ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

Quantitative Assessment of Music Aesthetic Education Immersion in Virtual Reality (VR) Scenes: Based on Synchronization Analysis of Physiological Signals

Yuan Tan[1,*]

[1] Academy of Film and Television Arts, Hunan Mass Media Vocational and Technical College, Changsha 410100, China

SUMMARY: *In order to quantify the immersive experience in virtual reality music aesthetic education scene, an evaluation model based on physiological signal synchronization analysis was proposed. In this paper, EEG, ECG, EMG, respiration and head pose data of 36 participants were synchronously collected during 180 VR music interactions, forming 1440 samples under four immersion levels. The model takes music events as time anchors, constructs features by timing alignment, phase consistency, coupling strength and synchronization representation, and combines context reweighting, relationship aggregation, gated screening and dual-branch output to complete immersion level recognition and continuous scoring. The accuracy, macro-F1, mean absolute error and Pearson correlation were used as evaluation indicators in the experiment, and compared with single EEG classifier, early stitching model, convolution loop fusion model and time domain statistical synchronization model. The results show that the proposed model achieves 92.8% classification accuracy, 90.6% macro-F1 value, 0.214 mean absolute error and 0.873 correlation, which shows cross-subject stability and scene adaptability in the immersion evaluation of VR music aesthetic education.*

KEYWORDS: *Virtual reality; Physiological signal synchronization; Multi-modal feature fusion; Quantitative assessment of immersion*

1 Introduction

The development of virtual reality technology has promoted the digital reconstruction of immersive art experience, and music aesthetic education activities have gradually shifted from traditional listening space to an interactive, perceptible and quantifiable computing environment. The linkage of head-mounted display, spatial audio, motion capture and physiological acquisition devices enables learners 'attention allocation, emotional fluctuations and sensory input in the virtual scene to be continuously recorded, and provides a more fine-grained data basis for immersion state analysis. Music aesthetic education emphasizes the collaborative generation of perception, understanding and aesthetic response. It is difficult to describe the instantaneous changes in the experience process or reveal the dynamic coupling relationship between different modal responses by solely relying on subjective questionnaires. Advances in computer technology in signal processing, pattern recognition, and multimodal modeling have made it possible to shift immersion from empirical description to data-driven evaluation.

*tanyuan2025@126.com
https://doi.org/10.65102/is2026744
Published: 30 April, 2026

Focusing on the correlation between music-induced emotions and neural responses, Cui X et al. studied the framework of music emotion recognition and analysis based on EEG signals, indicating that EEG features have high sensitivity to the representation of music experience [1]. Kang T K proposed a short-time emotion recognition method with multiple physiological signals, and verified the joint value of indicators such as ECG and dermatogram in state discrimination [2]. Miyamoto K et al. studied the online EEG emotion prediction and music generation mechanism, showing that the real-time computing model can make a rapid response to the change of emotional state [3]. Wang X et al. summarized the development path of deep learning EEG emotion recognition, indicating that time series modeling and feature expression enhancement have become important technical directions in this field [4]. Li Q et al. proposed a multi-physiological signal emotion recognition method, which further showed that heterogeneous signal fusion was helpful to enhance classification stability [5].

Fu B et al. studied the EEG and eye movement fusion network and demonstrated the supporting effect of cross-modal feature alignment on the accuracy of emotion discrimination [6]. Wei Y et al. proposed the Transformer Capsule Network for EEG emotion recognition, which provides a new network implementation for complex spatio-temporal dependency modeling [7]. Tang J et al. studied a hierarchical multi-modal fusion method with scene adaptation and contrast alignment capabilities, indicating that synchronous modeling can improve recognition consistency in different contexts [8]. Pei G et al. proposed the EEG emotion computing framework in virtual reality environment, and established a balanced implementation path between recognition accuracy and computational efficiency [9]. Bastida L et al. studied the virtual reality emotion experiment scheme integrating physiological and facial analysis, and showed that multi-source observation in immersive scenes is more suitable for describing individual experience changes [10].

Existing research provides an important basis for physiologically driven emotion recognition and experience analysis in virtual environments. However, the synchronization modeling for quantitative assessment of immersion in VR music aesthetic education still needs to be further refined, especially to map the temporal consistency, coupling strength and immersion level between multimodal physiological responses into a unified computational representation space. Based on this, this paper constructs a physiological signal synchronization analysis model around the VR music aesthetic education scene, takes EEG, ECG, electrodermal and respiratory signals as joint input, and forms a quantitative assessment path of immersion through time alignment, rhythm coupling, phase congruency measure and feature fusion module.

The innovation of this paper is reflected in three aspects: first, a synchronous representation framework of multi-modal physiological signals is established for the process of music aesthetic education experience, so that the continuous perception response can be transformed into computable time sequence characteristics. The second is to construct a quantitative evaluation model of immersion, and complete the mapping between synchronization features and immersion levels through machine learning. Thirdly, the stability and generalization ability of the method in cross-individual scenarios are verified from the two levels of model effect and feature combination.

This paper is divided into five parts. The first part is the introduction. The second part is related work. The third part constructs a quantitative evaluation model of VR music aesthetic education immersion based on physiological signal synchronization analysis. Section IV analyzes the model experimental setup, the evaluation results, and the comparison results of different feature combinations. Section 5 is the conclusion.

2 Related work

Immersion recognition in virtual reality environment is gradually shifting from subjective evaluation to multi-modal computational analysis. The mapping relationship between physiological signals, interactive behaviors and scene stimuli has become an important research direction in recent years. Joo J H et al. studied immersive emotion analysis in VR environment and proposed an emotion distortion suppression method based on sensor collaboration, which integrates multi-source perception information to reduce state deviation and maintain high stability of recognition results under complex interaction conditions [11]. Arslan E et al. proposed a biological signal classification method for emotional intelligent virtual environment, which embed emotion recognition into the virtual interaction system, making physiological response the direct input of environmental understanding and state discrimination. Relevant results show that the biological signal model has good adaptability to the emotional stratification in the virtual scene [12]. Alharbi H combined the interpretable feature selection and deep learning recognition method of eye movement and physiological data, and incorporated visual attention allocation and physiological changes into the analysis framework, so that the feature selection process had a clear calculation basis, and provided a new implementation path for the determination of the key channel of immersion state [13].

Linares-Vargas B G P et al. conducted a systematic review on interactive virtual reality environment and emotion research, pointing out that immersive experience analysis is shifting from static emotion classification to continuous interaction modeling, and the joint use of scene feedback, perceived load and physiological coupling indicators has become the main trend in this field [14]. Fauveau V et al. studied the comprehensive physiological and psychological response assessment method in virtual reality experience, and proposed to integrate the changes of heart rate, electrodermal and psychological scales through a unified observation framework to depict the overall response structure in the virtual environment, indicating that virtual experience is not the result of single-channel perception, but a dynamic process of multi-dimensional responses [15]. Daşdemir Y proposed a high performance emotion estimation method based on salient channel pair selection, which highlights the most discriminative channel combination under virtual reality conditions to reduce the impact of redundant inputs on computational efficiency. This study shows that channel compression and discrimination preserving can be achieved simultaneously [16]. In order to more clearly show the technical path and application scenarios of the existing research, the related work is summarized in Table 1.

Table 1: Comparison of related studies on VR and physiological signals emotion computing

Researcher	Data Type	Main Method	Application Scenario	Main Performance
Joo J H et al. [11]	Multisensor physiological data	Sensor collaboration and distortion suppression	VR emotion analysis	Good stability under complex interactions
Arslan E E et al. [12]	Biosignals	Emotion classification in virtual environments	Intelligent virtual interaction	Good performance in emotion stratification
Alharbi H [13]	Eye-tracking and physiological data	Explainable feature selection and deep learning	VR recognition	Improved feature interpretability
Linares-Vargas B G P et al. [14]	Review data	Summary of interactive VR emotion research	VR emotion research	Clarifies the trend of continuous interaction modeling
Fauveau V et al. [15]	Heart rate, electrodermal activity, and psychological scales	Comprehensive response assessment	VR experience research	Supports overall response analysis
Daşdemir Y [16]	Physiological channel pairs	Salient channel selection	VR emotion estimation	Balances accuracy and efficiency
Wang D et al. [17]	EEG	Vision Transformer	Music-induced emotion	Strengthens long-sequence association
Qiao Y et al. [18]	EEG	Temporal convolutional attention network	Music emotion recognition	Improves rhythmic representation capability
Lee J H et al. [19]	EEG and audiovisual features	Contrastive learning	Multimodal emotion recognition	Enhances cross-modal consistency
Tschacher W et al. [20]	Physiological synchrony data	Audience synchrony analysis	Live music experience	Relates to experiential attitudes

In terms of music-induced emotion computing, Wang D et al. studied the music emotion classification based on EEG signals, and proposed the introduction of vision Transformer modeling method, so that the EEG response under music stimulation can complete the feature correlation in a long time range, and the relevant results show that the EEG pattern in music scene has a strong ability to distinguish emotions [17]. Qiao Y et al. proposed an EEG music emotion recognition method based on temporal convolutional attention network, which improved the model's accuracy in describing music-driven emotion changes by strengthening

the correspondence between local rhythm fluctuations and key frequency band responses [18]. Lee J H et al. studied the emotion recognition method based on the fusion of EEG and audio-visual features, and proposed a cross-modal modeling framework based on contrastive learning, which enabled the alignment of EEG responses and external audio-visual stimuli in a unified representation space. Relevant results show that multi-modal consistency learning can enhance the discrimination boundary of emotion recognition [19]. Tschacher W et al. studied the physiological synchronization phenomenon in classical music scene and analyzed the synchronization between audience and experience attitude, and the results showed that there was a clear correlation between the degree of physiological synchronization and audience experience evaluation [20].

