Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 20 No. 4 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijim.v20i04.60519

PAPER

Immersive Music Therapy Using Virtual Reality and Mobile Technologies

Sai Wang (✉)

Hebei University of Environmental Engineering, Qinhuangdao, China

18233545656@163.com

ABSTRACT

The global prevalence of mental disorders such as depression and anxiety continues to rise. Traditional music therapy faces challenges such as a lack of immersion, delayed emotional response, and rigid rehabilitation pathways, making it difficult to meet the demand for personalized treatment. The immersive characteristics of virtual reality (VR) and the portability of interactive mobile technologies provide technical support to overcome these limitations, with accurate emotion recognition being a key prerequisite for personalized intervention. This paper aims to construct an immersive music therapy environment based on the deep integration of VR and interactive mobile technologies and proposes a dual-level fusion method for multimodal emotion recognition, combining feature-level and model-level integration to dynamically optimize the psychological rehabilitation path. Methodologically, we first design a "hardware collaboration-software adaptation-interactive feedback loop" architecture that integrates physiological signal acquisition, VR scene rendering, and mobile interaction control modules. Multimodal data, including EEG, ECG, facial expression images, and subjective emotional ratings, are collected. These are then aligned and weighted across modalities at the feature level to extract high-level features, and deep learning models with attention mechanisms at the model level are used for precise emotion classification. Finally, based on real-time emotion recognition results, an optimization algorithm driven by reinforcement learning is developed to dynamically adjust music parameters and VR scene elements. The study confirms that the integration of VR and interactive mobile technologies can break through the limitations of traditional therapy scenarios. The dual-level fusion strategy provides higher accuracy and robustness for emotion recognition, while the dynamic optimization mechanism offers personalized solutions for psychological rehabilitation, with significant academic innovation and clinical application potential.

KEYWORDS

virtual reality (VR), interactive mobile technologies, immersive music therapy, multimodal emotion recognition, feature-level fusion, model-level fusion, psychological rehabilitation path optimization

Wang, S. (2026). Immersive Music Therapy Using Virtual Reality and Mobile Technologies. *International Journal of Interactive Mobile Technologies (iJIM)*, 20(4), pp. 120–134. https://doi.org/10.3991/ijim.v20i04.60519

Article submitted 2025-10-22. Revision uploaded 2025-12-12. Final acceptance 2022-12-14.

1 INTRODUCTION

The global mental health crisis continues to worsen. According to a report from the World Health Organization (WHO), approximately 280 million people worldwide suffer from depression, and 260 million people are affected by anxiety disorders [1, 2]. These mental health disorders have become one of the leading causes of disability-adjusted life years lost globally, resulting in social and economic losses amounting to trillions of dollars each year [3, 4]. Against the backdrop of increasing life pressures, the imbalance between the supply and demand for traditional psychological treatment resources has become more apparent. Developing efficient, convenient, and personalized treatment technologies has become an urgent need in the global public health field [5, 6]. Music therapy, as a core non-pharmacological intervention, is widely applied in psychological rehabilitation. However, traditional models face significant bottlenecks: treatment scenarios are limited to fixed clinics, making it difficult to create a deeply immersive emotional resonance environment [7]; emotional assessments rely on patient self-reports and therapists' experiences, which are delayed and highly subjective [8]; rehabilitation pathways are based on general guidelines, lacking individual dynamic adaptability, which severely restricts their clinical effectiveness and promotion [9].

The immersive characteristics of virtual reality (VR) can create highly realistic treatment scenarios and reduce patients' psychological defenses [10]; interactive mobile technologies break spatial limitations, enabling portability and real-time treatment [11]; multimodal emotion recognition integrates objective data to provide core support for personalized treatment. The fusion of these three elements offers a new path to overcome these bottlenecks [12]. However, existing research has clear deficiencies: the integration of VR and mobile technologies is mostly functional addition, lacking the "immersion-interaction-data" closed-loop architecture [13]; feature-level fusion in multimodal emotion recognition faces the challenge of data heterogeneity, while model-level fusion is highly complex and lacks adaptation to patients with psychological disorders [14]; rehabilitation path optimization relies on historical data, lacking real-time emotional feedback, and core intervention variable regulation is insufficiently refined [15]. Currently, the field has yet to form an integrated solution of "VR-interactive mobile fusion environment + high-precision multimodal emotion recognition + dynamic rehabilitation path optimization," which provides the core entry point for this study.

This study sets three main objectives: ① to construct a VR-interactive mobile fusion music therapy environment with both high immersion and portability; ② to propose a feature-level and model-level dual-fusion multimodal emotion recognition method; ③ to design a rehabilitation path dynamic optimization mechanism based on real-time emotional feedback. The innovations are reflected in three aspects: ① Architectural innovation, achieving the organic unity of immersion experience and mobile control through device protocol adaptation and data synchronization; ② Methodological innovation, using cross-modal feature alignment and attention-weighted fusion strategies to improve emotion recognition accuracy and robustness; ③ Mechanism innovation, constructing a closed-loop optimization model driven by reinforcement learning to dynamically adjust treatment parameters and achieve personalized rehabilitation.

The structure of the paper is as follows: Chapter 2 describes the architecture design and core module implementation of the immersive treatment environment; Chapter 3 details the dual-level fusion method for multimodal emotion recognition; Chapter 4 builds a reinforcement learning-driven rehabilitation path optimization mechanism; Chapter 5 verifies the effectiveness of the solution through clinical experiments; Chapter 6 discusses the research value, limitations, and future directions; Chapter 7 summarizes the core findings and prospects for application.

