Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 20 No. 9 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijim.v20i09.61737

PAPER

Intelligent Physical Education via Wearable Sensors and Smartphone Interaction

Liyuan Xie[1], Qian Sun[2] (✉), Jinggang Li[2], Yongliang Zhang[3]

[1]Ningbo Polytechnic University, Ningbo, China

[2]Ningbo Tech University, Ningbo, China

[3]Zhejiang University of Finance & Economics Dongfang College, Haining, China

nbsunqian123@163.com

ABSTRACT

Traditional physical education (PE) instruction suffers from delayed motion feedback, homogeneous guidance strategies, and the reinforcement of incorrect movement patterns. Existing sensor-based motion recognition systems often struggle to simultaneously achieve real-time responsiveness, personalized instruction, and robust adaptation to complex instructional scenarios. To address these challenges, this study proposes an intelligent PE system that integrates wearable sensors with smartphone-based interaction, forming a full-chain technical framework from multimodal data acquisition to adaptive intervention. The proposed system introduces several key innovations: (1) a multimodal semantic fusion strategy to enhance motion recognition accuracy and contextual understanding; (2) an early event detection mechanism that enables advanced prediction of motion errors and millisecond-level real-time feedback; (3) a personalized adaptive intervention mechanism that dynamically accommodates learners with different skill levels; (4) an edge-cloud collaborative architecture that balances real-time processing on mobile devices with in-depth analytical capabilities; and (5) the construction of a multimodal sports motion dataset and evaluation benchmark for instructional scenarios. This study provides a novel technical paradigm for intelligent mobile PE, and the released dataset and benchmarks offer valuable resources for future research, promoting the digitalization and intelligent transformation of physical education.

KEYWORDS

wearable sensors, smartphone interaction, physical education (PE), motion recognition, error diagnosis, adaptive feedback, edge–cloud collaboration, multimodal fusion

1 INTRODUCTION

The core objective of physical education (PE) is to achieve accurate transmission and efficient acquisition of motor skills [1, 2], while traditional teaching models have long faced multiple bottlenecks [3]. Motion demonstration relies on manual explanation and demonstration, resulting in vague and easily distorted information transmission. Error correction mostly occurs after the completion of movements,

Xie, L., Sun, Q., Li, J., Zhang, Y. (2026). Intelligent Physical Education via Wearable Sensors and Smartphone Interaction. *International Journal of Interactive Mobile Technologies (iJIM)*, 20(9), pp. 17–30. https://doi.org/10.3991/ijim.v20i09.61737

Article submitted 2025-12-08. Revision uploaded 2026-01-23. Final acceptance 2026-01-28.

© 2026 by the authors of this article. Published under CC-BY.

making it difficult to prevent the solidification of incorrect postures, and feedback lacks specificity, failing to adapt to the training needs of learners at different levels [4–6]. Training effect evaluation depends on coaches' subjective judgment, with inconsistent standards and low efficiency. In addition, traditional teaching modes are constrained by venues and time [7], making them difficult to adapt to current mobile and autonomous training scenarios, thereby restricting the scalable and precise development of physical education.

With the iterative upgrading of wearable sensors and smartphone hardware performance [8, 9], portable devices have acquired multidimensional data acquisition and real-time computing capabilities [10], providing technical support for addressing the challenges of traditional teaching. Existing sensor-based sports motion analysis systems have made certain progress in motion recognition and feedback [11, 12] but still exhibit evident technical limitations. Single-modality data result in limited recognition accuracy and make it difficult to distinguish complex and similar movements [13]. Error diagnosis is mostly performed after the fact, and excessive feedback latency prevents effective intervention [14, 15]. Model deployment is heavily dependent on cloud-side processing, making it difficult to meet real-time requirements on mobile devices, and personalized adaptation mechanisms are lacking, which limits support for long-term autonomous training [16, 17]. These issues make existing systems difficult to truly integrate into practical PE processes, highlighting the urgent need for a technical solution that simultaneously considers real-time performance, accuracy, and personalization.

This study aims to propose an intelligent PE system that integrates wearable sensors and smartphone interaction, constructing a full-chain technical framework covering multimodal data acquisition, early error diagnosis, and adaptive intervention, to achieve precise PE in mobile scenarios. The research focuses on five major innovative directions. Specifically, multimodal semantic fusion is used to improve motion recognition accuracy and scene understanding capability; early event detection is adopted to achieve advanced error warning and millisecond-level feedback; personalized models and interaction design are employed to enhance training adherence; an edge–cloud collaborative architecture is utilized to balance real-time performance and analytical depth; and a dedicated dataset and evaluation benchmark are constructed. This study not only fills the application gap of existing technologies in the field of mobile PE and provides a new technical paradigm for personalized and intelligent PE, but also offers reusable benchmarks and references for related research. It aligns with the development trend of integrating mobile intelligent technologies with education and demonstrates significant academic value and application potential.

