Original Paper

Emotion- and Performance-in-the-Loop: Multimodal Difficulty-Adaptive Exergame System—Controlled Between-Participants Study

Liang Zhao[1,2], MEng; Dan Liu[3], Prof Dr; Hanli Zhang[1,2], MEng; Guanci Yang[3], Prof Dr; Donghua Zheng[4], PhD; Ling He[3], PhD

[1]School of Mechanical Engineering, Guizhou University, Guiyang, Guizhou, China

[2]North Alabama International College of Engineering and Technology, Guizhou University, Guiyang, Guizhou, China

[3]Key Laboratory of Advanced Manufacturing Technology, Ministry of Education, Guizhou University, Guiyang, Guizhou, China

[4]Department of Rehabilitation Medicine, The Third People's Hospital of Guizhou Province, Guiyang, Guizhou, China

Corresponding Author:
Dan Liu, Prof Dr
Key Laboratory of Advanced Manufacturing Technology, Ministry of Education
Guizhou University
No. 515, Jiaxiu South Road
Huaxi District
Guiyang, Guizhou, 550025
China
Phone: 86 189 8418 7485
Email: dliu@gzu.edu.cn

Abstract

Background: Existing exergames predominantly adjust difficulty based on explicit performance metrics, neglecting the player's implicit affective state, which often leads to early disengagement due to boredom or frustration.

Objective: To address this gap, we present a novel exergame system that implements joint emotion- and performance-driven dynamic difficulty adjustment (DDA), using a closed-loop controller that reasons over both motor performance and affective state.

Methods: The system integrates a computer-vision-based motion evaluation pipeline, a multimodal emotion recognition module that fuses facial action units (AUs), deep facial features, and motion-compensated heart rate variability (HRV) via an adaptive entropy-weighted decision-level fusion strategy, and a closed-loop difficulty controller. A controlled between-participants experiment (N=20) compared the adaptive system against a fixed-difficulty baseline during a 15-minute dance-exergame session.

Results: Emotion recognition accuracy reached 81.07% (macro-F_1: 81.02%) for 7 emotional states. The adaptive group demonstrated significantly lower cognitive load (NASA-TLX [National Aeronautics and Space Administration Task Load Index], P<.05), higher positive affect (Positive and Negative Affect Schedule [PANAS], P<.05), greater system usability (System Usability Scale [SUS], P<.05), improved motion accuracy (82.5% vs 74.8%, P<.05), and a physiologically healthier training load indicated by higher HRV root mean square of successive differences (RMSSD; 34.6 vs 24.3 ms, P<.05).

Conclusions: The joint emotion-and-performance difficulty-adaptation mechanism is feasible and significantly improves user experience and training outcomes compared to a fixed-difficulty baseline.

(JMIR Serious Games 2026;14:e101438) doi: 10.2196/101438

KEYWORDS

affective computing; dynamic difficulty adjustment; exergame; multimodal fusion; emotion recognition

XSL•FO
RenderX

Introduction

In recent years, sedentary lifestyles have become increasingly common worldwide, leading to a rise in chronic health issues such as obesity and cardiovascular diseases. In fact, insufficient physical activity has been identified as the fourth leading risk factor for mortality worldwide [1]. In this context, motion-based video games (exergames) have attracted research attention as an emerging way to combine physical exercise with engaging gameplay, using bodily movement as the primary means of game interaction so that exercise becomes more enjoyable and sustainable. Prior studies exploring the potential of fitness games to promote physical activity and health have shown that such games can increase users' activity levels, enhance motivation to exercise, and yield certain health benefits [2,3].

Despite the initial success of fitness games in motivating users to move, keeping players engaged over the long term remains a major challenge. One key factor is how to maintain sustained user engagement. In traditional games with fixed difficulty, there is a common phenomenon that if the game is too easy, players feel it is monotonous; if too difficult, players become frustrated—neither extreme provides an optimal experience [4]. This aligns with the psychological theory of "flow" that players experience immersive enjoyment only when the challenge level matches their skill level [5]. For exergames, without personalized difficulty adjustment mechanisms, players of different abilities may quickly experience the aforementioned boredom or frustration in a fixed-intensity exercise game, reducing their willingness to continue playing [6]. This necessitates that the game dynamically adjust exercise intensity and task difficulty according to the player's ability level, always keeping the player in the "optimal challenge zone" that is neither too easy nor overly difficult.

One effective way to improve the effectiveness and appeal of exergames is to implement dynamic difficulty adjustment (DDA) [7]—also referred to academically as dynamic game balancing [8], game difficulty adjustment [9], or game balancing [10]. DDA means the game automatically adjusts its difficulty based on the player's real-time performance in order to maintain a balance between challenge and skill and thereby enhance player engagement [11]. Over the past decade or more, researchers have explored various strategies for implementing DDA. Most current DDA methods rely solely on the player's external behavioral performance as the basis for adaptation. For example, the system can judge whether the current difficulty is appropriate based on metrics such as the player's score, number of failures, or reaction time, and accordingly decrease or increase the game challenge. Such performance-based adaptive techniques avoid the drawbacks of fixed difficulty to some extent, helping players of different levels receive an appropriate challenge [12]. However, this 1D difficulty adjustment neglects internal feedback such as the player's emotional and physiological states. Lyons [13] note that engagement and enjoyment in exergames are fostered not just by challenge but by how that challenge feels to the player. Recognizing this limitation, a substantial body of work has explored personalized and physiologically adaptive exergames. Martin-Niedecken et al [14,15] developed the ExerCube and related dual-flow fitness games, implementing real-time in-game adaptation based on both performance and physiological (eg, heart rate) signals. Göbel et al [16] and Hardy et al [17] proposed personalized exergames and formal adaptation models that integrate users' vital parameters into gameplay for sustainable training. More recently, Dill et al [18] reviewed sensing modalities in exergames and the accuracy of motion-quality and human-pose-estimation measurements, highlighting how motion data is leveraged to assess exergame efficacy.

The concept of using physiological signals for closed-loop difficulty adaptation in exercise and rehabilitation contexts has a longer history than is often recognized. Seminal work in this area was conducted by Koenig et al [19], who developed a real-time closed-loop system to control cognitive load during robot-assisted gait training. Their system used psychophysiological measurements (heart rate variability [HRV], skin conductance, respiration, and skin temperature) combined with task performance data to estimate whether patients who had stroke were cognitively overloaded, appropriately challenged, or underchallenged. An adaptive classifier was trained to recognize these states in real time, and the difficulty of a virtual task was automatically adjusted to maintain patients within an optimal cognitive engagement zone. Similarly, Novak et al [20] implemented a biocooperative feedback loop for upper extremity rehabilitation that adjusted task difficulty based on multiple psychophysiological measurements [21]. Their study found that psychophysiological measurements alone yielded modest accuracy (approximately 60%), but when combined with task performance and biomechanical data, accuracy improved substantially to 82%-89%. This finding highlighted the value of multimodal fusion and suggested that physiological signals are most effective when used as supplementary information rather than as a sole input. Recently, more and more studies have begun attempting difficulty adaptation that combines emotional and physiological feedback. Subsequent research has expanded affective and physiological closed-loop control to broader exercise contexts. Darzi and Novak [22] explored physiological linkage (covariation of participants' physiological responses) for automated affect classification and difficulty adaptation in competitive exergame scenarios, demonstrating that synchronization between participants' physiological states can provide additional information beyond individual responses. Other studies have investigated affective adaptation in virtual reality exergames [23], multimodal engagement recognition combining physiological signals with performance metrics [24], and deep learning approaches to physiological-based difficulty adjustment [25]. However, despite these advances, several gaps remain. First, most prior work in rehabilitation-focused affective computing has been conducted with clinical populations using specialized equipment in supervised settings, limiting generalizability to home-based fitness scenarios with healthy users and consumer-grade devices. Second, previous systems have primarily addressed low-intensity, controlled movements (eg, reaching tasks, slow gait) rather than high-intensity, whole-body dynamic exercises such as dance, where motion artifacts and rapid cardiovascular changes pose additional challenges for emotion recognition. Third, difficulty adaptation mechanisms in rehabilitation contexts have typically focused

XSL•FO
RenderX

on a single dimension (cognitive load or assistance level), whereas fitness exergames require joint consideration of both emotional state and physical performance to maintain motivation and training effectiveness.

This paper centers on the concept of "emotion- and performance-in-the-loop" adaptive difficulty for exercise games. First, we construct an emotion-in-the-loop closed-loop adaptive framework for exergames, expanding game difficulty adjustment from relying solely on behavioral performance to a dynamic mechanism that comprehensively senses both the user's external and internal state. Second, we investigate a multimodal user emotion recognition method using facial expressions and physiological signals, and, combined with motion recognition techniques to capture user performance, we simultaneously characterize physical performance and emotional state in real time, providing a unified and continuous state representation for high-level difficulty decision-making. Finally, we developed a prototype exergame system that implements the above methods for adaptive difficulty, and through controlled user experiments we evaluated its effects on user engagement, subjective experience, and system usability.