2 CONSTRUCTION OF IMMERSIVE MUSIC THERAPY ENVIRONMENT

This study is based on the hierarchical architecture model of multimodal interaction and rehabilitation path optimization in the immersive music therapy environment shown in Figure 1. A six-level collaborative architecture is designed: Technology Layer – Perception Interaction Layer – Data Processing Layer – Algorithm Decision Layer – Application Layer – Feedback Optimization Layer. This architecture achieves the organic integration of immersive experience and real-time interaction through precise hardware and software adaptation. The technology layer adopts a 5G + Wi-Fi 6 dual-mode transmission protocol, and the hardware integrates the Oculus Quest 2 VR headset, Android/iOS mobile terminals, and multimodal physiological data collection devices. The software relies on Unity3D, Flutter, and microservice cloud platforms for cross-terminal collaboration. The perception interaction layer synchronously collects physiological signals such as electroencephalography (EEG), electrocardiography (ECG), skin conductance, and facial expression behavior characteristics. It also enables bi-directional interaction of scene interaction and parameter adjustment through VR headsets, motion controllers, and mobile terminals. The data processing layer is deployed on edge cloud nodes and provides high-quality data for subsequent analysis through filtering, ICA denoising, and normalization algorithms. The algorithm decision layer adopts a feature-level and model-level dual-fusion multimodal emotion recognition method, combined with reinforcement learning algorithms, to dynamically output optimized music parameters, VR scenes, and interaction tasks. The application layer provides treatment plan recommendations, multimodal data visualization, and HL7 standard data interface functions. The feedback optimization layer closes the feedback loop of algorithm decision results to the interaction and application layers while continuously optimizing the intervention strategy based on the patient's subjective Self-Assessment Manikin (SAM) scores. This six-level architecture forms a complete closed loop of Collection – Transmission – Processing – Decision – Application – Feedback, providing systemic support for the technical implementation and clinical application of immersive music therapy.

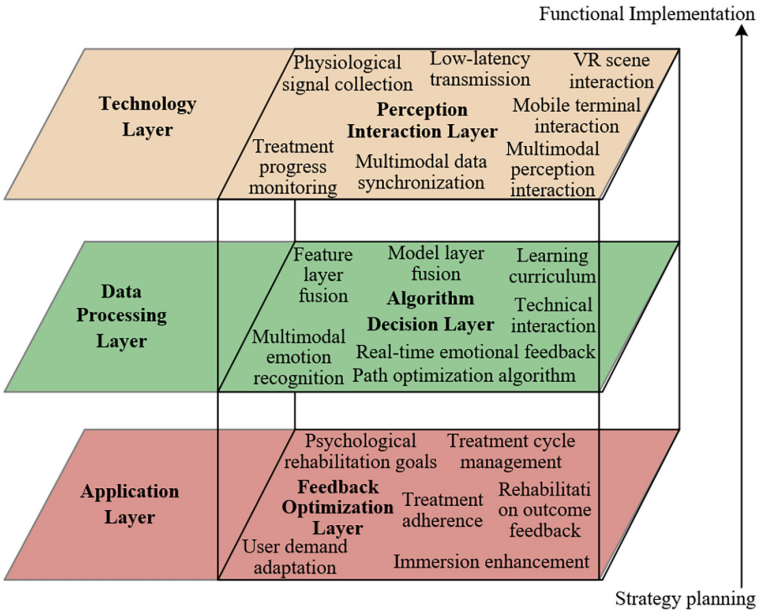


Fig. 1. Hierarchical architecture model of multimodal interaction and rehabilitation path optimization in immersive music therapy environment

Core modules collaborate to ensure the immersion and reliability of treatment. VR scenes are designed with clinical requirements for both relaxing and stimulating systems. Environmental psychology is referenced to optimize scene parameters, and detail-level technologies are used to improve rendering efficiency. A field-of-view adaptive adjustment is applied to control the incidence of dizziness below 5%. The music library integrates international therapeutic music resources and establishes a database of parameters such as rhythm, pitch, and volume. Scene-music mapping algorithms are used to ensure smooth transitions when switching between parameters. The core functions of the mobile app include intelligent treatment plan recommendations, real-time display of emotional and physiological indicators, and slider-based parameter adjustment, compatible with HL7 standards for data integration with clinical systems. Physiological signal preprocessing targets noise removal: a 50Hz notch filter is used to eliminate power line interference, a 4–30 Hz low-pass filter extracts effective EEG signals, independent component analysis (ICA) separates eye movement and muscle artifacts, and the data is finally normalized using Z-score normalization, mapping multimodal data to the same feature space to provide high-quality input for subsequent emotion recognition.

3 MULTIMODAL EMOTION RECOGNITION METHOD

3.1 Emotion modal data collection plan

The emotion modal data collection in this study strictly follows clinical experimental protocols and data standardization principles. The experimental subjects are patients diagnosed with depression/anxiety disorder according to the diagnostic criteria of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM-5*). Inclusion criteria: age 18–55 years, primary school education or above, and ability to cooperate with the collection process. Exclusion criteria: comorbidity of severe physical diseases, history of epilepsy, allergy to VR devices, or cognitive dysfunction. The data collection process is divided into three stages: the baseline testing stage, where the participant's basic physiological signals and facial expression data are collected for five minutes in a quiet environment, along with completing a basic information questionnaire; the emotion induction stage, where different valence images from the International Affective Picture System are selected and combined with a self-constructed standardized emotional music library, presented through an immersive VR scene to induce the target emotion; and the data recording stage, where multimodal data is continuously collected until the presentation of the inducing materials is complete. The collected data includes three types: physiological data collected using specialized equipment, EEG recorded through a 16-channel amplifier at key electrode points in the international 10–20 system (Fp1, Fp2, etc.), ECG to record heart rate, RR interval, and other indicators, and skin conductance response to record peak and rise slope; behavioral data captured by the VR headset's built-in high-definition camera to capture facial expression images, with 68 facial feature points extracted based on the Dlib library; subjective data collected through a self-assessment model, where the participant rates each emotional induction material on a 1–9 scale for pleasure and arousal levels, providing complementary verification of objective data and subjective experiences.