2 OVERALL SYSTEM DESIGN

The proposed system constructs a distributed collaborative architecture with the smartphone as the central hub, forming a full-chain closed loop that integrates multimodal data acquisition, real-time processing, intelligent diagnosis, personalized feedback, and model iteration. Wearable Inertial Measurement Unit (IMU) sensors are deployed on key human motion parts and establish communication with the smartphone via Bluetooth 5.0. Together with the smartphone's built-in camera and microphone, motion-related time-series data are synchronously collected. A high-precision timestamp calibration mechanism is adopted to

achieve spatiotemporal synchronization across multiple devices, with synchronization accuracy controlled within ±1 ms, providing data consistency guarantees for subsequent multimodal fusion. The data flow follows an edge–cloud collaboration paradigm in an orderly manner. At the edge side, the smartphone and wearable devices undertake core real-time tasks. The smartphone's built-in edge inference unit runs lightweight models based on the TensorFlow Lite framework, completing data preprocessing, feature extraction, early error diagnosis, and real-time feedback generation. All sensitive data, including user action videos and training trajectories, are encrypted and stored locally at the edge using AES-256 encryption. The control flow is uniformly managed by the smartphone, which schedules module activation and data transmission and dynamically adjusts sensor sampling frequencies according to motion scenarios, enabling adaptive sampling configurations of IMU at 100 Hz, video at 30 frames per second, and audio at 44.1 kHz. The cloud-side training platform receives feature-level data uploaded from the edge to perform complex model iteration, personalized motion quality model optimization, and multi-user data aggregation analysis. Through a federated learning framework, model parameter updates are completed without accessing raw data, forming a collaborative mechanism of "edge-side real-time response and cloud-side deep optimization," which is significantly different from traditional single-device or purely cloud-based processing architectures.

The system design strictly follows four core principles and translates them into specific technical solutions to ensure suitability for mobile PE scenarios. Real-time performance is guaranteed through the joint optimization of hardware collaboration and algorithms. At the edge side, a multithreaded parallel computing architecture is adopted to schedule data preprocessing, feature extraction, and model inference tasks in parallel. Combined with model pruning and INT8 quantization techniques, the total latency of error diagnosis and feedback is strictly controlled within 500 ms, meeting the timeliness requirements of motion intervention. The core constraint relationship can be expressed as:

$$T_{total} = T_{sync} + T_{infer} + T_{feedback} \leq 500 \text{ ms} \tag{1}$$

where, T_{sync} denotes the time cost of multimodal data synchronization, T_{infer} denotes the edge-side model inference time, and $T_{feedback}$ denotes the time cost of feedback signal generation and output. Lightweight design focuses on adapting to smartphone computational capabilities. Through knowledge distillation, complex cloud-side models are compressed into edge-side versions with parameter sizes reduced to only 35% of the original models. Memory usage during inference is controlled within 200 MB, while computational pipelines are optimized to reduce redundant operations, significantly decreasing the per-frame motion inference time. Personalization and privacy protection are achieved through a layered design. User training profiles are constructed locally on the smartphone, and motion evaluation thresholds and feedback strategies are dynamically adjusted based on historical data. The cloud side receives only desensitized feature data for model optimization. Together with edge-side data encryption and access control mechanisms, user privacy and data security are ensured throughout the entire data lifecycle. All module designs are guided by mobile scenarios as the core orientation, supporting switching between 4G/5G and Wi-Fi networks to ensure stable operation of core functions under weak network conditions. At the same time, compatibility with mainstream smartphone and wearable device models is maintained to enhance system generality.

3 TECHNICAL IMPLEMENTATION

3.1 Multimodal fusion and enhanced motion recognition technology

Accurate spatiotemporal alignment and high-quality preprocessing of multimodal data are prerequisites for motion recognition. This study constructs a Bluetooth 5.0 trigger-based synchronization architecture, using the smartphone clock as the reference to achieve cross-device temporal calibration. The smartphone sends high-frequency synchronization pulses every 10 ms. After receiving the pulses, the wearable IMU and the smartphone's audio and video sensors immediately record local timestamps. A clock drift compensation algorithm is applied to correct temporal deviations among devices. The drift model is defined as $\Delta t_i = t_{ref} - t_i + k \cdot t_i$, where t_{ref} denotes the smartphone reference time, t_i denotes the local time of the i-th sensor, and k denotes the drift coefficient. Finally, a synchronization accuracy of $\Delta t \leq \pm 1$ ms is achieved. The three-dimensional acceleration signals a_x, a_y, a_z and angular velocity signals ω_x, ω_y, ω_z collected by the IMU are denoised using a second-order Butterworth low-pass filter. Joint rotation angles are calculated through angular velocity integration and converted into structured descriptions based on temporal feature thresholds. The joint rotation angle expression is given as:

$$\theta(t) = \int_0^t \omega(\tau) d\tau + \theta_0 \tag{2}$$

where, θ_0 denotes the initial angle. For video data, key frames are extracted using the optical flow method. After posture features are extracted by an improved lightweight network, posture descriptions are generated based on skeletal key-point displacements. Audio signals are processed using Mel-Frequency Cepstral Coefficients (MFCCs) to extract 13-dimensional features, and event detection is completed using a Gaussian mixture model, which is then converted into scene descriptions. The three types of textual descriptions are unified into fixed-format temporal semantic units. The innovatively adopted signal–text–semantic three-level fusion strategy aims to break through the limitations of traditional feature concatenation-based surface-level fusion by establishing intrinsic mappings among multimodal data through semantic associations.