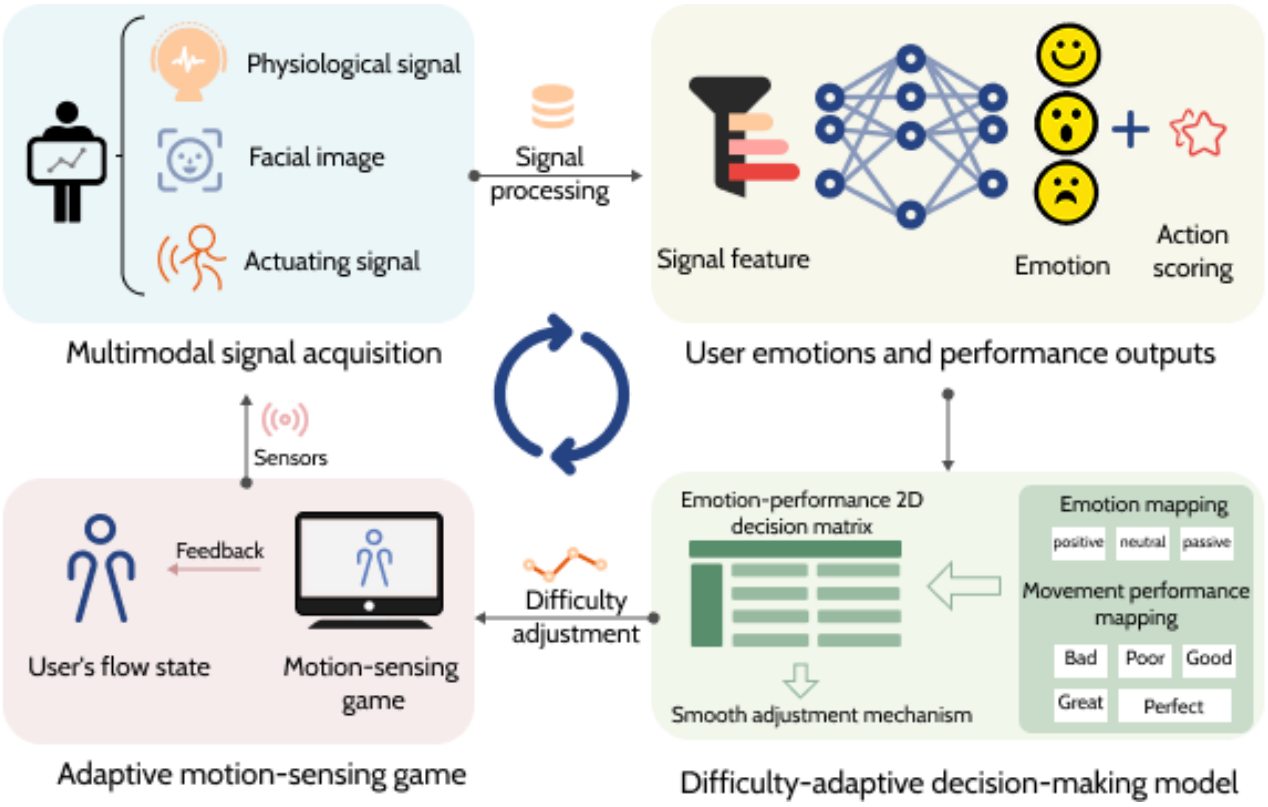
Methods

Adaptation Framework

Figure 1 shows the proposed "emotion- and performance-in-the-loop" adaptive framework, which consists of multiple interlinked functional modules. The multimodal signal acquisition module uses various sensors to collect the user's physiological, visual, and motion signals in real time. Through timestamps and synchronization mechanisms, it ensures temporal alignment of multisource data, providing real-time, synchronized raw data for subsequent processing. The signal processing and feature extraction module preprocesses each modality's data (eg, filtering and denoising) and extracts key features (eg, HRV, facial expression features, and kinematic features). After normalization, these features are fused into a unified high-dimensional feature vector, forming a comprehensive representation of the user's state. The user emotion and performance output module integrates the recognition results into a 2D emotion-performance state representation, where the emotional dimension and performance dimension form an intuitive coordinate of the user's state, providing a basis for subsequent adaptive decision-making.

Ultimately, the system evaluates the user's motion performance in real time based on information from the action evaluation module and simultaneously captures the user's emotional state (eg, enjoyment or frustration). Both are used as feedback signals for dynamic game difficulty modulation. By incorporating the user's emotions into the control loop, the exergame system can more comprehensively sense the user's immediate state. It can lower the difficulty when the user shows negative emotions such as frustration to alleviate discouragement and appropriately increase difficulty when the user is performing well and feeling positive to intensify the training stimulus.

Figure 1. Emotion- and performance-in-the-loop adaptive difficulty adjustment framework for exergames.



XSL•FO
RenderX

Game and Interaction

Before describing the sensing and adaptation mechanisms, we outline the game itself. The exergame follows a motion-imitation paradigm common in rhythm-based fitness games. The player stands 1.5-2.5 m in front of a camera and display. An on-screen reference performer demonstrates full-body choreography—arm raises, side steps, squats, jumps, and torso rotations—synchronized to music at 100-140 beats per minute. The player attempts to reproduce each movement in real time. The system compares the player's skeletal keypoints to the reference pose frame by frame and immediately displays 1 of 6 accuracy ratings (miss, ok, yeah, good, great, and perfect) together with a running score and combo counter. A typical session lasts 10-15 minutes and comprises 3-5 songs, each presenting 40-80 consecutive moves. The camera is positioned to capture the full body from head to knees, and players are instructed to remain facing the camera throughout the session. Two parameters govern how demanding the game is: the similarity thresholds that separate the 6 rating bands (eg, at baseline, ≥90% similarity yields perfect, ≥80% yields great) and the playback tempo of the choreography and music, expressed as a percentage of the original speed. These are exactly the 2 parameters the adaptive controller manipulates (described below). Because both engagement and physical exertion fluctuate rapidly in this whole-body, fast-paced setting, it constitutes a demanding testbed for closed-loop difficulty adaptation.

Multimodal User State Sensing

Achieving the above affective closed-loop adaptation hinges on sensing the user's state accurately and in real time. We use a multimodal sensing approach, simultaneously monitoring the player's performance and emotional state to obtain a comprehensive quantitative view of their status. The user state monitoring module of our system consists of two parts: (1) motion tracking via computer vision to assess the player's performance; (2) emotional state detection via physiological signals and facial expressions to infer the player's psychological state.

Motion Recognition and Evaluation

In the motion analysis channel, the system uses Google MediaPipe's lightweight BlazePose model to extract the user's skeletal keypoints and their confidence scores in real time, forming a time series of human poses. As shown in Figure 2 [26], the BlazePose model can obtain coordinates for 33 keypoints on the human body.

Figure 3 illustrates the multidimensional motion similarity evaluation process based on MediaPipe. First, to eliminate individual differences and motion variations, the raw coordinates undergo normalization preprocessing: using the hip midpoint (equation 1) and shoulder midpoint (equation 2) as references for translation, using trunk length (equation 3) for scaling normalization, and then aligning the body's principal axis to vertical using rotation matrix *R*. This produces a standardized pose sequence (equation 4), where *s* is the scale factor (equation 5).

$$c_h = \frac{P_{23} + P_{24}}{2} \quad (1)$$

$$c_s = \frac{P_{11} + P_{12}}{2} \quad (2)$$

$$L_t = \left\| c_s - c_h \right\|_2 \quad (3)$$

$$P_j^{nom} = R\frac{P_j - c_h}{s} \quad (4)$$

$$s = \max\left(L_t^{nom},\ D_{max}\right) \quad (5)$$

The module tracks whole-body, standing dance-type movements (eg, arm swings and raises, trunk rotation, leg raises, and knee flexion). Reliable tracking requires the whole body to remain within the depth camera's field of view under adequate lighting; accuracy may degrade for very fast movements or severe self-occlusion (eg, limbs occluding one another or rapid turns away from the camera). Transient occlusion and dropped frames are mitigated by keypoint interpolation and by inferring the region of interest (ROI) from the previous frame, and an auto-tracking pan or tilt unit keeps the user centered in the view.

XSL•FO
RenderX

Figure 2. The 33 skeletal keypoint indices in the BlazePose model.

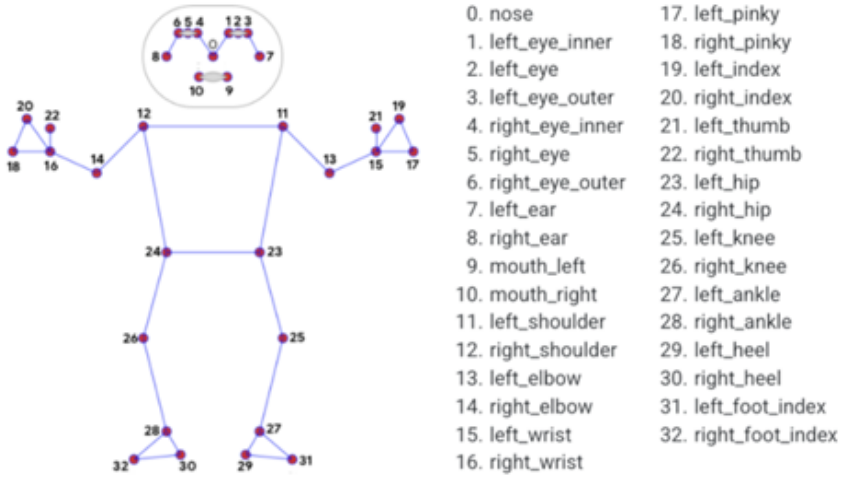
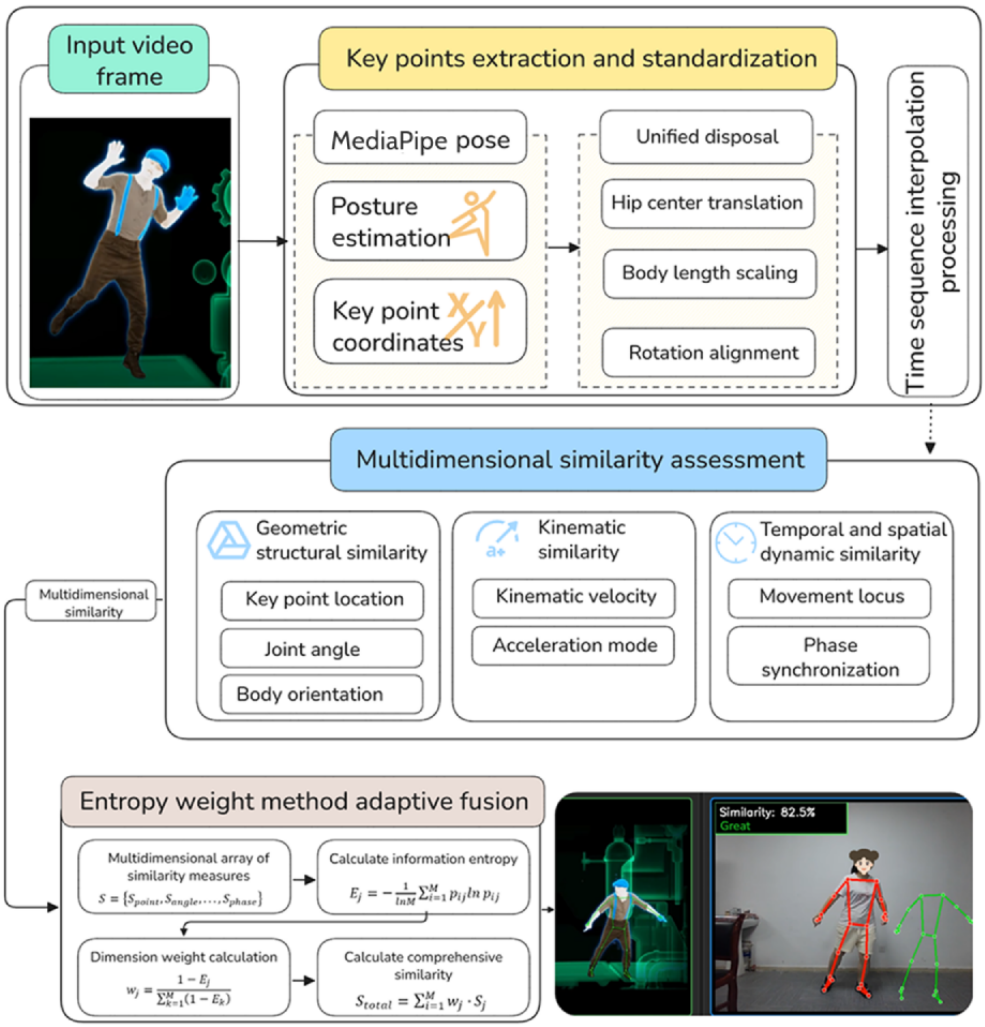


Figure 3. Multidimensional motion evaluation method based on MediaPipe.