3.2 Feature-level fusion strategy

The goal of single-modal feature extraction is to accurately capture the emotional representations of each modality. Differentiated extraction schemes are designed for

different data types. In terms of physiological features, EEG signals extract time-domain and frequency-domain features based on preprocessed data. Time-domain features, such as mean, standard deviation, and kurtosis, reflect the signal amplitude distribution characteristics. Frequency-domain features, calculated through fast Fourier transform, include power spectral density of α waves (8–13Hz), β waves (14–30Hz), and θ waves (4–7Hz), representing the brain's electrical activity rhythms under different emotional states. ECG features focus on heart rate and heart rate variability indicators, extracting core parameters such as the mean RR interval, standard deviation, and root mean square of adjacent RR interval differences. GSR features extract signal peak values, rising slopes, and mean amplitudes, reflecting skin conductance responses triggered by emotional arousal levels. In terms of behavioral features, a lightweight CNN architecture is used to extract features from preprocessed facial expression images. Depth wise separable convolutions are employed to reduce the model parameters while retaining key facial expression features, ultimately outputting a 256-dimensional facial depth feature vector. Subjective features directly quantify the pleasure and arousal ratings from the self-assessment model into a 2-dimensional feature vector, achieving numerical representation of subjective emotional experience.

The focus of feature-level fusion is to address the data heterogeneity problem of multimodal data and construct high-level feature representations for cross-modal collaboration. The first step is to perform feature alignment: for time-series physiological data like EEG and ECG, frame-level synchronization is performed based on timestamps; all single-modal features are Z-score normalized, mapping feature values to the [0,1] range to eliminate dimensional differences. Based on this, a dynamic weighted fusion strategy is designed. The contribution of each single-modal feature to emotion recognition is calculated through a validation set experiment, and higher weights are assigned to modalities with higher contributions. The specific fusion formula is defined as follows:

$$V = \omega_1 E_{EEG} + \omega_2 E_{ECG} + \omega_3 E_{GSR} + \omega_4 E_{Face} + \omega_5 E_{SAM} \tag{1}$$

where, V is the fused feature vector, ω_i is the weight of the i-th single-modal feature, and F_{EEG}, F_{ECG}, ... correspond to each single-modal feature vector. This strategy not only avoids information redundancy caused by simple feature concatenation but also strengthens the representation ability of effective information through dynamic weight allocation, providing high-recognition multimodal base features for subsequent model-level fusion.

3.3 Model-level fusion strategy

To address the limitations of feature-level fusion in mining complex intermodal associations and eliminating redundant information, this study constructs a multi-model ensemble architecture based on a cross-modal attention mechanism, achieving deep collaborative learning and precise representation of multimodal emotional information. Based on the characteristics of different modal data, a heterogeneous base model system is designed: physiological features with strong temporal dependence are input to a bidirectional long short-term memory network (BiLSTM) to capture the temporal evolution patterns of emotional states; deep facial expression features are input to a lightweight convolutional neural network (CNN) model to further extract local key emotional representations; and subjective features, along with intermediate features from the above modalities, are integrated

and input to a multi-layer perceptron (MLP) to adapt to the nonlinear mapping requirements of high-dimensional heterogeneous data. The core innovation lies in the introduction of a cross-modal attention mechanism. By constructing an attention weight matrix, the importance coefficient of each base model's output result is dynamically learned. For example, in anxiety emotion recognition, higher weight is assigned to the EEG feature model, which contributes more, while also uncovering the potential intermodal associations, achieving adaptive fusion of multi-model outputs and generating a more compact and information-rich shared emotional feature vector.

To ensure the generalization ability and classification performance of the model, a multi-dimensional regularization collaborative optimization mechanism is established to effectively suppress overfitting. During model training, L_2 regularization is applied to penalize the weight parameters of each base model, reducing model complexity by adding a weight squared term to the loss function. The weight squared term expression is given by:

$$L_{total} = L_{CE} + \lambda \sum \|W\|_2^2 \tag{2}$$

where L_{CE} is the cross-entropy loss, λ is the regularization coefficient, and W is the model weight matrix. Dropout layers are embedded in both CNN and MLP layers, with a deactivation probability of 0.2 to randomly mask some neurons and avoid over-reliance on local features. Additionally, early stopping is employed to monitor the validation set loss, and training is terminated when the loss does not decrease for 10 consecutive training epochs, retaining the model parameters with the best generalization performance. The optimizer used for model optimization is Adaptive Moment Estimation (Adam), with an initial learning rate of 0.001, and a cosine annealing strategy is used to dynamically adjust the learning rate, improving training stability and convergence speed. Finally, the fused shared features are input into a Softmax classifier, which outputs the probability distribution of four emotional states: joy, calmness, anxiety, and depression, completing the emotion recognition task.