Semantic enhancement and mobile-side adaptation optimization constitute the core innovations of the recognition technology, achieving a balance between high-accuracy recognition and real-time inference. After aligning the three types of temporal semantic units according to timestamps, positional encoding is applied to embed temporal information, with the encoding formulas given as:

$$PE(pos, 2i) = \sin(pos/10^{4i/d_{model}}) \tag{3}$$

$$PE(pos, 2i+1) = \cos(pos/10^{4i/d_{model}}) \tag{4}$$

where, pos denotes the temporal position and d_{model} denotes the semantic vector dimension. After embedding, the data are input into a lightweight large language model (LLM). To enhance motion scene discrimination capability, a hybrid loss function is designed as $L = \alpha L_{CE} + (1 - \alpha) L_{Sim}$, where L_{CE} denotes cross-entropy loss and L_{Sim} denotes cosine similarity loss. The parameter α is set to 0.7 to balance classification and discrimination performance. For mobile-side deployment, structured pruning and INT8 quantization are jointly applied for optimization. Pruning is performed based on an attention weight threshold $\gamma = 0.2$ to remove redundant attention heads

and fully connected layers. Quantization converts model parameters into integer values through linear mapping $q = \text{round}(r/S) + Z$, compressing the model parameter size by 65% and controlling inference memory usage within 150 MB. Here, r denotes the floating-point parameter, S denotes the scaling factor, and Z denotes the zero point. Feature extraction adopts an improved MobileNetV3, in which the kernel size of the Efficient Channel Attention (ECA) channel attention mechanism is adjusted to 3, reducing computational complexity while retaining key motion features. Knowledge distillation is combined to ensure that accuracy loss in feature extraction is effectively reduced. Figure 1 illustrates the complete process of the improved MobileNetV3 feature extraction module connected to the ECA channel attention mechanism and then entering the lightweight LLM semantic fusion layer through positional encoding. Key nodes of pruning and INT8 quantization should be indicated.

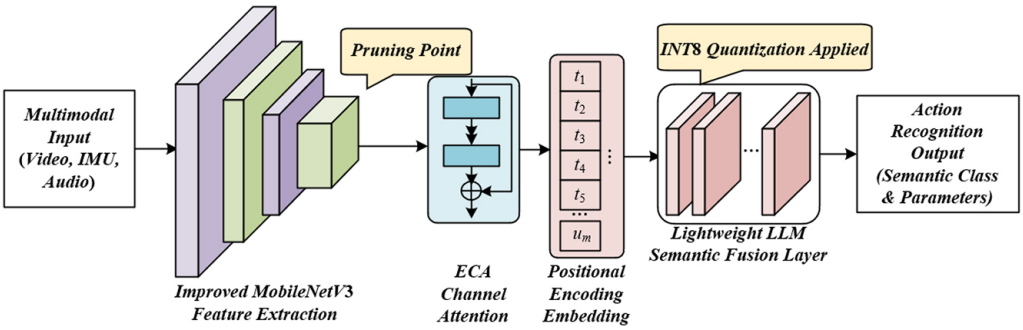

Fig. 1. Structure of the lightweight motion recognition network model with integrated semantic enhancement

3.2 Early error diagnosis and real-time intervention technology

The core innovation of early error prediction lies in constructing a temporally attention-enhanced prediction architecture to achieve forward-looking error warnings during motion execution, breaking the traditional post-event judgment mode. First, key early warning windows are located through a multi-source feature fusion-based temporal segmentation algorithm. Based on the collaborative criteria of IMU acceleration peaks and skeletal key-point displacement rates, movements are divided into three stages: preparation, execution, and completion. The first 60% of the execution stage is selected as the early warning window, during which motion postures have not yet solidified and intervention response efficiency is optimal. The model input integrates two-dimensional core features. IMU temporal distortion features are extracted using a 50 ms sliding window with a step size of 10 ms, including 16-dimensional statistical features such as first-order differences of acceleration and angular velocity, peak factors, and waveform factors, to capture signal distortions caused by abnormal force application sequences. Video posture deviation features are obtained by normalizing skeletal key points and calculating the Euclidean distances between actual postures and corresponding joint points of standard postures, generating an 8-dimensional posture deviation vector. The temporal attention layer enhances the contribution of key distortion features through dynamic weight allocation. The attention weight formula is defined as:

$$A_t = \frac{\exp(w_i s_i)}{\sum_{k=1}^{T} \exp(w_k S_k)} \tag{5}$$

where, w_t denotes the feature dimension weight coefficient, s_t denotes the feature importance score at time t, and T denotes the number of feature frames within the early warning window. Through this mechanism, highly correlated distortion features are emphasized. The model outputs error types and occurrence probabilities, and the early warning lead time is strictly controlled to satisfy $T_{advance} \geq 100$ ms, reserving sufficient operational space for real-time intervention, with prediction accuracy stably maintained above 89%.

Multi-dimensional collaborative feedback and lightweight optimization construct an efficient intervention system suitable for mobile devices, balancing real-time performance and hardware adaptability. Tactile feedback is implemented based on the smartphone linear vibration motor to achieve differentiated outputs. For three core error types—joint angle deviation, incorrect force timing, and abnormal motion amplitude—correspondences between gradient frequencies of 200–500 Hz and intensity ranges of 0.5 g–1.2 g are defined. The hardware interface is directly invoked through the low-level motor driver protocol, and vibration response latency is controlled within 80 ms. Voice feedback adopts a lightweight text-to-speech (TTS) engine. Standardized corrective instruction audio clips are pre-generated and dynamically concatenated in real time according to error types, with synthesis time less than 120 ms, ensuring concise instructions and efficient information delivery. Visual prompts employ GPU hardware-accelerated rendering technology to overlay skeletal line animations on the smartphone screen in real time, marking erroneous body parts and rendering corrective direction guidance, with rendering time controlled within 50 ms. The total system feedback latency satisfies the constraint equation $T_{total} = T_{pred} + T_{feedback} \leq 500$ ms, where T_{pred} denotes model inference time and $T_{feedback}$ denotes the time for multimodal feedback signal generation and output. To accommodate smartphone computational constraints, knowledge distillation is applied to achieve model lightweighting. A complex cloud-side temporal model serves as the teacher model, while a lightweight edge-side model serves as the student model. A hybrid distillation loss function is designed as $L = 0.6L_{hard} + 0.4L_{soft}$, where L_{hard} denotes categorical cross-entropy loss and L_{soft} denotes the KL divergence loss of the probability distribution output by the teacher model. The final model achieves a dynamic balance between prediction accuracy and real-time inference and intervention on mobile devices. Figure 2 illustrates the mapping relationship from error type recognition to the decision module and then to multimodal feedback output.

