# Standardized Teaching Aid System for Piano Performance Movements Integrating Posture Estimation and Temporal Feature Extraction

**Jing Gao, Huaning Sun**
School of Music, Linyi University, Linyi 276000, Shandong, China
Corresponding author: Huaning Sun (e-mail: huaning666@163.com)

**ABSTRACT** Current piano performance training primarily relies on subjective observation and experience-based assessment, limiting the objective evaluation of posture accuracy and temporal coordination. This study proposes a standardized teaching aid system that integrates an improved posture estimation network with temporal feature extraction to achieve intelligent analysis and real-time feedback for piano performance movements. The system extracts upper-limb and finger keypoints through a lightweight posture estimation framework enhanced by spatial attention mechanisms and constructs spatiotemporal representations using a Temporal Transformer with embedded temporal position encoding. A quantitative evaluation network is further developed to assess spatial coordination, temporal consistency, and fingering stability, generating standardized scores and dynamic feedback signals in a closed-loop teaching framework. Experimental results demonstrate a keypoint localization stability rate of 96.0%, feedback accuracy of 97.2%, and response delays ranging from 0.48 s to 0.77 s across different performance scenarios. The proposed framework exhibits strong robustness and adaptability for complex motion analysis. It is particularly applicable to intelligent sensing environments supported by wireless communication infrastructures and antenna-enabled edge devices, where reliable real-time transmission of visual and motion data is essential for responsive feedback and human–machine interaction. This study provides an effective engineering solution for intelligent motion assessment, standardized skill training, and data-driven interactive learning systems.

**INDEX TERMS** posture estimation, temporal transformer, motion analysis, wireless sensing networks

## I. INTRODUCTION

IN piano teaching practice, the standardization of performance movements directly affects tone quality and technical expression. However, current teaching still relies mainly on teachers' visual judgment and experience-based guidance, lacking quantitative basis for the accuracy and timing coordination of movements. Since performance posture involves complex multi-joint coordination and high-speed fingering changes, it is difficult to identify subtle posture deviations and rhythmic differences through manual analysis alone, affecting the objectivity and systematic nature of teaching [1,2]. Establishing an intelligent analysis system based on visual recognition and movement modeling is of great significance for realizing the scientification and standardization of piano teaching [3,4]. At the same time, the flexibility and stress transmission laws in human performance movements [5,6] have certain similarities with the elastic response and deformation distribution of fiber texture in textile materials. This cross-domain relation between movement-induced force transmission and texture deformation patterns provides a clearer basis for posture modeling.

Several technical challenges exist in the current action standardization analysis. Variations in lighting, occlusion interference, and scale differences in local finger movements in visual input can easily lead to keypoint drift during posture estimation [7,8]. The time dependence of high-frequency movements during performance is complex, and traditional frame-level analysis struggles to characterize continuity [9,10]. At the evaluation level, the lack of unified quantitative indicators for action coordination and temporal consistency makes it difficult for the model to consistently output reliable teaching feedback [11,12]. These issues collectively limit the accuracy and adaptability of piano performance movement analysis systems. Existing research has improved action recognition performance through pose detection, skeleton modeling, and deep learning methods. Convolutional neural networks have been used to extract features of the upper limbs and fingers of performers, and heatmap regression methods have improved the accuracy

of keypoint localization [13,14]. Recurrent neural networks and long short-term memory structures have been used to describe temporal changes, and some works have applied attention mechanisms in the temporal dimension to enhance cross-frame dependency modeling [15,16]. However, these methods mostly focus on spatial posture estimation and are insufficient in capturing the global correlation of temporal structures, making it difficult to effectively integrate multi-dimensional features for prescriptive analysis [17,18]. At the same time, scoring mechanisms often rely on a single indicator, resulting in a large feedback delay, which cannot meet the real-time requirements of teaching scenarios [19,20]. To address these shortcomings, it is necessary to construct a unified analysis framework that takes into account both spatial and temporal information to achieve efficient linkage between action recognition, evaluation, and feedback [21,22].

The lightweight encoder and spatial attention pipeline are adjusted to accommodate the fine-grained structural distribution of piano performance postures. The redesigned texture-guided refinement layer and the anatomically constrained attention weighting process form a feature extraction pattern distinct from existing lightweight pose estimators. These adjustments establish the actual improvement claimed by the posture estimation module in this study.

To address the lack of integrated analysis caused by the separation of posture recognition from temporal feature modeling, this study proposes a teaching aid system that unifies these two components within a single framework. The study employs a modified lightweight posture estimation network to extract upper limb and finger keypoints, while implementing spatial attention to increase the response of the hand region. Action sequences are then established through creating a skeleton topology, then combined to create a unified posture matrix through temporal interpolation and scale normalization. At the temporal level, this article adopts Transformer with embedded temporal position encoding to model cross-frame dependencies, utilizing multi-head self-attention to obtain joint spatial-posture and temporal-rhythm representation. Then, this article uses the multi-dimensional mapping network to transfer action features into specific score evaluation on spatial coordination, temporal coherence, and fingering stability, respectively. Moreover, this article maps the rate of change of scores into dynamic feedback signals, offering more visual feedback and numerical feedback to reflect the deviation of user's performance in certain movement, which provides realtime correction and detailed guidance in teaching process. In this way, this article builds a whole closed-loop process from posture extraction and temporal modeling to score feedback, and gives quantifiable intelligent support for standardized teaching of piano performance movements.