$$V = \sum_i \omega_i F_i$$

$$Loss = L_{CE} + \lambda \|W\|^2$$

4 PSYCHOLOGICAL REHABILITATION PATH OPTIMIZATION

The precise construction of the rehabilitation path is based on the dissection of core elements and scientific initial configuration, laying the foundation for subsequent dynamic optimization. This study identifies five core elements of the rehabilitation path, forming a multidimensional collaborative intervention system: the treatment goal distinguishes between short-term and long-term levels, with the short-term focusing on immediate relief of anxiety/depression emotions and the long-term aiming for comprehensive restoration of psychosocial functions; the treatment cycle is set to 8–12 weeks according to clinical guidelines, with two treatments per week, each lasting 45 minutes; the music parameter combinations cover three core dimensions: rhythm, pitch, and volume; VR scene types

are adapted to the treatment stages, initially using a relaxing natural scene and switching to an encouraging dynamic scene based on emotional improvement; interactive tasks include lightweight tasks such as scene exploration and virtual instrument playing to enhance patient engagement. The initial path generation relies on the core recommendations from the *China Depression Disorder Prevention and Treatment Guidelines* (2023 Edition) and the *Anxiety Disorder Diagnosis and Treatment Guidelines* (2021 Edition), combined with the consensus experience of three senior experts in psychotherapy. The initial path is divided into three levels—mild, moderate, and severe—based on the patient's initial Self-Rating Depression Scale (SDS)/Self-Rating Anxiety Scale (SAS) scores. A differentiated initial path library is constructed; for example, in the case of severe patients, the initial path focuses on low-stimulation relaxing music and forest VR scenes, gradually increasing the intervention intensity.

The reinforcement learning-driven optimization model provides intelligent technical support for the dynamic adjustment of the rehabilitation path. Its core lies in achieving adaptive iteration of the path through environmental interaction and reward feedback. This study deeply binds the core components of reinforcement learning with the treatment scene: the agent is defined as the immersive music therapy system, responsible for perceiving the patient's state and executing path adjustment actions; the environment includes the VR immersive scene and the patient's real-time state, forming a dynamic interaction carrier; the state space is represented by a high-dimensional vector, including key physiological indicators such as multimodal emotion recognition results, treatment progress ratio, real-time SDS/SAS scores, and heart rate variability, forming a 32-dimensional state vector; the action space focuses on fine-tuning the three main intervention variables, with music parameters supporting rhythm adjustments of ±5 BPM, pitch adjustments of 1 scale, and volume adjustments of ±5 dB, while VR scenes provide four scene switching options, and interactive tasks include updates to three difficulty levels; the reward function design adopts a deviation quantification mechanism, defined as:

$$R = 1 - \left| S_{real} - S_{tar} \right| / S_{\max} \tag{3}$$

where S_{real} is the real-time emotional scale score of the patient, S_{tar} is the target score for the corresponding treatment stage, and S_{max} is the maximum score. When S_{real} approaches S_{tar}, the reward value approaches 1; otherwise, it decreases. The optimization algorithm uses Deep Q-Network (DQN) to build a "convolutional layer-fully connected layer" neural network architecture, with the convolutional layer extracting high-dimensional features from the state space. The fully connected layer fits the action value function, and the experience replay mechanism stores historical interaction data and randomly samples for training. The target network periodically updates the policy, ensuring the model converges to the optimal treatment path.

The dynamic execution mechanism ensures the effectiveness and individual adaptation of the optimized path through real-time feedback loops and personalized adjustments. This study designs a four-stage closed-loop execution process of "Emotion Recognition – Path Evaluation – Parameter Adjustment – Effect Monitoring," setting every two treatment cycles as an adjustment window. First, based on the multimodal emotion recognition results and SDS/SAS retest data, the effectiveness of the current path is evaluated. Then, the reinforcement learning model generates

the optimal adjustment plan, simultaneously updating the music parameters, VR scenes, and interactive tasks. The adjusted path is executed in subsequent treatments, while real-time monitoring of the patient's physiological indicators and emotional state provides data support for the next round of adjustments. To account for individual patient differences, an individual difference factor γ is introduced to dynamically optimize the weight distribution of the reward function. For elderly patients or those with a long course of illness, γ is set to a lower value to slow down the adjustment rate of intervention intensity, while for younger or more responsive patients, the γ value is increased to accelerate the path optimization. This mechanism achieves the precise transformation of the rehabilitation path from "generalized" to "personalized," ensuring that the intervention strategy highly matches the individual characteristics of the patient and the treatment dynamics. Figure 2 presents the multimodal data interaction and rehabilitation path optimization process flowchart for the immersive music therapy environment, visually demonstrating the complete closed-loop process of "Patient Treatment Interaction – Multimodal Data Collection – Emotion Recognition and Path Optimization – Treatment Plan Adjustment Feedback."

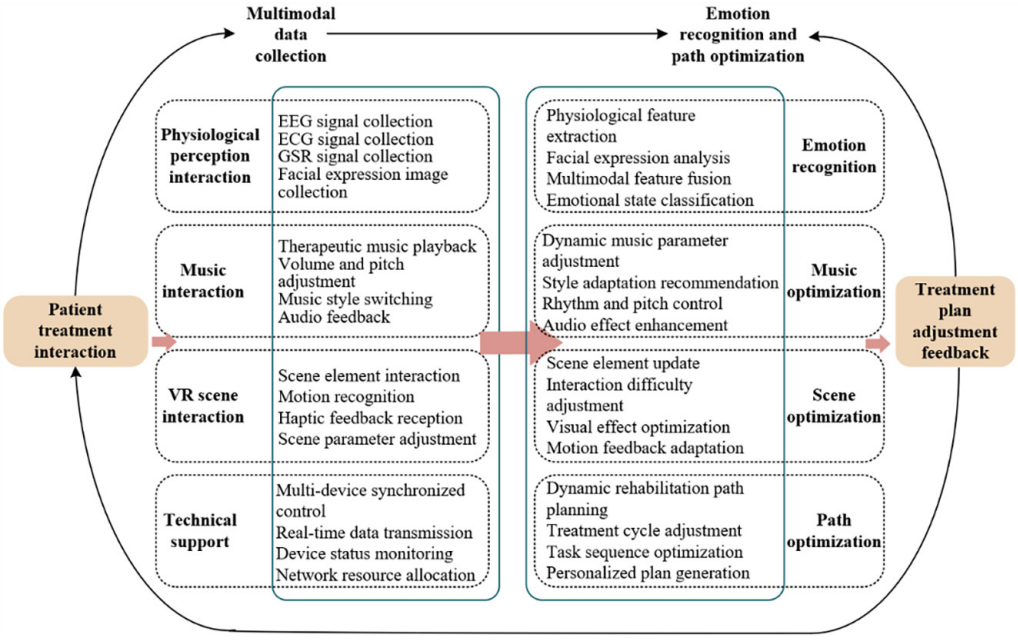


Fig. 2. Multimodal data interaction and rehabilitation path optimization process flowchart for immersive music therapy environment