Emotion Recognition and Fusion

In the emotion recognition channel, we adopt the foundational account of 7 basic emotions by Ekman [27]. Two considerations underlie the choice of a categorical model over continuous valence-arousal regression. First, the controller consumes only the sign of valence, and categorical posteriors pooled into a weighted valence score are comparatively robust to head motion, partial occlusion, and nonfrontal viewing that characterize vigorous exergame play, conditions under which direct dimensional regression degrades substantially [28,29]. Confusions within the negative valence classes leave the resulting decision unchanged, which is desirable for closed-loop control. Second, arousal is dominated by physical exertion in

XSL•FO
RenderX

this setting and is therefore weakly diagnostic of affective state, leaving valence as the actionable dimension [30,31]. This is a pragmatic engineering choice validated by the downstream effectiveness of the adaptation, not a claim that categorical models are theoretically superior. The system infers the user's emotions through 2 modalities: facial expression analysis and physiological signal processing (Figure 4). For the image modality, first, we investigate emotion classification methods based on facial keypoints and facial action units (AUs) using images captured by a depth camera. Then, to mitigate accuracy loss from single-modality data dropout, we further develop a method to extract HRV features from a wrist-worn photoplethysmography (PPG) sensor. We also design a motion-compensation technique for the physiological signals to suppress interference from physical activity. Finally, we design a fusion strategy for the image and physiological modalities to obtain the user's emotional state changes during exercise.

Figure 5 shows the principle of microexpression emotion recognition from facial images, consisting of 4 steps: face detection and image preprocessing, facial feature extraction using a DeepFace framework, AU modeling, and temporal analysis. First, each RGB (red, green, and blue) video frame is processed with face detection and image preprocessing. A Haar cascade classifier is used for rapid face localization, combined with depth data from the camera for validation, ensuring reliable face capture even with head movement or lighting changes. The located and aligned normalized face image is then input for expression feature extraction. Facial emotion classification is performed with the DeepFace framework, invoking its emotion module with the default pretrained convolutional backend and default detector settings; the 7 output classes are happy, sad, angry, surprise, fear, disgust, and neutral (as shown in Figure 6). After multiple convolutional and fully connected layers, the network's last fully connected layer (FC) outputs a vector of 7 values, which are raw outputs called logits. Logits are unnormalized scores spanning the real number range; they represent the model's "raw evidence strength" for each emotion category but cannot be directly interpreted as probabilities. Therefore, a Softmax function is used to convert the logits output by the activation function into a probability distribution between 0 and 1 that sums to 1. The Softmax function is given by equation 6, where z_k is the logits value for class k. Combining the above, after the 16-layer deep convolutional network, the Softmax classifier outputs a probability distribution over 7 basic emotions (anger, disgust, fear, happiness, sadness, surprise, and neutral).

$$P(y = k \mid x) = \frac{exp(z_k)}{\Sigma exp(z_j)} \qquad (6)$$

On the other hand, based on the coordinates of facial keypoints, we compute the intensity of specific facial AUs to capture subtle local muscle movements (Figure 7). For example, a happy emotion can be characterized by the combination of AU6 and AU12; a sad emotion by AUs 1, 7, and 15; and a neutral emotion is not directly tied to any specific AU combination and is determined when no significant AU activation is present. Deep learning features emphasize global expression semantics, while AU intensity features reflect local microexpression changes; the 2 are complementary. The system fuses these 2 types of features and inputs them into an emotion classifier, which outputs a preliminary emotion recognition result and its probability distribution based on facial expression. This approach, which fuses deep semantic features with geometric AU features, helps recognize brief and subtle microexpressions during physical exercise, improving the robustness of visual emotion recognition.

Figure 8 illustrates the principle of using HRV as an emotional physiological feature. The interbeat interval (RR) series was obtained from a wrist-worn PPG sensor based on the MAX30102 module (sampled at 100 Hz). RR intervals were extracted by peak detection on the PPG waveform, from which time-domain HRV features (mean RR interval, SD of Normal-to-Normal intervals [SDNN], root mean square of successive differences [RMSSD], and pNN50 [percentage of successive interbeat intervals differing by more than 50 ms]) and frequency-domain features (low frequency [LF] and high frequency [HF] power and the ratio of low-frequency to high-frequency power [LF/HF]) were computed. HRV indices were computed over a 30-second sliding window of RRs. This length was chosen so the control loop responds within a song rather than across a session. Time-domain indices and RMSSD in particular have been validated at ultrashort window lengths, whereas frequency-domain indices require longer recordings and were therefore not used by the controller [32,33]. These features reflect the activity level of the autonomic nervous system and are closely related to emotional stress and arousal [34]. To eliminate the influence of exercise on emotional physiological features, we use a motion-aware quality-weighting scheme. During vigorous whole-body exercise, wrist-PPG HRV measurement is degraded by 2 factors: motion artifacts in the signal and exercise-induced tachycardia that reflects exertion rather than affect. The wrist unit's 3-axis accelerometer yields a per-sample magnitude $a = (a_x^2 + a_y^2 + a_z^2)^{\frac{1}{2}}$; over a 6-second window, we compute $\sigma(a)$ and normalize it to a motion level $m \in [0,1]$, empirically calibrated so that light arm swings and vigorous jumping span the interval. A quality term $q = 1 - 0.5m$ is attached to the HRV channel, raised by 0.2 (capped at 1.0) when a clean beat series is recovered. This quality term q weights the HRV channel in the decision-level fusion with the facial channel (described below). When no clean beat series is available, m enters a fallback arousal estimate at 10% weight. We deliberately do not apply independent component analysis (ICA), Kalman, or adaptive-filter reconstruction, as these add latency incompatible with a per-move control loop; recent work indicates that under high-motion conditions, artifact detection with selective weighting performs at least as well as reconstruction for downstream inference. The quality-weighted HRV features obtained through this process more accurately reflect emotion-induced physiological responses [35,36]. Finally, the corrected HRV features are input into a pretrained Sigmoid emotion mapping model. Thus, the system obtains the emotion classification result and probability distribution based on physiological signals.

XSL•FO
RenderX

To enhance recognition robustness in the exercise setting, the decision-level fusion module applies 3 steps sequentially: confidence evaluation, adaptive entropy-weighted fusion, and temporal smoothing.

Regarding confidence evaluation, for the probability vector from the facial channel and the probability vector from the physiological channel, we compute their confidences (equation 7) separately. Here, $C_{entropy}$ represents the information entropy of a probability distribution, reflecting classification certainty; $C_{quality}$ represents a modality-specific signal quality metric, which is normalized and used in weighting. This step ensures that a modality with higher reliability carries greater weight in the fusion.

Regarding adaptive fusion based on entropy weighting, first, we use the confidence values to adjust and normalize the probability vectors. Next, for each emotion category e_i, we calculate the consistency entropy of the decisions across modalities. The smaller the entropy E_i (equations 8,9), the more consistent the modalities are for category e_i, and thus the higher the fusion weight assigned to that category. The final fused probability P_{fused} is computed accordingly (equation 10), where λ_m is the base weight for a modality, which can be dynamically adjusted based on the current exercise intensity.

Regarding temporal smoothing, to suppress single-frame recognition fluctuations, we apply an exponential smoothing filter to the fusion results. Based on the interaction pace of exergames and multiple test experiments, we found that setting the smoothing factor γ to 0.7 achieves a balance between stability and responsiveness (equation 11). After smoothing, we take the category with the highest probability as the final emotion output, providing the subsequent difficulty adaptation module with continuous emotional state feedback.

$$C_{\mathrm{mod}ality} = \beta \cdot C_{entropy} + (1-\beta) \cdot C_{quality} \quad (7)$$

$$P'_{\mathrm{mod}ality}(e_i) = \frac{C_{\mathrm{mod}ality} \cdot P_{\mathrm{mod}ality}(e_i)}{\sum_i C_{\mathrm{mod}ality} \cdot P_{\mathrm{mod}ality}(e_i)} \quad (8)$$

$$E_i = -\sum_m p_m^{(i)} \ln p_m^{(i)} \quad (9)$$

$$P_{fused}(e_i) = w_i \cdot \sum_{m \in M} \lambda_m P'_m(e_i) \quad (10)$$

$$P_{smoothed}^{(t)}(e_i) = \gamma \bullet P_{fused}^{(t)}(e_i) + (1-\gamma) \bullet P_{smoothed}^{(t-1)}(e_i) \quad (11)$$

After these 3 steps, the system outputs the final smoothed emotional probability distribution. The emotion category with the highest probability is taken as the recognized result at that moment, and its probability value is also used as an indicator of emotion intensity. This provides a precise, continuous, and robust emotion feedback input for the subsequent difficulty adaptation algorithm.

XSL•FO
RenderX

Figure 4. Workflow of emotion recognition based on facial images and physiological signals. CNN: convolutional neural network; HRV: heart rate variability.