## II. CONSTRUCTION PRINCIPLES OF STANDARDIZED PIANO PERFORMANCE MOVEMENT TEACHING AID SYSTEM

### A. PERFORMANCE POSTURE INFORMATION EXTRACTION BASED ON KEYPOINT DETECTION

The process of extracting piano playing posture information in this article is based on the detection of keypoints. The improved lightweight posture estimation network can accurately identify the upper limb and finger joints in the spatiotemporal domain [23,24]. The improvement is reflected in the reconstruction of the lightweight encoder path and the refinement of the spatial attention process. The encoder inserts a texture-guided refinement layer that adjusts the activation distribution according to the gradient density of local skin and fingertip regions, forming a feature enhancement structure that departs from standard lightweight encoders. The depthwise separable convolution blocks are arranged in a progressive receptive field expansion chain that preserves sensitivity to subtle structural variations of narrow joint areas. The spatial attention mechanism integrates an anatomical continuity constraint that stabilizes the response of channels associated with the hand–wrist axis, reducing fluctuations that often occur when fast fingertip transitions appear in lightweight architectures. These structural adjustments constitute the basis of the improved lightweight posture estimation network adopted in this study. The input frame sequence is denoted as $I = \{I_t | t = 1, \ldots, T\}$, where $T$ is the number of video frames. The network structure consists of a feature encoding module, a spatial attention module, and a coordinate regression module. The encoding module extracts local texture and fabric features based on depthwise separable convolution, enabling the system to have stronger structural perception ability when recognizing subtle movements of skin, clothing, and hand surfaces [25,26]. This texture-sensitive feature extraction method is also commonly used for surface defect detection and fabric texture recognition in textiles. Assuming the convolution kernel weight is $W_c$, and the input feature is $X_t$, the output feature mapping $F_t$ is defined as:

$$F_t = \sigma(W_c * X_t) \tag{1}$$

In Formula (1), "*" represents the convolution operation; $\sigma(\cdot)$ is a non-linear activation function. The spatial attention module adaptively weights the channel-dimensional features to form a saliency map $A_t$, and its calculation process is as follows:

$$A_t = \text{softmax}\left(\frac{Q_t K_t^T}{\sqrt{d}}\right) \tag{2}$$

In Formula (2), $Q_t = W_q F_t$, and $K_t = W_k F_t$. $W_q$ and $W_k$ are linear mapping matrices, and $d$ is the feature dimension. Attention weights are applied to $F_t$ to generate enhanced features $\bar{F}_t = A_t V_t$, where $V_t = W_v F_t$, which are used to highlight the structural information of the hand and forearm regions.

The attention mapping incorporates an inter-joint modulation factor that aligns the saliency response with the

displacement trends of adjacent anatomical nodes, producing a stabilized activation pattern that differs from common lightweight attention formulations.

The coordinate regression module predicts the two-dimensional coordinates of each keypoint using Gaussian heatmap regression. Let the actual heatmap of the $j$-th keypoint be $H_j$, and the predicted heatmap be $\hat{H}_j$. The loss function is defined as:

$$L_p = \frac{1}{J}\sum_{j=1}^{J}\left\|H_j - \hat{H}_j\right\|_2^2 \tag{3}$$

In Formula (3), $J$ denotes the total number of keypoints, and the notation $\|\cdot\|_2$ denotes the Euclidean norm applied to the pixel-wise difference between the predicted and ground-truth heatmaps. High-precision joint positioning is achieved by minimizing $L_p$. The model outputs a set of keypoints $P_t$, describing the attitude structure at time $t$ in two-dimensional coordinates.

In the skeleton data generation period, the frame keypoints establish a skeleton topology graph $G = (V, E)$ based on the physiological connections, where $V$ is the node set corresponding to each keypoint in $P_t$, and $E$ is the edge set representing the anatomical constraint connections between keypoints. Finally, the skeleton sequence $S = \{G_t | t = 1, \ldots, T\}$ is further sampled and normalized in terms of temporal interpolation and scale normalization to eliminate the variance of viewpoint and distance, obtaining a spatiotemporally consistent set of pose representation matrices.

Although the pose representation is constructed on two-dimensional keypoint coordinates, the original motion of piano performance contains vertical displacement during key pressing and rotation of the wrist around three-dimensional axes. The projection of these depth-related components onto a single image plane introduces variations that originate from changes in perspective and the relative position between the camera and the performer. These factors alter the apparent length and orientation of limb segments and compress depth cues into planar trajectories, which introduces structural distortion that is not fully recoverable through normalization. The lightweight encoder extracts spatial texture information from image regions that have already lost depth variation, and the subsequent skeleton topology preserves only the planar joint configuration. This leads to deviations between the actual spatial configuration and the inferred representation whenever vertical motion amplitude or wrist rotation intensity increases. Such deviations influence the stability of the temporal sequence matrix used for subsequent dynamic modeling and form a systematic source of uncertainty throughout the entire evaluation process.

$$M = [N(P_1), N(P_2), \ldots, N(P_t)] \tag{4}$$

$N(\cdot)$ denotes the normalization function that normalizes the coordinates of keypoints into the same coordinate system in Formula (4). $M$ represents the input of temporal modeling, including the spatial displacement information of arm and finger during performance, and is also the basic data representation for the action dynamic modeling.

### B. ACTION DYNAMIC MODELING BASED ON TEMPORAL TRANSFORMER

The dynamic changes in performance movements are modeled using temporal features to achieve global dependency representation analysis [27,28]. Using the pose matrix $M$ output from the previous stage as input, a temporal Transformer model with embedded temporal position encoding is constructed to capture the temporal continuity and cross-frame dependencies of the movements. The input matrix is linearly mapped to generate a feature sequence $Z = \{z_1, z_2, \ldots, z_t\}$, where $z_t \in R^d$ is the pose feature vector at time step $t$, and $d$ represents the feature dimension. To preserve temporal position information, a learnable position encoding vector $E_t$ is applied and added to the input features to form a sequence representation $X_t = z_t + E_t$.

