International Scientific Technical and Economic Research (ISSN: 2959-1309)
Vol. 4, No. 2, (June 2026), pp. 51-77
Journal homepage: https://www.istaer.online

RESEARCH ARTICLE DOI: 10.71451/ISTAER2615

Research on Transformer-Based Action Sequence Modeling of Intangible Cultural Heritage Shadow Play Using Attention Mechanisms

Yuxiao Liu[1], Shuolei Feng[2], Mengyu Liu[1]*

[1]*Art and Design, Beijing City University, Shunyi District, Beijing, China*

[2]*Department of Information Science, Beijing City University, Shunyi District, Beijing, China*

Abstract: Shadow puppet movements are characterized by long-range spatiotemporal dependence, pronounced stylization, and complex control and transmission relationships, these characteristics pose two major challenges to digital modeling: capturing long-range dependencies and preserving artistic style expression. This paper proposes an improved Transformer model incorporating a multi-level attention mechanism for modeling and generating action sequences of intangible cultural heritage shadow play. The model designs three types of collaborative attention modules: spatial attention introduces bone adjacency priors to enhance structural rationality; temporal attention captures cross-frame long-range dependencies; and style-aware attention adjusts local computations via global feature statistics to preserve genre-specific performance styles. Furthermore, an enhanced architecture alternately stacking graph convolution and Transformer is adopted, and sparse and hierarchical modeling strategies are introduced to reduce computational complexity from quadratic to approximately linear in sequence length. The experimental results show that the average joint position error of the proposed method in motion prediction tasks is 31.4, which is 11.8 lower than that of the standard Transformer; Style loss decreased by 24.6%; Under the extreme condition of 50% missing key points, the error ratio is 1.31, which is significantly better than the comparison method. The proposed model provides effective technical support for the digital protection and intelligent inheritance of intangible cultural heritage.

Keywords: Intangible cultural heritage digitization; Shadow play; Action sequence modeling; Transformer; Multi-level attention mechanism

How to Cite: Liu, Y., Feng, S., & Liu, M. (2026). Research on Transformer-Based Action Sequence Modeling of Intangible Cultural Heritage Shadow Play Using Attention Mechanisms. *International Scientific Technical and Economic Research, 4*(2), 51–77. https://doi.org/10.71451/ISTAER2615

Article history: Received: 15 Jan 2026; **Revised:** 24 Feb 2026; **Accepted:** 28 Mar 2026; **Published:** 08 Apr 2026

1. INTRODUCTION

As an important representative of China's traditional intangible cultural heritage, shadow play integrates sculpture, singing, and manipulation techniques. Its core artistic charm lies in

*Corresponding author: Mengyu Liu, Art and Design, Beijing City University, Shunyi District, Beijing, China. Email: daimou971231@naver.com

the performer's delicate manipulation of shadow figures using bamboo sticks, presenting dynamic action narratives on a two-dimensional screen. However, with the aging of the older generation of artists and the aesthetic changes of the young audience, the living transmission of shadow play faces severe challenges [1]. Digital modeling of shadow play movements can not only record and reproduce this ancient art with high fidelity, but also open up a new path for the intelligent protection, interactive display and virtual inheritance of intangible cultural heritage [2]. Compared with conventional human body movements, shadow puppet movements have distinct uniqueness: it is essentially an indirect control. The bamboo sticks and shadow puppet characters form a unique transmission relationship. The movement style is highly stylized and the genre differences are significant. At the same time, continuous performance is often accompanied by long-time posture maintenance and instant explosive action switching. These characteristics put forward higher requirements for digital modeling.