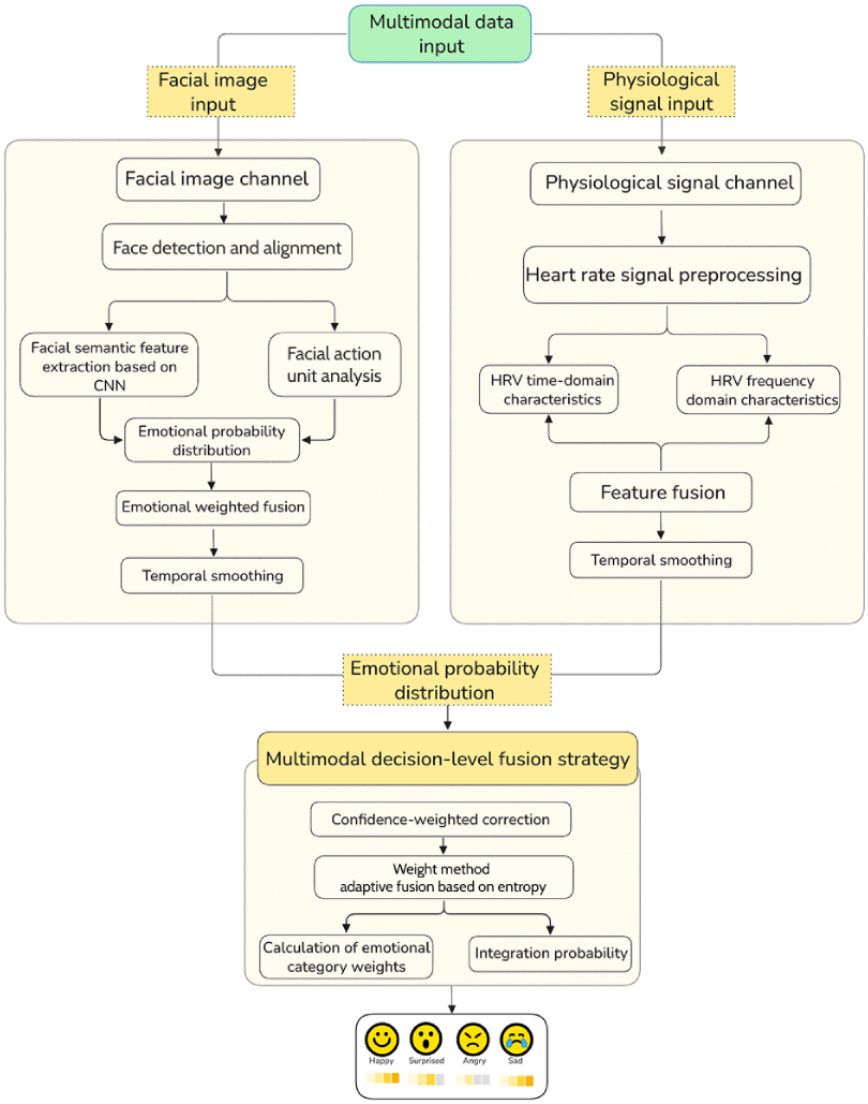
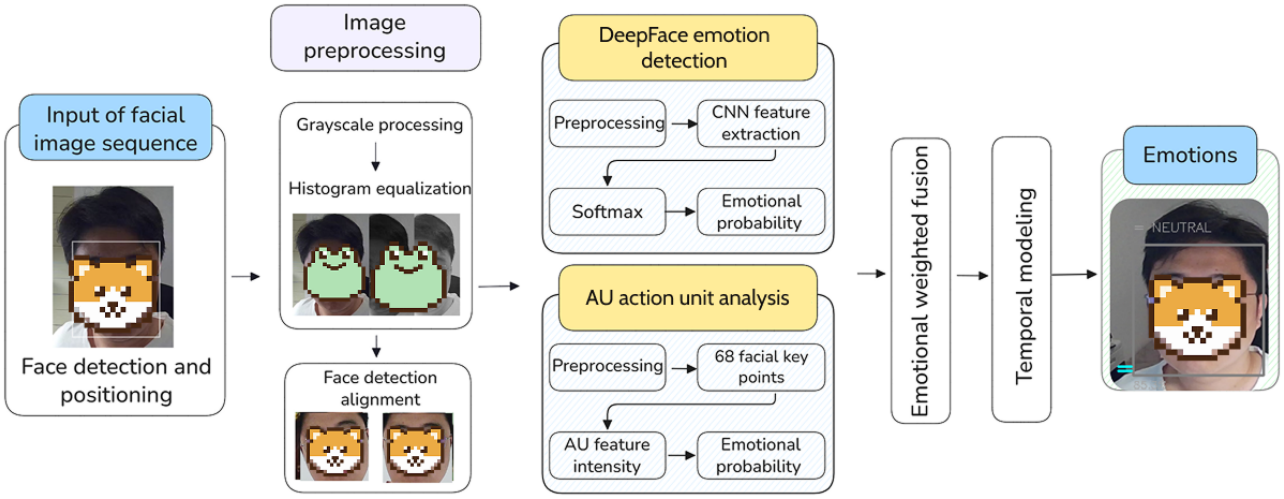


Figure 5. Facial image-based emotion recognition method. AU: action unit; CNN: convolutional neural network.



XSL•FO
RenderX

Figure 6. DeepFace emotion recognition module architecture. ReLU: rectified linear unit; RGB: red, green, and blue.

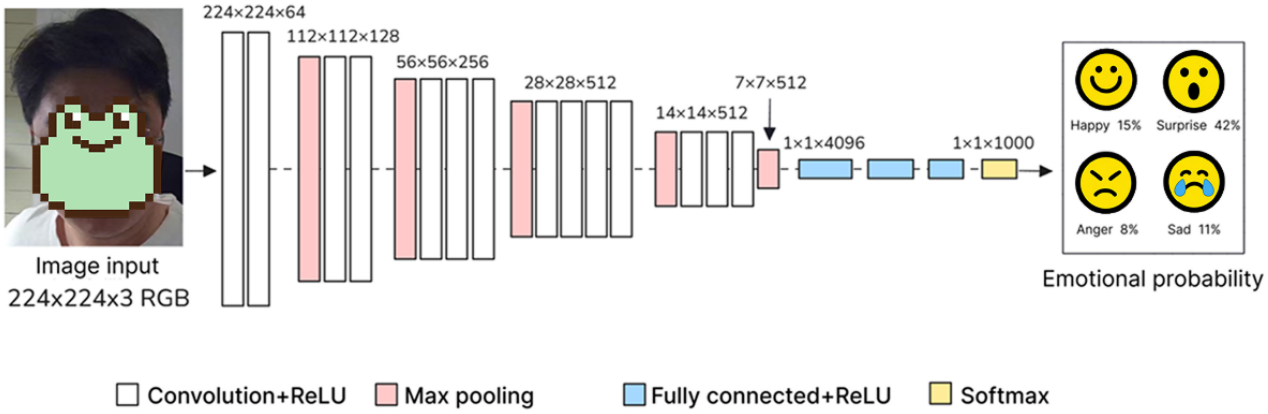


Figure 7. Relationship between facial AUs and emotions. AU: action unit.

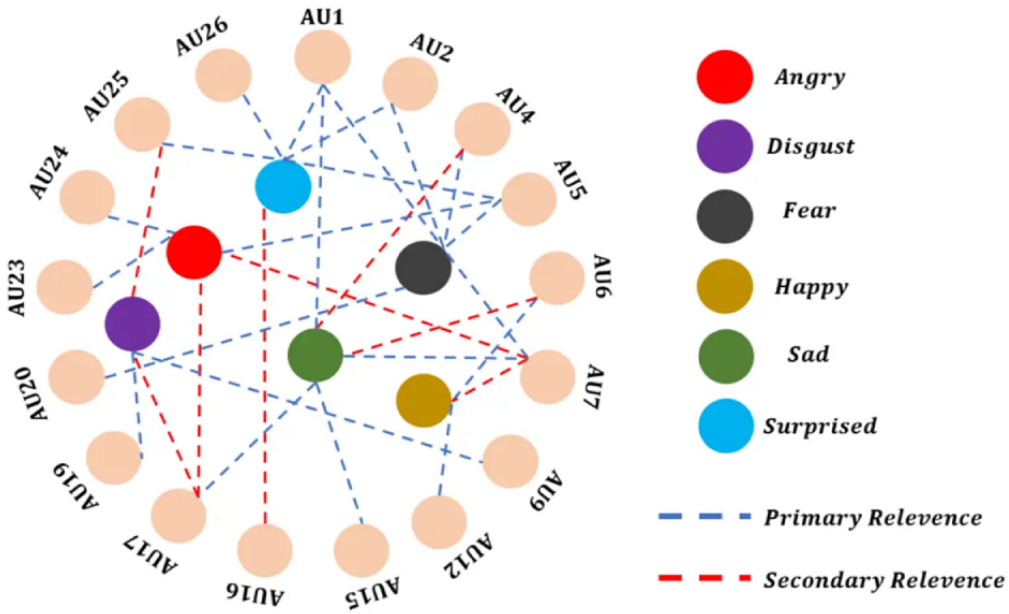
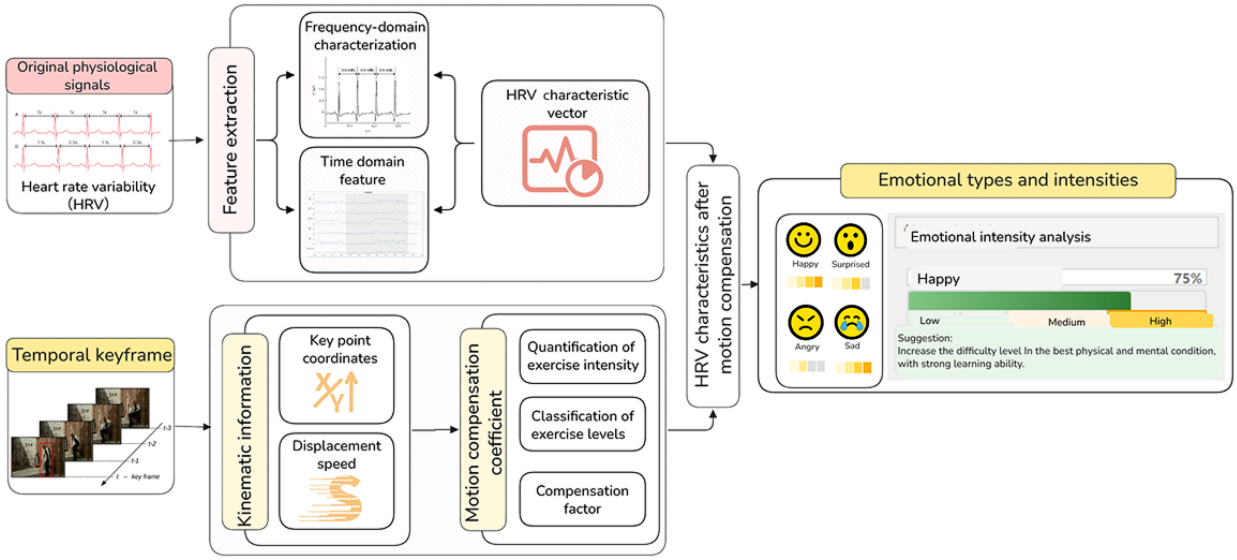


Figure 8. Physiology-based emotion recognition method. HRV: heart rate variability.



XSL•FO
RenderX

Difficulty Adaptation

The difficulty controller operates on a hand-crafted, interpretable rule base rather than a learned policy. Machine learning is confined to the upstream modules for emotion recognition and motion scoring; the adaptation rules themselves are explicitly designed and grounded in Csikszentmihalyi's flow theory.

The controller first abstracts the continuous 7-class emotion output into 3 discrete valence levels—negative, neutral, and positive. This is achieved in 3 steps.

- Each of the 7 emotion classes is assigned a fixed valence weight: happy +1.0, surprise +0.5, neutral 0, sad −0.6, disgust −0.7, fear −0.8, and angry −0.9.
- The classifier's posterior probabilities are combined with these weights into a single score $V = \sum_i p_i w_i$, which lies in [−1, +1]. For example, posteriors of 0.60 happy, 0.30 neutral, and 0.10 sad give V = 0.60(+1.0) + 0.30(0) + 0.10(−0.6) = 0.54.
- V is thresholded into the 3 levels used by the controller: positive (V ≥ +0.3), neutral (−0.3 < V < +0.3), and negative (V ≤ −0.3). In the example, V = 0.54 yields positive valence. Averaging over posterior probabilities rather than taking the argmax class implies that misclassifications among negative classes—which share the same sign—do not affect the final predicted label.

In parallel, the motion-evaluation module produces a composite performance score, denoted as OverallScore, which is a weighted combination of 6 submetrics: average accuracy (weight 0.4), one minus the miss rate (0.2), perfect rate (0.2), consistency (0.1), score trend (0.05), and recent improvement (0.05).

