fusion

The Transformer model has shown significant advantages in multi-modal emotion recognition. It excels in integrating modal information due to its ability to focus on continuous emotion data on a global scale and emphasize regions contributing to emotion recognition using a self-attention mechanism. Wang et al. [29] proposed an emotion Transformer Fusion (ETF) based on a pure attention mechanism model, which combines a Transformer encoder with an attentional mechanism-based fusion to exploit the parallelism and simplicity of emotion recognition with EEG and eye movement signals, achieving 90.02% accuracy on a homebrew dataset, which significantly improves the ability to discriminate between anger, surprise, and neutral emotions. Fu et al. [30] proposed a new multi-modal feature fusion neural network model. The multiscale feature fusion module therein utilizes cross-channel soft attention to adaptively select information from multiple event space scales for feature acquisition and effective fusion, achieving 87.32% accuracy on the SEED-IV dataset. These studies demonstrate the capability of the Transformer-based architecture for multi-modal emotion recognition using EEG and eye movement signals. Attention-based EEG-eye-movement fusion has also been extended through deep feature extraction and adaptive attention fusion [28]. Despite the significant progress of Transformer-based architectures in multi-modal emotion recognition, several critical limitations remain when these models are applied to immersive VR scenarios. First, most existing approaches treat multi-modal fusion as a static or symmetric process, assuming fixed correlations between EEG and eye movement signals. However, in VR environments, emotional responses are highly dynamic and context-dependent, where the relative contribution of neural activity and visual behavior may vary over time. Second, many Transformer-based fusion models rely on heavy architectures with a large number of parameters, which increases training complexity and limits their applicability to real-time or adaptive VR systems. Moreover, existing methods often focus on improving classification accuracy, while paying less attention to modeling the temporal interaction and mutual influence between internal neural states and external behavioral cues.

To address these challenges, our proposed cross-attention multi-modal fusion framework explicitly models the bidirectional temporal interaction between EEG signals and eye-tracking signals. By allowing each modality to dynamically attend to the other across time, our approach captures fine-grained complementary information that static fusion or self attention mechanisms fail to exploit. This design is particularly

suitable for VR-based emotion systems, where users' affective states continuously evolve in response to immersive stimuli.

3 Approach

Our approach constructed a deep learning model as shown in Fig. 1, which mainly contains feature extraction, dimension alignment, cross-attention fusion, and classification. In the feature extraction stage, EEG features were extracted by the 3D-CNN-Transformer, and the Transformer extracted eye movement features. In the dimension alignment stage, a multi-layer perceptron (MLP) was used to convert various features into a unified dimension, laying the foundation for subsequent fusion operations. The cross-attention fusion mechanism established cross-attention connections between features of different modalities of EEG and eye movement. It dynamically adjusted the importance of features of various modalities to improve the performance of emotion recognition. Finally, the emotion classification was achieved using a fully connected layer and the SoftMax function.

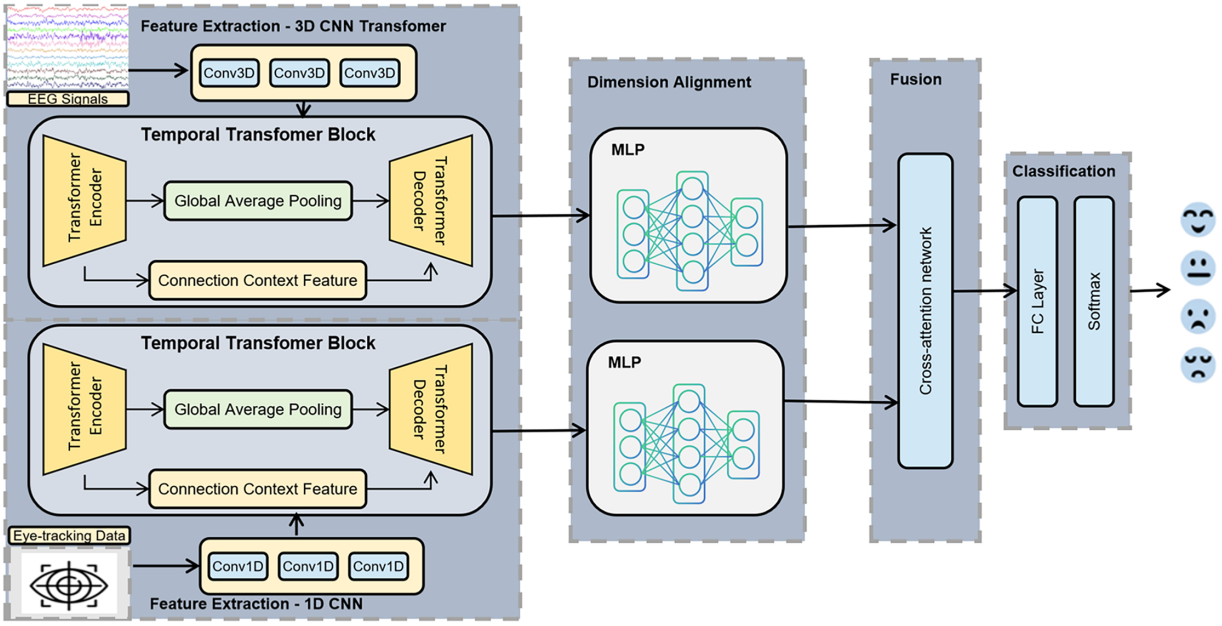

Figure 1: The framework of the proposed model. The model we proposed consists of four parts: feature extraction, dimension alignment, feature fusion, and a classifier. The feature extraction module is divided into EEG data feature extraction and eye movement data feature extraction, both of which include a convolution module and a Temporal Transformer Block; feature alignment is achieved by a multi-layer perceptron (MLP); feature fusion uses the cross-attention fusion method proposed in this paper; and after feature fusion, full connection is performed for classification.

3.1 Feature Extraction

For multi-modal feature fusion to achieve good emotion recognition, the prerequisite is to extract high-quality emotion features on a single modality. The feature extraction task is to obtain high-quality features from the EEG and eye movement features, respectively. For the problem of insufficient EEG feature dimension extraction, we proposed an EEG emotion feature extraction method based on 3D-CNN-Transformer, with the framework shown in Fig. 1, including 3D feature organization, spatial-frequency-spectral convolution, and temporal coding and decoding, in which the 3D feature organization aims at extracting EEG spatial-frequency-domain-temporal three-dimensional features as the inputs. The spatial-frequency-spectral convolution used three convolutional layers to extract spatial-frequency domain features from each time

slice. Temporal coding and decoding connected four Transformer encoding layers and four Transformer decoding layers through a U-shaped structure to obtain contextual features.

Fig. 2 shows the 3D feature generation process. To fully utilize the time-domain information of EEG data, the T-second EEG signal was segmented into N time slices using a 0.5-s non-overlapping sliding window. Since the EEG signals were acquired from multiple channels, each corresponding to a specific brain region, this paper retained the channels of the EEG signals, thus enabling the preservation of critical spatial information. In addition, studies have shown that the high-frequency portion of the EEG signal has a more significant impact on emotion recognition than other frequency portions [8,10]. Therefore, we divided the EEG signal into multiple frequency bands and extracted features from each band. For each frame in the sample, it was decomposed into five frequency bands based on the Fourier transform, which were delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–42 Hz). Since Power Spectral Density (PSD) and Differential Entropy (DE) features were effective in EEG emotion recognition [8,27], we extracted PSD and DE features separately for each frame in all five frequency bands of each channel.