The above research has formed a number of technical paths such as virtual reality, physiological signals, musical emotion recognition and cross-modal learning. Compared with the existing work, the research on quantitative assessment of immersion in VR music aesthetic education emphasizes the temporal consistency between physiological responses and the correspondence between coupling strength and immersion level. It is necessary to map heterogeneous signals such as EEG, ECG, dermatogram and respiration into a unified computational representation space, and complete interpretable immersion state discrimination through synchronous feature fusion. From the perspective of method evolution, related research has expanded from single EEG classification to multi-source signal collaborative modeling, from off-line recognition to scene correlation analysis, and from accuracy comparison to interpretable feature selection and synchronization relationship characterization. This evolutionary path illustrates that computational evaluation of immersive experiences is more suitable to be built on top of a multimodal synchronous analysis framework.

Therefore, the linkage modeling of music stimulation, virtual scene feedback and physiological synchronization indicators can better reflect the continuous immersion change and its internal structure in aesthetic education experience. This also makes the subsequent model design have a more explicit data foundation, and forms a more stable evaluation and calculation modeling process. Based on this idea, this paper integrates synchronous representation, feature fusion and immersion level discrimination into a unified technology chain to adapt to the continuous quantitative assessment in VR music aesthetic education scenarios.

3 Quantitative evaluation model of VR music aesthetic education immersion based on physiological signal synchronization analysis

3.1 Synchronous representation framework of multimodal physiological signals in VR music aesthetic education scene

In the virtual reality music aesthetic education scene, immersion is not an instantaneous strong response triggered by a single stimulus, but a continuous experience formed by spatial audio, visual flow, interactive action and body rhythm. If only the local amplitude of EEG or heart rate is retained, only fragmentary excitation changes can be obtained, which is difficult to describe the learners' overall collaborative trajectory between music entry, melody advancement, visual switching and action feedback. Therefore, this paper takes synchronization representation as the underlying modeling step, and maps EEG, ECG, skin conductance, respiration and head posture onto the same time reference axis. Then, the

immersion related features are extracted from three levels: event consistency, segmental association and segment structure stability. The goal of the framework is not to simply stack multiple sources of signals, but to transform the common responses of different physiological channels to the same musical event into a synchronized representation that can be calculated, compared, and fed into subsequent models.

To illustrate more clearly how multi-modal data can be sliced, aligned and output under the same music event, Fig. 1 shows the overall processing path of the synchronous representation framework in this section. The leftmost part of the figure is the original input layer, which receives the sequences of EEG, ECG, dermatogram, respiration and head posture output synchronously by the head display device and the physiological acquisition module. The first layer in the middle is the event slice layer, in which the segment is divided according to the music paragraph boundary, the shot change point and the interaction timestamp. The second layer is the synchronization calculation layer, in which the peak delay, phase congruency, rhythm coupling and segment stability indicators are extracted. On the right is the synchronous vector output layer, which generates a segment-level multimodal representation oriented to immersion evaluation.

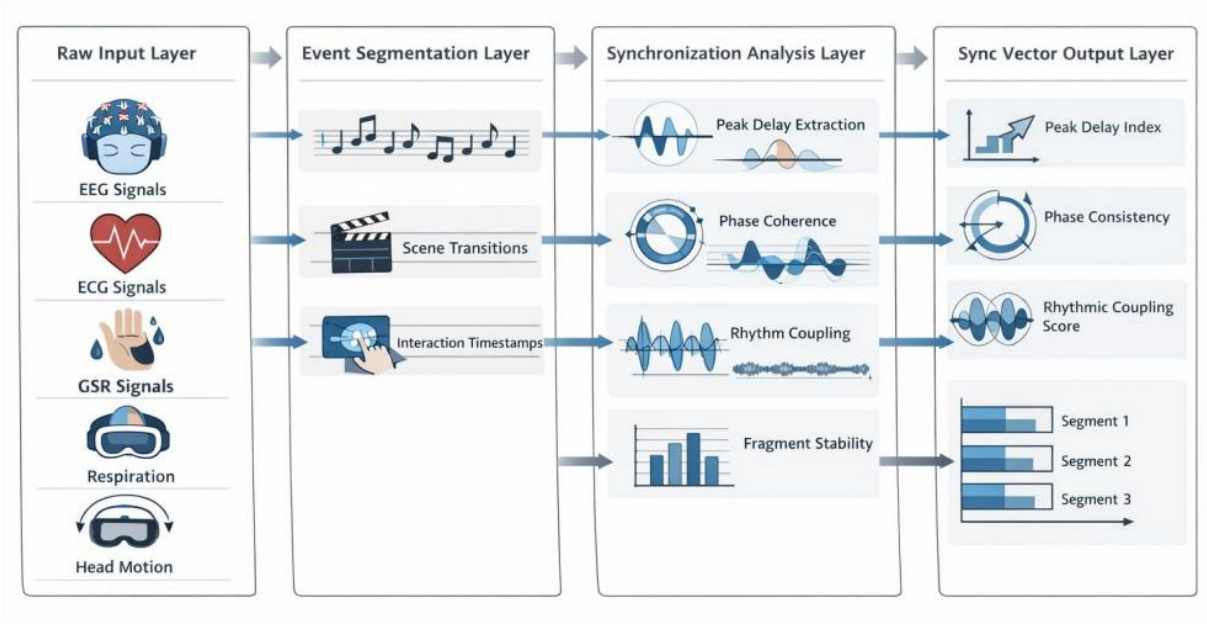

Figure 1: Total flow chart of synchronous representation of multimodal physiological signals in VR music aesthetic education scenario

In order to write the window partition and multi-modal time alignment process driven by music events into a unified and reproducible calculation form, this paper jointly expresses the event slicing function and the modal synchronization mapping as equation (1).

$$\Omega_r^{(m)} = \left\{ x^{(m)}(t) \mid t \in [\tau_r - \delta_l, \tau_r + \delta_r] \right\}, \quad \widetilde{\Omega}_r^{(m)} = \mathcal{R}_\kappa\left(\Omega_r^{(m)}\right) \tag{1}$$

Here, $\Omega_r^{(m)}$ represents the original window of the m mode signal under the r music event, τ_rr represents the time anchor of the r event, δ_l and δ_r represent the forward and backward retention length of the event respectively, $R_\kappa(\cdot)$ represents the mapping operator that resampling and retuning are performed according to the uniform sampling rate κ, $\Omega_r^{(m)}$ is the aligned mode window. The function of this formula is to bundle the time span of different modalities into the same segment boundary, so that the subsequent synchronization

calculation is based on the standardized event window instead of relying on the heterogeneous sampling Settings of the original device.

In order to reduce the scale deviation caused by individual baseline drift on multi-modal synchronous calculation, and make various physiological signals more prominent in the change trend rather than the absolute amplitude, this paper adopted the normalized representation after baseline correction, as shown in equation (2).

$$\hat{x}_r^{(m)}(t) = \frac{x_r^{(m)}(t) - \mu_b^{(m)}}{\sigma_b^{(m)} + \varepsilon} \tag{2}$$

Here, $\hat{x}_r^{(m)}(t)$ represents the value of the m mode signal in the r segment at time t, $\mu_b^{(m)}$ and $\sigma_b^{(m)}$ represent the mean and standard deviation of the mode on the resting baseline segment, ε is the minimal constant to prevent the denominator from being zero, and $x_r^{(m)}(t)$ is the normalized signal. This processing allows amplitude differences across subjects to be compressed to a stable range, allowing synchronization features to reflect event-driven changes more than individual physiological baseline differences.

In order to further describe the phase consistency of EEG, ECG, ECG and respiration to music beat changes in the same window, this paper introduces the phase synchronization index as the core component of the multi-modal synchronization vector, and its calculation form is shown in Formula (3).

$$\Psi_r^{(p,q)} = \left| \frac{1}{T_r} \sum_{t=1}^{T_r} \exp\left(j\left(\phi_r^{(p)}(t) - \phi_r^{(q)}(t) \right) \right) \right| \tag{3}$$

Here, $\Psi_r^{(p,q)}$ represents the phase synchronization index between mode p and mode q in the r segment, T_r represents the window length, $\phi_r^{(p)}(t)$ and $\phi_r^{(q)}(t)$ represent the instantaneous phases of the two types of modes at time t, respectively, and j is an imaginary unit. The closer the value of this equation is to 1, the more stable the phase synergy relationship between the two types of signals within the event window. For VR music aesthetic education scene, this quantity can better describe whether different physiological systems form a consistent response around the same rhythm event.