As shown in Table 1, this continuous score is subsequently discretized into 5 performance levels: poor, below average, average, good, and excellent. The core adaptation decision is made by indexing a predefined 3 × 5 decision matrix with the 3 valence levels and 5 performance levels. Each cell of the matrix specifies 1 of 3 possible actions: decrease, maintain, or increase difficulty. The matrix follows flow-theory principles: when performance is low and affect is negative, the challenge is deemed too high, triggering a difficulty decrease; conversely, when performance is high and affect is positive, the challenge is deemed too low, prompting an increase. All boundary states default to maintain, which provides an observation window that effectively prevents decision oscillation. The matrix deliberately does not raise difficulty on negative valence alone. The costs are asymmetric: raising difficulty on a player who is in fact frustrated prompts immediate withdrawal (stopping or ending the song), whereas failing to challenge a bored player degrades engagement gradually and leaves later opportunities to adjust. With a 3-level valence signal that cannot distinguish frustration from boredom, the conservative policy has the better worst case. The performance axis, not the emotion axis, separates the 2: negative valence with low performance is treated as overload and decreases difficulty, while negative valence with sustained high performance is treated as possible disengagement and returns toward baseline.

Difficulty in this game is governed by 2 factors: how strictly a movement must match the reference to achieve a given rating and the pace at which targets are presented. These correspond to 2 sets of parameters that the controller manipulates—similarity thresholds and tempo. The 6 rating bands are based on similarity thresholds; by default, ≥90% similarity gives a "perfect" and ≥80% gives a "great." Increasing all thresholds by a constant amount makes high ratings harder to obtain. Tempo is set as a percentage of the original choreographed speed and is altered only at song boundaries to prevent midsong rhythm disruption. A difficulty-decrease action, for instance, would reduce every threshold by 3 percentage points and drop the tempo from 100% to 90%.

Once an action is determined, the controller adjusts these parameter sets accordingly. The raw adjustment magnitude per update is given by:

$$\Delta = \text{clamp}(\beta \cdot I, \Delta_{\min}, \Delta_{\max}) \quad (12)$$

where I is the decision intensity and β is the gain coefficient. Through offline replay of user sessions, the constants were calibrated to β = 0.03, $\Delta_{\min}$ = 0.01, and $\Delta_{\max}$ = 0.15, ensuring that each change is large enough to have a noticeable effect yet small enough to remain imperceptible to the user. To avoid abrupt transitions, each threshold τ is updated via first-order smoothing:

$$\tau_{\text{current}} \leftarrow \text{Lerp}(\tau_{\text{current}}, \tau_{\text{target}}, \alpha) \quad (13)$$

with a smoothing factor α=0.5. Additionally, a minimum interthreshold spacing is enforced at every step to prevent rating overlap.

Finally, the controller evaluates the fused state continuously, but it applies at most one adjustment per cooldown interval. This cooldown mechanism, together with temporal smoothing of the emotion output, effectively suppresses transient fluctuations and system oscillation, ensuring that all difficulty changes remain gradual and stable throughout the entire session.

Table 1. Emotion-performance decision matrix for difficulty adjustment.

Emotional valence	Poor	Below average	Average	Good	Excellent
Negative	Decrease	Decrease	Decrease	Maintain	Maintain
Neutral	Decrease	Decrease	Maintain	Maintain	Increase
Positive	Decrease	Maintain	Maintain	Increase	Increase

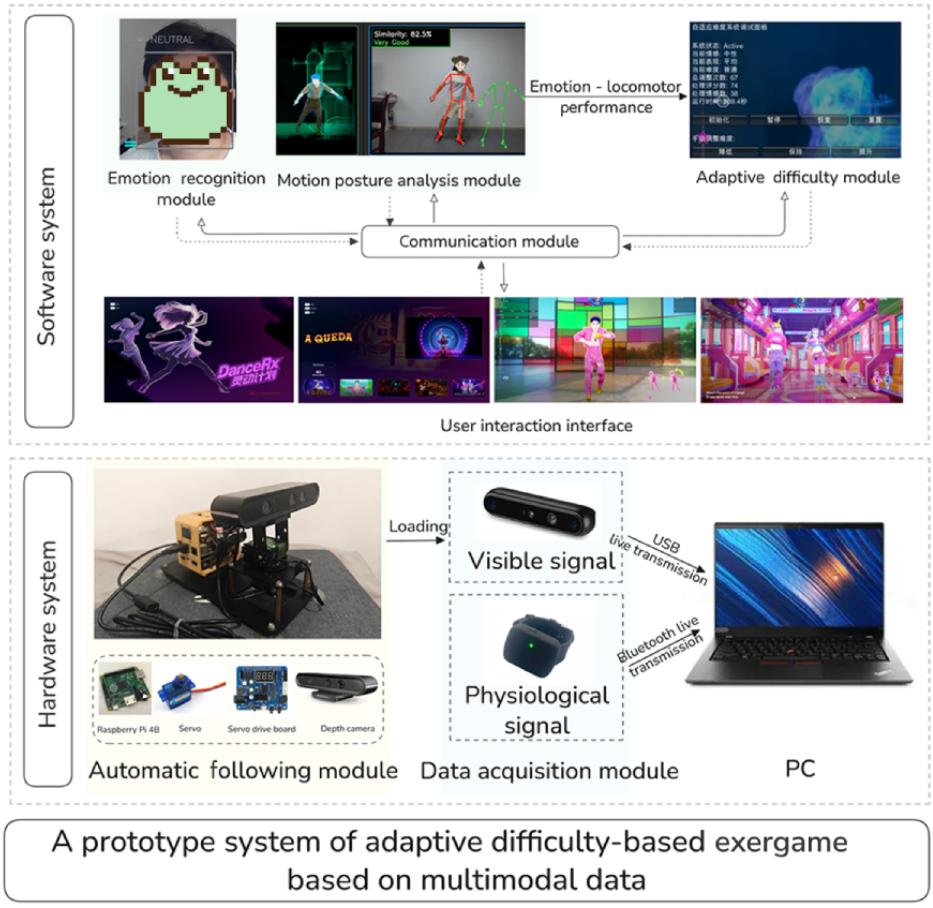
XSL•FO
RenderX

Implementation

The system is implemented in Python, using MediaPipe for pose estimation, the DeepFace framework (v0.0.75+) for facial emotion classification, and NumPy/SciPy for beat detection and HRV computation. The wrist unit (MAX30102-based PPG sensor with three-axis accelerometer) streams at 50 Hz over serial. On a gaming laptop (16-core CPU, 32 GB RAM, dedicated GPU), the game renders at 60 fps, and the full sensing-to-adaptation path runs at a measured end-to-end latency of 125 ms (SD 17.24; 95% of frames under 156.3 ms).

Based on the above requirements, we designed the overall system architecture (Figure 9) using an open-source Just Dance (Ubisoft)–style project (based on camialtr), rather than as a modification of the commercial Ubisoft title. On this open-source dance game, we added (1) a multimodal affective-computing-driven DDA system and (2) an improved interaction and feedback system; baseline functionality already provided by the open-source project is not claimed as a contribution. Just Dance is referenced only as a gameplay paradigm; no proprietary Ubisoft code or assets were used. The architecture consists of a hardware system and a software system. The hardware system includes an auto-tracking module, a data acquisition module, and a PC; the software system includes a front-end interaction module responsible for user interface rendering and interaction, a back-end decision module carrying the core intelligent algorithms, and a data communication module to ensure efficient data flow.

Figure 9. System architecture design.



Experimental Evaluations

We conducted a controlled between-participants experiment in which participants were randomly assigned to either the joint emotion-and-performance adaptive condition or a fixed-difficulty control condition. The study had three objectives that together evaluate the joint emotion-and-performance controller: (1) to evaluate the accuracy of the emotion-recognition module under physical activity; (2) to evaluate the reliability of the motion-performance evaluation, which provides the performance input to the joint controller; and (3) to compare the joint emotion-and-performance adaptive mechanism against a fixed-difficulty baseline on users' subjective experience, engagement, and training outcomes.

We recruited 20 healthy young volunteers (sedentary university students and office workers), 10 male and 10 female, ages 18-35 years. All participants had no injuries or cardiovascular conditions affecting exercise, normal vision and hearing, and no extreme aversion to exergames. The 20 participants were randomly divided 1:1 into an adaptive difficulty group (n=10) and a fixed difficulty group (n=10). The adaptive group used our proposed difficulty adaptation mechanism based on dual feedback of emotion and performance; the fixed group used the same exergame but with difficulty fixed at a preset level, with no adjustment based on user state during training. The 2 groups had identical training task content, experimental environment, and feedback interface to ensure the comparison's credibility and validity.

XSL•FO
RenderX

The experiment was conducted in a quiet indoor environment, using a hardware setup integrating a depth camera and a wrist-worn heart rate sensor. On the software side, Unity3D presented the exergame interface, and a Python backend processed multimodal data in real time. Each participant completed a single 15-minute training session, during which the system recorded skeletal sequences, real-time scores, emotion recognition results, and timestamps; if a participant quit early, their duration was recorded. Immediately after training, we collected questionnaire responses and interview feedback.

Different data collection and labeling methods were used for the 3 objectives.

First, regarding motion recognition or evaluation module performance evaluation, we extracted motion segments from the recorded videos and invited exercise rehabilitation experts to manually score them on a 6-level scale (miss to perfect), forming a test set of about 1200 samples. We computed the overall accuracy and macro-F_1 between the algorithm's evaluations and the expert ratings and used a confusion matrix to observe confusion between levels. Medical rehabilitation experts manually rated the recorded movements on the 6-level scale to establish a human-rating gold standard against which the automatic ratings were compared.

Second, we randomly sampled frames containing faces from the experiment videos. Five annotators—all university faculty conducting research in facial expression analysis and affective computing—completed a calibration session on the 7-category scheme before labeling. The 1400 frames were evenly split among the annotators (280 each), so that every frame received 1 expert label as ground truth. To assess interrater reliability, a randomly selected subset of 500 frames was independently reannotated by all 5 raters, yielding Fleiss $\kappa = 0.71$—indicating substantial agreement, comparable to values reported for in-the-wild facial affect corpora [37,38]. We used overall accuracy and macro-F_1 as evaluation metrics and generated a confusion matrix to analyze the error distribution of our model's predictions against the expert labels.

Third, regarding user experience and training effect evaluation, subjectively, we used the NASA-TLX (National Aeronautics and Space Administration Task Load Index; cognitive load), Positive and Negative Affect Schedule (PANAS; positive or negative affect), and System Usability Scale (SUS; system usability) scales. Objective metrics included the action accuracy (the correctness of performing required movements, computed by the system), the duration of gameplay (in minutes that each user kept playing or number of levels completed), and HRV (RMSSD, reflecting physiological load during training). Through between-group comparisons and semistructured interviews, we holistically evaluated the effect of the adaptive mechanism.

Ethical Considerations

Ethics Approval

This study was approved by the Human Medical Experiment Ethics Sub-Committee of Guizhou University (HMEE-GZU-2025-T042).

Informed Consent

All participants received a written information sheet describing the purpose of the study, the procedure, the data to be collected, and their right to withdraw at any time without penalty. Written informed consent was obtained from every participant before the experiment began.