5 EXPERIMENT DESIGN AND RESULT ANALYSIS

5.1 Experiment design

This experiment strictly follows the protocol of a randomized controlled clinical trial. A total of 120 patients diagnosed with depression/anxiety were selected as experimental subjects, with the sample size determined using GPower3.1 software. The inclusion criteria are as follows: meeting the diagnostic criteria of the *DSM-5*, SDS score ≥ 53 or SAS score ≥ 50, age between 18–60 years, ability to operate

VR devices, and voluntary signing of the informed consent form. Exclusion criteria include the presence of severe physical illnesses, a history of epilepsy, cognitive dysfunction, or having received other psychological interventions in the past three months. Patients were randomly assigned to the experimental group and the control group, each with 60 individuals, using a random number table. Baseline data comparison showed no significant differences between the two groups in terms of age, gender ratio, disease duration, and initial SDS/SAS scores ($P > 0.05$), indicating comparability between the groups. In terms of equipment, the experimental group used the Oculus Quest 3 VR headset, a 16-channel EEG amplifier, a portable ECG sensor, and a self-developed therapeutic interactive app, with data processing supported by Alibaba Cloud's edge computing nodes. The control group used a professional audio system and a tablet with a standard music library installed. The experimental treatment cycle lasted for 10 weeks, with both groups undergoing treatment three times a week, each session lasting 45 minutes. Data collection occurred before treatment, every two weeks during treatment, and after treatment in stages. Equipment calibration was completed before treatment, real-time operational status was recorded during treatment, and subjective rating data was collected post-treatment.

5.2 Experiment results analysis

Table 1. Immersive treatment environment performance test results

Evaluation Dimension	Specific Indicator	Unit	Experimental Group Test Value	Industry Reference Standard	Compliance Status
Immersion	IPQ Scale – Spatial Presence	Points (1–5)	4.2 ± 0.3	≥3.5 Points	Compliant
	IPQ Scale – Sense of Reality Loss	Points (1–5)	4.0 ± 0.4	≥3.0 Points	Compliant
	IPQ Scale – Involvement	Points (1–5)	4.3 ± 0.3	≥3.5 Points	Compliant
	IPQ Total Score	Points (1–5)	4.2 ± 0.3	≥3.5 Points	Compliant
Real-Time Performance	EEG Signal Collection Delay	*ms*	12.3 ± 1.5	≤20 *ms*	Compliant
	Multimodal Data Transmission Delay	*ms*	28.5 ± 2.1	≤50 *ms*	Compliant
	VR Scene Switch Delay	*ms*	35.2 ± 3.0	≤50 *ms*	Compliant
	Music Parameter Adjustment Delay	*ms*	18.7 ± 1.8	≤30 *ms*	Compliant
Stability	Continuous Operation Time	*h*	72	≥48 *h*	Compliant
	Device Failure Rate	%	0.8	≤2%	Compliant
	Data Loss Rate	%	0.3	≤1%	Compliant
User Adaptability	Dizziness Rate	%	4.2	≤10%	Compliant

To verify whether the constructed VR-interactive mobile integrated treatment environment meets clinical needs in terms of immersion, real-time performance, and stability, a performance test experiment was conducted. As shown in Table 1, the experimental group exceeded industry reference standards in all core performance indicators:

- Immersion: The IPQ scale scores in the three dimensions of spatial presence, sense of reality loss, and involvement, as well as the total score, all exceeded the qualified threshold of 3.5 points. Specifically, spatial presence and involvement scored above 4.0 points, indicating that the VR scene effectively creates an immersive therapeutic atmosphere.
- Real-Time Performance: Key processes, including EEG signal collection and data transmission delays, were controlled within 50ms. Music parameter adjustments and scene switching had delays of 18.7ms and 35.2ms, respectively, meeting the technical requirements for real-time emotional feedback and dynamic path adjustments.
- Stability: The device was continuously operated for 72 hours, with a failure rate of only 0.8% and a data loss rate of 0.3%. The dizziness rate was as low as 4.2%, demonstrating the clinical applicability of the environment.

In conclusion, the core performance of the constructed environment meets the standards, providing reliable technical support for the subsequent implementation of personalized treatments.

Table 2. Performance comparison of different emotion recognition methods

Recognition Method		Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	AUC Value
Single Modality Methods	EEG Recognition Alone	72.3 ± 1.5	71.8 ± 1.7	70.5 ± 1.6	71.1 ± 1.5	0.782
	Facial Expression Recognition Alone	75.6 ± 1.3	74.9 ± 1.4	73.2 ± 1.5	74.0 ± 1.4	0.805
Single Fusion Methods	Feature Layer Fusion Only	83.5 ± 1.2	82.9 ± 1.3	81.7 ± 1.2	82.3 ± 1.2	0.868
	Model Layer Fusion Only	84.2 ± 1.1	83.5 ± 1.2	82.4 ± 1.1	82.9 ± 1.1	0.875
	91.6 ± 0.8	91.6 ± 0.8	90.8 ± 0.9	90.2 ± 0.8	90.5 ± 0.8	0.943
Ablation Experiments	Removing Feature Layer Fusion	80.3 ± 1.4	79.6 ± 1.5	78.5 ± 1.4	79.0 ± 1.4	0.841
	Removing Model Layer Fusion	79.5 ± 1.3	78.9 ± 1.4	77.8 ± 1.3	78.3 ± 1.3	0.836