In order to associate the energy coupling relationship within the segment with the event context and avoid the single amplitude feature weakening the cross-channel correlation information, this paper further constructs the cross-modal coupling matrix, which is shown in Formula (4).

$$C_r^{(p,q)} = \frac{\sum_{t=1}^{T_r} \left(\hat{x}_r^{(p)}(t) - \bar{x}_r^{(p)} \right) \left(\hat{x}_r^{(q)}(t) - \bar{x}_r^{(q)} \right)}{\sqrt{\sum_{t=1}^{T_r} \left(\hat{x}_r^{(p)}(t) - \bar{x}_r^{(p)} \right)^2} \sqrt{\sum_{t=1}^{T_r} \left(\hat{x}_r^{(q)}(t) - \bar{x}_r^{(q)} \right)^2}} \tag{4}$$

Here,$C_r^{(p,q)}$ represents the coupling coefficient between mode p and mode q in the r segment, and $\bar{x}_r^{(p)}$ and $\bar{x}_r^{(q)}$ represent the average values of the two types of modes within the current window, respectively. This formula promotes cross-modal collaboration from "whether it changes at the same time" to "whether the direction of change and the amplitude of change have structural consistency". In the immersion representation, if the EEG arousal is

increased, the electrodermal response is enhanced, the respiratory rhythm is tightened, and the heart rate change also maintains a consistent trend, the multiple entries in the coupling matrix will show a more obvious concentrated structure.

To illustrate how the relationship between rhythm synchronization and scene segmentation is encoded in the representation layer, Fig. 2 presents the way in which the stream of musical stimuli is connected to the stream of physiological responses. The lower part of the figure shows the music stimulus flow, including rhythm strength change, melody advancement position, visual focus switch and interactive feedback nodes. Above is the physiological response flow, which contains EEG frequency band fluctuations, heart rate interval changes, skin electrical peak density and respiratory cycle changes. The intermediate alignment node is responsible for binding the external stimulus and the internal response in the same time segment, and generating the cross-modal synchronization relationship according to the rhythm coupling strength in the event window, which is used for subsequent synchronization feature calculation and segment-level representation output.

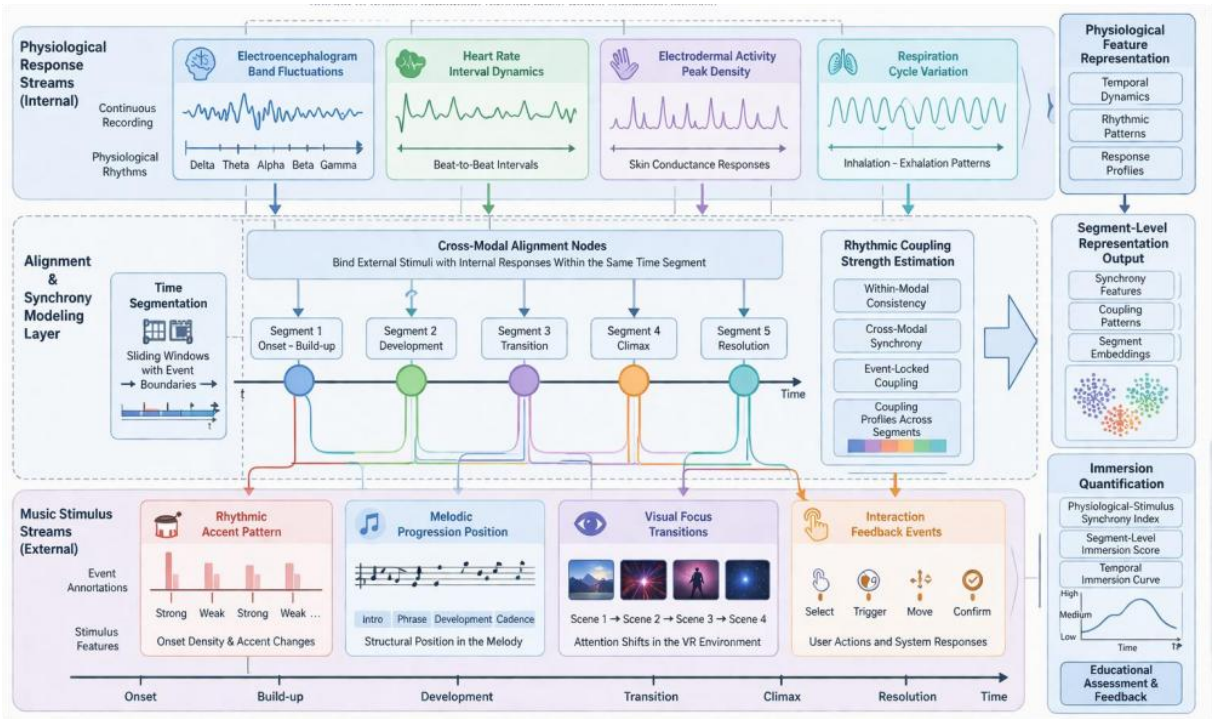

Figure 2: Graph of synchronization between musical stimulus segments and circadian rhythms

In order to write the temporal consistency within segments and the transition smoothness between segments into the unified feature at the same time, this section constructs the synchronization weight with event conditions, as shown in Formula (5).

$$\omega_r = \frac{\exp(\eta_1 \bar{\Psi}_r + \eta_2 \bar{C}_r + \eta_3 g_r)}{\sum_{u=1}^{R} \exp(\eta_1 \bar{\Psi}_u + \eta_2 \bar{C}_u + \eta_3 g_u)} \tag{5}$$

Here, ω_r represents the event weight of the r segment, $\bar{\Psi}_r$ represents the average phase synchronization value of all modal pairs within this segment, $\bar{C}_r$ represents the average coupling strength, g_r represents the scene stimulus intensity or interaction density index, and η_1 to η_3 are the weight parameters. This formula makes the synchronization representation

no longer just a parallel accumulation, but a differentiated weighting of different segments according to the event intensity and synchronization stability, so as to improve the sensitivity of the model to the key immersion segments.

In order to avoid the synchronization features only staying on the single level of segment statistics, this paper organizes the segment level, scene level and subject level information linkage into hierarchical representation, and the specific form is shown in Formula (6).

$$s_r = [u_r \,\|\, \omega_r \,\|\, \Delta u_r \,\|\, e_r], \quad S = \frac{1}{R}\sum_{r=1}^{R} \omega_r \, s_r \tag{6}$$

Here,u_r represents the synchronization subvector of the r segment, Δu_r represents the differential feature with neighboring segments, e_r represents the event label embedding, $\|$ represents the vector concatenation, and S represents the scene-level synchronization representation. This formula encodes local changes, event information and segment weights in a linkage, so that the synchronous representation not only retains short-term reactions, but also describes the stage advancement in the complete aesthetic education experience.

In order to show how the synchronization representation results gradually converge from the low-level segments to the high-level scene representation, Fig. 3 gives the organization of the hierarchical synchronization vectors. The bottom layer in the figure is the original segment feature unit, which contains the time-corrected EEG, ECG, ECG and respiratory segments. The middle layer is the synchronous sub-vector generation layer, and the joint encoding of phase congruency, rhythm coupling strength and segment difference information is completed in this layer. The top layer consists of segment-level, scene-level, and subject-level hierarchical aggregated outputs, which are used to form a unified synchronous representation for immersion assessment. This figure emphasizes the hierarchical progression from local event response to the overall immersion state representation.