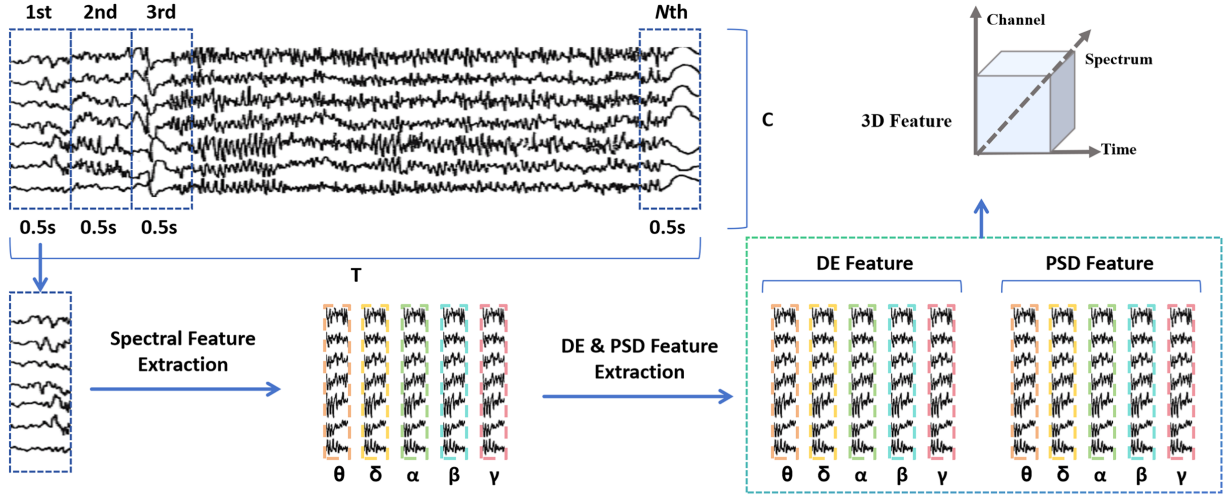

Figure 2: 3D feature generation process.

EEG signals contain information between multiple channels and frequency domains. Therefore, the correlation features between different channels and frequency domains must be considered in the model construction process. The main task of spatial-frequency domain convolution is to extract spatial-spectral features from each channel and frequency domain. This model used three consecutive convolutional layers; the number of filters in the three convolutional layers were 8, 16, and 32, respectively, and the size of the convolutional kernel of each convolutional layer is set to 3×3. A bottleneck structure was used between each convolutional layer to extract spatial-frequency-domain features better.

To more effectively decode emotions from EEG time series, we proposed a Temporal Transformer Block (TTB) that incorporates a multi-communicative attention mechanism and a residual attention layer, as shown in Fig. 3a, and employed a U-shaped structure similar to the U-Net to connect the underlying features to the higher-level features effectively, as shown in Fig. 3b, thereby helping the model to better capture the contextual information in the time series.

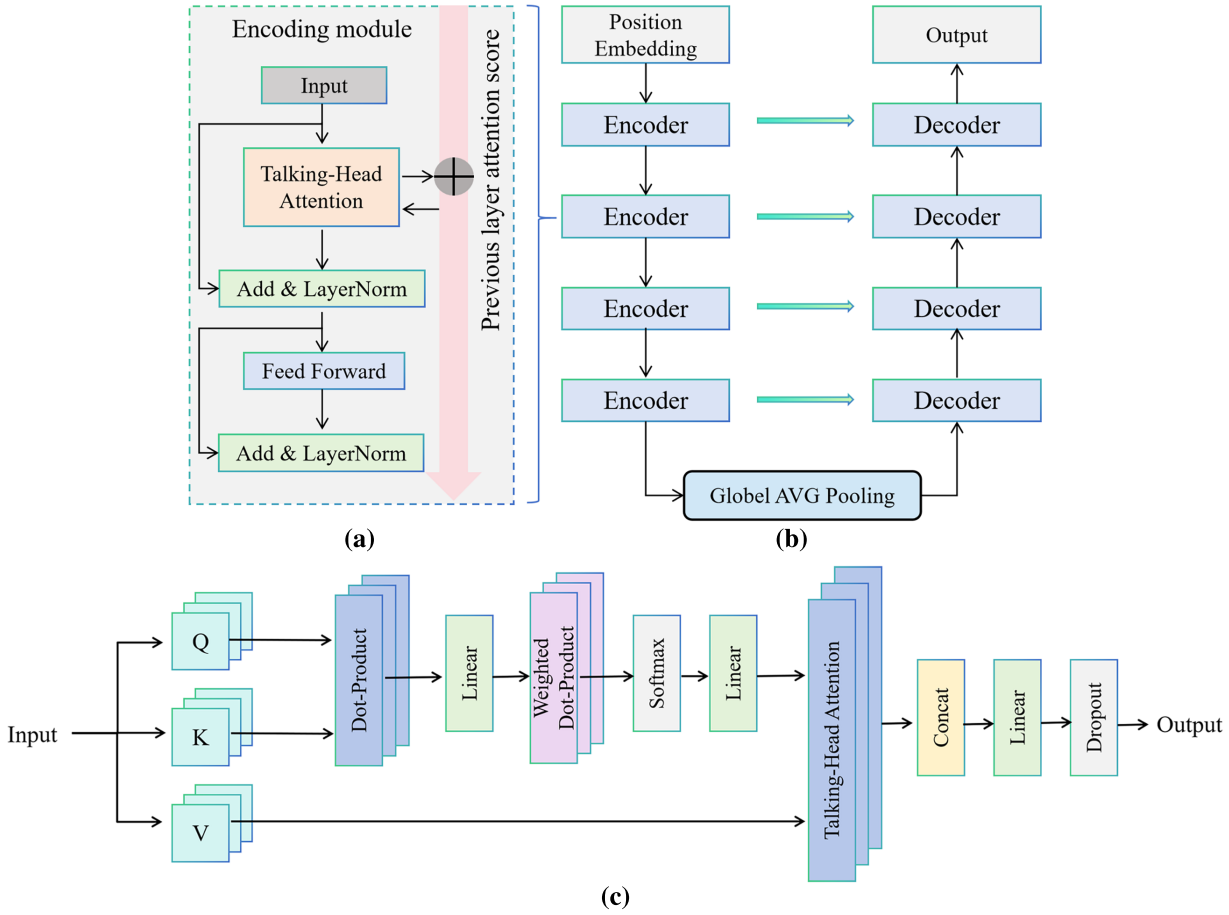

Figure 3: Architecture of the temporal transformer block: (**a**) encoding module with Talking-Head Attention and residual attention scores; (**b**) U-shaped temporal encoder-decoder structure; (**c**) Talking-Heads Attention block.

The traditional Multi-Heads Attention mechanism used in Transformer [31] features different attention heads that are independent of each other and perform separate fixation computations, with no explicit mechanism for dynamically adjusting the importance of the heads relative to one another. This limitation restricts the modeling capabilities in addressing long-distance dependencies, such as those found in emotional EEG. Therefore, we introduced Talking-Heads Attention (THA), as shown in Fig. 3c. It was characterized by the introduction of a learnable head selection mechanism, which allows the attention heads to dynamically "talk" with each other; the calculation process is shown in Eq. (1):

$$THA(Q, K, V) = \text{Concat}\left(W_1 \cdot head_1, W_2 \cdot head_2, \cdots, W_h \cdot head_h\right) W^O \tag{1}$$

where h denotes the number of attention heads, the query, key, and value of the $i - th$ attention head, λ is a scaling parameter, and W_{i} is the attention head's weight corresponding to that position.

Compared with conventional multi-head attention, in which different attention heads are computed independently and only concatenated at the output stage, Talking-Head Attention allows information exchange across attention heads through learnable head-wise transformations. This design enables the model to dynamically adjust the relative importance of different attention patterns. For emotional EEG signals, discriminative information is often distributed across multiple channels, frequency bands, and temporal segments. Different attention heads may focus on different spatial-spectral-temporal dependencies, while the

interaction among heads can help reduce redundant attention responses and enhance complementary feature representations. Therefore, Talking-Head Attention is suitable for improving the representation ability of EEG temporal modeling and can contribute to more robust emotion recognition performance.