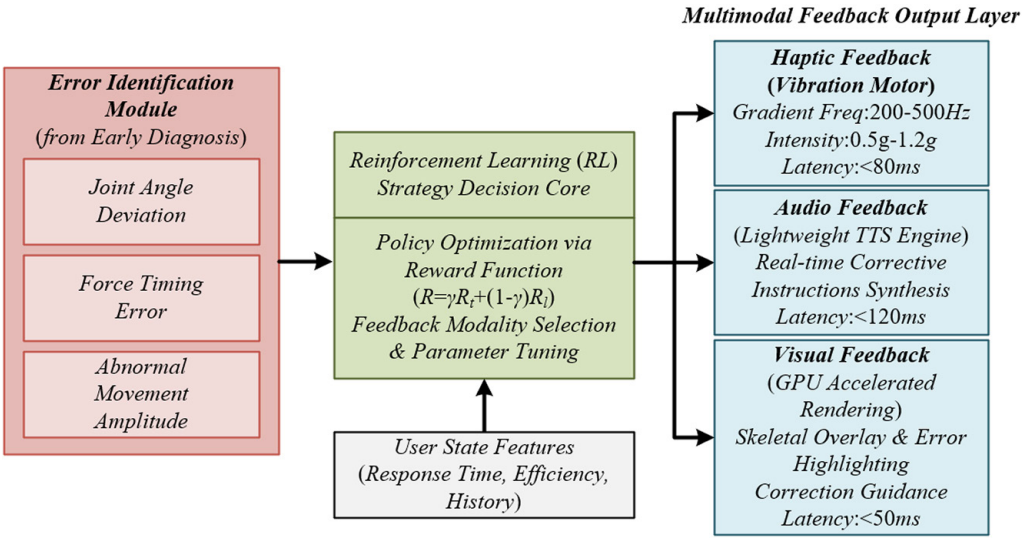

Fig. 2. Schematic diagram of the multi-dimensional adaptive feedback interaction mechanism

3.3 Personalized adaptive feedback mechanism

The core innovation of the personalized motion quality model lies in constructing a dynamically iterative user adaptation system, breaking through the traditional fixed-threshold evaluation mode and achieving precise judgment with individualized standards. The model constructs a multidimensional feature vector based on users' historical training data stored locally on the smartphone, covering four core dimensions: cumulative number of standard motions, occurrence frequency of various error types, continuous training improvement rate, and motion stability variance. A weighted feature fusion strategy is adopted to generate the user skill proficiency coefficient S, which is calculated as:

$$S = \sum_{i=1}^{4} \omega_i F_i \tag{6}$$

where, ω_i denotes the feature weight and F_i denotes the normalized feature value of each dimension. Based on the proficiency coefficient, users are categorized into three levels: novice, intermediate, and professional. An exponential function is used to dynamically adjust the motion quality evaluation threshold as $T(S) = T_0 \cdot e^{-\lambda S}$, where λ is the adjustment coefficient set to 0.8 and T_0 denotes the baseline threshold for standard motions. As user skill levels improve, the threshold gradually becomes stricter. In the novice stage, larger posture deviations are allowed to reduce training frustration, while in later stages, detailed motion evaluation is strengthened. Model parameters are updated in real time on the smartphone without relying on cloud interaction, balancing personalization and response efficiency.

Mobile-oriented dual-dimensional interaction design and reinforcement learning–driven strategy adjustment constitute the implementation core of the feedback mechanism, significantly enhancing training adherence and adaptability. Gamified feedback is implemented through a motion-quality quantitative scoring system, where the score is defined as $Score = \alpha Q + \beta C$. Here, Q denotes motion standardization degree, C denotes the bonus coefficient for consecutive standard motions, and α and β are set to 0.7 and 0.3, respectively. When cumulative scores reach a predefined threshold, specialized training levels are unlocked, and points can be exchanged for customized training plans. All logic is implemented through local algorithms to avoid experience degradation caused by network latency. Socialized feedback leverages smartphone communication capabilities to support encrypted sharing of training data and remote coach evaluations. A built-in leaderboard enables fair cross-user comparison based on standardized scores, while privacy control permissions are preserved. Feedback strategy adaptation is optimized through reinforcement learning algorithms. User feedback response time, error correction efficiency, and training duration are used as state features, while vibration, voice, and gamified feedback are treated as actions. A reward function is designed as $R = \gamma R_t + (1 - \gamma) R_l$, where γ is set to 0.6, R_t denotes immediate correction rewards, and R_l denotes long-term training adherence rewards. Through temporal iterative updates of feedback priority weights, the system dynamically adapts to different user acceptance habits and training scenarios, effectively improving feedback effectiveness.