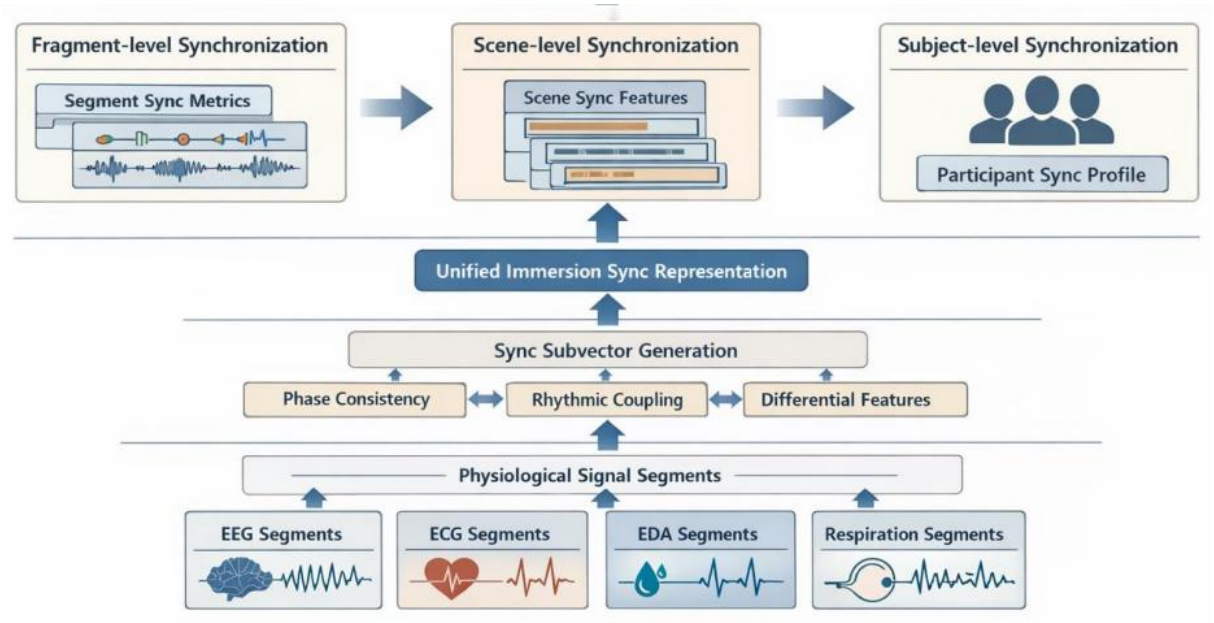

Figure 3: Hierarchical synchronization representation with scene-level sink structure diagram

In order to transform the final output synchronization vector into a unified encoding that can be directly fed into the evaluation model, this section performs another linear

renormalization and nonlinear compression of the scene-level synchronization representation, and the expression form is shown in Equation (7).

$$z = \tanh(PS + b) \tag{7}$$

Here, S is the scene-level synchronization representation, P represents the mapping matrix, b represents the bias vector, and z represents the synchronization embedding of the final output. The function of this formula is to compress the multi-source synchronization features into a stable dimension, which not only retains the original segment structure information, but also provides a unified input for the subsequent quantitative assessment of immersion.

Synthesizing the above process, the computational chain from the original multimodal physiological signal to the unified synchronous embedding has been constructed. The framework takes music event as the time anchor, phase congruency, coupling strength and segment difference as the core representation unit, and organizes heterogeneous data such as EEG, ECG, ECG and respiration into comparable, computable and inputable synchronous representations. The scene-level synchronous embedding thus formed not only preserves the detailed changes of local physiological responses, but also preserves the continuous advancement relationship in the complete experience, thus providing a stable data foundation and a unified feature entry for subsequent quantitative immersion evaluation.

3.2 Synchronous feature fusion module for immersion quantitative assessment

After the multi-modal synchronization representation is completed, the key to quantitative assessment of immersion is no longer to read a single synchronization index, but to integrate features from different sources, different scales, and different time sensitivities into a discriminative fusion representation. The synchronization feature itself already contains temporal consistency, rhythm coupling and segment transition information, but if it is directly fed into the classifier as a common vector, the modulation effect of music event strength, visual transition density and interactive action complexity on immersion state is often weakened. Therefore, this section constructs a synchronization feature fusion module for quantitative assessment of immersion. The synchronization embedding, event context and segment differential information are jointly input in the structure, and the fusion calculation is completed by attention allocation, gated screening, relationship aggregation and bi-objective constraints, so that the model can maintain stable discrimination under cross-subject conditions.

To illustrate more clearly how synchronous features accomplish reweighting, cross-aggregation, and branch output inside the fusion module, Fig. 4 shows the overall structure of the module in this section. The leftmost part of the figure is the input, which receives the scene-level synchronization embedding, segment difference vector and event label embedding output from the previous section. The first layer in the middle is a context-aware reweighting layer, in which the weight of synchronization features is adjusted according to the intensity of music clips, visual switching density and interactive feedback frequency. The second layer is the relationship aggregation layer, in which the dependency calculation and information propagation between modalities are completed. On the right is the two-branch output layer, one for immersion level classification and the other for continuous immersion score regression.

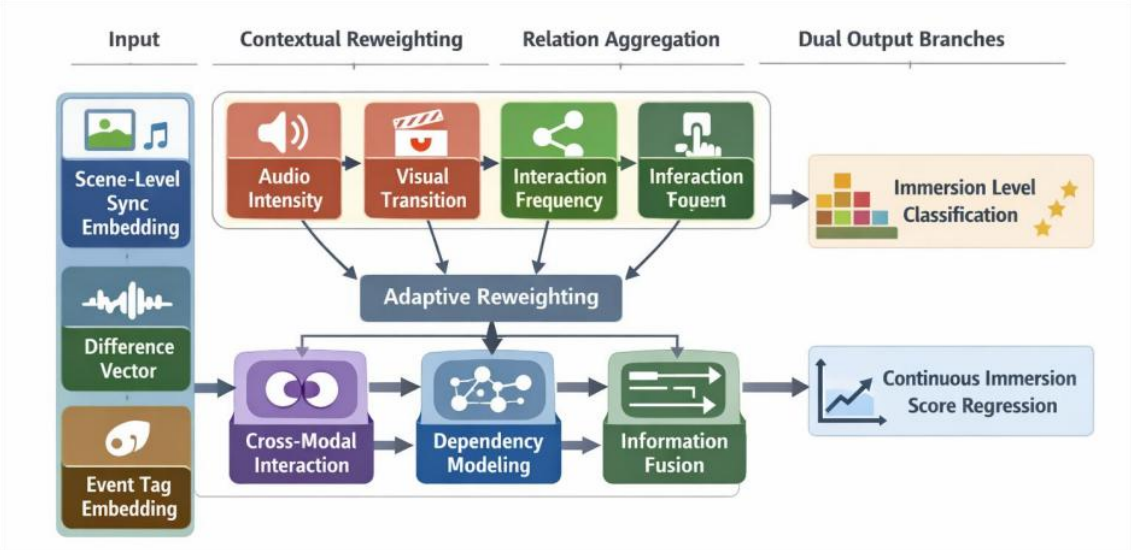

Figure 4: Overall structure diagram of synchronous feature fusion module

In order to make the synchronous features redistribute their contributions according to the event context after entering the fusion module, and highlight the components more relevant to the immersion change, this paper defines the context-aware attention weight as Equation (8).

$$\alpha_i = \frac{\exp(q^\top \tanh(Uz_i + Vc))}{\sum_{k=1}^{K} \exp(q^\top \tanh(Uz_k + Vc))} \tag{8}$$

Here, α_i represents the attention weight of the i synchronization feature subunit, z_i represents the i input subvector, c represents the context vector consisting of music intensity, visual switching and interaction frequency, U and V are mapping matrices, and q is the scoring vector. This formula is used to adaptively weight different synchronization features according to the context changes in the VR music scene, so that the features more related to the immersion state receive higher weights.

In order to explicitly propagate the dependencies between multimodal synchronous subvectors, instead of relying only on static concatenation to form a single input, this section writes the relation aggregation in schema propagation form, as shown in Equation (9).

$$g_i = \sigma\left(\sum_{j=1}^{K} \rho_{ij}\, Mz_j + d\right), \quad \rho_{ij} = \frac{\exp(z_i^\top z_j)}{\sum_{u=1}^{K} \exp(z_i^\top z_u)} \tag{9}$$

Here, g_i represents the i relation representation after aggregation, ρ_{ij} represents the correlation coefficient between input cell i and j, M is the transformation matrix, d is the bias term, and $\sigma(\cdot)$ is the nonlinear activation function. This formula enables the information exchange between different synchronization features according to their similarity. Therefore, the model no longer treats the EEG phase, heart rate coupling and electrocortical response as isolated quantities, but as the correlation structure that together constitutes the immersion experience.

In order to suppress redundant feature channels in the fusion process and retain the synchronization components that are more sensitive to immersion changes, this section further introduces the conditional gating mechanism, whose specific form is shown in Equation (10).

$$r = \sigma(Ag + Bc + h), \quad \tilde{g} = r \odot g \tag{10}$$

Here, g represents the overall feature after relation aggregation, c represents the context vector, A and B are mapping matrices, h is the bias term, r is the gating coefficient, $\odot$ represents element-wise multiplication, and g is the filtered fusion representation. The function of this formula is to let the model automatically suppress the components that are weakly related to the current segment immersion state, so that the output is more focused on the effective synchronization structure.

To illustrate how the classification branch and the regression branch operate collaboratively around the same fusion representation, Fig. 5 illustrates the two-branch output mechanism. In the figure, the left side is the gated fusion vector, and the upper middle branch is the level classification branch, which outputs discrete categories such as low immersion, medium immersion and high immersion. The middle and lower branch was the continuous scoring branch, and the normalized immersion score was output. The right side is the joint decision side, which performs consistency constraints and result integration on the two types of outputs.