In Transformer, applying the residual structure to the attention layers is adding residual connections between neighboring attention layers, as shown in Fig.3a, where adding a jump edge connecting the multi-communicating attention modules in the neighboring layers means that the output of the previous layer (usually the attention scores after a SoftMax operation) was used as an additional input to the current layer. This connection created a "short circuit" that makes it easier to pass information from one layer of the network to another without having to go through all the intermediate layers. This helped maintain the message's integrity as it travels and reduces the loss of information due to too many layers. More intuitively, a parameter Prev, the attention score of the previous layer, was added as an additional input to the attention of the current layer's multiple exchanges with the following formula:

$$Res - THA(Q, K, V, Prev) = \text{Concat}\left(W_1 \cdot head_1, W_2 \cdot head_2, \cdots, W_h \cdot head_h\right) W^O \tag{2}$$

where, $head_i = Res - Attention(QW_i^Q, KW_i^K, VW_i^V, Prev_i)$, $Res - Attention$ added the residual scores on top of the $Prev_i$, and then calculates the weighted sum, with the following formula:

$$Res - Attention\left(Q', K', V', Prev'\right) = \text{softmax}\left(\frac{Q'K'T}{\sqrt{d_k}} + Prev'\right) V' \tag{3}$$

Finally, the new attention score was passed to the next layer, and this structure enabled the model to utilize both the computational results of the current layer and the information of the previous layer. The effective utilization of information between different layers enabled the model to capture critical features in the EEG signals more accurately, thereby improving the expressiveness and robustness of the model.

The eye movement data is a time series. Considering the advantages of the Temporal Transformer Block model in time-series data processing, especially its multi-head communicative attention mechanism that can effectively capture temporal dynamic features, we also choose to use this model for the implementation, which contains position coding, multi-head communicative attention module, feed-forward neural network, and a global average pooling layer to assist in extracting the eye movement temporal feature dependencies in the signal.

3.2 Dimension Alignment

After extracting the EEG and eye movement multi-modal features, respectively, because of the difference in the dimensionality of the two types of features, it is necessary first to perform dimensional alignment before multi-modal feature fusion to facilitate the subsequent fusion operation. We converted the two types of features to the same dimensionality employing a Multilayer Perceptron (MLP), which is a feed-forward neural network consisting of multiple fully connected layers, each of which uses a nonlinear activation function, ReLU, that maps the features to a new space, thus aligning the dimensionality of the original features.

Assuming that the original EEG feature vector dimension is d_{eeg} and the original eye movement vector dimension is d_{eye}, the goal is to map both to dimension d_t. For the EEG and eye movement features, a multilayer perceptron was designed for mapping, respectively. Let each MLP have L layers (here let $L = 3$), for each layer, the MLP is computed as follows:

$$h_{eeg}^{(l)} = f\left(W_{eeg}^{(l)} \cdot h_{eeg}^{(l-1)} + b_{eeg}^{(l)}\right) \tag{4}$$

where $h_{eeg}^{(l-1)}$ is the output of the layer $l-1$, for $l=1$, there are the original EEG features $h_{eeg}^{(0)} = x_{eeg}$, $W_{eeg}^{(l)}$ is the weight matrix of layer l, $b_{eeg}^{(l)}$ is the bias vector of layer l, and f is the activation function, in this paper, ReLU is used. The output of the last layer L is the dimensionally-aligned EEG features $h_{eeg}^{(L)} = x'_{eeg}$ with dimensionality d_t. Similarly, a similar operation is performed for the original eye-movement features, and the output of the last layer L is the dimensionally aligned eye-movement features $h_{eye}^{(L)} = x'_{eye}$, whose dimension is also d_t. Finally, the dimensionally aligned EEG and eye movement features are obtained, which is convenient for subsequent feature fusion operations.

3.3 Cross-Attention Based Fusion

Both EEG and eye movement data contained information about emotional states, but the correlation between the two modalities is not obvious. Synchronization between EEG signals and eye-movement data in different emotional states is an important prerequisite to reveal their interactions in the temporal dimension, which in turn enhances the interpretation of emotional states. Considering that both EEG and eye movement signals have certain dynamic properties, the emotional state is more affected by the interaction between them. For this reason, we proposed trans-attentive networks that integrate EEG and eye movements. By designing in the time dimension, the network can capture the dynamic interaction information between the two modalities, adaptively selecting and weighting the EEG and eye movement features at different time points, so as to better describe the correlation between them. The internal structure of the trans-attentive network is shown in Fig. 4. Its computation is divided into two steps: attention calculation and feature fusion.

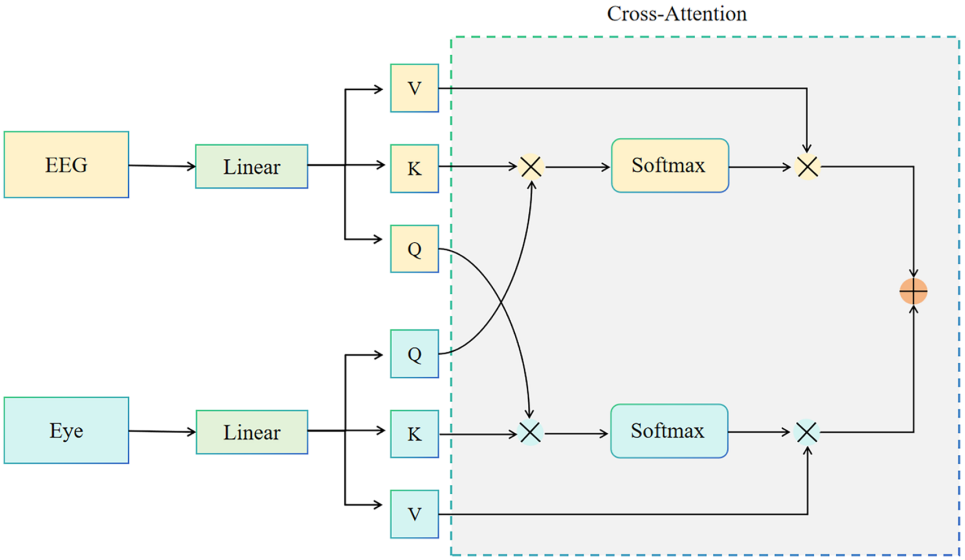

Figure 4: Cross-attention network.

3.3.1 Attention Calculation

First, the attentional weights between two input sources (EEG data and eye movement data) need to be computed, i.e., to determine the importance of one input source to the other. For the EEG to eye-movement cross-attention, it is defined as follows:

$$W_{eeg\to eye} = \text{softmax}\left(\frac{Q_{eeg}K_{eye}^{T}}{\sqrt{d_k}}\right) \tag{5}$$

where Q_{eeg} is the EEG feature matrix after linear transformation as the query matrix, K_{eye} is the eye movement feature matrix after linear transformation as the key matrix, and d_k is the scaling factor of the feature dimension, which is used to prevent the problem of vanishing gradient due to too large a dot product.

The same method was used to calculate the attentional weights in the other direction, which were defined as follows for eye movement to EEG cross-attention:

$$W_{eye\to eeg} = \text{softmax}\left(\frac{Q_{eye}K_{eeg}^{T}}{\sqrt{d_k}}\right) \tag{6}$$

Attentional weights are obtained by calculating the correlation between the query vectors of the EEG data and the key vectors of the eye movement data to introduce information about each other.