The existing motion sequence modeling methods exposed obvious deficiencies in dealing with complex intangible cultural heritage movements such as shadow play [3],[4]. Although recurrent neural network-based models can capture temporal dependencies, their recursive structure suffers from the vanishing gradient problem, making it difficult to effectively model long-range dependencies in continuous actions lasting tens of seconds, and they are particularly unable to accurately correlate similar action patterns that are far apart [5],[6]. Although the graph convolution network can use the bone structure information, its receptive field is limited by the number of graph convolution layers, and its modeling ability is limited to the global space-time dependence. It often divides the movements into the combination of local joint movements, which is difficult to restore the overall rhythmic sense of shadow play movements [7]. More importantly, the existing methods generally lack the ability to explicitly model the artistic style. They pursue numerical accuracy of the motion trajectory rather than fidelity of the performance style [8]. As a result, the generated actions are “similar in form” but “different in essence”, which cannot reflect the subtle differences in motion amplitude and rhythm between different shadow puppet genres. In addition, when the standard sequence model is faced with a long shadow play sequence, the computational complexity increases squarely with the sequence length, which is difficult to achieve efficient reasoning in practical applications [9].

In recent years, the Transformer architecture has achieved great success in natural language processing and computer vision by virtue of its self-attention mechanism, offering a new paradigm for action sequence modeling [10],[11],[12],[13],[14]. Its core advantage is that it can directly calculate the dependence between any two positions in the sequence, which is naturally suitable for capturing the long-range patterns in shadow play movements, such as the posture echo of the opening appearance and the end stop, and the periodicity of repetitive dance steps. At the same time, the interpretability of attention mechanism also provides the possibility of analyzing the movement patterns acquired by the model [15],[16],[17]. However, the direct application of Transformer in shadow puppet movement modeling faces three major challenges: first, the standard self-attention treats all time steps equally, lacks a priori guidance of bone spatial structure, and is easy to ignore the physical connection constraints between joints; Second, the computational complexity of attention increases with the square of sequence length, which makes it difficult to deal with hundreds of frames of continuous action in shadow play; Third, the standard Transformer does not have the ability of explicit style modeling, and it is difficult to distinguish the performance characteristics of different schools.

To solve the above problems, this paper proposes a set of improved Transformer model with multi-level attention mechanism, and its core contributions are reflected in four aspects. First, a multi-level attention fusion mechanism is designed, including spatial attention based on bone adjacency priors to enhance structural rationality, temporal attention to capture long-range dependencies for modeling long-sequence action patterns, and an innovative style-aware attention. By extracting global action features to adjust local attention calculation, the model can learn and maintain the unique performance styles of different shadow puppet genres. Secondly, a graph structure enhanced Transformer architecture is proposed, which alternately

stacks graph convolution operation and self-attention mechanism, so that the model has the ability of bone topology constraint and global dependency modeling at the same time, and the generated joint trajectory is more consistent with the physical linkage of shadow puppet characters in structure. Thirdly, aiming at the challenge of long sequence modeling, the sparse and hierarchical modeling strategy for long sequence is introduced. By sparse attention, only the interaction between key time steps is calculated, and combined with hierarchical structure, the collaborative modeling of local details and global semantics is realized, which reduces computational complexity from quadratic to approximately linear. Fourth, through systematic experimental verification, on the self-built large-scale shadow play action data set, the method was comprehensively evaluated from multiple dimensions such as motion prediction accuracy, style fidelity, model generalization ability and computational efficiency. The experimental results fully proved the advancement and practical value of this method in the task of digital modeling of intangible cultural Heritage shadow play action.

2. RELATED WORK

Motion sequence modeling is a basic problem in the field of computer vision and graphics. For a long time, three technical routes have been formed, represented by cyclic neural network, graph convolution network and Transformer. Recurrent neural networks and their variants, such as LSTM and GRU, transmit temporal information frame by frame through hidden states, and once dominated early action recognition and prediction tasks [18],[19]. The advantage of this kind of method is that it can deal with variable length sequences naturally, but its recursive structure determines that the information must be transferred gradually along the time chain [20]. When the length of the sequence exceeds tens of frames, the long-distance dependent information is easy to be attenuated or lost in the transmission process, that is, the so-called vanishing gradient problem. Although LSTM alleviates this dilemma to a certain extent through the gating mechanism, its essential sequence dependence makes it difficult to parallelize the training, and it is still difficult to accurately capture the movement echo relationship that may span hundreds of frames in shadow play [21]. The graph convolution network starts from the perspective of spatial structure, modeling the human skeleton as a graph structure, and learning the spatial characteristics of actions by transferring information between adjacent joints [22],[23],[24]. Based on this, spatiotemporal graph convolution network expands the convolution operation of time dimension and becomes one of the mainstream methods of bone motion recognition. However, the receptive field of convolution is limited by the kernel size and the number of convolutional layers. Although the receptive field can be expanded by stacking multiple layers, excessive deepening of the network leads to oversmoothing, that is, the characteristics of different joints tend to be homogeneous. This makes the ability of graph convolution network in modeling the remote spatial dependence between non adjacent joints in shadow play limited, for example, the cooperative motion mode between wrist and ankle is difficult to be effectively captured.