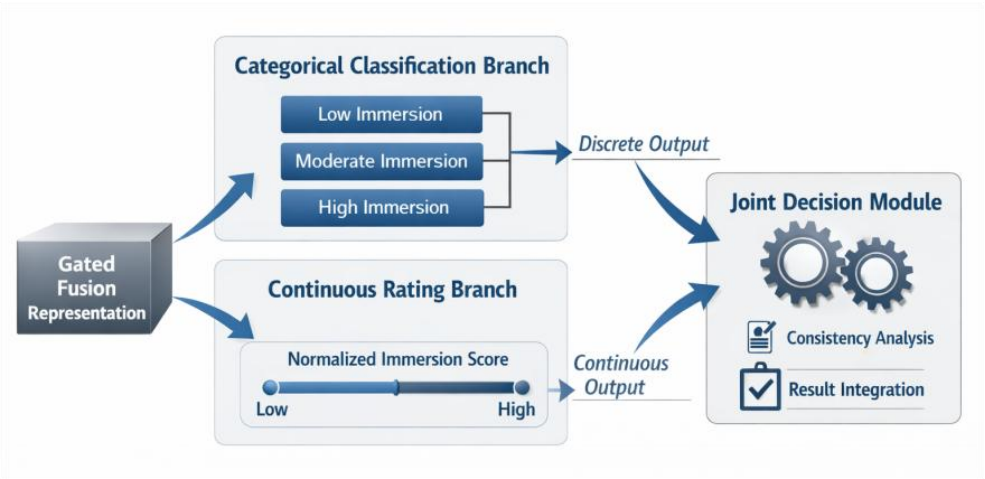

Figure 5: Double-branch structure diagram of immersion level classification and continuous rating

In order to map the gated fusion vector into rank probability and continuous immersion score, this section defines the classification output and regression output respectively, as shown in Equation (11).

$$y^c = \mathrm{softmax}(F\tilde{g} + b_c), \quad y^r = w_r^T \tilde{g} + b_r \tag{11}$$

Here, y^c represents the classification probability vector of the immersion level, y^r represents the continuous immersion score, F is the classification mapping matrix, b_c is the classification bias, w_r and b_r are the weights and biases of the regression branches, respectively. This formula makes the same fusion representation play a role in both discrete discrimination and continuous estimation directions, thereby improving the expression ability of the model for fine-grained changes in immersion state.

In order to enhance the consistency of the fusion representation between different subjects and make the segments of the same immersion level maintain a more stable geometric structure in the feature space, this section adds a contrastive consistency constraint above the output layer, as shown in Equation (12).

$$\mathcal{L}_{\mathrm{con}} = -\log \frac{\exp\left(\mathrm{sim}(\tilde{g}_a, \tilde{g}_p)/\tau\right)}{\exp\left(\mathrm{sim}(\tilde{g}_a, \tilde{g}_p)/\tau\right) + \sum_n \exp\left(\mathrm{sim}(\tilde{g}_a, \tilde{g}_n)/\tau\right)} \tag{12}$$

Here, $\tilde{g}_a$ represents the fusion representation of anchor samples, $\tilde{g}_p$ represents the positive sample representation of the same level, $\tilde{g}_n$ represents the negative sample representation of different levels, $\mathrm{sim}(\cdot,\cdot)$ is the similarity function, and τ is the temperature parameter. This formula pushes samples of the same immersion level closer in the feature space, and separates samples of different levels at the same time, so as to reduce the interference of individual differences on the discrimination boundary.

In order to keep the distribution of the fusion module stable under cross-subject conditions and avoid individual baseline differences from re-entering the high-level feature space, this section introduces a regularization constraint on the subject offset, as shown in Equation (13).

$$\mathcal{L}_{\mathrm{sub}} = \frac{1}{B}\sum_{b=1}^{B} \left\| \tilde{g}_b - \bar{g}_{(s_b)} \right\|_2^2 \tag{13}$$

where B represents the number of samples in a batch, $\tilde{g}_b$ represents the fusion representation of the b sample, $\bar{g}_{(s_b)}$ represents the mean representation of the subject to which the sample belongs in the current batch, and $\|\cdot\|_2$ represents the two-norm. This formula does not eliminate individual differences, but limits the high-level features to be overly dominated by irrelevant subjects, so that the fusion representation focuses more on the immersion state itself.

To fully show how the fusion representation is computed in a closed-loop from the input to the target constraint, Fig. 6 shows the training path of the synchronous feature fusion module. In the figure, the synchronization vector and context input are shown on the left, reweighting, relational aggregation, gated screening, dual-branch output and consistency constraint are shown in the middle, and the total objective function and parameter writeback are shown on the right. This path shows that quantitative assessment of immersion is not a single mapping, but a steady-state learning under the joint action of multiple constraints.

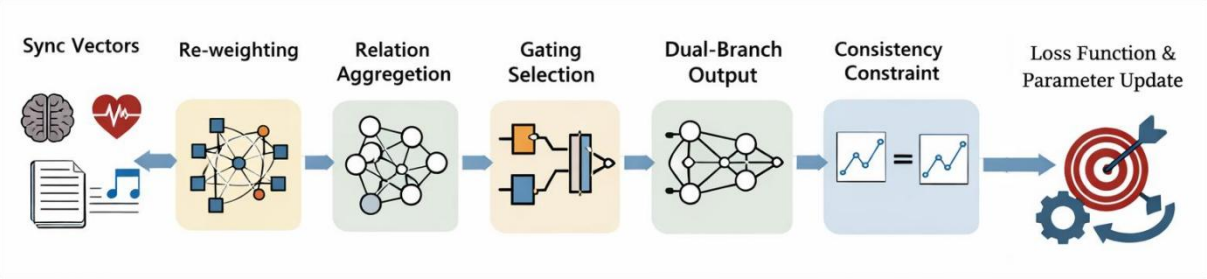

Figure 6: Joint training flowchart of the synchronous feature fusion module

In order to incorporate classification error, regression error, contrastive consistency and subject constraint into the same training objective, this paper defines the total loss of the synchronous feature fusion module as Eq. (14).

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\mathrm{cls}} + \lambda_2 \mathcal{L}_{\mathrm{reg}} + \lambda_3 \mathcal{L}_{\mathrm{con}} + \lambda_4 \mathcal{L}_{\mathrm{sub}} \tag{14}$$

Here, L_{cls} represents the classification cross-entropy loss, L_{reg} represents the regression mean square error loss, L_{con} represents the contrastive consistency loss, L_{sub} represents the subject constraint loss, and λ_1 to λ_4 are the various weight coefficients. This overall objective enables the fusion module to maintain the fine-grained expression of continuous ratings while retaining the ability of discrete rating discrimination, and improves the cross-subject robustness through additional constraints.

In summary, the synchronous feature fusion module is not a simple concatenation processing of the output vectors of the previous stage, but forms a high-level discriminant structure for immersive quantitative assessment under the combined effect of context awareness, relationship aggregation, gated screening and dual-branch constraints. This module is able to further highlight the key components closely related to the immersion level change and weaken the destabilizing effects caused by subject differences and scene disturbances on the basis of retaining the original information of synchronization features. After this layer of processing, the model has a clearer level discrimination ability and a more stable continuous scoring ability, and also establishes a complete calculation foundation for the effect verification and comparative analysis in the subsequent experiments.

4 Analysis of immersion evaluation results of VR music aesthetic education based on computational model

4.1 Experimental Setup

In order to test the quantitative assessment ability of the model in VR music aesthetic education scene, the self-built multimodal synchronization dataset was used as the data source. Data acquisition was completed in the immersion laboratory, and the experimental environment was composed of a head-mounted display terminal, a spatial audio playback system, a physiological acquisition device and an action recording module. The subjects were 36 college students with normal hearing and vision. Each subject completed 5 rounds of VR music interaction tasks, and 180 groups of scene sessions were formed in total.

EEG,ECG,ECG, respiration and head posture signals are recorded synchronously, and the nodes of music passage switching, visual focus change and interactive feedback are written into the event stream according to the timestamp mechanism. To ensure comparability between different channels, the sampling rate of EEG was set to 512Hz, the sampling rate of ECG, dermatogram and respiration was uniformly resampled to 256Hz, and the head pose sequences were aligned with frame-level time markers. After the original signal enters the computation flow, bandpass filtering, power frequency suppression, artifact removal and baseline correction are performed first, and then window slicing is performed with music events as anchors.

Each session was divided into eight valid analysis segments, resulting in 1440 labeled samples. The immersion level was jointly calibrated according to the subjective scale score, length of stay, gaze stability and physiological synchronization strength, and was divided into four categories: low immersion, low immersion, high immersion and high immersion. In order to reduce the influence of class distribution shift on the training results, the experiment used stratified sampling method to divide the training set, validation set and test set by 8∶1∶1, containing 1152, 144 and 144 samples respectively. Before model training, all synchronization features are standardized by Z-score, and the segment-level difference component is constructed by sliding window difference to enhance the ability to describe the continuous change of immersion state. The core experimental parameter Settings are shown in Table 2.