3.3.2 Feature Fusion

The purpose of feature fusion is to apply the attentional weights to the original feature vectors to generate the fused feature representation. In the EEG-eye-movement multi-modal emotion recognition task, for the feature vectors of EEG V_{eeg} and the fused attentional weights $W_{eeg\to eye}$, the fused EEG feature representation can be obtained by weighted summation with the following formula:

$$F_{eeg} = V_{eeg} \cdot W_{eeg\to eye} \tag{7}$$

For the feature vectors of eye movements V_{eye} and the fused attentional weights $W_{eye\to eeg}$, the fused eye movement feature representations can likewise be obtained by weighted summation:

$$F_{eye} = V_{eye} \cdot W_{eye\to eeg} \tag{8}$$

The fused EEG features and eye movement features are then connected using Concat to obtain the EEG-eye movement fusion features $F_{eeg-eye}$, so that the EEG data and eye movement data can be fused through the cross-attention mechanism, and the importance of the EEG data to the eye movement data, as well as the importance of the eye movement data to the EEG data, is decided according to the specific attentional weights, thus improving the performance of emotion recognition.

3.4 Emotion Classification

After feature extraction and cross-attention fusion, higher-quality fused features are obtained, which are fed into the classification module to classify emotions using the fully connected layer and SoftMax operations. The fully connected layer can transform high-dimensional feature vectors into output vectors corresponding to the number of sentiment categories, and SoftMax transforms them into probability distributions. Finally, cross-entropy loss is used to compare the predicted labels with the real labels to get the recognition results. The formula is as follows

$$L = -\frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M} y_n^m \log(\widehat{y}_n^m) \tag{9}$$

where N is the batch size and M is the number of categories given by the particular dataset and task. y_n^m and $\widehat{y}_n^m$ are the true label and the predicted probability of the corresponding category, respectively. Finally,

the classification probability of each category is obtained, and the category with the highest probability is considered as the final prediction.

4 Experiment and Results

4.1 Dataset and Experimental Settings

In terms of publicly available datasets, we selected SEED-IV [32]. This multi-modal emotion recognition dataset combines data from two modalities: EEG and eye movement, and contains four basic emotion categories: happy, neutral, sad, and fear. A 64-channel ESI NeuroScan system and an SMI eye-tracker were used to acquire EEG and eye-movement data for this dataset, comprising a total of 15 participants, including seven males and eight females.

In addition, we collected EEG data and eye movement data for different emotions in the VR environment to produce a private dataset, VR-EED. Fifteen participants were recruited for the formal emotion experiment, including 9 males and 6 females. The participants were aged between 20 and 24 years, with a mean age of 22.5 years and a standard deviation of 0.9 years. Of these, 60% reported having used VR before. All participants had no history of psychiatric illness and had normal or corrected vision. Before the experiment, participants' vertigo under VR was assessed on a 7-point Likert scale, and all were found to be free of significant vertigo to rule out the influence of VR vertigo on the experiment. This work was conducted in collaboration with Zhejiang Hospital, China. This study was approved by the Ethics Committee of Zhejiang Hospital (Approval No. 2025-C-095). All participants voluntarily participated in the study and provided written informed consent before data collection. All participants followed the same set of experimental methods, using the Borecon Neusen W 64-channel EEG device and the HTC VIVE VR device (embedded with an eye tracker) for data acquisition. For EEG data, 21 channels (Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, T6, Fz, Cz, Pz) were selected. For eye movement data, three types of features were selected: 3D gaze point, 2D gaze point, and pupil area. The entire experiment was conducted in a quiet lab, with only the participants and researchers present. Finally, standard preprocessing procedures were applied. First, the raw EEG was re-referenced to the common average and down-sampled from 1000 Hz to 100 Hz to reduce computational complexity. Second, a fourth-order Butterworth band-pass filter with cut-off frequencies of 0.5–45 Hz was applied to suppress baseline drift and high-frequency noise while preserving the emotion-related rhythms across the delta–gamma range used in our feature extraction (Section 3.1). Third, ocular (blink), muscular (EMG), and head-movement artifacts were removed using independent component analysis (ICA) with subsequent visual inspection, thereby improving the signal-to-noise ratio [33].

In the VR-EED dataset, emotional responses were induced using emotionally labeled movie clips. The stimulus materials were selected according to their emotion labels, including happy, sad, and neutral states. Each experiment contained 24 movie clips, and each clip lasted approximately 2 min. To improve the diversity of the stimulus materials, each participant completed three experiments at different time intervals, and each experiment used a different set of 24 movie clips. Each trial consisted of a 5-s start cue, a 2-min video-watching stage, and a 45-s self-assessment stage. During the video-watching stage, participants wore the HTC VIVE headset and viewed the emotional stimuli in the VR environment, while EEG and eye-movement signals were recorded simultaneously.

The batch size of our model was set to 64, the fixed learning rate was 0.0001, the weight decay was 0.000001, the number of epochs was set to 200, the cross-entropy loss function was used, and the Adam optimization algorithm was chosen. Five-fold cross-validation was used in the experiment by dividing the data into five subsets, using four of them as the training set each time and the remaining one as the test set. Accuracy (Acc) and standard deviation (Std) were used as evaluation metrics.

Computational complexity analysis. To further evaluate the computational feasibility of the proposed model, we estimated its complexity in terms of parameter count, FLOPs, and average inference latency per 0.5-s EEG–eye-movement time slice. The estimation was based on the network configuration described in Section 3, including the 3D-CNN-based EEG feature extractor, the Temporal Transformer Block, the MLP-based dimension alignment layers, the bidirectional cross-attention fusion module, and the final fully connected classifier. The proposed cross-attention fusion model contains approximately 3.5–4.5 million trainable parameters and requires approximately 0.75–0.80 GFLOPs per 0.5-s time slice. The estimated inference latency is approximately 6–18 ms on a desktop GPU and approximately 70–120 ms on a CPU-only platform, depending on the hidden dimension setting and hardware configuration.

4.2 Procedures

The experiment triggered specific emotional responses by showing participants 24 emotionally labeled movie clips in VR, and participants were required to wear an HTC VIVE headset to watch the corresponding movie clips, each of which lasted approximately two minutes.

Each participant performed three experiments at different time intervals, and each experiment used a completely different set of 24 stimulus materials to ensure the variety and generalization of the results. Each experiment consisted of a 5-s start cue, a 2-min video, and a 45-s self-assessment. The specific experimental process is shown in Fig. 5. Through experiments, we created the dataset VR-EED.