The introduction of the Transformer architecture has brought a paradigm shift to action sequence modeling. The self-attention mechanism enables the model to directly calculate the interaction weight between any two positions in the sequence within a single layer, which fundamentally solves the problem of long-distance dependency modeling. In the field of motion recognition, the work of motion Transformer and video Transformer has proved the superiority of Transformer in capturing global spatio-temporal patterns [25],[26]. In the task of motion generation and prediction, the Transformer based method has also made significant progress, which can generate more coherent and realistic human motion than the cyclic neural network. However, standard Transformer has two inherent limitations when applied to motion modeling. First, the computational complexity of self-attention increases squarely with the length of the sequence, which limits its application in real-time or interactive scenes. Second, the Transformer treats the input as a set of unordered tokens. Although the timing information is retained through the position coding, for the spatial input with a clear topological structure,

such as bones, the Transformer lacks a priori guidance for the physical connection relationship, which leads to the need for the model to learn the correlation between joints from the data by itself. The learning efficiency is low and it is easy to over fit.

The development of attention mechanism itself in time sequence and action modeling is also worth paying attention to. The early attention mechanism is often used as a supplementary module of the recurrent neural network, which is used to give different weights to different positions of the input sequence to help the model focus on key frames or key regions. With the rise of Transformer, self-attention has become the mainstream modeling paradigm, and researchers have further explored a variety of improved attention mechanisms [27]. Restricted self-attention reduces complexity by limiting each position's attention range, e.g., only computing attention within a local window, but this sacrifices the ability to model long-range dependencies. Axial attention decomposes two-dimensional attention into row and column directions, which reduces the complexity but maintains the global receptive field [28]. In the field of motion modeling, someone proposed a cross time step sparse attention strategy, which only calculated the attention of the sampled key frames, effectively reducing the amount of calculation. However, most of the existing methods focus on the optimization of computational efficiency, ignoring the potential of attention mechanism in style modeling. Style information is usually represented by global statistical features rather than local temporal or spatial relationships. How to design attention mechanisms to enable it to perceive and maintain the overall style attributes of the sequence is an issue that has not been fully explored.

In the field of digital protection of intangible cultural heritage, scholars at home and abroad have carried out a lot of research work. The early digitization of intangible cultural heritage mainly focused on the three-dimensional reconstruction and high-precision archiving of static cultural relics, and then gradually extended to the recording and reproduction of performance intangible cultural heritage. For traditional drama and dance intangible cultural heritage, researchers have attempted to use motion capture technology to record performer action data and then play back and display them using 3D animation technology [29],[30]. However, high-precision optical motion capture equipment is expensive and requires a controlled environment. Although the video based attitude estimation algorithm is convenient, the extracted key point sequence often contains noise and missing values. At the level of motion modeling, most of the research focuses on the direct recording and simple interpolation playback of the original motion data, lacking the ability to understand and model the inherent regularities and artistic styles of movement. Some studies try to use statistical models or hidden Markov models to identify and classify specific action patterns, but these models have limited expression ability and are difficult to describe the richness and style diversity of intangible cultural heritage actions. In recent years, deep learning methods have been applied to the modeling of intangible cultural heritage movements, such as generating traditional dance movements using generative confrontation networks, or predicting opera movements using recurrent neural networks. However, most of these works follow the general action modeling framework, and there is no special design for the particularity of intangible cultural heritage action, especially in style expression and long-range structure modeling.