The temporal dependency modeling process is implemented through a multi-head self-attention mechanism [29,30]. For the h-th attention head, the calculation process is as follows:

$$H_t^h = \text{softmax}\left(\frac{Q_t^h (K_t^h)^T}{\sqrt{d_h}}\right) V_t^h \tag{5}$$

In Formula (5), $Q_t^h = W_q^h X_t$, $K_t^h = W_k^h X_t$, and $V_t^h = W_v^h X_t$. $W_q^h$, $W_k^h$, and $W_v^h$ are the linear transformation matrices of the $h$-th attention head, and $d_h$ is the feature dimension of a single head. The softmax function guarantees the normalization of the attention weights so that the model applies the weights in a global way over the entire feature sequence. The outputs of several attention heads are concatenated and linearly transformed to obtain the fused global temporal feature $H_t$.

To enhance the nonlinear expressive power of the model, each coding layer of the Transformer is followed by a feedforward network module. This module consists of two linear transformation layers and an activation function, calculated as follows:

$$Y_t = \sigma(W_1 H_t + b_1) W_2 + b_2 \tag{6}$$

In Formula (6), $W_1$, $W_2$ represent the weight matrices of the feedforward network; $b_1$, $b_2$ represent the bias terms; $\sigma(\cdot)$ represents the activation function. The residual connection and layer normalization structure ensure stable gradient propagation and reduce feature drift.

The temporal Transformer, after being stacked into multiple layers, outputs a global feature representation $F = \{f_1, f_2, \ldots, f_t\}$, where each layer $f_t \in R^d$ corresponds to the spatiotemporal joint representation at time $t$. To extract the overall dynamic trend of the performance movements,

the feature sequences are temporally aggregated, and average pooling is used to obtain a global action description vector.

$$\bar{f} = \frac{1}{T}\sum_{t=1}^{T} f_t \tag{7}$$

In Formula (7), $T$ is the total number of frames. The vector $\bar{f}$ represents the comprehensive characteristics of the performance action in the time dimension, covering the coordination of arm and finger movements and the rhythmic change pattern, providing a stable temporal characteristic basis for subsequent quantitative evaluation of the standardization of the action.

### C. DESIGN OF QUANTITATIVE EVALUATION AND FEEDBACK MECHANISM FOR ACTION STANDARDIZATION

The quantitative evaluation of action standardization achieves a comprehensive score for piano performance movements by constructing an evaluation network based on deep feature mapping [31,32]. The input is the global action feature vector $\bar{f} \in R^d$ output by the temporal Transformer. This feature is mapped to the scoring space through a multilayer perceptron to characterize the degree of standardization of the action in the dimensions of spatial coordination and temporal consistency. The computation process of the evaluation network is defined as follows:

$$s = \sigma(W_3\bar{f} + b_3) \tag{8}$$

In Formula (8), $W_3$ is the linear mapping matrix; $b_3$ is the bias vector; $\sigma(\cdot)$ is the Sigmoid activation function; $s \in (0, 1)$ represents the continuous output of the normalization score. To achieve multidimensional index fusion, feature branches $s^s$, $s^t$, $s^f$ are established in three subspaces: spatial pose, temporal coherence, and finger stability. The final quantitative score is obtained through weighted fusion.

$$S = \alpha_s s^s + \alpha_t s^t + \alpha_f s^f \tag{9}$$

In Formula (9), $\alpha_s$, $\alpha_t$, $\alpha_f$ are empirically determined weight coefficients that satisfy $\alpha_s + \alpha_t + \alpha_f = 1$, which are used to balance the contribution ratio of each sub-feature in the overall score.

The training objective of the prescriptive evaluation network is to minimize the mean squared error loss between the predicted score and the human expert score, defined as:

$$L_s = (1/N)\sum_{i=1}^{N}(S_i - \hat{S}_i)^2 \tag{10}$$

In Formula (10), $N$ represents the sample size; $\hat{S}_i$ represents the standard score given by the expert; $S_i$ represents the model prediction value. By minimizing $L_s$, the system gradually calibrates the prediction distribution during the training phase, ensuring that the output score is consistent with the professional judgment.

In the teaching feedback phase, the system calculates a dynamic scoring curve $S(t)$ from the real-time input action feature sequence $\{f_1, f_2, \ldots, f_t\}$ using a model. To visualize the feedback, the system generates a prompt signal based on the rate of change of the score $d_s/d_t$, defining a feedback function:

$$F(t) = \text{sign}\,(d_s/d_t) \cdot |d_s/d_t| \tag{11}$$

In Formula (11), $\text{sign}(\cdot)$ represents sign function, which is used to capture the directional alignment of an individual's performance in a scored outcome, and $|\cdot|$ refers to the magnitude of the change. The sign and magnitude of $F(t)$ correspond to the trend of improvement and the intensity of the deviation from the performance action, respectively. The information is displayed in real time on the user interface, in color and numbers, so that the performer can receive directional corrective prompts if and when their performance action deviates from the standard.

This article combines scoring networks with feedback to complete the entire closed-loop process from extracting time features to quantifying action standardization and then outputting teaching feedback [33]. The model output score and real-time feedback together constitute the core support part of standardized action teaching, which can provide a quantitative basis for the standardization and refinement of piano playing posture teaching process [34,35].

## III. EXPERIMENT AND RESULTS DISCUSSION ON THE STANDARDIZATION OF PIANO PERFORMANCE MOVEMENTS

### A. EFFECT OF KEYPOINT LOCALIZATION

In piano performance action recognition experiment, to explore the keypoint localization accuracy and stability of model under different performance techniques, five kinds of representative performance postures are randomly chosen for comparison analysis in this section. According to the keypoint coordinates output by the posture estimation network, the experiment calculates and analyzes the quartile of 4 quantitative indexes of recall, precision, F1 score, and matching stability, to evaluate the completeness of keypoint detection and the temporal consistency of model keypoint localization under different action complexity. The test scene includes various performance modes from normal sitting postures to rapid fingering, and explores the impact of movement amplitude, occlusion degree, and finger dexterity on the keypoint detection results. The experimental results are shown in Figure 1.