Privacy and Confidentiality

All data were pseudonymized at the point of collection: each participant was assigned a study ID, and the mapping between study IDs and identities was kept separately from the research data. Video and physiological recordings were stored on a password-protected local server accessible only to the research team and were not transmitted to any third party. For publication, all images of participants have been blurred so that individuals are not identifiable.

Participant Compensation

Participants received small gifts for their time. No financial compensation was provided.

Results

Motion Evaluation Module Performance

On the offline motion scoring test set, the system's motion evaluation module achieved an overall accuracy of 78.33% and a macro-F_1 score of 78.55% for the expert 6-level ratings. Per-class precision, recall, and F_1 scores, along with the confusion matrix, are provided in Figures S1 and S2 in Multimedia Appendix 1. This result indicates that the algorithm achieved nearly 80% accuracy under strict scoring consistency criteria and is able to reasonably reproduce the outcomes of manual evaluation. The confusion matrix shows that most of the model's misclassifications occurred between adjacent motion rating levels. For example, for motions that experts labeled as "good" (moderately high) vs the next higher "perfect," the model sometimes confused them, but it almost never misjudged an extreme "miss" (failed) as "perfect" or vice versa. This pattern of errors is consistent with the continuous gradation of human motion quality—boundaries between adjacent levels are fuzzy and involve some subjective judgment, whereas the difference between extreme levels is clear and easy for the model to distinguish. The partial confusion among intermediate levels also reflects the subjectivity of manual scoring and the unclear boundaries of motion completion. Techniques such as introducing exponential smoothing and fine-tuning scoring thresholds could further improve the consistency and robustness of the model's judgments.

Emotion Recognition Module Performance

On the constructed offline test set, the emotion recognition module correctly classified 1135 of 1400 (81.07%) samples,

XSL•FO
RenderX

with a macro-F_1 of 81.02%, indicating high emotion classification performance. Per-class precision, recall, and F_1 scores, along with the confusion matrix, are provided in Figures S3 and S4 in Multimedia Appendix 1. The confusion matrix shows that certain negative emotions (such as anger, disgust, fear, and sadness) have some degree of mutual confusion, meaning the model sometimes misclassifies 1 negative emotion as another. This error distribution aligns with the fact that human facial expressions are objectively difficult to distinguish for subtle negative emotions. The main causes of misclassification include (1) some participants exhibited only slight facial changes when experiencing negative emotions, or there were brief occlusions of expressions during data capture; and (2) individual differences in facial muscle movement patterns cause significant intraclass variation for the same emotion category. Despite these challenges, the model still achieved good accuracy for most emotions, especially neutral and positive emotions being recognized reliably. In summary, the emotion recognition module attained high accuracy in offline annotated tests, and its multimodal fusion approach effectively reduced the interference of physical exercise on emotion judgments, providing reliable and continuous emotional state input for subsequent difficulty adaptation.

Compared to emotion recognition accuracies reported in the literature on public datasets (eg, Database for Emotion Analysis Using Physiological Signals [DEAP], typically in the 60%-85% range), our system's performance is at the higher end for this field.

User Experience and Training Effect Results

Subjective Evaluation Results

Statistical analysis of the posttraining questionnaire data (independent-samples *t* test) showed that the adaptive group outperformed the fixed difficulty group on all subjective evaluation dimensions, with differences being statistically significant. Table 2 summarizes the 2 groups' average scores and significance levels for the NASA-TLX, PANAS, and SUS measures. The results indicate that the adaptive group reported significantly lower subjective cognitive load than the fixed group (mean 45.0 vs 55.2, $P<.05$), suggesting that using the emotion-adaptive mechanism can reduce users' mental and physical burden during the training task. In terms of emotional experience, the adaptive group's positive-affect score was significantly higher and its negative-affect score significantly lower than the fixed group ($P<.05$ for both), indicating that the adaptive mechanism kept users in a more positively valenced emotional state after training and reduced the occurrence of negative emotions (such as frustration and tension). Additionally, the adaptive group's system usability satisfaction (SUS score) was clearly better than the fixed group's (mean 80.5, SD 6.4 vs 69.8, SD 7.5, $P<.05$), meaning users found the adaptive system easier to use, were more willing to continue using it, and gave higher overall experience ratings.

Participants' subjective feedback further supports the questionnaire results. Most participants in the adaptive group noted in interviews that when they felt fatigue or frustration during training, the system detected it and automatically lowered the game difficulty, making them feel less pressure and more willing to complete the entire training session. In contrast, some participants in the fixed difficulty group reported that since the difficulty never changed, when prolonged play caused fatigue or frustration, there was no relief, making them more inclined to quit. These subjective evaluations and feedback indicate that the adaptive difficulty mechanism effectively improved users' emotional experience and engagement, enhancing the system's friendliness and motivational effect.

Overall, the adaptive group was significantly better than the fixed group in aspects such as "more positive emotional experience," "higher concentration," and "greater overall satisfaction." This shows that the emotion-and-performance adaptive exergame outperformed the fixed-difficulty system on subjective usability in this study. We note that this advantage reflects the difference between adaptive and nonadaptive difficulty control and cannot be attributed specifically to the affective component.

Table 2. Comparison of subjective evaluation results between adaptive and fixed groups.

Measure	Adaptive group (mean, SD)	Fixed group (mean, SD)	*P* value
NASA-TLX[a]	45.0 (8.5)	55.2 (9.4)	<.05
PANAS[b] positive	34.6 (4.8)	28.5 (6.2)	<.05
PANAS negative	14.3 (5.1)	21.7 (7.0)	<.05
SUS[c]	80.5 (6.4)	69.8 (7.5)	<.05

[a]NASA-TLX: National Aeronautics and Space Administration Task Load Index.

[b]PANAS: Positive and Negative Affect Schedule.

[c]SUS: System Usability Scale.

Objective Performance Metrics

In addition to subjective evaluations, we objectively recorded users' motion and physiological data to quantitatively assess differences in training performance between the 2 groups. The main objective metrics include action accuracy (the percentage of correctly completed required movements), game duration (the length of time each user kept playing or the number of levels completed in each session), and HRV (measured by RMSSD, used to reflect physiological load during training).

XSL•FO
RenderX

Table 3 shows that the adaptive group outperformed the fixed group in terms of movement quality and training persistence. The adaptive group's average action accuracy reached 82.5% (SD 5.2), significantly higher than the fixed group's 74.8% (SD 7.9, $P<.05$). This suggests that by adaptively adjusting difficulty based on both motor performance and emotional state, users completed movements with higher accuracy—likely because difficulty adaptation reduced errors caused by tasks being overly difficult. At the same time, all adaptive group participants were able to persevere through the full 15-minute training (average 15.0, SD 0.0 minutes), whereas the fixed group's average play time was slightly shorter, and some participants did not persist to the end (average 13.5, SD 1.8 minutes; $P<.05$). This means the adaptive difficulty mechanism increased users' willingness and tolerance to complete the training, allowing them to finish the intended duration without excessive fatigue.

In terms of physiological load, comparing HRV metrics reveals differences in physiological responses between the 2 groups during training. The adaptive group's average HRV (RMSSD, in ms) during training was 34.6 ms, significantly higher than the fixed group's 24.3 ms ($P<.05$). Generally, higher HRV often indicates better autonomic balance and lower stress load [39,40]. This suggests that adaptive group users maintained a steadier HRV trend during training. In other words, the emotion-adaptive mechanism may have prevented excessive spikes or dramatic fluctuations in heart rate, helping them maintain moderate arousal, and thus it had a certain advantage in managing users' physiological load. In contrast, fixed difficulty may have caused some users to experience overly high tension or fatigue responses, leading to reduced HRV. These HRV values were obtained from wrist PPG during vigorous exercise and are interpreted as a within-study, between-group difference rather than a clinical-grade HRV measurement.

Application testing results indicate that compared to a fixed difficulty design, the adaptive system can significantly improve users' emotional experience and subjective satisfaction and reduce cognitive load; at the same time, it improves movement execution quality and persistence during training and shows advantages in users' physiological load. These findings are of great significance for designing user-centered, motion-based interventions: by introducing emotion sensing and adaptive regulation, exergame training systems can better maintain user engagement and exercise effectiveness. End-to-end latency averaged 125.0 (SD 17.24) ms, with 95% of frames processed within 156.3 ms, which meets the requirements of our system.

In summary, the experimental results of this study show that the proposed emotion recognition and motion evaluation modules achieved high accuracy and real-time performance in actual exercise scenarios, providing a solid foundation for implementing emotion-in-the-loop DDA. Meanwhile, compared to a traditional exergame with fixed difficulty, the multimodal affective computing-driven adaptive exergame demonstrated significant advantages in subjective aspects such as reducing user cognitive load, enhancing positive emotional experience, and improving system usability, and also clearly outperformed the fixed system in objective outcomes such as higher motion accuracy, longer training persistence, and better maintenance of moderate physiological load.

Table 3. Comparison of objective training outcomes between adaptive and fixed groups.

Metric	Adaptive group, mean (SD)	Fixed group, mean (SD)	*P* value
Action accuracy (%)	82.5 (5.2)	74.8 (7.9)	<.05
Duration (min)	15.0 (0.0)	13.5 (1.8)	<.05
HRV[a] RMSSD[b] (ms)	34.6 (9.5)	24.3 (7.8)	<.05

[a]HRV: heart rate variability.

[b]RMSSD: root mean square of successive differences.

Comparison With Prior Work

Table 4 summarizes how the present system compares with representative prior work along 4 dimensions: adaptation signal, exercise context, sensing modality, and validation outcomes. Against performance-driven adaptation in rehabilitation [19,21], the present loop adds a channel that registers disengagement in a player whose performance is still adequate—a state invisible to a performance-only controller. Against affect-driven adaptation in seated games [25,30], the present setting introduces physical exertion as a confound those systems do not face and adds a performance channel that separates task-difficulty mismatch from exertion-related discomfort. Against the affect-vs-performance comparison reported in prior work [22], the null result there is informative precisely because it concerns substitution: if each signal is systematically blind to states the other detects, neither should be expected to dominate, and the productive question becomes how to combine them—which is the question this study addresses.

XSL•FO
RenderX

Table 4. Comparison with prior work.

Study	Adaptation signal	Setting	Sensing	Reported outcome
Koenig et al [19,21]	Performance	Clinical gait or limb rehabilitation	Robotic or force	Improved participation vs fixed
Martin-Niedecken et al [14,15]	Performance (adaptive)	Whole-body "ExerCube"	Multisensor	Better training outcomes
Chanel and Lopes [25]	Affect (physiology)	Seated game	ECG[a], GSR[b], physiological DL[c]	Effective affective DDA[d]
Darzi and Novak [22]	Affect and performance	Competitive or VR[e]-ish	Affect + task	No significant difference between signals
This work	Affect and performance (joint)	Vigorous whole-body dance	RGB[f] camera + wrist PPG[g]	Benefit vs fixed across subjective, motor, and physiological

[a]ECG: electrocardiogram.