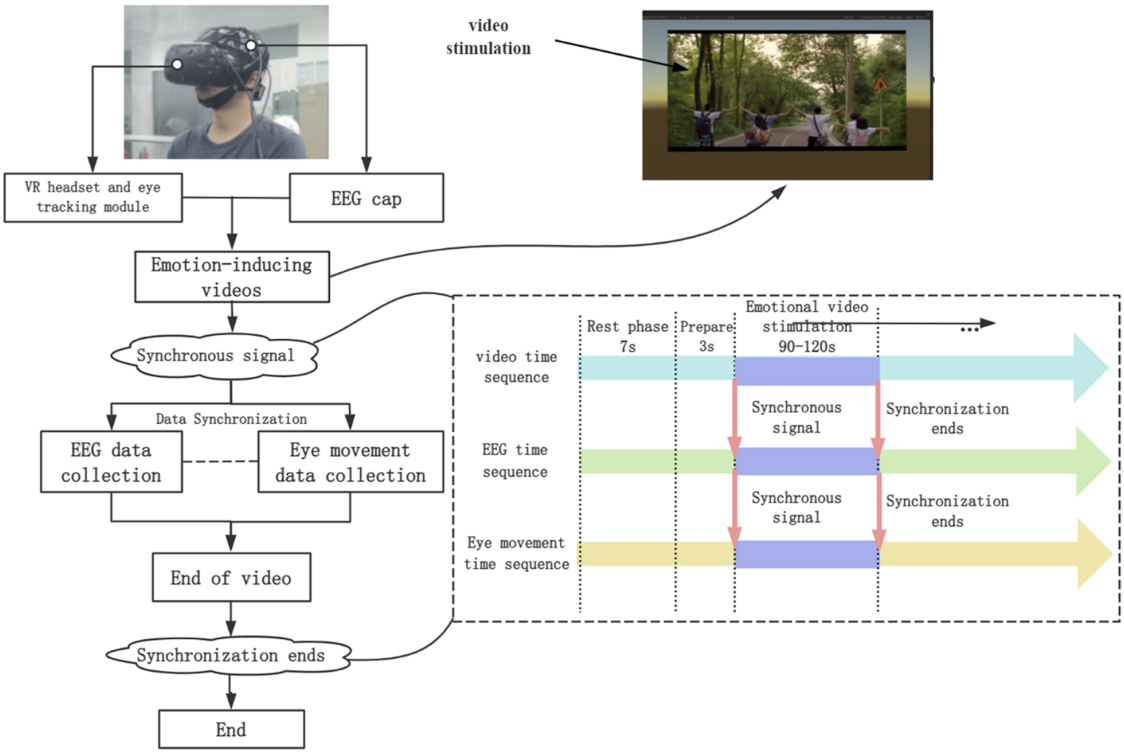

Figure 5: Emotion induction experiment process in VR.

4.3 Results and Analysis

To validate the effectiveness of the proposed cross-attention fusion method model, we conducted experiments on a single participant on both the SEED-IV dataset and our own collected dataset, VR-EED. Our current evaluation demonstrates within-dataset effectiveness on both a public dataset and a self-collected VR dataset, rather than full cross-dataset generalization. Table 1 gives the emotion recognition accuracy and standard deviation of the cross-attention fusion method for single subjects per session on the SEED-IV dataset. Comparing the average accuracies of the same subjects under different sessions, it can be found that the average accuracy of 15 subjects on 3 sessions was 90.04% (*STD* = 4.86%), the highest average accuracy of 93.64% (*STD* = 4.49%) was found in P12, and the lowest average accuracy of 84.92% (*STD* = 7.17%), these differences may stem from individual differences in emotional expression and physiological responses between subjects. Comparing the average accuracy of different subjects within the same session reveals that session 1 has the highest average accuracy of 91.15% (*STD* = 4.23%), while session 3 has the lowest average accuracy of 89.18% (*STD* = 5.35%). This difference may be due to the emotional evocation of video clips under different session abilities being different. Nevertheless, the difference between the average accuracy rates of the three sessions remains within 2%, indicating that the cross-attention fusion model can maintain a more stable performance across different sessions.

Table 1: The results of the cross-attention fusion method on the SEED-IV dataset. Bold numerical values in the AVG row indicate average results across all participants.

Participant No.	Session 1		Session 2		Session 3		AVG	
	Acc (%)	Std (%)	Acc (%)	Std (%)	Acc (%)	Std (%)	Acc (%)	Std (%)
P1	93.55	3.45	89.13	5.13	91.92	4.47	91.53	4.35
P2	90.14	4.38	89.74	6.52	87.30	5.45	89.06	5.45
P3	92.62	4.81	90.48	5.13	87.55	5.39	90.22	5.11
P4	86.84	5.24	89.68	5.22	86.53	6.93	87.68	5.80
P5	85.53	6.35	83.31	7.56	85.90	7.59	84.92	7.17
P6	91.34	3.47	89.67	4.21	85.92	6.73	88.98	4.80
P7	86.71	4.63	90.05	6.63	93.28	4.11	90.01	5.12
P8	85.67	5.30	96.67	2.76	91.18	3.61	91.17	3.89
P9	95.39	3.53	91.15	4.07	92.39	4.14	92.98	3.91
P10	96.15	2.75	83.33	6.51	86.87	6.85	88.78	5.37
P11	98.29	1.62	90.33	4.79	85.43	5.64	91.35	4.02
P12	95.23	3.25	91.87	5.94	93.81	4.27	93.64	4.49
P13	84.26	6.79	93.25	2.39	89.85	5.21	89.12	4.80
P14	91.79	4.12	92.70	4.74	90.45	5.36	91.65	4.74
P15	93.74	3.79	85.56	3.58	89.25	4.45	89.52	3.94
AVG	**91.15**	**4.23**	**89.80**	**5.01**	**89.18**	**5.35**	**90.04**	**4.86**

On the homemade VR-EED dataset, due to the manual introduction of eye-movement features, it is necessary to first filter the eye-movement features and combinations to get the most discriminative features for emotion recognition. Fig. 6 demonstrates the average classification accuracy corresponding to different combinations of eye movement features, where F denotes 3D gaze point features, F' denotes 2D gaze point

features, and P denotes pupil size. From the figure, it can be learned that multiple feature combinations can achieve better results compared to single features in most cases, and the average classification accuracy of feature combinations based on 3D gaze point, 2D gaze point, and pupil size reaches the highest of 77.52%, which proves the superiority of eye movement multi-feature fusion in emotion recognition. In the comparison of single features, the classification accuracy based on pupil size is significantly higher than that based on gaze point only, and the combination of features, including pupil size, also shows significant advantages in classification performance, which further suggests that pupil features may carry richer information in emotion recognition. The advantage of pupil-size features is also supported by physiological evidence. Bradley et al. [34] reported that pupil dilation is closely related to emotional arousal and autonomic activation, and that emotionally arousing pleasant and unpleasant stimuli can elicit larger pupil responses. Therefore, the improved performance obtained by incorporating pupil-size features is physiologically reasonable, as pupil dynamics may provide complementary information related to autonomic activity and affective processing.

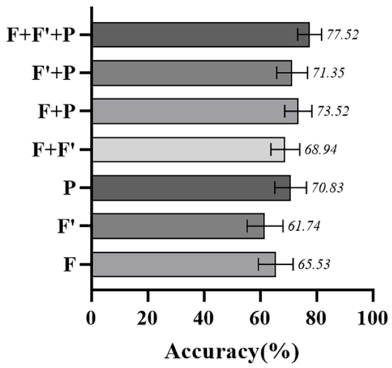

Figure 6: Average accuracy under different combinations of eye movement features.

It is seen that the eye movement features based on the combination of 3D gaze point, 2D gaze point, and pupil size features gave the best classification results, so these three eye movement feature combinations were selected. Fig. 7 shows the classification results of the proposed cross-attention fusion method on our self-collected dataset, VR-EED, and compares it with the unimodal EEG results. The results show that our proposed method achieves an average accuracy of 97.17%, which is an improvement of 1.9% compared to the accuracy of 95.22% obtained from single modal EEG, indicating its effectiveness. Overall, after adding eye movement features, the accuracy of most participants was improved, with P3 showing the most significant improvement of 5.69%, indicating that there is a good complementarity between the EEG and eye movement signals; however, there are also some individual participants who showed a slight decrease in accuracy, such as P7, which showed a decrease of 0.49% compared with the single modal. This decrease may be attributable to lower eye-tracking signal quality for P7, or there is some redundancy between the eye movement signal and the EEG signal in terms of emotional expression, failing to achieve the expected effect when fusing the information.