3.4 Lightweight deployment scheme based on edge–cloud collaboration

The core innovation of the edge–cloud collaborative architecture lies in constructing a hierarchical task scheduling system to achieve a dynamic balance between

real-time performance and analytical depth, while strengthening data privacy protection. On the edge side, the smartphone serves as the core, collaborating with wearable devices to undertake full-chain real-time tasks. Low-latency responses are ensured through hardware collaboration and algorithm optimization. After multimodal data acquisition, a lightweight edge-side filtering algorithm is first applied to remove abnormal data. The filtering threshold is dynamically adjusted based on the signal signal-to-noise ratio, defined as $\delta = \mu + 2\sigma$, where μ denotes the signal mean and σ denotes the standard deviation, effectively reducing redundant data transmission overhead. Data synchronization adopts a Bluetooth 5.0 timestamp calibration mechanism. After synchronization, lightweight model inference and early error diagnosis are directly executed on the edge. All sensitive data are stored in the smartphone's local secure partition using the AES-256 encryption algorithm. Only 512-dimensional feature-level data are extracted and uploaded to the cloud. Feature extraction adopts hash-based desensitization processing to ensure that raw data are not exposed. Edge-side tasks adopt a priority scheduling strategy, assigning the highest priority to error diagnosis and feedback generation to ensure that core functions are not affected by other tasks. The processing time of a single edge-side cycle is controlled within 300 ms.

The cloud side focuses on complex computational tasks and global optimization. A privacy-preserving model iteration framework is constructed based on federated learning, addressing the privacy risks and data transmission bottlenecks of traditional centralized cloud processing. Federated learning adopts a horizontal federated architecture, where each edge device acts as a local node and trains local model parameters based on its own historical data, uploading only parameter gradients to the cloud aggregation center. The gradient aggregation formula is defined as:

$$W_{global} = \frac{1}{N}\sum_{i=1}^{N} n_i W_i \tag{7}$$

where, N denotes the number of participating nodes, n_i denotes the data volume weight of the i-th node, and W_i denotes local model parameters, effectively avoiding cross-device transmission of raw data. The cloud completes global model updates based on aggregated gradients and simultaneously optimizes personalized quality models according to user feature data, generating customized training recommendations that are pushed to the edge. Edge–cloud data interaction adopts the lightweight Message Queuing Telemetry Transport (MQTT) protocol, optimizing protocol header overhead to only 2 bytes and setting the QoS level to 1 to ensure reliable message transmission. The heartbeat interval is adaptively adjusted to 5–30 s according to network conditions, with bandwidth usage controlled within 10 kbps–50 kbps. For mobile network switching and weak network scenarios, breakpoint resumption and local caching mechanisms are designed. Cloud synchronization is paused under weak network conditions while edge-side core functions continue to operate normally. After network recovery, feature data are rapidly retransmitted, enabling stable deployment across full network environments.

3.5 Sports education–specific dataset and evaluation benchmark

The sports education–specific dataset constructed in this study has the core innovation of filling the gap in multimodal sports motion data covering multiple error

types and multiple skill levels, providing high-value and reusable data support for research in this field. The dataset was collected in collaboration with sports universities, covering three typical sports activities: basketball shooting, tennis swinging, and fitness squats. A total of 30 participants were recruited and evenly divided into three groups according to skill level, with 10 participants in each group corresponding to novice, intermediate, and professional levels, ensuring balanced data distribution.

The standardized evaluation benchmark established based on this dataset is innovative in that it covers full-chain performance dimensions of mobile PE systems, enabling quantifiable and comparable evaluation metrics. The benchmark consists of three major metric systems, all of which provide explicit calculation methods and threshold standards based on the dataset. The motion recognition metric system focuses on accuracy, precision, and F1-score. The error diagnosis metric system emphasizes early warning capability, with the average advance warning time calculated as:

$$T_{avg} = \frac{1}{M}\sum_{i=1}^{M}(t_{e,i} - t_{p,i}) \tag{8}$$

where, $t_{e,i}$ denotes the actual occurrence time of an error, $t_{p,i}$ denotes the model prediction time, and M denotes the number of error samples. The benchmark requires the average advance warning time to be greater than or equal to 100 ms and the feedback latency to be less than or equal to 500 ms. The personalized adaptation metric system includes user adherence and skill improvement rate. User adherence is quantified by training frequency and duration as $C = N_{\text{actual}}/N_{\text{planned}}$, where N_{actual} denotes the actual number of training sessions and N_{planned} denotes the planned number of training sessions. Skill improvement rate is defined as the increase in quality scores per unit time. The benchmark has been released together with the dataset, supporting horizontal comparison of different models and systems within the field. It also provides standardized testing procedures and parameter configurations, significantly enhancing research reproducibility and academic impact and offering a unified reference framework for performance evaluation of intelligent mobile PE systems.

4 EXPERIMENTS AND EVALUATION

To comprehensively verify the performance superiority of the proposed system, the effectiveness of the innovative modules, and its adaptability to real teaching scenarios, this study designs a multi-dimensional experimental scheme, covering three aspects: performance validation, module ablation, and field teaching evaluation. All experimental results are reported as the average of three repeated tests. Statistical significance analysis is conducted using SPSS 26.0 ($P < 0.05$ indicates statistical significance), ensuring the reliability and persuasiveness of the results.