As shown in Figure 1, the model performs stably overall in the keypoint localization task, with the highest accuracy under the standard sitting posture condition: recall 94.1%, precision 95.0%, F1 score 94.5%, and matching stability 96.0%, indicating that the system extracts spatial features of common performance movements quite effectively. The

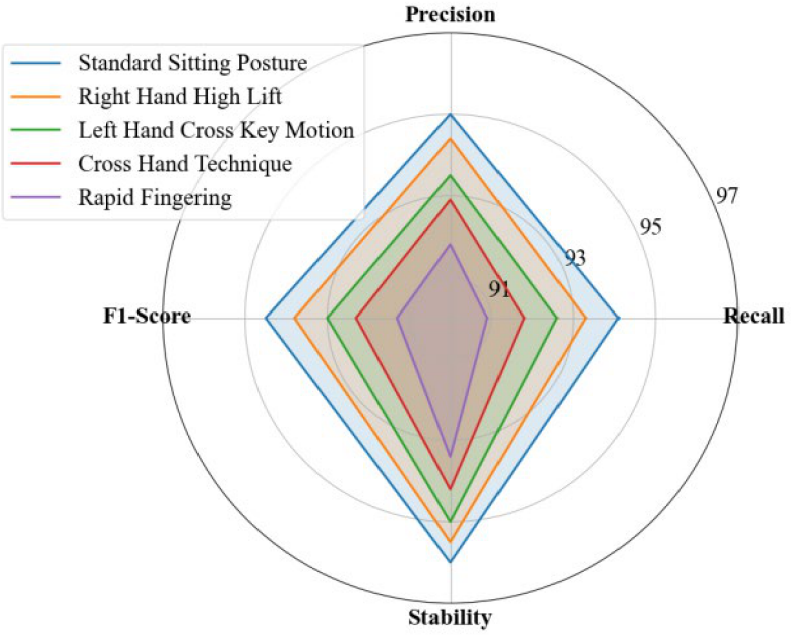


**FIGURE 1.** Keypoint positioning performance of different piano performance techniques

F1 scores for the right hand raised high and the left hand crossing keys are 93.8% and 93.0%, respectively, showing a slight decrease in localization performance. This is mainly due to keypoint occlusion caused by arm raising and hand crossing, leading to a decrease in response accuracy during the heatmap regression stage. When both hands are crossed, the precision drops to 92.9%, and the matching stability is 94.2%, indicating that overlapping limbs cause confusion of local keypoints in temporal tracking. The four indicators are lowest in the rapid fingering scenario: F1 score 91.3% and matching stability 93.4%, reflecting that the shortened time interval of high-frequency fingering movements causes excessively frequent changes in the position of keypoints, weakening the model's ability to model continuous features. The overall trend shows that the keypoint localization performance gradually decreases as the complexity of the performance movement increases, but it still remains at a stable level above 90%, which verifies the system's reliable localization capability in complex piano performance scenarios.

The noticeable drop of the F1 score in the rapid fingering condition arises from limitations that originate at both the pose estimation stage and the temporal modeling stage. The shortened displacement interval of fingertip trajectories introduces motion blur in consecutive frames, which narrows the spatial distinction among adjacent keypoints during heatmap regression and reduces the discrimination strength of local textures used by the lightweight encoder. The temporal Transformer accumulates these frame-level uncertainties in its global attention process, since the attention weights distribute across a sequence that no longer retains stable inter-frame correspondences under high-frequency motion. The temporal position encoding preserves ordering information, yet the encoder still receives feature sequences whose short-term continuity is degraded by rapid velocity fluctuations. These factors jointly restrict the Transformer from restoring full temporal coherence when fast fingering movements continuously exceed the temporal resolution fixed by the frame rate. This interaction explains why the improved system maintains overall stability but still shows weakened performance in rapid fingering.

### *B. PERFORMANCE VERIFICATION OF TEMPORAL FEATURE MODELING*

To verify the temporal feature modeling capability of the Temporal Transformer in piano performance action analysis, this section quantitatively evaluates the model performance based on experimental results from different training rounds, focusing on four dimensions: temporal relevance, temporal consistency, action continuity, and temporal structure retention rate. The model's performance metrics from training rounds 10 to 100 are standardized statistically to reflect its convergence characteristics and dynamic changes during temporal learning. Figure 2 shows the trend changes of each performance metric at different training stages.

Figure 2 shows that temporal correlation is 77.3% at 10 epochs, stabilizing at 93.7% at 100 epochs, indicating that the model quickly establishes temporal dependencies between action frames in the early training stages and stabilizes in the later stages. Temporal consistency increases from 74.0% to 81.6% in the first 20 epochs, then the increase slows down, reaching 89.3% at 100 epochs, indicating that the model's overall coordination of temporal features gradually improves. The growth curves of action continuity and

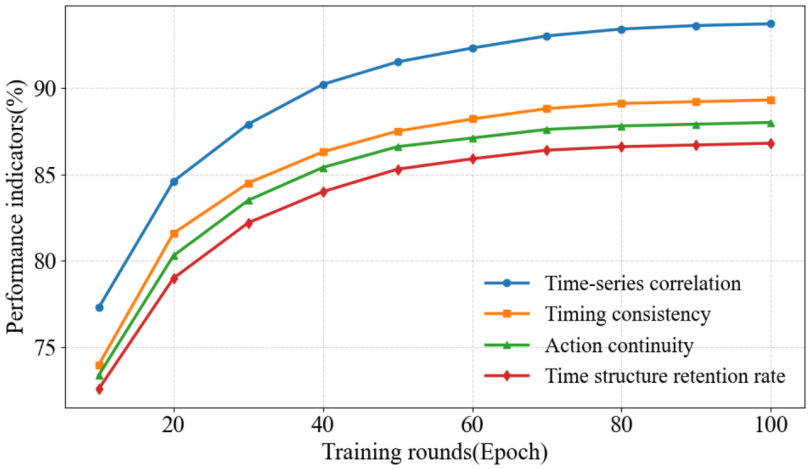


**FIGURE 2.** Convergence curve of time series feature modeling performance

temporal structure retention are relatively flat. The former maintains a stable increase in the middle stage, eventually reaching 88.0%, while the latter remains at 86.8% in the convergence stage. The difference between the two indicates that there are still slight differences in the model's representation of action transitions under complex rhythm changes. The overall trend reflects that Transformer has high learning stability and convergence performance in temporal feature modeling, and the convergence of indicators in the later training stages verifies that the model has sufficient learning and stable convergence in temporal feature representation.