[b]GSR: galvanic skin response.

[c]DL: deep learning.

[d]DDA: dynamic difficulty adjustment.

[e]VR: virtual reality.

[f]RGB: red, green, and blue.

[g]PPG: photoplethysmography.

Discussion

Impact of Recognition Accuracy on User Experience

The accuracy of the user state recognition modules is critical to the effectiveness of the adaptive system. If emotion or action recognition is inaccurate, the game's difficulty adjustments may be based on incorrect information, thereby affecting the user experience. Overall, our multimodal recognition modules achieved high accuracy in the experiment, ensuring that in most cases the system can accurately perceive the player's state and respond appropriately—this is the foundation for the adaptive mechanism's positive effects. Our method has only small errors, mainly between adjacent levels, and rarely misclassifies a "very poor" movement as "perfect"; thus, it did not cause obvious bias in user feedback. In addition, the smoothing and cooldown strategies adopted by the system alleviated the negative impact of momentary recognition errors to some extent; this prevents occasional recognition mistakes from immediately triggering drastic difficulty swings, thereby maintaining consistency and stability in the game experience.

For emotion recognition, its accuracy determines whether difficulty adjustments truly align with the user's psychological state. Emotional state is a more latent signal than performance; if the system misreads the player's emotion, the adaptation may be counterproductive. For example, if the player actually feels frustrated but the system fails to detect it and continues to increase difficulty, it may exacerbate the frustration; conversely, if the player is actually feeling fine but is misclassified as negative, the system might lower the difficulty too early, potentially making the player bored. Our results show that multimodal fusion enabled the emotion recognition to correctly distinguish positive vs negative emotion in the vast majority of cases, allowing the system to promptly decrease game intensity when it detects the user's psychological stress increasing and to appropriately raise the challenge when the user is emotionally relaxed and happy. This reliable emotion-based adjustment keeps players always in the "neither too frustrated nor too relaxed" flow channel. Therefore, the high accuracy of the emotion recognition module is one of the important factors in the success of the adaptive system. Although some subtle emotion classifications are still not entirely accurate, our difficulty decision focuses more on the overall emotional valence (positive vs negative) rather than the specific emotion category. Thus, confusion among negative emotions has a relatively limited impact on the final adjustment decisions—for instance, misclassifying anger as fear still effectively signals a negative state and thus would result in a difficulty reduction to alleviate stress, so it does not substantively change the adaptation strategy. Overall, the current performance of the recognition modules is sufficient to support the basic needs of the emotion-in-the-loop adaptive framework, but further improving accuracy and robustness will help refine the system's understanding of user state, thereby optimizing the interactive experience on a more granular level.

For motion evaluation, errors are mainly concentrated in differentiating adjacent levels, stemming in part from the subjective fuzziness of human rating standards. Different experts may have slightly varying criteria for "good" vs "great" movements, so the model's judgment sometimes falls on different sides of such a boundary. However, these errors usually do not lead to serious consequences, because whether a movement is rated 1 level higher or lower, the feedback and rewards the system gives differ only in degree and do not fundamentally alter the training process. Moreover, in very rare cases, skeleton keypoint tracking errors or frame drops could cause motion scoring deviations—for example, a momentary camera tracking failure could cause the motion similarity to drop abnormally. But in our system design, we have mitigated sensor errors as much as possible through measures such as keypoint interpolation and anomaly detection. By contrast,

XSL•FO
RenderX

errors in emotion recognition largely stem from limitations of the multimodal signals themselves. First, microexpressions have such small magnitude that the signal-to-noise ratio for facial feature extraction is reduced. When a player only shows subtle hints of displeasure or fatigue in their expression, it is difficult even for the human eye to detect in time, so the model inevitably misses some instances. Second, during exercise, facial occlusion or expression distortion can occur—for example, intense movement can cause the face to turn away from the camera, or sweat may affect facial capture—which will cause the visual channel to temporarily fail. Although we use pose estimation to assist face ROI localization and fuse HRV at the decision level, emotion recognition is still affected when visual information is missing. Third, physiological signals are also greatly affected by exercise, which is a big challenge. Intense exercise causes a significant increase and fluctuation in heart rate, and this physiological response from physical exertion can present an HRV pattern similar to that caused by negative emotions and high arousal. If the model fails to sufficiently distinguish between the physiological effects of exercise load and emotional stress, it might misinterpret normal exercise-induced heart rate changes as emotional fluctuations. We have partially alleviated this issue by applying exercise intensity corrections to the HRV features, and purely exercise-induced emotion misclassifications were not common in our experiments. However, in extreme cases (such as a few participants with abnormally high exercise heart rates), the physiological channel can still produce “false alarms.” The above error sources indicate that the system still has room for improvement in handling edge cases and individual differences.

Strategies and Future Work

Our 3-level valence mapping cannot distinguish frustration from boredom, or exertion-related discomfort from task-driven negative affect; a dimensional model or a finer-grained categorical model may support adaptations this mapping cannot express. To address the above recognition errors and limitations, we have implemented and plan to implement a series of strategies for improvement. First, multimodal feature fusion will continue to be a key approach for improving accuracy. Our system has already validated the effectiveness of decision-level fusion; next, we can explore tighter fusion methods, for example, implementing feature-level or model-level fusion inside the model, using deep neural networks to automatically learn correlated features of visual and physiological signals. Introducing appropriate deep models, while ensuring real-time performance, could further improve the ability to distinguish microexpressions and physiological patterns.

Second, dynamic weighting mechanisms are another direction to enhance robustness. Currently, we use fixed-rule confidence weighting for fusion; in the future, fusion weights could be adjusted adaptively based on environmental and signal quality conditions. For example, if it is detected that the user’s face is occluded or expressions are subtle, the system could automatically increase the weight of physiological signals in emotion determination; conversely, if intense motion causes large fluctuations in the HRV signal, the system could temporarily rely more on facial expression cues. Through such adaptive weight adjustments, we can maximize the advantages of each modality and minimize the impact of single-channel failures on overall recognition.

Third, temporal smoothing and sequence modeling remain effective means to reduce transient errors. We already apply exponential smoothing at the output layer to filter high-frequency noise; going forward, we could consider more complex temporal models (such as using a sliding window or a recurrent neural network such as long short-term memory [LSTM]) to model the state sequence over time. This could not only smooth out noise but also use the evolving patterns of emotions and motions to improve accuracy. For example, recognizing the persistence of an emotion to avoid treating a brief facial twitch as an emotional change, or analyzing the trend of motion performance to judge whether an error is accidental or due to lack of skill.

Fourth, in terms of hardware and sensing, we could introduce more diverse signal sources to enhance the system’s perception abilities. For instance, adding physiological indicators such as skin conductance or respiration rate can provide additional emotional cues; incorporating voice or posture inputs can further enrich the emotion recognition’s input dimensions; adding inertial sensors for motion evaluation could improve the capture accuracy of fast 3D movements.

Apart from improving the recognition algorithms, our system also has room to expand in personalization and adaptivity. Currently, the model uses a unified set of recognition parameters for all users, without fully considering individual differences. In the future, personalized calibration could be used to improve recognition accuracy: for example, when a user first uses the system, we could collect their resting heart rate and neutral facial baseline as individual references, thereby using deviation-based thresholds for recognition (such as adjusting microexpression sensitivity based on the user’s typical expressiveness or calibrating emotion thresholds according to their baseline HRV when calm). Simultaneously, an online learning mechanism could be introduced, allowing the system to continuously learn the user’s behavior and emotion patterns during interaction and gradually optimize its model parameters. This continual adaptation would help the system track changes in the user’s state characteristics over time (for example, as the training progresses, the user’s fitness improves or emotional response patterns change), ensuring that recognition accuracy does not decline during long-term use.

Finally, from the user experience perspective, we should also consider feedback strategies when recognition errors occur. Even the best algorithm cannot be 100% accurate, so when the system is uncertain about the user’s state, it can adopt conservative adjustments or ask for confirmation to reduce the risk of misadjustment. For example, if the emotion model’s confidence is very low, the system could postpone changing the difficulty or provide a simple interface for the user to self-correct their emotional state. Such human-in-the-loop methods can tolerate algorithm errors to a certain extent, ensuring the user is not adversely affected by system misjudgments.

XSL•FO
RenderX

Acknowledgments

We thank Dr Jin Yan for technical assistance, assistance with data collection, and critical review of the manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (NSFC; 62566009) and the Scientific and Technological Innovation Talent Project of Guizhou High-level Talent Cultivation Support Program (KJLYRC-[2026]006). The funders had no involvement in this study's design, data collection, analysis, interpretation, or the writing of this paper.

Data Availability

All data generated or analyzed during this study are included in this published paper.

Generative AI Use

The authors declare that generative AI was not used in the preparation of this manuscript.

Authors' Contributions

Conceptualization: LZ, DL, GY, DZ
Data curation: HZ
Funding acquisition: DL
Investigation: LZ, DZ
Supervision: DL, GY
Visualization: HZ
Writing – original draft: LZ
Writing – review & editing: HZ, DL, GY, DZ, LH
All authors contributed to the study conception and design. All authors reviewed the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Per-class precision, recall, and F1 scores, and the confusion matrix data.
[DOCX File , 1286 KB-Multimedia Appendix 1]

References

1. Kohl HW, Craig CL, Lambert EV, Inoue S, Alkandari JR, Leetongin G, et al. Lancet Physical Activity Series Working Group. The pandemic of physical inactivity: global action for public health. Lancet. 2012;380(9838):294-305. [doi: 10.1016/S0140-6736(12)60898-8] [Medline: 22818941]
2. Huber T, Mertes S, Rangelova S, Flutura S, André E. Dynamic difficulty adjustment in virtual reality exergames through experience-driven procedural content generation. 2021. Presented at: 2021 IEEE Symposium Series on Computational Intelligence (SSCI); December 5-7, 2021:1-8; Orlando, FL. [doi: 10.1109/ssci50451.2021.9660086]
3. Peng W, Lin JH, Crouse J. Is playing exergames really exercising? A meta-analysis of energy expenditure in active video games. Cyberpsychol Behav Soc Netw. 2011;14(11):681-688. [doi: 10.1089/cyber.2010.0578] [Medline: 21668370]
4. Larche CJ, Dixon MJ. The relationship between the skill-challenge balance, game expertise, flow and the urge to keep playing complex mobile games. J Behav Addict. 2020;9(3):606-616. [FREE Full text] [doi: 10.1556/2006.2020.00070] [Medline: 33027060]
5. Csikszentmihalyi M, Abuhamdeh S, Nakamura J. Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi. Dordrecht. Springer Netherlands; 2014:227-238.
6. Chen J. Flow in games (and everything else). Commun ACM. 2007;50(4):31-34. [doi: 10.1145/1232743.1232769]
7. Hunicke R. The case for dynamic difficulty adjustment in games. 2005. Presented at: ACE '05: Proceedings of the 2005 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology; June 15-17, 2005:429-433; Valencia Spain. [doi: 10.1145/1178477.1178573]
8. Adams E. Fundamentals of Game Design. London. Pearson Education; 2014.
9. Mueller F, Vetere F, Gibbs M, Edge D, Agamanolis S. Balancing exertion experiences. 2012. Presented at: CHI '12: CHI Conference on Human Factors in Computing Systems; May 5-10, 2012:1853-1862; Austin, TX. [doi: 10.1145/2207676.2208322]