Table 2: Experimental parameter Settings

Module	Parameter Item	Value	Setting Description
Dataset	Number of subjects / sessions / samples	36 / 180 / 1440	Covers cross-subject and multi-round scene interaction
Signal Acquisition	Sampling rate of EEG / other physiological signals	512 Hz / 256 Hz	Balances temporal detail preservation and computational cost
Data Split	Training / validation / test set	8:1:1	Maintains consistent category distribution
Representation Layer	Synchronization representation dimension	128	Preserves cross-modal coupling information
Training Layer	Batch size / learning rate / optimizer	32 / 0.0005 / AdamW	Maintains training stability and convergence efficiency
Loss Layer	Classification weight / regression weight / contrastive coefficient / subject coefficient	1.0 / 0.6 / 0.2 / 0.1	Balances level discrimination and continuous scoring

The parameter Settings in Table 2 take into account both computational efficiency and evaluation stability. The dimension of synchronization representation is set to 128, which is able to retain cross-modal phase relationships and coupling strength information. The batch size of 32 and the learning rate of 0.0005 were used in the training phase, and AdamW was selected as the optimizer to ensure the convergence stability of the multi-branch fusion module. Four types of losses, classification, regression, contrast and subject constraint, jointly participate in the optimization, so that the model maintains a good balance between rank discrimination and continuous scoring.

In the training phase, five-fold cross-validation is performed within the training set to alleviate the over-fitting, and the macro-average F1 value and the mean absolute error of the validation set are used as the model selection criteria. The experimental platform uses Python 3.10, PyTorch 2.2 and CUDA 12.1, the running environment is Intel Core i7-12700K processor, NVIDIA RTX 3090 graphics card and 32GB memory, and the operating system is Windows 11. Finally, the generalization performance evaluation is completed on the independent test set.

4.2 Effect analysis of VR music aesthetic education immersion quantitative assessment model

After completing the experimental setup and model training, it is necessary to further investigate the quantitative assessment performance of the built model in VR music aesthetic education scene. This part mainly analyzes four aspects: level recognition results, continuous score fitting effects, cross-subject stability, and output under different scene complexities, so as to verify the suitability of synchronous representation and fusion module in continuous experience evaluation. The overall analysis focuses not only on the discriminative quality of discrete grade labels, but also on the degree of fit between continuous immersion scores and manual calibration, so as to more comprehensively reflect the computational performance of the model in practical applications.

The classification results under different immersion levels are shown in Table 3. The specific performance of the four levels in precision, recall and F1 value is listed in the table,

which is used to further analyze whether the model maintains a balanced output among the levels, especially to test the recognition stability in medium and high immersion states.

Table 3: Classification results for different immersion levels

Immersion Level	Precision / %	Recall / %	F1-score / %
Low Immersion	88.6	87.0	87.8
Relatively Low Immersion	90.5	89.4	89.9
Relatively High Immersion	91.8	91.2	91.5
High Immersion	93.7	92.7	93.2

As can be seen from Table 3, the F1 values of the four levels are 87.8%, 89.9%, 91.5% and 93.2%, respectively, showing a gradual improvement trend with the increase of immersion level as a whole. The high immersion category maintains the highest level in both metrics of precision and recall, indicating that the model is easier to capture stable immersion structures when the musical passage enters the climax interval, visual feedback is concentrated, and physiological synchronization is enhanced. Although the F1 value of the low immersion category is relatively low, it still maintains above 87%, indicating that the model does not neglect the recognition of the weaker state due to the strong response of the medium and high immersion segments. On the whole, the difference of each level index is not large, indicating that the model maintains a good balance among the four levels of states, and there is no significant bias to a certain category.

In order to show the correspondence between the predicted labels and the manually calibrated labels more intuitively, Fig. 7 shows the heatmap of the confusion matrix for the four-level immersion state.

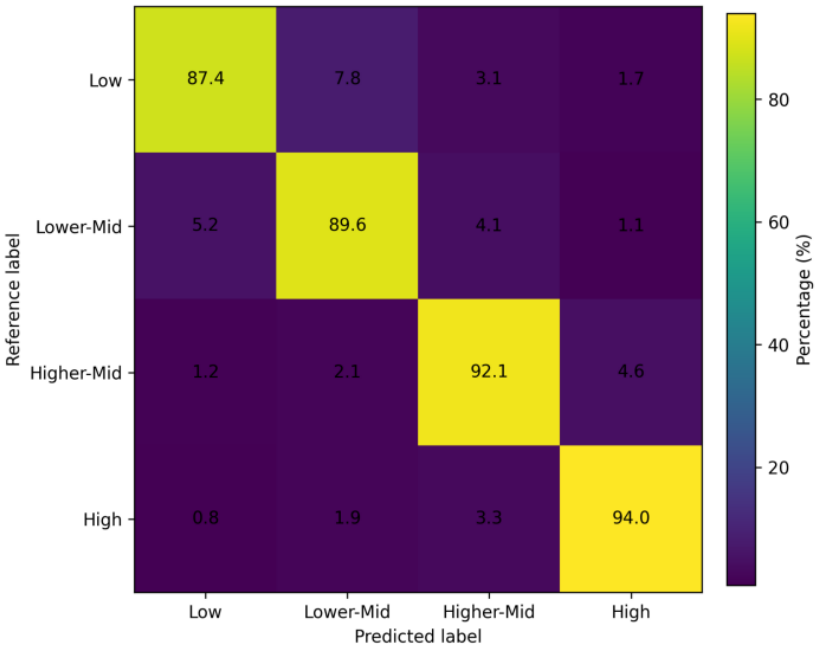

Figure 7: Heatmap of confusion matrix for recognition of four levels of immersion

As shown in Fig. 7, 87.4% of the low immersion samples were correctly identified, and the proportion of the main diagonal of the low immersion samples was 89.6%, and the proportion of the main diagonal of the high immersion and high immersion samples reached 92.1% and 94.0%, respectively. The misclassification was mainly concentrated between

adjacent levels, in which the proportion of deviation from low immersion to low immersion was 7.8%, and the proportion of deviation from high immersion to high immersion was 4.6%, and there was no large misclassification across more than two levels. Such distribution indicates that the model has been able to extract stable level boundary information from the EEG phase congruency, heart rate rhythm change, skin electricity peak density and respiratory synergy.

In addition to showing the evaluation performance under different complexity scenarios, Table 4 lists the error, correlation and inference delay indicators in three types of scenarios: low complexity, medium complexity and high complexity.

Table 4: Evaluation results for different scenario complexities

Scene Complexity	Mean Absolute Error	Correlation Coefficient	Single-Sample Inference Latency / ms
Low Complexity	0.198	0.891	18.7
Medium Complexity	0.214	0.873	21.4
High Complexity	0.236	0.854	24.9

As can be seen from Table 4, with the increase of scene complexity, the mean absolute error increases from 0.198 to 0.236, the correlation coefficient decreases from 0.891 to 0.854, and the single-sample inference delay increases from 18.7ms to 24.9ms. Although the error increases in high-complexity scenes, the overall change is small, indicating that the model can still maintain a relatively stable evaluation ability under the conditions of music passage switching, visual dynamic enhancement, and interactive events increasing. From the perspective of music structure location, the average recognition accuracy of the front entry region is 90.9%, the middle advance region is 92.6%, the climax region is 94.3%, and the end drop region is 93.1%. This change is basically consistent with the distribution trend of physiological synchronization intensity, indicating that the model can not only judge the overall immersion level, but also better reflect the dynamic ups and down of music aesthetic education experience in different structural segments.

To further examine the output consistency in the cross-subject condition, Fig. 8 presents a boxplot of the distribution of correlation coefficients across 36 subjects.

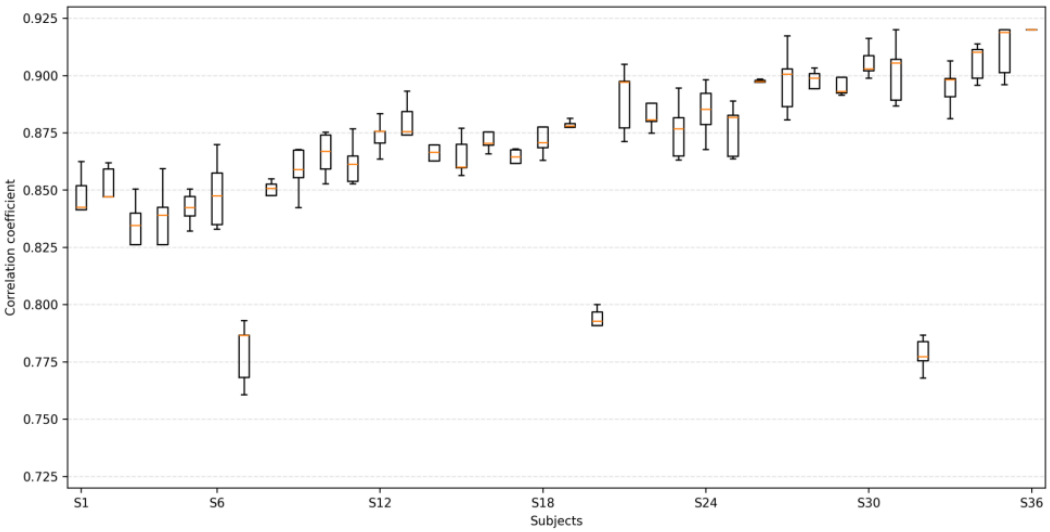

Figure 8: Boxplots of correlation coefficients for subject-level immersion scores

As shown in Fig. 8, the correlation coefficients of most subjects are distributed between 0.82 and 0.90, with a median of 0.871, and only three subjects are below 0.80. The corresponding sample review showed that these three subjects had more frequent head pose changes during the interaction, and the local action amplitude was larger, which led to some disturbance of the synchronization structure. Despite this, its overall output remains in the acceptable range without a large drift.