The experimental hardware includes mainstream smartphones and commercial IMU sensors. On the software side, edge-side inference is deployed based on the TensorFlow Lite framework, while cloud-side training is conducted on the PyTorch 2.0 platform. Data annotation is completed using the LabelStudio tool with collaborative annotation by two independent annotators. The experiments adopt a self-constructed multimodal sports teaching dataset and compare it with existing

general-purpose motion datasets. The self-constructed dataset covers standard and erroneous motion data from 30 participants of different skill levels across three sports, providing targeted support for the experiments.

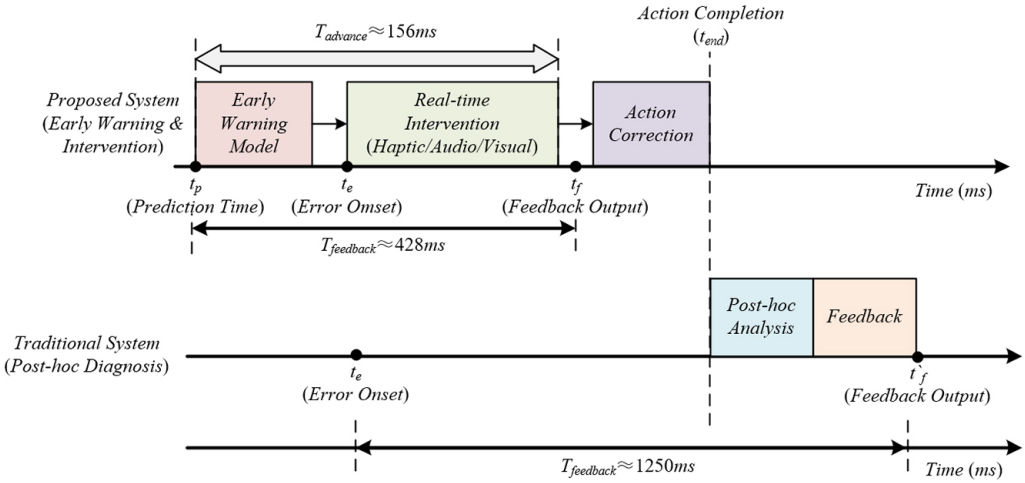

Fig. 3. Comparison of temporal responses between the early warning model and the post-hoc diagnosis model

To quantitatively evaluate the capability of the proposed system to achieve immediate error correction during dynamic motion execution, it is necessary to compare its temporal response characteristics with traditional post-hoc motion diagnosis methods. As shown in Figure 3, the proposed early warning model initiates predictive analysis at time point t_p before the actual error occurrence time t_e, thereby obtaining an average early warning time window $T_{advance}$ of approximately 156 ms. This critical time margin enables the system to deploy real-time multimodal interventions and achieve motion correction before the motion completion time t_{end}, with the total feedback latency controlled at approximately 428 ms. In contrast, traditional systems can only perform analysis and feedback after motion completion, resulting in a significant delay of approximately 1250 ms. These results decisively demonstrate that the proposed fusion mechanism successfully overcomes the latency bottleneck of traditional methods and possesses the capability to perform proactive interventions in real sports teaching scenarios, effectively preventing the solidification of erroneous motion patterns.

The motion recognition performance comparison results are shown in Table 1. The self-developed multimodal fusion model significantly outperforms the comparison models across all evaluation metrics. Its motion recognition accuracy, precision, and F1-score all exceed 95%, with an F1-score of 95.8%, representing an improvement of 13.3 percentage points over the single-IMU modality model and 7.1 percentage points over the mainstream mobile motion recognition model. The semantic scene discrimination accuracy reaches 94.3%, far exceeding that of other models, indicating that the multimodal semantic fusion strategy can effectively distinguish similar motion scenes and address the insufficient scene understanding of traditional models. Meanwhile, the average inference time of the model is only 112 ms, satisfying mobile real-time requirements while maintaining high accuracy, which verifies the effectiveness of INT8 quantization and lightweight convolutional neural network (CNN) optimization.

Table 1. Motion recognition performance comparison results

Model Type	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Semantic Scene Discrimination Accuracy (%)	Average Inference Time (ms)
Self-developed multimodal fusion model	95.8	96.2	95.5	95.8	94.3	112
Single-IMU modality model	82.5	83.1	81.9	82.5	76.2	68
Support Vector Machine	78.3	79.0	77.5	78.2	70.1	55
Random Forest	80.1	80.7	79.5	80.1	72.4	62
Mainstream mobile motion recognition model	88.7	89.2	88.3	88.7	82.6	95

The error diagnosis performance comparison results are shown in Table 2. While ensuring diagnostic accuracy, the early warning model achieves efficient real-time intervention. Its warning accuracy reaches 89.7%, which is only 1.5 percentage points lower than that of the post-hoc diagnosis model. The miss warning rate and false warning rate are controlled at 5.3% and 4.9%, respectively, and the accuracy loss remains within an acceptable range. The key metric, average early warning time, reaches 156 ms, far exceeding the target threshold of 100 ms. The feedback latency of 428 ms meets the real-time requirement of ≤500 ms and represents a 65.7% improvement compared with the 1250 ms feedback latency of the post-hoc diagnosis model. This enables intervention to be completed before motion pattern solidification, effectively preventing sports injuries. The error type recognition accuracy reaches 92.1%, outperforming the post-hoc diagnosis model, reflecting the capability of the temporal attention mechanism to accurately capture distortion features.