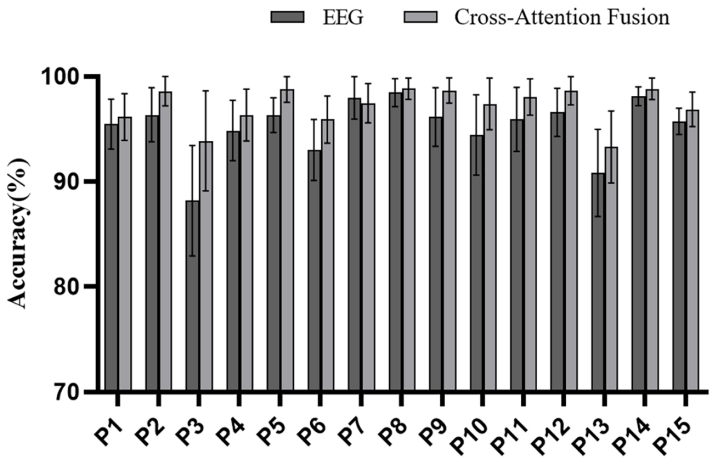

Figure 7: Per-participant classification accuracy of unimodal EEG vs. the proposed cross-attention fusion on the VR-EED dataset.

4.4 Comparison to State-of-the-Art Approaches

To verify the advancement of the proposed cross-attention fusion strategy, we selected representative EEG and eye movement fusion algorithms in recent years for comparison, and the results on the SEED-IV dataset and the VR-EED dataset are shown in Table 2. Among the compared methods, MFFNN [30] constructs a dual-branch feature extraction module to ensure that the features of the two modalities can be extracted while being temporally aligned, and introduces a multi-scale feature fusion module to achieve effective fusion of features of different spatial scales; and DCCA [35], a deep-network-based extension that nonlinearly learns individual representations of each modality, was capable of and learns a shared representation by maximizing the correlation between two modalities, achieving accuracies of 87.45% and 91.12%; PCA + KNN [36] re-quantifies emotional EEG and eye-movement features for classification using statistical tests, achieving accuracies of 88.89% and 90.25%. Our proposed approach introduced a trans-attentive mechanism that can adaptively select and weigh EEG and eye movement features at different time points to capture the correlation between the two modalities more effectively. Experimental results show that our approach achieves a classification accuracy of 90.04% on the SEED-IV dataset, which is 1.15% better than the best result (PCA + KNN), and this significant performance improvement fully validates the superiority of our proposed fusion strategy.

Table 2: Accuracy comparison of the existing methods (%). Bold numerical values indicate the best performance on each dataset, and the bold method name denotes the proposed method.

References	Method	SEED-IV	VR-EED
Fu et al. [30]	MFFNN	87.32	89.53
Liu et al. [35]	DCCA	87.45	91.12
Goshvarpour and Goshvarpour [36]	PCA + KNN	88.89	90.25
Gong et al. [37]	CoDF-Net	89.01	88.21
Ours	**Cross-Attention Fusion**	**90.04**	**97.17**

5 Discussion

5.1 Results and Analysis of Fusion Strategy

We explored multi-modal fusion experiments under different fusion strategies and compared the results with single-modal results. The network structure used in single-modal fusion is the same as the corresponding modal network structure in multi-modal fusion. The results are shown in Table 3, where the feature layer fusion extracted 3D features for EEG data and eye movement features based on the combination of gaze features and pupil size features for eye movement data, which are normalized respectively and then fused by splicing, and the final classification results are obtained by classifying them using fully-connected layer and SoftMax for classification to get the final classification results; the decision layer fusion, on the other hand, first obtained the same EEG features and eye movement features through the classifiers fully connected layer and SoftMax respectively to get the corresponding classification results and then obtained the final classification results through weighted average.

Table 3: Results of multi-modal individual subjects under different fusion strategies on the SEED-IV dataset. Bold numerical values in the AVG row indicate average results across all participants.

Participant No.	EEG	Eye Movement	Feature Fusion	Decision Fusion	Ours
P1	81.46	73.34	83.34	79.28	91.53
P2	75.02	76.53	79.25	77.29	89.06
P3	80.53	78.37	84.78	85.87	90.22
P4	82.45	73.85	85.17	83.39	87.68
P5	82.04	78.45	88.96	82.98	84.92
P6	83.35	80.53	89.73	84.35	88.98
P7	81.42	76.47	85.63	82.53	90.01
P8	81.83	74.36	83.36	79.04	91.17
P9	80.36	81.46	84.81	82.74	92.98
P10	83.72	78.57	86.74	85.86	88.78
P11	82.68	77.74	85.25	83.04	91.35
P12	83.46	78.35	87.89	84.75	93.64
P13	82.47	77.64	84.73	84.26	89.12
P14	79.56	80.74	83.63	81.07	91.65
P15	80.32	76.36	85.28	83.64	89.52
AVG	**81.38**	**77.52**	**85.24**	**82.67**	**90.04**

As can be seen in Table 3, the multi-modal fusion strategy shows significant advantages in the emotion recognition task. Overall, the average classification results of both feature layer fusion and decision layer fusion are higher than the single model pure EEG or pure eye movement results, which fully demonstrates the effectiveness of multi-modal data fusion in emotion recognition. Specifically, in single modal, the average classification result obtained by pure EEG is 81.38%, which is 3.86% higher than that of pure eye movement, which indicates that EEG data contains richer emotion-related information; however, it is worth noting that not all subjects' EEG classification results are higher than that of eye movement, e.g., the eye movement results of the P2, P9, and P14 subjects are slightly higher than that of the EEG results, which indicates that different modal data sources between individuals perceive the same task differently, which also reflects the necessity of multi-source data fusion. In multi-modality, the classification accuracy obtained by fusion at the feature level is 85.24%, which is 2.57% higher than that of fusion at the decision level, indicating that

the fusion of EEG and eye movement signals at the feature level can more fully utilize the complementary information between the two signals. EEG signals reflect the temporal and spatial dynamics of neural activity in the brain, while eye movement signals provide important information about visual attention and cognitive processing. Through feature layer fusion, these two signals can complement each other to capture changes in cognitive and perceptual processes more comprehensively. Our proposed cross-attention fusion method obtained the best results among different fusion strategies, reaching 90.04%, which is 4.80% higher than the feature layer fusion results and 7.37% higher than the decision layer fusion results, indicating that there is a dynamic interaction characteristic between EEG and eye movement features, and the correlation between EEG and eye movement data can be improved by capturing the correlation between EEG and eye movement data through the use of cross-attention to improve the emotion recognition accuracy.

To determine whether the differences in emotion recognition accuracy were statistically significant across the five levels of the "fusion strategy" factor (EEG-only, Eye-only, Feature Fusion FF, Decision Fusion DF, and cross-attention fusion), we conducted significance tests using One-way Analysis of Variance (ANOVA) [38] and post-hoc comparisons with Least Significant Difference (LSD) [39]. At a confidence level of 0.05, all data passed the variance chi-square test. The results, shown in Fig. 8, indicate a significant difference in accuracy at the 0.05 confidence level ($F = 2.329$, $p < 0.05$) across the five levels. Pairwise comparisons reveal that EEG-only results differ significantly from feature layer fusion results ($p < 0.05$) and are not significantly different from decision layer fusion results ($p = 0.061$); eye movement-only results differ significantly from both feature layer and decision layer fusion results ($p < 0.05$); and our proposed cross-attention fusion model shows significant differences compared to all other levels. All comparisons (EEG-only, eye movement-only, feature layer fusion, decision layer fusion) show significant differences ($p < 0.05$), demonstrating that our fusion strategy is effective, yielding significant improvements in emotion recognition accuracy over single-modal and traditional multi-modal fusion approaches.

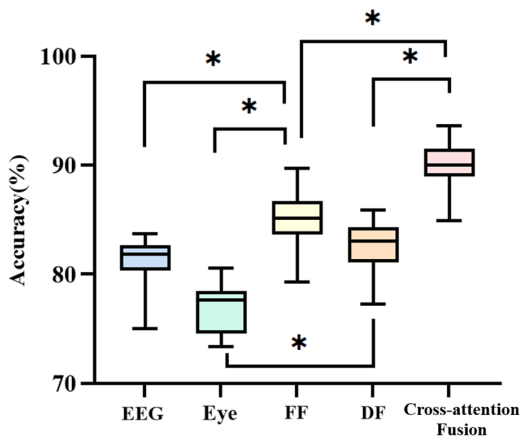

Figure 8: Statistical comparison of classification accuracy across the five fusion strategies (EEG-only, Eye-only, feature-layer fusion, decision-layer fusion, and the proposed cross-attention fusion), using one-way ANOVA with LSD post-hoc tests. * denotes $p < 0.05$.