The temporal indicators reported in Figure 2 are derived from the comparison between the temporal embeddings produced by the Transformer during training and the reference temporal patterns constructed from expert-annotated sequences. These reference patterns record the temporal evolution of joint coordination extracted from expert performance data through the same posture estimation and skeleton generation pipeline. The indicators quantify the alignment between the internal temporal representation and the expert-derived temporal dynamics that serve as supervisory signals during optimization. The values describe the stability of the internal representation within the temporal modeling component and do not correspond to external behavioral labels of performance quality. Their convergence reflects the formation of a consistent temporal embedding space that supports the evaluation network, rather than the real-world accuracy of the complete system.

The internal temporal metrics record the extent to which the model has formed stable temporal dependencies during the training phase and serve as an intermediate measurement that ensures the reliability of the temporal feature space used by the scoring module. These metrics do not assess the behavior of the full detection, evaluation, and feedback process in real performance scenarios. The feedback accuracy reported later in the teaching response experiment is computed from direct comparisons between the system's output scores and expert assessments on complete performance sequences. The two groups of metrics therefore represent distinct stages of the system, with the temporal indicators reflecting internal learning stability and the feedback accuracy reflecting observable performance in real conditions.

### C. STABILITY OF ACTION SCORING

To test the system's spatial pose recognition performance for different types of piano movements, this article selects eight typical performance tasks in the same sampling environment, including two kinds of movement modes: posture change and rhythm control. Based on the keypoint coordinates output by the improved pose estimation network, this article further analyzes the recognition accuracy and stability of the system under different types of movements. All the test data are standardized, and the scores of different types of movements are counted. The range and statistical dispersion of the scores are calculated. The experimental results are shown in Figure 3, where the performance change of the model under different movement complexities can be visually observed.

From Figure 3, it can be seen that the median value of the standard sitting posture is 0.941, and the range is relatively small, indicating that the system output is relatively stable when the standard posture is adopted. The median value of right-hand high lifting and left-hand key crossing are 0.917 and 0.900, respectively, and the range is slightly larger. This indicates that when the height of the limbs changes and the span of the keys increases, the fluctuation of the score occurs. The median value of hand crossing and rapid fingering is 0.891 and 0.874, respectively. The score drops significantly, indicating that there is a certain degree of deviation in the modeling of the continuity of the action of the system. The instability of the system output is caused by the overlap of hands and the high frequency of key touches. The median value of arpeggio legato and broken chords are 0.926 and 0.910, respectively; the score

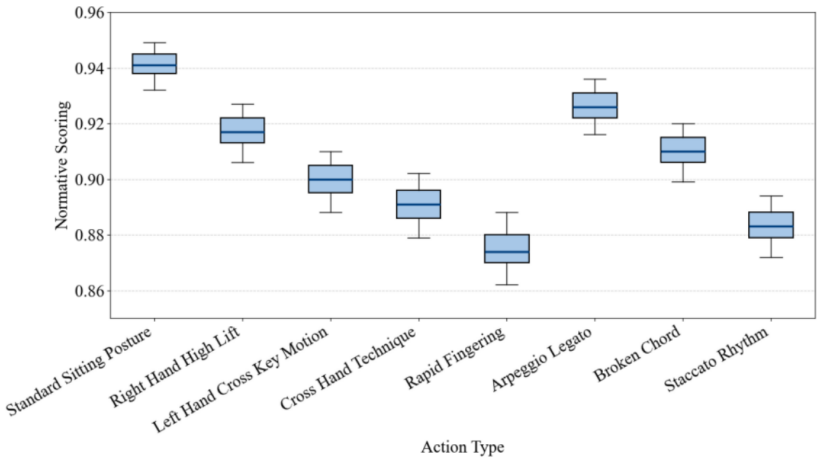


**FIGURE 3.** Stability analysis of piano movement scoring

is concentrated; the dispersion is low, indicating that when recognizing the continuity of fingering, the system output is still relatively stable and high. The median value of the staccato rhythm is 0.883, and the lower limit is slightly low. This is because the amplitude changes of the staccato key touch affect the temporal response of the model. From the distribution characteristics, it can be seen that the fluctuation range of the system score is not large for most action types; the output results are relatively stable; the consistency of cross-action is good.

### *D. TEACHING FEEDBACK RESPONSE PERFORMANCE TEST*

In the system performance verification experiment, in order to verify the response characteristic and realtime stability of the teaching feedback module under different performance movements, this article chooses eight typical piano performance postures, including various technical movements from common sitting postures to high-frequency fingerings, and records the average value of the feedback delay and feedback accuracy of the system under continuous interactive conditions. In the experiment, the input video frame rate and model inference parameters are kept unchanged to eliminate the impact of external variables on the response results and reflect the dynamic response characteristics of the system under different action scenarios. The experimental results are shown in Figure 4.