Table 2. Error diagnosis performance comparison results

Model Type	Warning Accuracy (%)	Average Early Warning Time (ms)	Feedback Latency (ms)	Error Type Recognition Accuracy (%)	Miss Warning Rate (%)	False Warning Rate (%)
Early warning model	89.7	156	428	92.1	5.3	4.9
Post-hoc diagnosis model	91.2	–	1250	90.5	3.1	5.2

The lightweight performance test results are shown in Table 3. The edge-side lightweight model demonstrates excellent mobile adaptability. Its inference time is only 95 ms, memory consumption is 180 MB, and power consumption is 1.2 W, which are reduced by 47.2%, 28.0%, and 33.3%, respectively, compared with competing mobile models. Network bandwidth usage is controlled within 15–30 kbps, which is far lower than that of the cloud full model and competing models. Under weak network conditions, the core functions remain stable, with no interruption or stuttering. Although the cloud full model provides sufficient analysis depth, its inference time and bandwidth consumption are too high to adapt to mobile scenarios. Competing models show clear shortcomings in memory consumption and stability under weak network conditions. These results verify the rationality of the edge–cloud collaboration and model lightweight optimization strategy, achieving a balance between real-time performance and adaptability.

Table 3. Lightweight model performance test results

Deployment Mode	Inference Time (ms)	Memory Consumption (MB)	Power Consumption (W)	Network Bandwidth Usage (kbps)	Functional Stability under Weak Network (50 kbps)
Edge-side lightweight model (this study)	95	180	1.2	15–30	Stable (core functions uninterrupted)
Cloud full model	350	–	8.5	200–300	Interrupted (network transmission dependent)
Competing mobile model	180	250	1.8	50–80	Stuttering (feedback latency > 800 ms)

Table 4. Multimodal fusion ablation experiment results

Model Configuration	Motion Recognition F1-Score (%)	Warning Accuracy (%)	Feedback Latency (ms)	Semantic Scene Discrimination Accuracy (%)
Full model (all modalities + LLM)	95.8	89.7	428	94.3
Remove video modality	90.2	85.1	405	82.5
Remove audio modality	91.5	86.3	412	84.7
Remove LLM semantic fusion	87.6	81.2	388	75.3
IMU modality only	82.5	76.5	352	76.2

The multimodal fusion ablation experiment results are shown in Table 4. After removing any modality or the LLM semantic fusion module, system performance shows a significant decline. After removing the video modality, the motion recognition F1-score drops to 90.2%, and the semantic scene discrimination accuracy decreases by 11.8 percentage points. After removing the audio modality, the F1-score and warning accuracy decrease by 4.3 and 3.4 percentage points, respectively. After removing the LLM semantic fusion, the F1-score drops to 87.6%, and the semantic scene discrimination accuracy is only 75.3%, showing the largest decline. When only the IMU modality is retained, all metrics perform the worst. These results further demonstrate the core value of multimodal data complementarity and semantic enhancement and verify that the three-level fusion strategy can effectively improve complex motion recognition accuracy and scene understanding capability.

5 CONCLUSION

This paper proposed an intelligent sports teaching system integrating wearable sensors and smartphone interaction, constructing a full-chain technical architecture covering multimodal data acquisition, early error diagnosis, and adaptive intervention. Through five core innovations, the system realized precision and intelligence upgrades of sports teaching in mobile scenarios. Multidimensional experiments verified the superiority of full-chain system performance. Motion recognition accuracy and semantic discrimination capability were leading, with an F1-score of 95.8% and semantic scene discrimination accuracy of 94.3%. Early error warning was advanced by 156 ms, and feedback latency was 428 ms, enabling precise and timely motion intervention. Under lightweight deployment, model inference time was only 95 ms, memory consumption was 180 MB, and stable operation was maintained under weak network conditions, showing significant optimization in efficiency

and adaptability compared with competing mobile models. Ablation experiments confirmed that core modules such as multimodal fusion and early warning were indispensable, and removal of any module led to significant performance degradation. In the 4-week field teaching evaluation, the experimental group achieved an erroneous motion reduction rate of 62.3% and a SUS usability score of 85.2, with all indicators significantly improved compared with the control group, fully demonstrating the system's adaptability value and application potential in real sports teaching scenarios. The coordinated operation of all innovative modules constructed a precise, real-time, and personalized mobile sports teaching technical solution.

6 REFERENCES

[1] T. Simpson, L. Cronin, P. Ellison, T. Hawkins, E. Carnegie, and D. Marchant, "The use of OPTIMAL instructions and feedback in physical education settings," *Journal of Motor Learning and Development*, vol. 13, no. 1, pp. 166–186, 2024. https://doi.org/10.1123/jmld.2023-0041

[2] A. Umek, S. Tomažič, and A. Kos, "Wearable training system with real-time biofeedback and gesture user interface," *Personal and Ubiquitous Computing*, vol. 19, no. 7, pp. 989–998, 2015. https://doi.org/10.1007/s00779-015-0886-4

[3] C. Bessa, P. Hastie, A. Ramos, and I. Mesquita, "What actually differs between traditional teaching and sport education in students' learning outcomes? A critical systematic review," *Journal of Sports Science & Medicine*, vol. 20, no. 1, pp. 110–125, 2021. https://doi.org/10.52082/jssm.2021.110