To verify the recognition performance of the proposed feature-layer-model-layer two-level fusion method and the necessity of the two-level fusion strategy, comparative and ablation experiments were conducted. As shown in Table 2, the proposed method significantly outperformed the single-modality and single-fusion methods in all core indicators: the accuracy reached 91.6%, which is an improvement of 16.0 percentage points compared to the best single-modality method and

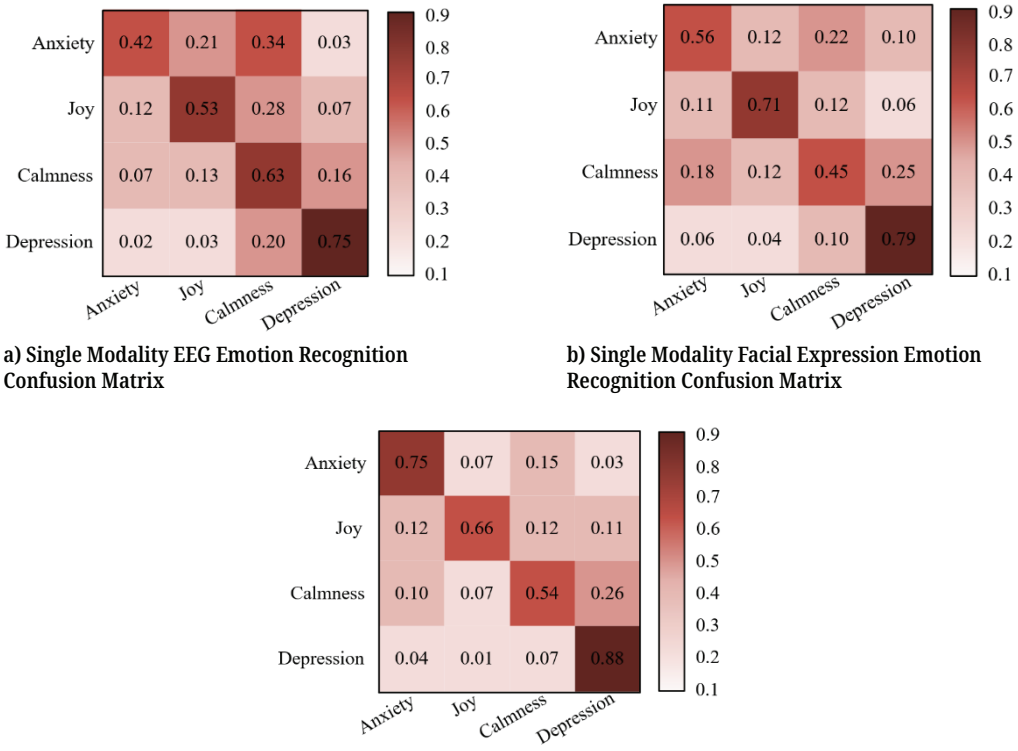
7.4 percentage points compared to the best single-fusion method (model layer fusion); the F1 score and AUC value reached 90.5% and 0.943, respectively, reflecting the method's advantage in classification accuracy and generalization ability. The ablation experiment results showed that removing either the feature-layer or model-layer fusion resulted in a noticeable decrease in recognition performance. The accuracy dropped to 80.3% and 79.5%, and the F1 score dropped by more than 11 percentage points, indicating that the cross-modality information integration at the feature layer and the attention-weighted integration at the model layer form a synergistic effect, both of which are essential. In conclusion, the proposed two-level fusion method can achieve accurate emotion state recognition for patients with psychological disorders and provide reliable data input for dynamic optimization of rehabilitation pathways.

Table 3. Comparison of rehabilitation effect scores between the experimental group and the control group (x ± s, points)

Evaluation Indicator	Group	Pretreatment	4 Weeks of Treatment	8 Weeks of Treatment	10 Weeks of Treatment	Change Rate from Baseline After Treatment	Intergroup Difference (P Value)
Depression Scale (SDS)	Experimental Group	62.5 ± 7.3	53.2 ± 6.8	45.1 ± 5.9	40.2 ± 5.2	–35.7%	<0.01
	Control Group	63.1 ± 7.5	58.6 ± 7.1	52.3 ± 6.5	48.5 ± 6.1	–23.1%	
Anxiety Scale (SAS)	Experimental Group	58.6 ± 6.9	49.5 ± 6.2	42.3 ± 5.5	37.8 ± 4.9	–35.5%	<0.01
	Control Group	59.2 ± 7.2	54.8 ± 6.8	49.1 ± 6.2	45.3 ± 5.8	–23.5%	
Quality of Life Scale – Physical Dimension	Experimental Group	58.2 ± 6.5	64.5 ± 6.1	70.3 ± 5.8	75.6 ± 5.3	+29.9%	<0.01
	Control Group	57.8 ± 6.7	60.2 ± 6.3	63.5 ± 5.9	66.8 ± 5.6	+15.6%	
Quality of Life Scale – Psychological Dimension	Experimental Group	55.3 ± 7.1	62.1 ± 6.5	68.4 ± 6.0	74.2 ± 5.5	+34.2%	<0.01
	Control Group	54.9 ± 7.3	58.6 ± 6.8	62.3 ± 6.2	65.1 ± 5.8	+18.6%	
Quality of Life Scale – Social Relationship Dimension	Experimental Group	59.1 ± 6.8	65.3 ± 6.2	71.5 ± 5.9	76.8 ± 5.4	+29.9%	<0.01
	Control Group	58.7 ± 7.0	61.5 ± 6.5	65.2 ± 6.0	68.9 ± 5.7	+17.4%	
Quality of Life Scale – Environmental Dimension	Experimental Group	60.5 ± 6.6	66.8 ± 6.1	72.4 ± 5.8	77.3 ± 5.2	+27.8%	<0.01
	Control Group	60.1 ± 6.8	63.2 ± 6.3	66.5 ± 5.9	69.8 ± 5.5	+16.1%	