The feedback delay for normal sitting posture is 0.48s, and the feedback accuracy is 97.2%, which shows that the model achieves an ideal output in terms of both delay and accuracy under the condition of static posture recognition. The delay for right-hand high-raise and arpeggio legato is 0.55s and 0.53s, respectively, and the accuracy is still larger than 96%, showing that the computational load is relatively light when the amplitude of limb movement is light and the rhythm is continuous. The delay for left-hand key crossing and staccato rhythms increases to around 0.63 to 0.66s, and the accuracy decreases to 95.1 % to 94.8 %, respectively. This shows that when the posture crossing happens and transient key touches emerge besides main playing actions, the output accuracy of the system decreases because more complex information of keypoint recognition is extracted from the video frame. Therefore, the delay of feedback generation increases accordingly. The delay for crossed hands and rapid fingering reaches 0.69s and 0.77s, and the accuracy decreases to 94.3% and 93.5%, respectively. This shows that the computational coupling pressure on the model increases when more occlusions emerge and the frames are transitioned at a high frequency. The trend of output delay shows that the system can maintain an stable output when the delay is smaller than 0.8s and the accuracy is larger than 93% under different action conditions, and the real-time performance and recognition reliability of teaching feedback mechanism show high consistency with the adaptability.

### *E. EVALUATION OF THE OVERALL EFFECTIVENESS OF TEACHING SUPPORT*

This section evaluates the applicability and stability of the integrated posture estimation and temporal feature extraction model into different piano performance tasks by comprehensive evaluation of eight typical performance movements after training of the system is finished from multiple aspects. The experimental evaluation is designed based on five indicators: posture standardization, rhythm consistency, fingering stability, feedback responsiveness, and learning improvement rate. The distribution of normalized score in different movements is plotted to analyze the synergistic performance of teaching aid system in handling complex movement and real-time feedback in Figure 5.

As shown in Figure 5, when executing the standard sitting posture and arpeggio legato tasks, the overall posture regularity of the system is still above 0.94, which proves that the feature extraction is accurate for stable posture recognition and continuous fingering tracking. The rhythmic consistency of the right-hand high lift and arpeggios are 0.915 and 0.902

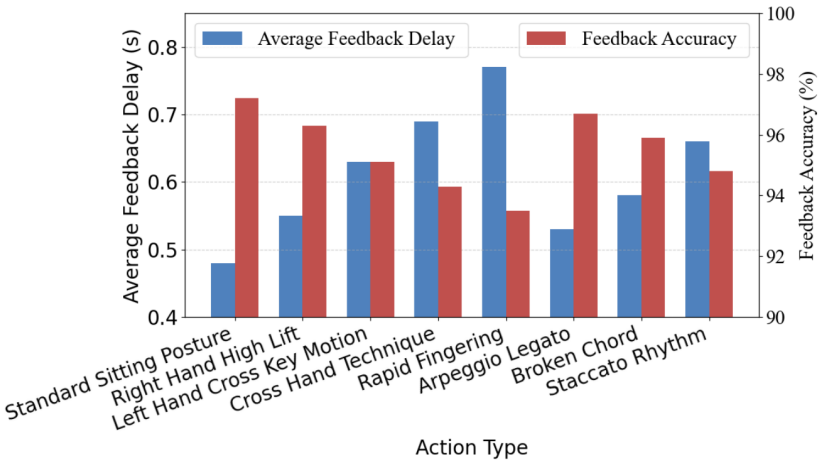


**FIGURE 4.** Comparison of teaching feedback performance under different performance movements

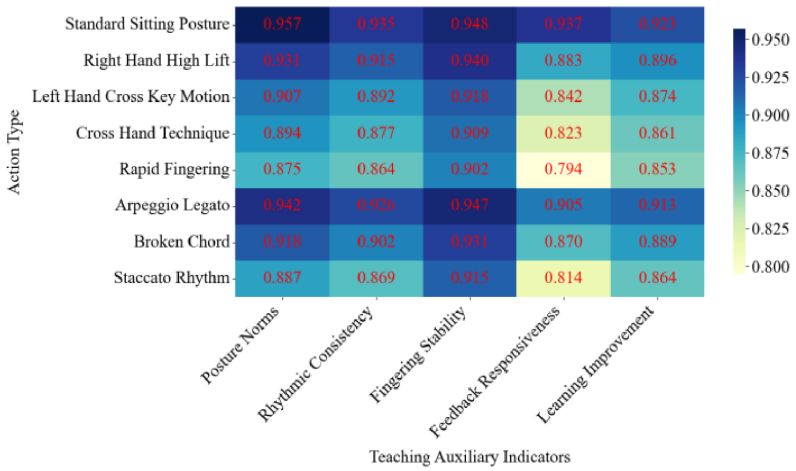


**FIGURE 5.** Effect of the integrated teaching support system

at medium tempos, which proves that the system has a certain ability to resist time deviation for fast fingering and staccato rhythms. However, the feedback responsiveness of fast fingering and staccato rhythms is less than 0.82, which means that when the model infers at a high frequency input, there is a certain delay in the speed of response, but the overall speed of response is still acceptable. In summary, the differences among indicators are consistent with the influence of performance complexity on the recognition and feedback performance of the system, which proves that the constructed model has good robustness and teaching aid value in multi-action scenarios.

## IV. CONCLUSION

In this study, this article builds a teaching tool system for piano performance movements based on a combined strategy of posture estimation and temporal Transformer. This novel integration of spatial and temporal analysis, while focused on music, offers a methodological template that could be adapted for other industry to objectively analyze and standardize complex sequences of manual actions. With the support of keypoint detection and global dependency modeling, a common representation of spatial posture attribute and temporal sequence attribute is learned. In the experiment, compared with standardized sitting posture situations, F1 score of spatiotemporal collaborative analysis and standardized quantitative evaluation system reaches 94.5%; the degree of temporal continuity is further enhanced to 89.3% in the global dependency modeling of temporal feature; standardized score of posture over 0.94 can be achieved in the comprehensive evaluation teaching rubric for the same graded tasks, which reflects certain stability in the multidimensional indicators. This article conducts spatiotemporal collaborative analysis and standardized quantitative evaluation of piano performance movements, and provides real-time feedback and objective evidence in the teaching process. The limitations observed in high-frequency fingering originate from the mismatch between the temporal sampling density of the video input and the rapid changes of fine-grained joint positions. The