Wearing a VR head-mounted display introduces additional sources of EEG contamination compared with conventional 2D setups, mainly: (i) mechanical pressure from the headset strap and facial interface, which is concentrated on frontal (Fp1, Fp2, F7, F8) and occipital (O1, O2) sites and may increase electrode impedance or cause low-frequency drift; (ii) motion artifacts induced by head movements during immersion; and (iii) ocular artifacts from blinks and saccades that couple into frontal channels. To mitigate these effects, our acquisition and preprocessing pipeline incorporated several quality-control steps: the 21-channel

montage followed the standard 10–20 system, participants were seated and instructed to minimize head motion during the 2-min clips, and signals were band-pass filtered to retain emotion-related frequency components while suppressing drift and high-frequency noise, followed by artifact removal targeting blink, muscle, and head-movement contamination. Notably, because the eye tracker is embedded inside the HMD, our setup is free of the camera-occlusion problem that affects external eye-tracking under headsets; moreover, the synchronized eye-movement signals provide ocular references that aid identification of EOG-related EEG artifacts. We did not collect a within-subject 2D control under identical stimuli, so a strictly controlled SNR comparison is not available in this study. As indirect evidence, our model attains comparable high accuracy on the 2D-acquired SEED-IV dataset (90.04%) and the VR-acquired VR-EED dataset (97.17%), suggesting that VR acquisition does not substantially degrade the discriminative information carried by EEG.

5.2 Limitations and Future Work

Although the proposed multi-modal cross-attention fusion approach demonstrates strong performance in emotion recognition within VR environments, several limitations remain. First, the size and demographic diversity of the self-built VR-EED dataset remain limited. This relatively narrow age range and limited demographic diversity may restrict the generalizability of the findings to broader populations, such as adolescents, older adults, or users with different levels of VR experience. Due to the limited number of participants and the exploratory nature of the current VR-EED dataset, LOSO evaluation was not used as the main validation protocol in this study. Second, although EEG and eye movement signals provide complementary information, the quality of the data can vary significantly across participants due to sensor noise, eye-tracking precision, and individual differences in physiological responses, potentially affecting robustness. Third, as shown in the computational complexity analysis in Section 4.1, the proposed model contains approximately 3.5–4.5 million trainable parameters and requires approximately 0.75–0.80 GFLOPs per 0.5-s EEG–eye-movement time slice. Although the latency is acceptable for offline analysis and GPU-based inference, further lightweight optimization, such as pruning, quantization, or knowledge distillation, is still needed for real-time deployment on CPU-only or resource-constrained standalone VR devices. Finally, interpretability remains an important issue for deep multi-modal emotion recognition models. In this study, the proposed cross-attention mechanism provides a feature-level interpretation by adaptively weighting the interaction between EEG and eye-movement representations. Since the EEG representations contain channel-, frequency-, and time-dependent information, the learned attention weights can help indicate how multi-modal neural and behavioral cues contribute to emotion classification. Nevertheless, future work should further combine attention analysis with visualization-based methods to obtain more fine-grained interpretations of EEG regions, frequency bands, and temporal segments, thereby improving the neurophysiological transparency of the proposed framework.

Future work will therefore focus on several directions. Expanding the datasets with more participants, more diverse cultural backgrounds, and longer-term data collection can improve model robustness and generalizability. Incorporating additional modalities, such as facial expressions, speech, or physiological signals like ECG and GSR, may provide richer information for emotion recognition. Moreover, exploring lightweight and efficient architectures will make real-time applications more feasible, especially in VR-based rehabilitation or adaptive learning systems.

5.3 Implication of Design

The findings of this study highlight key implications for the design of VR-based multi-modal emotion recognition systems. VR provides an immersive and ecologically valid environment for emotion induction, while the present experimental results mainly demonstrate the effectiveness of the proposed EEG-eye

movement cross-attention fusion model within VR settings. The proposed cross-attention fusion method therefore enables more accurate and reliable recognition of users' emotions in VR. This capability provides a foundation for developing emotion-aware VR applications across diverse domains. For instance, in education, emotion recognition can support personalized learning experiences that adapt to students' affective states. In healthcare, it can assist with early screening and intervention for autism spectrum disorders, or support rehabilitation programs targeting emotional trauma and stress recovery. Together, these insights underscore the potential of designing next-generation VR systems that are both immersive and emotionally adaptive.

6 Conclusion

In this study, we use VR to enhance human presence and evoke stronger emotions. While considering EEG signals, we incorporate eye movement data and introduce a multi-modal cross-attention fusion method for emotion recognition. The experimental results reveal that our proposed emotion recognition approach achieves an average accuracy of 90.04% on the public dataset SEED-IV and 97.17% on our self-collected dataset VR-EED, representing significant improvements over single-modal data. Our research offers insights and guidance for emotion recognition in VR and investigates the potential of multi-modal data to improve emotional detection. In the future, we plan to include additional emotion-related signals and examine their usefulness for enhancing emotion recognition. Ultimately, our study advances the development of emotion-centered applications such as personalized learning and mental therapy in VR, aiming to improve user experience.

Acknowledgement: The authors would like to thank all the volunteers who participated in the experiments.

Funding Statement: This research work was supported in part by the Zhejiang Provincial Natural Science Foundation of China under grant number LZ26F020007, and the National Natural Science Foundation of China under grant numbers 62172368 and 61772468.

Author Contributions: Study conception and design: Danyi Sheng, Shiwei Cheng, Yang Liu, Junjie Wu; data collection: Danyi Sheng; analysis and interpretation of results: Danyi Sheng, Junjie Wu; Writing—original draft: Yang Liu, Danyi Sheng; Writing—review & editing: Junjie Wu, Yang Liu, Shiwei Cheng. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data will be available on reasonable request.

Ethics Approval: This work was conducted in collaboration with Zhejiang Hospital, China. This study was approved by the Ethics Committee of Zhejiang Hospital (Approval No. 2025-C-095). All participants voluntarily participated in the study and provided written informed consent before data collection. The data collection procedures complied with institutional ethical, privacy, and safety guidelines. All collected data were anonymized before analysis.

Conflicts of Interest: Given his role as Editorial Board Member of this journal, Shiwei Cheng had no involvement in the peer review of this article and had no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to another journal editor. The authors declare no other conflicts of interest.

References

1. Quan XL, Zeng ZG, Jiang JH, Zhang YQ, Lv BL, Wu DR. Physiological signals based affective computing: a systematic review. Acta Autom Sin. 2021;47(8):1769–84.
2. Picard RW. Affective computing. Cambridge, MA, USA: MIT Press; 1997.