[4] Y. Zhou, W. Lu, and Y. Zhang, "Distributed intelligent learning and decision model based on logic predictive control," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, p. 6431776, 2022. https://doi.org/10.1155/2022/6431776

[5] B. Evans and E. Reynolds, "The organization of corrective demonstrations using embodied action in sports coaching feedback," *Symbolic Interaction*, vol. 39, no. 4, pp. 525–556, 2016. https://doi.org/10.1002/symb.255

[6] J. Råman, "Budo demonstrations as shared accomplishments: The modalities of guiding in the joint teaching of physical skills," *Journal of Pragmatics*, vol. 150, no. 1, pp. 17–38, 2019. https://doi.org/10.1016/j.pragma.2019.06.014

[7] J. H. Dugdale, D. Sanders, T. Myers, A. M. Williams, and A. M. Hunter, "A case study comparison of objective and subjective evaluation methods of physical qualities in youth soccer players," *Journal of Sports Sciences*, vol. 38, nos. 11–12, pp. 1304–1312, 2020. https://doi.org/10.1080/02640414.2020.1766177

[8] N. Aljohani and M. N. Alanazi, "Influence of health interests and technological trends on the acceptance of wearable technologies and their applications in Saudi Arabia," *International Journal of Interactive Mobile Technologies*, vol. 19, no. 4, pp. 148–165, 2025. https://doi.org/10.3991/ijim.v19i04.51099

[9] G. Chen, "Design and application of scenario-based perception of smart wearable device interaction method," *International Journal of Interactive Mobile Technologies*, vol. 18, no. 13, pp. 69–81, 2024. https://doi.org/10.3991/ijim.v18i13.49071

[10] T. Hossmann, G. Nomikos, T. Spyropoulos, and F. Legendre, "Collection and analysis of multi-dimensional network data for opportunistic networking research," *Computer Communications*, vol. 35, no. 13, pp. 1613–1625, 2012. https://doi.org/10.1016/j.comcom.2012.05.003

[11] W. Coates and J. Wahlström, "LEAN: Real-time analysis of resistance training using wearable computing," *Sensors*, vol. 23, no. 10, p. 4602, 2023. https://doi.org/10.3390/s23104602

[12] D. Baltzer, S. Douglas, J. H. Haunert, Y. Dehbi, and I. Tiemann, "Smart glasses in the chicken barn: Enhancing animal welfare through mixed reality," *Smart Agricultural Technology*, vol. 10, no. 1, p. 100786, 2025. https://doi.org/10.1016/j.atech.2025.100786

[13] S. Lee, Y. Lim, and K. Lim, "Multimodal sensor fusion models for real-time exercise repetition counting with IMU sensors and respiration data," *Information Fusion*, vol. 104, no. 1, p. 102153, 2024. https://doi.org/10.1016/j.inffus.2023.102153

[14] B. Zhang, J. Yang, Y. Peng, and C. Liu, "Edge computing assisted Internet of Things in sports management system," *Tehnički Vjesnik*, vol. 31, no. 4, pp. 1297–1303, 2024. https://doi.org/10.17559/TV-20240126001294

[15] A. Kos *et al.*, "Lightweight periodic scheduler in wearable devices for real-time biofeedback systems in sports and physical rehabilitation," *Applied Sciences*, vol. 15, no. 12, p. 6405, 2025. https://doi.org/10.3390/app15126405

[16] J. Tan and J. Chen, "Generating context-specific sports training plans by combining generative adversarial networks," *PLoS ONE*, vol. 20, no. 1, p. e0318321, 2025. https://doi.org/10.1371/journal.pone.0318321

[17] W. Sun, "Predictive analysis and simulation of college sports performance fused with adaptive federated deep learning algorithm," *Journal of Sensors*, vol. 2022, no. 1, p. 1205622, 2022. https://doi.org/10.1155/2022/1205622

7 AUTHORS

Liyuan Xie is a Lecturer at P.E. Education, Ningbo Polytechnic University. She holds a master's degree from Ningbo University. Her research interests encompass physical education, sports coaching, sports rehabilitation, and sports nutrition. To date, she has published nine academic papers and was awarded the First Prize in the Ningbo Vocational Teaching Competence Competition. Furthermore, she has frequently coached students in the Zhejiang Provincial College Students' Aerobics Competition, leading them to achieve outstanding results (E-mail: 13858255889@163.com).

Qian Sun is currently a Lecturer working at the Institute of Qixin, Ningbo Tech University. She graduated from Ningbo University, and her research interests include physical education, sports training, and school sports, among others. She has published over 10 academic papers, including one in an SCI Zone 2 journal and two in Chinese CSSCI (Nanjing University Core) journals (E-mail: nbsunqian123@163.com).

Jinggang Li is currently an Associate Professor working at the Institute of Qixin, Ningbo Tech University. He is a doctor who graduated from Kyungnam University, and his research interests include physical education, sports training, and school sports, among others. He has published over 15 academic papers, including two SCI journals (E-mail: nbsunqian123@163.com).

Yongliang Zhang is a Lecturer at Zhejiang University of Finance & Economics at Dongfang College. He holds a master's degree from Ningbo University. His primary research interest is physical education. He has frequently coached students in the Zhejiang Provincial Collegiate Dance & Aerobics Competitions and Basketball Tournaments, consistently achieving outstanding results (E-mail: m18167316112@163.com).