To verify the clinical therapeutic advantages of the optimized rehabilitation path, a comparison of the rehabilitation effect differences between the experimental group and the control group was conducted. As shown in Table 3, both groups showed a reduction in SDS and SAS scores and an increase in quality-of-life scale scores after treatment. However, the experimental group showed significantly better improvements than the control group: after 10 weeks of treatment, the experimental group's SDS and SAS scores decreased by 35.7% and 35.5%, respectively, which was more than 12 percentage points higher than the control group; the total quality of life scale score and improvements in the physical, psychological, social relationship, and environmental dimensions all exceeded 27%, which was 10–16 percentage points higher than the control group. Statistical analysis showed that the experimental group showed significant improvements after four weeks of treatment ($P < 0.05$), and the trend of improvement continued to strengthen. After 10 weeks of treatment, the intergroup differences reached a significant level ($P < 0.01$). This indicates that the dynamically optimized rehabilitation path based on real-time emotional

feedback can more accurately adapt to patients' emotional changes. Combined with the immersive advantages of the VR-mobile integrated environment, it significantly enhances the improvement of depression, anxiety symptoms, and quality of life, with notable clinical therapeutic value.



a) Single Modality EEG Emotion Recognition Confusion Matrix

b) Single Modality Facial Expression Emotion Recognition Confusion Matrix

c) Proposed Two-Level Fusion Emotion Recognition Confusion Matrix

Fig. 3. Performance comparison of multimodal emotion recognition methods' confusion matrices

To verify the performance advantages of the proposed feature-layer-model-layer two-level fusion method in emotion recognition for patients with psychological disorders, this study compared the confusion matrices of single-modality and two-level fusion methods. From the results in Figure 3, it can be seen that in single-modality EEG recognition, the accuracy of depression recognition reached 0.75, but anxiety was only 0.42, indicating a clear category bias. In single-modality facial expression recognition, the recognition accuracy for joy was 0.71, and depression was 0.79, but the cross-category discriminability for anxiety (0.56) and calmness (0.45) was insufficient. In the proposed two-level fusion method, the recognition accuracy for anxiety (0.75), joy (0.66), calmness (0.54), and depression (0.88) all significantly improved, and the balance of recognition accuracy for each emotion category and the overall accuracy far exceeded the single-modality methods. This indicates that the two-level fusion strategy effectively integrates multimodal physiological and behavioral information, overcoming the dimensional limitations of single-modality data, and enables accurate recognition of complex emotional states in psychological disorder patients. This provides high-confidence data input for dynamic optimization of rehabilitation paths, fully demonstrating the method's technical adaptability and clinical value in psychological rehabilitation scenarios.

Table 4. User experience evaluation results of the experimental group

Evaluation Dimension	Specific Indicator	Unit	Test Value	Reference Standard	Compliance Status
Immersion Experience	IPQ Scale – Spatial Presence	Points (1–5)	4.2 ± 0.3	≥ 3.5 points	Compliant
	IPQ Scale – Loss of Reality	Points (1–5)	4.0 ± 0.4	≥ 3.0 points	Compliant
	IPQ Scale – Involvement	Points (1–5)	4.3 ± 0.3	≥ 3.5 points	Compliant
	IPQ Scale – Total Score	Points (1–5)	4.2 ± 0.3	≥ 3.5 points	Compliant
Treatment Adherence	Treatment Attendance Rate	%	96.2	≥80%	Excellent
	Interaction Task Completion Rate	%	94.5	≥85%	Excellent
	Treatment Plan Execution Rate	%	93.8	≥85%	Excellent
Device and Function Satisfaction	Device Operation Convenience Rating	Points (1–10)	8.6 ± 1.2	≥ 7 points	Satisfied
	Function Adaptability Rating	Points (1–10)	8.4 ± 1.3	≥ 7 points	Satisfied
	Overall Satisfaction Rating	Points (1–10)	8.5 ± 1.2	≥ 7 points	Satisfied
Adverse Reactions	Device Discomfort Incidence Rate	%	5.3	≤10%	Compliant
Immersion Experience	IPQ Scale – Spatial Presence	Unit	Test Value	Reference Standard	Compliance Status

To verify the user acceptance and clinical feasibility of the immersive treatment environment, a comprehensive evaluation of the user experience in the experimental group was conducted. As shown in Table 4, the experimental group performed excellently in core dimensions such as immersion, treatment adherence, and device satisfaction: The scores for the three dimensions of the IPQ scale and the total score all exceeded the passing threshold of 3.5, with spatial presence and involvement reaching scores above 4.0, indicating that the VR scene's immersive experience effectively enhances emotional involvement. The treatment attendance rate, interaction task completion rate, and plan execution rate all exceeded 93%, far surpassing the basic standard of 80%, showing that the optimized treatment path and immersive environment significantly improve patient treatment engagement. The device operation convenience, function adaptability, and overall satisfaction ratings all exceeded 8.4 points, and the device discomfort incidence rate was only 5.3%, meeting the safety and comfort requirements for clinical application. In conclusion, the constructed VR-interactive mobile fusion immersive treatment environment has high user acceptance, providing important support for the clinical promotion and popularization of the technology.