Taking these results together, it can be seen that the proposed model maintains good stability at three levels of rank recognition, continuous scoring and cross-subject generalization, and can provide a reliable baseline for the subsequent comparison of different models with different combinations of synchronous features.

4.3 Comparative analysis of different evaluation models and synchronization feature combinations

After verifying the overall effect of the model, it is necessary to further compare the differences between different evaluation models and different synchronous feature combinations to clarify whether the performance improvement comes from the model structure, synchronous modeling method or feature combination strategy. In this part, the single EEG classifier, the ECG early stitching model, the convolutional loop fusion model, the synchronization model using only time domain statistics, and the model in this paper are selected as control objects, and compared under uniform data division and training conditions. At the same time, around the combination of single-modal, multi-modal and complete synchronous features, the influence of different input configurations on immersion level recognition and continuous scoring is analyzed.

In order to observe the response differences of different combinations of synchronization features on the four levels of immersion, Fig. 9 illustrates the rank sensitivity heatmaps of the five types of feature combinations.

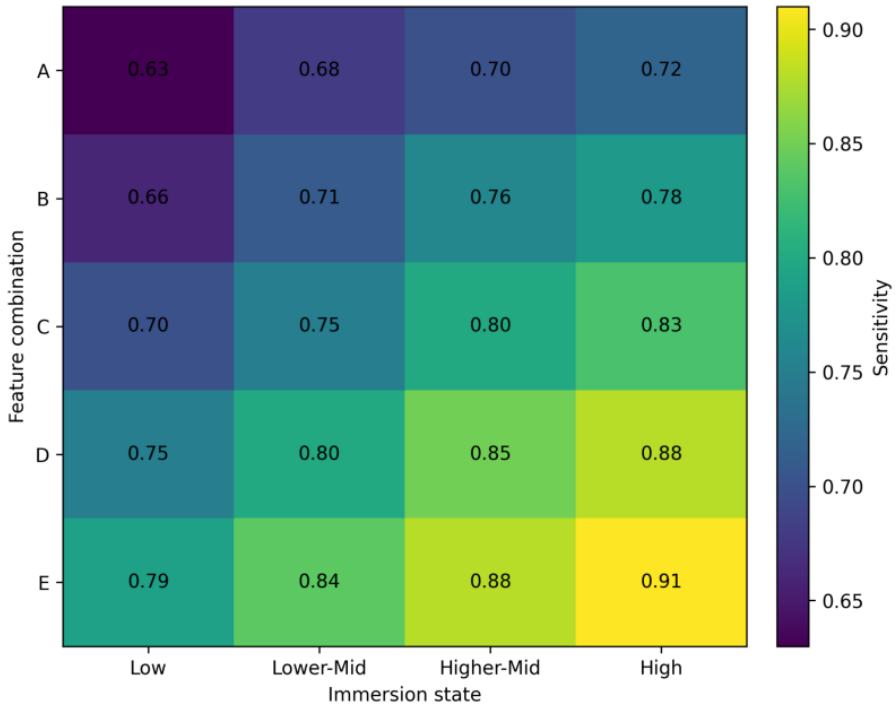

Figure 9: Rank sensitivity heatmaps for different combinations of synchronization features

As shown in Fig. 9, the response of combination A is weak in low immersion and low immersion regions, with average sensitivities of 0.63 and 0.68, respectively. After adding heart rate and respiration, the response of combination B increased to 0.71 and 0.76 in the lower and higher immersion regions. In combination C, the sensitivity of the high immersion region reaches 0.83 after adding the peak skin density. After the introduction of phase congruency and coupling strength in combination D, the response distribution of the four levels is significantly more balanced. The average sensitivity of combination E, the complete synchronization feature set, on four levels of immersion reaches 0.79, 0.84, 0.88 and 0.91, respectively. Such changes suggest that immersion assessment does not rely on a single channel, but more on a complete portrayal of cross-modal synergistic relationships. Especially in the higher immersion and high immersion interval, after adding the synchronization index, the level boundary becomes clearer, and the thermal distribution also changes from local highlighting to global enhancement.

The overall performance of the different evaluation models is shown in Table 5. In this table, a unified comparison of multiple models is performed from four dimensions of accuracy, macro average F1, mean absolute error and correlation coefficient to determine whether the proposed model maintains advantages in classification and continuous scoring tasks at the same time.

Table 5: Overall performance comparison of different evaluated models

Model	Accuracy / %	Macro-F1 / %	Mean Absolute Error	Correlation Coefficient
EEG-Only Classifier	84.9	82.6	0.318	0.732
Early Fusion Model of ECG and EDA	86.3	84.1	0.297	0.761
Convolutional-Recurrent Fusion Model	89.7	87.5	0.248	0.826
Temporal Statistical Synchronization Model	90.8	88.9	0.236	0.841
Proposed Model	92.8	90.6	0.214	0.873

Table 5 shows that the proposed model remains optimal in all four indicators. Compared with the single EEG classifier, the accuracy is increased by 7.9 percentage points, the macro-average F1 is increased by 8.0 percentage points, the mean absolute error is decreased by 0.104, and the correlation coefficient is increased by 0.141. Compared with the convolutional loop fusion model, the accuracy is increased by 3.1 percentage points, the mean absolute error is decreased by 0.034, and the correlation coefficient is increased by 0.047. Compared with the time domain statistical synchronization model, the accuracy, F1 and correlation coefficient of the proposed model are increased by 2.0 percentage points, 1.7 percentage points and 0.032 respectively, while the error is further compressed to 0.214. Such results show that the advantages of the proposed method are not only reflected in a single index, but also reflected in two types of tasks: rank classification and continuous estimation, indicating that the synchronization feature fusion module can more fully release the discriminative value of multimodal physiological synchronization information.

To further compare the error fluctuations of different feature combinations in the cross-subject condition, Fig. 10 shows the error distribution plots of the five class combinations over 36 subjects.

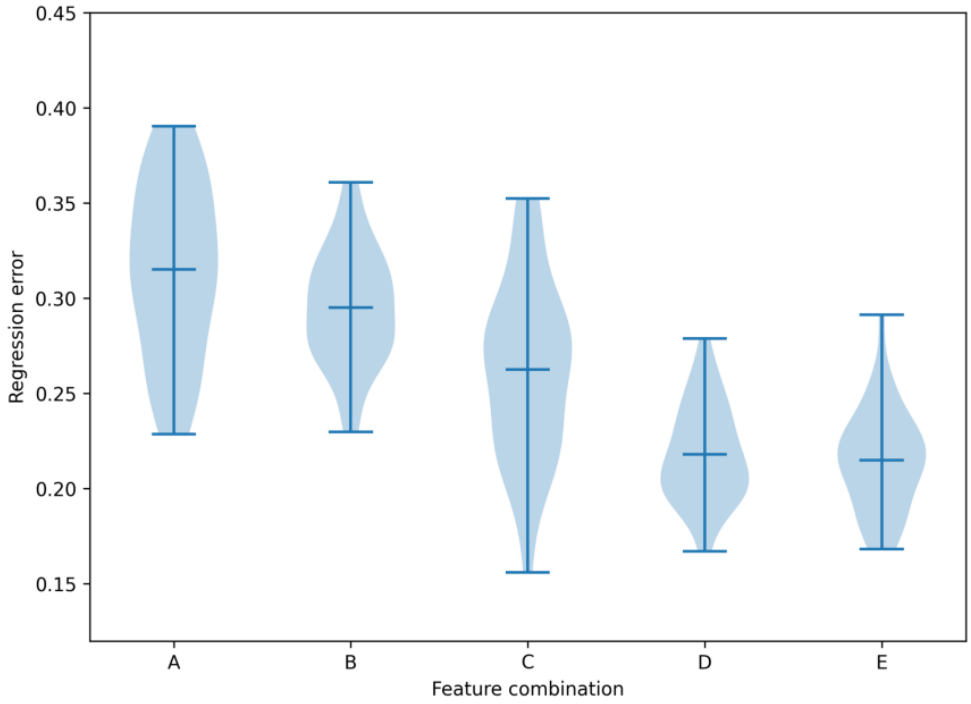

Figure 10: Subject error distribution plots for different combinations of synchronization features

As shown in Fig. 10, combination A and B have a large distribution span, with median errors of 0.318 and 0.287, respectively. The median error of combination C converges to 0.256. The combination D further decreased to 0.229. The median error of combination E, the complete synchronization feature set, is reduced to 0.214, and the number of long-tail samples is significantly reduced. This indicates that as the feature combination advances from single-modal to multi-modal synchronous structure, the adaptability of the model to different subjects is continuously enhanced. Especially in sessions with frequent interactive actions and visual focus switches, simple EEG or simple concatenation features are more susceptible to local disturbances, while complete synchronization features can maintain more stable regression output.