pose estimation network extracts spatial cues that depend on stable texture distribution, yet the fast transitions of fingertip regions compress the valid exposure interval and induce blurred boundaries, which weakens the reliability of the heatmap regression stage. The temporal Transformer receives these unstable spatial features as its input and forms global dependencies on sequences that carry reduced local continuity. This reduces the precision of temporal correlation modeling when motion velocity surpasses the representational granularity established during training. The joint effect of motion blur and insufficient temporal resolution leads to residual errors in the global attention fusion, which accumulate as the temporal depth increases. However, when facing the situation of complex movements and slight fluctuations in feedback accuracy for high-frequency fingering, the problem of model response delay increase exists. In addition, when there are occlusions, the temporal continuity representation is to be represented by optimizing scales of attention mechanisms and skeleton topology modeling. This article can also explore the introduction of mechanical parameters and physiological signals, and joint modeling in the feedback module, to expand its application potential in personalized teaching and sports health assessment. The use of two-dimensional keypoints forms a practical framework for the integrated analysis pipeline, yet the intrinsic three-dimensional nature of key pressing trajectories and joint rotations indicates that depth-aware estimation may refine the reliability of the standardization score. The spatial cues produced by monocular imaging retain only the projected contour of the movement, and depth variation is not explicitly represented. This constraint limits the accuracy of posture interpretation when vertical displacement changes the spatial relationship between forearm, wrist, and fingertip regions. A three-dimensional estimation framework provides depth-conditioned joint coordinates and reduces ambiguity introduced by viewpoint variance. Such an extension can support a more faithful reconstruction of performance movements and strengthen the interpretability of long-term temporal dependencies during evaluation, thus extending the system's utility into occupational health and personalized training for skilled manual laborers.

## AUTHOR CONTRIBUTIONS

Conceptualization –Jing Gao and Huaning Sun; methodology – Jing Gao and Huaning Sun; investigation – Jing Gao and Huaning Sun; writing-original draft preparation – Jing Gao and Huaning Sun. All authors have read and agreed to the published version of the manuscript.

## CONFLICTS OF INTEREST

The authors declare no conflict of interest.

## FUNDING

This research received no external funding.

## ACKNOWLEDGEMENTS

Not applicable.

## REFERENCES

[1] W. B. Verwey, "Chord skill: Learning optimized hand postures and bimanual coordination," Experimental Brain Research, vol. 241, no. 6, pp. 1643-1659, 2023, doi: 10.1007/s00221-023-06629-2.

[2] K. Ito, T. Watanabe, T. Horinouchi, T. Matsumoto, K. Yunoki, H. Ishida, et al., "Higher synchronization stability with piano experience: Relationship with finger and presentation modality," Journal of Physiological Anthropology, vol. 42, no. 1, p. 10, 2023, doi: 10.1186/s40101-023-00327-2.

[3] P. Yang, "Integrating intelligent algorithms in music education to analyze and improve posture and motion in instrumental training," Molecular & Cellular Biomechanics, vol. 22, no. 1, pp. 762-762, 2025, doi: 10.62617/mcb762.

[4] S. Zhang, P. Ni, J. Wen, Q. Han, X. Du, and J. Fu, "Intelligent identification of moving forces based on visual perception," Mechanical Systems and Signal Processing, vol. 214, no. 1, 111372, 2024, doi: 10.1016/j.ymssp.2024.111372.

[5] C. F. P. Monfredini, D. B. Coelho, A. J. Marcori, and L. A. Teixeira, "Control of interjoint coordination in the performance of manual circular movements can explain lateral specialization," Human Movement Science, vol. 90, no. 1, 103102, 2023, doi: 10.1016/j.humov.2023.103102.

[6] J. Yuk, N. M. Kitchen, A. Przybyla, R. A. Scheidt, and R. L. Sainburg, "Symmetry and synchrony of bimanual movements are not predicated on interlimb control coupling," Journal of Neurophysiology, vol. 131, no. 6, pp. 982-996, 2024, doi: 10.1152/jn.00476.2023.

[7] X. Bai, X. Wei, Z. Wang, and M. Zhang, "CONet: Crowd and occlusion-aware network for occluded human pose estimation," Neural Networks, vol. 172, no. 1, 106109, 2024, doi: 10.1016/j.neunet.2024.106109.

[8] X. Jiang, H. Tao, J. N. Hwang, and Z. Fang, "A multiscale coarse-to-fine human pose estimation network with hard keypoint mining," IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 54, no. 3, pp. 1730-1741, 2023, doi: 10.1109/TSMC.2023.3328876.

[9] Q. Cui, H. Sun, Y. Kong, X. Zhang, and Y. Li, "Efficient human motion prediction using temporal convolutional generative adversarial network," Information Sciences, vol. 545, no. 1, pp. 427-447, 2021, doi: 10.1016/j.ins.2020.08.123.

[10] M. Wu and P. Shi, "Human pose estimation based on a spatial temporal graph convolutional network," Applied Sciences, vol. 13, no. 5, p. 3286, 2023, doi: 10.3390/app13053286.

[11] D. Mao and S. Liu, "Quantitative Evaluation of Piano Performance Technique and Style Based on Continuous Discretization Method," Applied Mathematics and Nonlinear Sciences, vol. 9, no. 1, p. 1, 2024, doi: 10.2478/amns.2023.2.00093.

[12] H. Qi and C. She, "Research on Improving Piano Performance Evaluation Method in Piano Assisted Online Education," International Journal of Advanced Computer Science and Applications, vol. 14, no. 8, pp. 1-12, 2023, doi: 10.14569/IJACSA.2023.0140850.

[13] X. Li, Y. Guo, W. Pan, H. Liu, and B. Xu, "Human pose estimation based on lightweight multi-scale coordinate attention," Applied Sciences, vol. 13, no. 6, p. 3614, 2023, doi: 10.3390/app13063614.

[14] D. Rong and F. Gang, "Coordinate-Corrected and Graph-Convolution-Based Hand Pose Estimation Method," Sensors, vol. 24, no. 22, p. 7289, 2024, doi: 10.3390/s24227289.