XSL•FO
RenderX

10. Gerling KM, Miller M, Mandryk RL, Birk MV, Smeddinck JD. Effects of balancing for physical abilities on player performance, experience and self-esteem in exergames. 2014. Presented at: CHI '14: CHI Conference on Human Factors in Computing Systems; April 26 to May 1, 2014:2201-2210; Toronto, Ontario, Canada. [doi: 10.1145/2556288.2556963]
11. Fisher N, Kulshreshth AK. Exploring dynamic difficulty adjustment methods for video games. Virtual Worlds. 2024;3(2):230-255. [doi: 10.3390/virtualworlds3020012]
12. Danousis M, Goumopoulos C, Fakis A. Exergames in the GAME2AWE platform with dynamic difficulty adjustment. 2022. Presented at: International Conference on Entertainment Computing; November 1-3, 2022:214-223; Bremen, Germany. [doi: 10.1007/978-3-031-20212-4_17]
13. Lyons EJ. Cultivating engagement and enjoyment in exergames using feedback, challenge, and rewards. Games Health J. 2015;4(1):12-18. [FREE Full text] [doi: 10.1089/g4h.2014.0072] [Medline: 26181675]
14. Martin-Niedecken AL, Mahrer A, Rogers K, de Bruin ED, Schättin A. "HIIT" the ExerCube: comparing the effectiveness of functional high-intensity interval training in conventional vs. exergame-based training. Front Comput Sci. 2020;2:33. [doi: 10.3389/fcomp.2020.00033]
15. Martin-Niedecken AL, Rogers K, Turmo Vidal L, Mekler ED, Márquez Segura E. Exercube vs. personal trainer: evaluating a holistic, immersive, and adaptive fitness game setup. 2019. Presented at: CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems; May 4-9, 2019:1-15; Glasgow, Scotland, UK. [doi: 10.1145/3290605.3300318]
16. Göbel S, Hardy S, Wendel V, Mehm F, Steinmetz R. Serious games for health: personalized exergames. 2010. Presented at: Proceedings of the 18th ACM International Conference on Multimedia; October 25-29, 2010:1663-1666; Firenze, Italy. [doi: 10.1145/1873951.1874316]
17. Hardy S, Göbel S, Gutjahr M, Wiemeyer J, Steinmetz R. Adaptation model for indoor exergames. Int J Comput Sci Sport. 2012;11(1):73-85. [FREE Full text]
18. Dill S, Müller PN, Caserman P, Göbel S, Antink CH, Tregel T. Sensing in exergames for efficacy and motion quality: scoping review of recent publications. JMIR Serious Games. 2024;12:e52153. [FREE Full text] [doi: 10.2196/52153] [Medline: 39499916]
19. Koenig A, Novak D, Omlin X, Pulfer M, Perreault E, Zimmerli L, et al. Real-time closed-loop control of cognitive load in neurological patients during robot-assisted gait training. IEEE Trans Neural Syst Rehabil Eng. 2011;19(4):453-464. [doi: 10.1109/TNSRE.2011.2160460] [Medline: 21827971]
20. Novak D, Mihelj M, Ziherl J, Olenšek A, Munih M. Psychophysiological measurements in a biocooperative feedback loop for upper extremity rehabilitation. IEEE Trans Neural Syst Rehabil Eng. Aug 2011;19(4):400-410. [doi: 10.1109/TNSRE.2011.2160357] [Medline: 21708507]
21. Koenig A, Omlin X, Bergmann J, Zimmerli L, Bolliger M, Müller F, et al. Controlling patient participation during robot-assisted gait training. J Neuroeng Rehabil. 2011;8:14. [FREE Full text] [doi: 10.1186/1743-0003-8-14] [Medline: 21429200]
22. Darzi A, Novak D. Automated affect classification and task difficulty adaptation in a competitive scenario based on physiological linkage: an exploratory study. Int J Hum Comput Stud. 2021;153:102673. [FREE Full text] [doi: 10.1016/j.ijhcs.2021.102673] [Medline: 34092990]
23. Izountar Y, Benbelkacem S, Otmane S, Khababa A, Masmoudi M, Zenati N. VR-PEER: a personalized exer-game platform based on emotion recognition. Electronics. 2022;11(3):455. [doi: 10.3390/electronics11030455]
24. Mitsis K, Zarkogianni K, Kalafatis E, Dalakleidi K, Jaafar A, Mourkousis G, et al. A multimodal approach for real time recognition of engagement towards adaptive serious games for health. Sensors (Basel). 2022;22(7):2472. [FREE Full text] [doi: 10.3390/s22072472] [Medline: 35408088]
25. Chanel G, Lopes P. User evaluation of affective dynamic difficulty adjustment based on physiological deep learning. Cham. Springer International Publishing; 2020. Presented at: International Conference on Human-Computer Interaction; July 19-24, 2020:3-23; Copenhagen, Denmark. [doi: 10.1007/978-3-030-50353-6_1]
26. Bazarevsky V, Grishchenko I, Raveendran K, Zhu T, Zhang F, Grundmann M. BlazePose: on-device real-time body pose tracking. arXiv. Preprint posted online on June 17, 2020. [doi: 10.48550/arXiv.2006.10204]
27. Ekman P. An argument for basic emotions. Cogn Emot. 1992;6(3-4):169-200. [doi: 10.1080/02699939208411068]
28. Toisoul A, Kossaifi J, Bulat A, Tzimiropoulos G, Pantic M. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. Nat Mach Intell. 2021;3(1):42-50. [doi: 10.1038/s42256-020-00280-0]
29. Ryumina E, Markitantov M, Axyonov A, Ryumin D, Dolgushin M, Dresvyanskiy D, et al. Team RAS in 10th ABAW competition: multimodal valence and arousal estimation approach. arXiv. Preprint posted online on March 13, 2026. [doi: 10.48550/arXiv.2603.13056]
30. Potts D, Broad Z, Sehgal T, Hartley J, O'Neill E, Jicol C, et al. Sweating the details: emotion recognition and the influence of physical exertion in virtual reality exergaming. 2024. Presented at: CHI '24: CHI Conference on Human Factors in Computing Systems; May 11-16, 2024:1-21; Honolulu, HI. [doi: 10.1145/3613904.3642611]
31. Barathi SC, Proulx M, O'Neill E, Lutteroth C. Affect recognition using psychophysiological correlates in high intensity VR exergaming. Association for Computing Machinery; 2020. Presented at: CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; April 25-30, 2020:1-15; Honolulu, HI. [doi: 10.1145/3313831.3376596]

XSL•FO
RenderX

32. Jadram N, Nishikawa Y, Sugaya M. Stress evaluation during low-speed mobility rides using ultra-short-term heart rate variability indices. 2026. Presented at: International Symposium on Affective Science and Engineering; March 16-18, 2026; Japan. [doi: 10.5057/isase.2026-c000055]
33. Xing Y, Kawauchi Y, Dilrukshi EC, Nishiyama Y, Nomura S. A preliminary study on psychophysiological effect of heartrate feedback relaxing music. Int J Affective Eng. 2023;22:129-136. [doi: 10.5057/ijae.tjske-d-22-00073]
34. Melo H, Rangel NB, Fialho GL, De Carvalho CR, Lin K, Walz R. Ultra-short heart rate variability reliability for cardiac autonomic tone assessment in severe traumatic brain injury. Brain Med. 2024;1(1):46-53. [FREE Full text] [doi: 10.61373/bm024r.0070]
35. Gao R, Miller CS, Lin BTW, Schwarz CW, Jones MLH. Signal quality assessment and reconstruction of PPG-derived signals for heart rate and variability estimation in in-vehicle applications: a comparative review and empirical validation. Sensors (Basel). 2025;25(24):7556. [FREE Full text] [doi: 10.3390/s25247556] [Medline: 41471550]
36. Naeini E, Sarhaddi F, Azimi I, Liljeberg P, Dutt N, Rahmani AM. A deep learning–based PPG quality assessment approach for heart rate and heart rate variability. ACM Trans Comput Healthcare. Nov 2, 2023;4(4):1-22. [doi: 10.1145/3616019]
37. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159-174. [doi: 10.2307/2529310] [Medline: 843571]
38. Mollahosseini A, Hasani B, Mahoor MH. AffectNet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans. Affective Comput. 2019;10(1):18-31. [doi: 10.1109/taffc.2017.2740923]
39. Shaffer F, Ginsberg JP. An overview of heart rate variability metrics and norms. Front Public Health. 2017;5:258. [FREE Full text] [doi: 10.3389/fpubh.2017.00258] [Medline: 29034226]
40. Thayer J, Ahs F, Fredrikson M, Sollers JJ, Wager TD. A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health. Neurosci Biobehav Rev. 2012;36(2):747-756. [doi: 10.1016/j.neubiorev.2011.11.009] [Medline: 22178086]

Abbreviations

AU: action unit
DDA: dynamic difficulty adjustment
DEAP: Database for Emotion Analysis Using Physiological Signals
FC: last fully connected layer
HF: high frequency
HRV: heart rate variability
ICA: independent component analysis
LF: low frequency
LF/HF: ratio of low-frequency to high-frequency power
LSTM: long short-term memory
NASA-TLX: National Aeronautics and Space Administration Task Load Index
PANAS: Positive and Negative Affect Schedule
pNN50: percentage of successive interbeat intervals differing by more than 50 ms
PPG: photoplethysmography
RGB: red, green, and blue
RMSSD: root mean square of successive differences
ROI: region of interest
RR: interbeat interval
SDNN: SD of Normal-to-Normal intervals
SUS: System Usability Scale

Edited by R Baranyi; submitted 16.May.2026; peer-reviewed by M Krishnamoorthi, V Novak, S Göbel; comments to author 08.Jul.2026; revised version received 14.Oct.2026; accepted 14.Oct.2026; published 05.Nov.2026

Please cite as:
Zhao L, Liu D, Zhang H, Yang G, Zheng D, He L
Emotion- and Performance-in-the-Loop: Multimodal Difficulty-Adaptive Exergame System—Controlled Between-Participants Study
JMIR Serious Games 2026;14:e101438
URL: https://games.jmir.org/2026/1/e101438
doi: 10.2196/101438
PMID:

XSL•FO
RenderX

©Liang Zhao, Dan Liu, Hanli Zhang, Guanci Yang, Donghua Zheng, Ling He. Originally published in JMIR Serious Games (https://games.jmir.org), 05.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Serious Games, is properly cited. The complete bibliographic information, a link to the original publication on https://games.jmir.org, as well as this copyright and license information must be included.

XSL•FO
RenderX