3. Magdin M, Prikler F. Are instructed emotional states suitable for classification? Demonstration of how they can significantly influence the classification result in an automated recognition system. Int J Interact Multimed Artif Intell. 2019;5(4):141–7. doi:10.9781/ijimai.2018.03.002.
4. Magdin M, Sulka T, Tomanová J, Vozár M. Voice analysis using PRAAT software and classification of user emotional state. Int J Interact Multimed Artif Intell. 2019;5(6):33–42. doi:10.9781/ijimai.2019.03.004.
5. Chen J, Hu B, Moore P, Zhang X, Ma X. Electroencephalogram-based emotion assessment system using ontology and data mining techniques. Appl Soft Comput. 2015;30:663–74. doi:10.1016/j.asoc.2015.01.007.
6. He H, Tan Y, Ying J, Zhang W. Strengthen EEG-based emotion recognition using firefly integrated optimization algorithm. Appl Soft Comput. 2020;94(3):106426. doi:10.1016/j.asoc.2020.106426.
7. Soleymani M, Pantic M, Pun T. Multi-modal emotion recognition in response to videos. IEEE Trans Affect Comput. 2011;3(2):211–23. doi:10.1109/t-affc.2011.37.
8. Wang X, Ren Y, Luo Z, He W, Hong J, Huang Y. Deep learning-based EEG emotion recognition: current trends and future perspectives. Front Psychol. 2023;14:1126994.
9. Liu H, Lou T, Zhang Y, Wu Y, Xiao Y, Jensen CS, et al. EEG-based multimodal emotion recognition: a machine learning perspective. IEEE Trans Instrum Meas. 2024;73:1–29.
10. Pillalamarri R, Shanmugam U. A review on EEG-based multimodal learning for emotion recognition. Artif Intell Rev. 2025;58(5):131. doi:10.1007/s10462-025-11126-9.
11. Somarathna R, Vuilleumier P, Mohammadi G. EmoStim: a database of emotional film clips with discrete and componential assessment. IEEE Trans Affect Comput. 2024;15(3):1202–12.
12. Larsen OFP, Tresselt WG, Lorenz EA, Holt T, Sandstrak G, Hansen TI, et al. A method for synchronized use of EEG and eye tracking in fully immersive VR. Front Hum Neurosci. 2024;18:1347974. doi:10.3389/fnhum.2024.1347974.
13. Puviani L, Rama S, Vitetta GM. Computational psychiatry and psychometrics based on non-conscious stimuli input and pupil response output. Front Psychiatry. 2016;7(10):190. doi:10.3389/fpsyt.2016.00190.
14. Salvucci DD, Goldberg JH. Identifying fixations and saccades in eye-tracking protocols. In: Proceedings of the Symposium on Eye Tracking Research & Applications; 2000 Nov 6–8; Palm Beach Gardens, FL, USA. p. 71–8.
15. Alhargan A, Cooke N, Binjammaz T. Multi-modal affect recognition in an interactive gaming environment using eye tracking and speech signals. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction; 2017 Nov 13–17; Glasgow, UK. p. 479–86.
16. Somarathna R, Bednarz T, Mohammadi G. Virtual reality for emotion elicitation: a review. IEEE Trans Affect Comput. 2023;14(4):2626–45.
17. Mohammadi G, Vuilleumier P. A multi-componential approach to emotion recognition and the effect of personality. IEEE Trans Affect Comput. 2020;13(3):1127–39. doi:10.1109/taffc.2020.3028109.
18. Yannakakis GN, Melhart D. Affective game computing: a survey. Proc IEEE. 2023;111(10):1423–44. doi:10.1109/jproc.2023.3315689.
19. AlzeerAlhouseini AM, Al-Shaikhli I, Rahman A, Dzulkifli MA. Emotion detection using physiological signals EEG and ECG. Int J Adv Comput Technol. 2016;8(3):103–12.
20. Katsigiannis S, Ramzan N. DREAMER: a database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices. IEEE J Biomed Health Inform. 2017;22(1):98–107.
21. Felnhofer A, Kothgassner OD, Schmidt M, Heinzle AK, Beutl L, Hlavacs H, et al. Is virtual reality emotionally arousing? Investigating five emotion inducing virtual park scenarios. Int J Hum Comput Stud. 2015;82(4):48–56. doi:10.1016/j.ijhcs.2015.05.004.
22. Marín-Morales J, Higuera-Trujillo JL, Greco A, Guixeres J, Llinares C, Scilingo EP, et al. Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors. Sci Rep. 2018;8(1):13657. doi:10.1038/s41598-018-32063-4.
23. Diemer J, Alpers GW, Peperkorn HM, Shiban Y, Mühlberger A. The impact of perception and presence on emotional reactions: a review of research in virtual reality. Front Psychol. 2015;6(345):26. doi:10.3389/fpsyg.2015.00026.

24. Peng X, Huang J, Denisova A, Chen H, Tian F, Wang H. A palette of deepened emotions: exploring emotional challenge in virtual reality games. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; 2020 Apr 25–30; Honolulu, HI, USA. p. 1–13.
25. Jin S. Research on emotion recognition based on deep learning and eye gaze signals [master's thesis]. Guangzhou, China: South China University of Technology; 2020.
26. Qiu JL, Li XY, Hu K. Correlated attention networks for multi-modal emotion recognition. In: Proceedings of the 2018 International Conference on Bioinformatics and Biomedicine (BIBM); 2018 Dec 3–6; Madrid, Spain. p. 2656–60.
27. Abdulwahhab AH, Myderrizi I, Asaad AY. TSConv-GAT: a hybrid temporal-spatial encoder-decoder with mRMR graph topologies using GATv2Conv network for optimizing EEG emotion recognition. Comput Electr Eng. 2026;135:111175.
28. Yang Z, Li D, Hou F, Song Y, Gao Q. Deep feature extraction and attention fusion for multimodal emotion recognition. IEEE Trans Circuits Syst II Express Briefs. 2024;71(3):1526–30. doi:10.1109/tcsii.2023.3318814.
29. Wang Y, Jiang WB, Li R, Lu BL. Emotion transformer fusion: complementary representation properties of EEG and eye movements on recognizing anger and surprise. In: Proceedings of the 2021 International Conference on Bioinformatics and Biomedicine (BIBM); 2021 Dec 9–12; Houston, TX, USA. p. 1575–8.
30. Fu B, Gu C, Fu M, Xia Y, Liu Y. A novel feature fusion network for multi-modal emotion recognition from EEG and eye movement signals. Front Neurosci. 2023;17:1234162. doi:10.3389/fnins.2023.1234162.
31. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17); 2017 Dec 4–9; Long Beach, CA, USA. p. 5998–6008.
32. Zheng WL, Liu W, Lu Y, Lu BL, Cichocki A. Emotionmeter: a multi-modal framework for recognizing human emotions. IEEE Trans Cybern. 2018;49(3):1110–22.
33. Cheng S, Wang J, Zhang L, Wei Q. Motion imagery-BCI based on EEG and eye movement data fusion. IEEE Trans Neural Syst Rehabil Eng. 2020;28(12):2783–93. doi:10.1109/tnsre.2020.3048422.
34. Bradley MM, Miccoli L, Escrig MA, Lang PJ. The pupil as a measure of emotional arousal and autonomic activation. Psychophysiology. 2008;45(4):602–7. doi:10.1111/j.1469-8986.2008.00654.x.
35. Liu W, Qiu JL, Zheng WL, Lu BL. Comparing recognition performance and robustness of multi-modal deep learning models for multi-modal emotion recognition. IEEE Trans Cogn Dev Syst. 2021;14(2):715–29. doi:10.1109/tcds.2021.3071170.
36. Goshvarpour A, Goshvarpour A. Novel high-dimensional phase space features for EEG emotion recognition. Signal, Image Video Process. 2023;17(2):417–25. doi:10.1007/s11760-022-02248-6.
37. Gong X, Dong Y, Zhang T. CoDF-Net: coordinated-representation decision fusion network for emotion recognition with EEG and eye movement signals. Int J Mach Learn Cybern. 2023;15(4):1213–26. doi:10.1007/s13042-023-01964-w.
38. Heiberger RM, Neuwirth E. One-way ANOVA. In: R through excel: a spreadsheet interface for statistics, data analysis, and graphics. New York, NY, USA: Springer; 2009. p. 165–91.
39. Williams LJ, Abdi H. Fisher's least significant difference (LSD) test. In: Encyclopedia of research design. Thousand Oaks, CA, USA: SAGE Publications; 2010. p. 840–53.