In order to analyze the influence of the internal composition of the model on the performance, Table 6 presents the ablation experiment results. The table removes synchronization representation, context-aware weight, relationship aggregation, and two-branch constraints respectively, and observe the change of indicators after the missing of each part to verify the actual contribution of each module in the complete calculation chain.

Table 6: Results of model ablation experiments

Setting	Accuracy / %	Macro-F1 / %	Mean Absolute Error	Correlation Coefficient
Without Synchronization Representation	88.7	86.2	0.281	0.798
Without Context-Aware Weighting	89.9	87.4	0.247	0.822
Without Relational Aggregation	90.4	88.1	0.239	0.831
Without Dual-Branch Constraint	91.1	88.8	0.231	0.846
Full Model	92.8	90.6	0.214	0.873

Table 6 shows that after removing the synchronous representation, the accuracy drops to 88.7%, the macro-average F1 drops to 86.2%, and the mean absolute error rises to 0.281, indicating that the synchronous modeling itself is the basis for the performance improvement. After removing the context-aware weights, the accuracy and F1 drop to 89.9% and 87.4%, respectively, indicating that scene condition guidance has a practical effect on highlighting key segments. After removing relation aggregation, the correlation coefficient drops to 0.831, indicating that inter-modal dependency propagation can enhance the consistency of successive ratings. After removing the double branch constraint, the classification and regression indexes decreased simultaneously, which indicates that there is a complementary relationship between discrete rank discrimination and continuous score estimation. The complete model remains optimal on all indicators, indicating that synchronous representation, context-aware, relation aggregation, and two-branch output together constitute a stable quantitative evaluation chain.

Combined with the above results, it can be seen that multimodal physiological synchronization analysis not only improves the overall accuracy, but also reduces the continuous scoring error, and makes the grade boundary clearer and the output fluctuation smaller. Especially, when the complete synchronization features are involved in the calculation, the model is more stable for the recognition of middle and high immersion states, and the error control of low immersion states is more balanced. Together, these results show that multimodal physiological synchronization analysis is the key support to improve the quantitative assessment performance of VR music aesthetic education immersion, and the constructed fusion computing chain can stably transform this synchronization advantage into measurable and comparable numerical output.

5 Conclusion

Focusing on the quantitative assessment task of immersion in virtual reality music aesthetic education scene, this paper constructs a computational model based on physiological signal synchronization analysis. The model takes front-end synchronous representation and back-end feature fusion as the main line, integrates multi-source information such as EEG, ECG, ECG, respiration and head posture into a unified time frame, and completes immersion level recognition and continuous score estimation through context awareness, relationship aggregation and dual-branch output. The experimental results show that the classification accuracy of the proposed model on the test set reaches 92.8%, the macro-average F1 value reaches 90.6%, the mean absolute error is 0.214, and the correlation coefficient is 0.873. Compared with single EEG classifier, early concatenation model, convolutional loop fusion model and temporal statistical synchronization model, the proposed method shows more stable output in level boundary delineation, cross-subject consistency and complex scene adaptability, indicating that multi-modal physiological synchronization information can effectively support the calculation of immersion state in VR music experience.

From the results distribution, the sensitivity of the complete synchronization feature set is the most balanced on the four levels of immersion, and maintains a clearer discrimination boundary in the high immersion interval. Synchronization representation, context-aware weight, relationship aggregation and two-branch constraint together constitute the main sources of model performance improvement. Without any of the links, the accuracy, macro average F1 value and correlation coefficient all decline to varying degrees. This suggests that immersion assessment is not a task that can be accomplished with a single signal or single statistic, but requires a continuous computational chain between event slicing, temporal alignment, cross-modal coupling, and high-level discrimination. The limitations of this paper

are mainly reflected in two aspects. On the one hand, the sample size is still focused on limited subjects and fixed experimental environment, and there is still room for further expansion of scene distribution and learner types. On the other hand, although the current model can maintain basically stability for fast visual transitions and high-frequency motion disturbances, there is still room for compression in local regression accuracy. Future work will continue to expand the samples of multi-age and multi-experience layers, introduce more complex music structures and interaction mechanisms, and combine lightweight deployment and online adaptive update strategies to improve the migration ability, real-time computing ability, long-term operation stability and interpretability of the model in real teaching scenarios.

References

[1] Cui X, Wu Y, Wu J, et al. A review: Music-emotion recognition and analysis based on EEG signals[J]. Frontiers in neuroinformatics, 2022, 16: 997282.

[2] Kang T K. Emotion Recognition using Short-Term Multi-Physiological Signals[J]. KSII Transactions on Internet & Information Systems, 2022, 16(3).

[3] Miyamoto K, Tanaka H, Nakamura S. Online EEG-based emotion prediction and music generation for inducing affective states[J]. IEICE TRANSACTIONS on Information and Systems, 2022, 105(5): 1050-1063.

[4] Wang X, Ren Y, Luo Z, et al. Deep learning-based EEG emotion recognition: Current trends and future perspectives[J]. Frontiers in psychology, 2023, 14: 1126994.

[5] Li Q, Liu Y, Yan F, et al. Emotion recognition based on multiple physiological signals[J]. Biomedical Signal Processing and Control, 2023, 85: 104989.

[6] Fu B, Gu C, Fu M, et al. A novel feature fusion network for multimodal emotion recognition from EEG and eye movement signals[J]. Frontiers in Neuroscience, 2023, 17: 1234162.

[7] Wei Y, Liu Y, Li C, et al. TC-Net: A Transformer Capsule Network for EEG-based emotion recognition[J]. Computers in biology and medicine, 2023, 152: 106463.

[8] Tang J, Ma Z, Gan K, et al. Hierarchical multimodal-fusion of physiological signals for emotion recognition with scenario adaption and contrastive alignment[J]. Information Fusion, 2024, 103: 102129.

[9] Pei G, Shang Q, Hua S, et al. EEG-based affective computing in virtual reality with a balancing of the computational efficiency and recognition accuracy[J]. Computers in Human Behavior, 2024, 152: 108085.

[10] Bastida L, Sillaurren S, Loizaga E, et al. Exploring human emotions: a virtual reality-based experimental approach integrating physiological and facial analysis[J]. Multimodal Technologies and Interaction, 2024, 8(6): 47.

[11] Joo J H, Han S H, Park I, et al. Immersive emotion analysis in VR environments: A sensor-based approach to prevent distortion[J]. Electronics, 2024, 13(8): 1494.

[12] Arslan E E, Akşahin M F, Yilmaz M, et al. Towards emotionally intelligent virtual environments: classifying emotions through a biosignal-based approach[J]. Applied Sciences, 2024, 14(19): 8769.

[13] Alharbi H. Explainable feature selection and deep learning based emotion recognition in virtual reality using eye tracker and physiological data[J]. Frontiers in Medicine, 2024, 11: 1438720.

[14] Linares-Vargas B G P, Cieza-Mostacero S E. Interactive virtual reality environments and emotions: a systematic review[J]. Virtual Reality, 2024, 29(1): 3.

[15] Fauveau V, Filimonov A K, Pyzik R, et al. Comprehensive assessment of physiological and psychological responses to virtual reality experiences[J]. Journal of Medical Extended Reality, 2024, 1(1): jmxr. 2024.0020.

[16] Daşdemir Y. Virtual reality-enabled high-performance emotion estimation with the most significant channel pairs[J]. Heliyon, 2024, 10(20).

[17] Wang D, Lian J, Cheng H, et al. Music-evoked emotions classification using vision transformer in EEG signals[J]. Frontiers in Psychology, 2024, 15: 1275142.

[18] Qiao Y, Mu J, Xie J, et al. Music emotion recognition based on temporal convolutional attention network using EEG[J]. Frontiers in human neuroscience, 2024, 18: 1324897.

[19] Lee J H, Kim J Y, Kim H G. Emotion recognition using EEG signals and audiovisual features with contrastive learning[J]. Bioengineering, 2024, 11(10): 997.

[20] Tschacher W, Greenwood S, Weining C, et al. Physiological audience synchrony in classical concerts linked with listeners' experiences and attitudes[J]. Scientific Reports, 2024, 14(1): 16412.