6 CONCLUSION

This study focuses on the construction of an immersive music therapy environment combining VR and interactive mobile technology and optimization of psychological rehabilitation pathways. By designing a six-level collaborative architecture—comprising the Technology Layer, Perception Interaction Layer, Data Processing Layer, Algorithm Decision Layer, Application Layer, and Feedback

Optimization Layer—it achieved the modular construction of an immersive treatment environment and multimodal interaction. The study proposed a feature-layer-model-layer dual-fusion multimodal emotion recognition method, solving the challenges of accuracy and robustness in emotion recognition for patients with psychological disorders. A reinforcement learning-driven dynamic rehabilitation path optimization mechanism was built. With clinical experimental validation, the experimental group's depression and anxiety scale scores decreased by more than 35% from baseline, and the improvement in quality of life across various dimensions exceeded 27%, significantly outperforming the traditional treatment group. Technologically, the study has overcome three key issues: "synergy between immersive experience and mobile interaction," "fusion of multimodal data heterogeneity," and "personalized dynamic optimization of rehabilitation paths." Clinically, it offers a non-pharmacological intervention plan that is immersive, accurate, and personalized for patients with depression and anxiety disorders. This provides a new paradigm for the intelligent and digital development of psychological rehabilitation technologies.

7 REFERENCES

[1] Y. Deng *et al.*, "Global, regional and national burden of lung cancer attributable to $PM_{2.5}$ air pollution: Trends from 1990 to 2021 with projections to 2045," *Journal of Environmental Management*, vol. 390, p. 126216, 2025. https://doi.org/10.1016/j.jenvman.2025.126216

[2] M. Ahmad, N. Wahid, R. A. Hamid, S. Sadiq, and A. Mehmood, "Decision level fusion using hybrid classifier for mental disease classification," *Computers, Materials & Continua*, vol. 72, no. 3, pp. 5041–5058, 2022. https://doi.org/10.32604/cmc.2022.026077

[3] N. H. Shahimi, R. Lim, S. Mat, C. H. Goh, M. P. Tan, and E. Lim, "Association between mental illness and blood pressure variability: A systematic review," *Biomedical Engineering Online*, vol. 21, no. 1, p. 19, 2022. https://doi.org/10.1186/s12938-022-00985-w

[4] L. Liao, M. Du, and Z. Chen, "Air pollution, health care use and medical costs: Evidence from China," *Energy Economics*, vol. 95, p. 105132, 2021. https://doi.org/10.1016/j.eneco.2021.105132

[5] H. Alan, "A comprehensive evaluation of digital mental health literature: An integrative review and bibliometric analysis," *Behaviour & Information Technology*, vol. 44, no. 10, pp. 2282–2304, 2025. https://doi.org/10.1080/0144929X.2024.2303626

[6] S. Nepal *et al.*, "Capturing the college experience: A four-year mobile sensing study of mental health, resilience and behavior of college students during the pandemic," in *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 8, no. 1, 2024, pp. 1–37. https://doi.org/10.1145/3643501

[7] Q. Ding, "Evaluation of the efficacy of artificial neural network-based music therapy for depression," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, p. 9208607, 2022. https://doi.org/10.1155/2022/9208607

[8] M. Wang, G. Luo, and H. Chen, "Practice of music therapy for autistic children based on music data mining," *Mathematical Problems in Engineering*, vol. 2022, no. 1, p. 4576211, 2022. https://doi.org/10.1155/2022/4576211

[9] C. Y. Wu, "Music therapy music selection based on big data analysis," *Applied Mathematics and Nonlinear Sciences*, vol. 9, no. 1, pp. 1–12, 2024. https://doi.org/10.2478/amns-2024-0019

[10] M. Song and Y. M. Song, "Randomized controlled trials of digital mental health interventions on patients with schizophrenia spectrum disorder: A systematic review," *Telemedicine and e-Health*, vol. 29, no. 6, pp. 798–812, 2023. https://doi.org/10.1089/tmj.2022.0135

[11] C. Zhang, X. Wang, D. A. Juraev, R. F. Efendiev, and X.-G. Yue, "Risk research on blockchain technology in interactive mobile hospitals based on the entropy method," *International Journal of Interactive Mobile Technologies*, vol. 19, no. 10, pp. 152–162, 2025. https://doi.org/10.3991/ijim.v19i10.55483

[12] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition: Review of sensors and methods," *Sensors*, vol. 20, no. 3, p. 592, 2020. https://doi.org/10.3390/s20030592

[13] B. Giardulli *et al.*, "Real and perceived feet orientation under fatiguing and non-fatiguing conditions in an immersive virtual reality environment," *Virtual Reality*, vol. 27, no. 3, pp. 2371–2381, 2023. https://doi.org/10.1007/s10055-023-00809-9

[14] S. Wei, "The role of EEG-based emotional feedback in enhancing the effectiveness of Tai Chi sports programs for oral health promotion," *IEEE Access*, vol. 13, pp. 81063–81082, 2025. https://doi.org/10.1109/ACCESS.2025.3561179

[15] H. Ren, L. Cheng, J. Zhang, and Q. Wang, "Eye-tracking investigation of emotional feedback to southern Hebei courtyard gates," *Journal of Asian Architecture and Building Engineering*, vol. 24, no. 6, pp. 5062–5079, 2024. https://doi.org/10.1080/13467581.2024.2407589

8 AUTHOR

Sai Wang holds a master's degree from Hebei University of Environmental Engineering. She is an assistant researcher, specializing in the history of music education, the historical development of urban music culture, and the study of modern and contemporary composers. She has led and participated in 8 provincial and municipal projects, published 10 papers, and edited 2 textbooks (E-mail: 18233545656@163.com).