[15] C. Yang, F. Mei, T. Zang, J. Tu, N. Jiang, and L. Liu, "Human Action Recognition Using Key-Frame Attention-Based LSTM Networks," Electronics, vol. 12, no. 12, p. 2622, 2023, doi: 10.3390/electronics12122622.

[16] J. Tang, J. Wang, and J. F. Hu, "Predicting human poses via recurrent attention network," Visual Intelligence, vol. 1, no. 1, p. 18, 2023, doi: 10.1007/s44267-023-00020-z.

[17] H. Wang, Q. Shi, and B. Shan, "Three-dimensional human pose estimation with spatial–temporal interaction enhancement transformer," Applied Sciences, vol. 13, no. 8, p. 5093, 2023, doi: 10.3390/app13085093.

[18] Z. Chen, W. Huang, H. Liu, Z. Wang, Y. Wen, and S. Wang, "ST-TGR: Spatio-temporal representation learning for skeleton-based teaching gesture recognition," Sensors, vol. 24, no. 8, p. 2589, 2024, doi: 10.3390/s24082589.

[19] K. Wilson and P. E. Pfeiffer, "Feedback in augmented and virtual reality piano tutoring systems: a mini review," Frontiers in Virtual Reality, vol. 4, no. 1, 1207397, 2023, doi: 10.3389/frvir.2023.1207397.

[20] W. Luo and B. Ning, "Toward piano teaching evaluation based on neural network," Scientific Programming, vol. 2022, no. 1, 6328768, 2022, doi: 10.1155/2022/6328768.
[21] Y. Wang, H. Kang, D. Wu, W. Yang, and L. Zhang, "Global and local spatio-temporal encoder for 3D human pose estimation," IEEE Transactions on Multimedia, vol. 26, no. 1, pp. 4039-4049, 2023, doi: 10.1109/TMM.2023.3321438.
[22] X. Liu and H. Tang, "STRFormer: Spatial–temporal–retemporal transformer for 3D human pose estimation," Image and Vision Computing, vol. 140, no. 1, 104863, 2023, doi: 10.1016/j.imavis.2023.104863.
[23] S. Yang, D. He, Q. Li, J. Wang, and D. Li, "Hand pose estimation based on improved NSRM network," EURASIP Journal on Advances in Signal Processing, vol. 2023, no. 1, p. 8, 2023, doi: 10.1186/s13634-023-00970-y.
[24] X. Zou and X. Bi, "LCFFNet: A Lightweight Cross-scale Feature Fusion Network for human pose estimation," Neural Networks, vol. 183, no. 1, 106959, 2025, doi: 10.1016/j.neunet.2024.106959.
[25] M. Rizwan, S. Ul Haq, N. Gul, M. Asif, S. M. Shah, T. Jan, et al., "Appearance Based Dynamic Hand Gesture Recognition Using 3D Separable Convolutional Neural Network," Computers, Materials & Continua, vol. 76, no. 1, pp. 1-35, 2023, doi: 10.32604/cmc.2023.038211.
[26] R. Jain, R. K. Karsh, and A. A. Barbhuiya, "Encoded motion image-based dynamic hand gesture recognition," The Visual Computer, vol. 38, no. 6, pp. 1957-1974, 2022, doi: 10.1007/s00371-021-02259-3.
[27] D. Cao, W. Liu, W. Xing, and X. Wei, "Human pose estimation based on feature enhancement and multi-scale feature fusion," Signal, Image and Video Processing, vol. 17, no. 3, pp. 643-650, 2023, doi: 10.1007/s11760-022-02271-7.
[28] Y. Ma, Q. Shi, and F. Zhang, "A lightweight Context-aware Feature Transformer Network for human pose estimation," Electronics, vol. 13, no. 4, p. 716, 2024, doi: 10.3390/electronics13040716.
[29] B. Huang and X. Li, "Human Motion Prediction via Dual-Attention and Multi-Granularity Temporal Convolutional Networks," Sensors, vol. 23, no. 12, p. 5653, 2023, doi: 10.3390/s23125653.
[30] J. Ma, Y. Zhang, H. Zhou, H. Yang, and X. Wu, "Multi-granularity spatial temporal graph convolution network with consecutive attention for human motion prediction," Applied Soft Computing, vol. 165, no. 1, 112126, 2024, doi: 10.1016/j.asoc.2024.112126.
[31] K. Y. Lai, C. H. Hsu, Y. C. Lin, C. H. Tsai, C. F. Lin, L. C. Kuo, et al., "Practically Feasible Sensor-Embedded Kinetic Assessment Piano System for Quantifying Striking Force of Digits During Piano Playing," Journal of Medical and Biological Engineering, vol. 43, no. 6, pp. 749-757, 2023, doi: 10.1007/s40846-023-00835-7.
[32] J. Lin, B. Ding, Z. Song, Z. Li, and S. Li, "A Model of Multi-Finger Coordination in Keystroke Movement," Sensors, vol. 24, no. 4, p. 1221, 2024, doi: 10.3390/s24041221.
[33] Y. Liu, T. Zhang, Z. Li, and L. Deng, "Deep learning-based standardized evaluation and human pose estimation: A novel approach to motion perception," Traitement du Signal, vol. 40, no. 5, pp. 2313-2320, 2023, doi: 10.18280/ts.400549.
[34] L. Amadi and G. Agam, "PosturePose: Optimized Posture Analysis for Semi-Supervised Monocular 3D Human Pose Estimation," Sensors, vol. 23, no. 24, p. 9749, 2023, doi: 10.3390/s23249749.
[35] F. Roggio, B. Trovato, M. Sortino, and G. Musumeci, "A comprehensive analysis of the machine learning pose estimation models used in human movement and posture analyses: A narrative review," Heliyon, vol. 10, no. 21, e39977, 2024, doi: 10.1016/j.heliyon.2024.e39977.