Based on the above analysis, the existing methods have three common problems in modeling the action sequence of intangible cultural heritage skin shadow play. First, there is a lack of effective modeling ability for the long-range spatio-temporal dependence in motion. It is difficult to capture the motion patterns spanning dozens of frames and the collaborative relationship between non adjacent joints at the same time, whether it is the gradient attenuation problem of the recurrent neural network or the limited receptive field of the graph convolution network. Second, the explicit modeling of artistic style is generally ignored. The existing methods aim at minimizing the joint position error and pursue the numerical accuracy of the trajectory rather than the fidelity of the performance style. As a result, the generated movements lack the unique charm and expressiveness of traditional shadow puppets although the joint position is accurate. Third, it is difficult to achieve a balance between computational efficiency and modeling ability. Although the standard Transformer has strong global modeling ability, its

computational complexity is too high, and the lightweight model often sacrifices the expression ability. Aiming at the above three key problems, this paper proposes an improved Transformer model which integrates multi-level attention mechanism. While maintaining efficient calculation, it realizes the integrated modeling of shadow play action time-space dependence and artistic style.

3. DATA SET AND PREPROCESSING

In order to support the high-quality modeling of shadow play action sequence, this paper constructs a set of structured data processing process for intangible cultural heritage action expression, including action collection, key point extraction, bone modeling, time series standardization and multi-dimensional data enhancement. First of all, in the data acquisition stage, the traditional shadow play performances are recorded with multi-view HD camera equipment (resolution 1920 × 1080, frame rate 60fps), and the key point sequence is extracted with attitude estimation algorithm (such as OpenPose/HRNet). Let the original video sequence be $V = \{I_t\}_{t=1}^{T}$, where I_t represents the image of the t-th frame and T is the total number of frames. Through the key detection function $\Phi(\cdot)$, the set of action key points is obtained:

$$X_t = \Phi(I_t) = \left\{ x_t^{(i)} \in \mathbb{R}^2 \,\middle|\, i = 1,2,\dots,N \right\} \tag{1}$$

Where, $x_t^{(i)} = (u_t^{(i)}, v_t^{(i)})$ represents the two-dimensional coordinates of the i-th joint in the t-th frame, and N is the number of joint points (set as 18 key joints in this paper). The confidence weight $c_t^{(i)} \in [0,1]$ is further introduced to form a weighted key point representation:

$$\tilde{x}_t^{(i)} = c_t^{(i)} \cdot x_t^{(i)} \tag{2}$$

Used for subsequent outlier filtering and smoothing. In order to improve the reliability of data, the low confidence point ($c_t^{(i)} < 0.5$) is interpolated and repaired. The linear interpolation form is:

$$x_t^{(i)} = \frac{x_{t-1}^{(i)} + x_{t+1}^{(i)}}{2} \tag{3}$$

In order to visually display the statistical characteristics of data collection and key point extraction, **Table 1** shows the scale and quality distribution of collected data under different plays.

Table 1. Statistical information of shadow play action dataset

Play category	Number of videos	Average duration (s)	Total frames (×10³)	Keys/frame	Effective key point rate (%)	Average confidence
Martial arts	120	35.2	253.4	18	94.6	0.87
Literary drama	98	42.7	251.2	18	92.1	0.84
Mythological category	85	38.9	198.3	18	95.3	0.89
History class	76	40.5	184.6	18	93.8	0.86
Folk customs	64	33.1	127.1	18	91.7	0.82
Comprehensive category	52	36.8	114.9	18	93.2	0.85

From the perspective of data coverage, the dataset contains five types of plays, including martial arts, literary drama, and mythology, with a total of more than 110000 frames, ensuring the diversity of model training. In terms of data quality, the effective key point rate for all categories was above 91%, and the average confidence was higher than 0.82. Among them, mythological plays performed best, with the effective key point rate of 95.3%, and the average confidence was 0.89. These high confidence data provide a reliable basis for subsequent model training.

In the aspect of bone structure modeling, this paper uses graph structure to model the action, abstracts the human body (shadow figure) as an undirected graph $G = (V, E)$, where the node set V represents the joints, and the edge set E represents the bone connection relationship. The adjacency matrix is defined as:

$$A_{ij} = \begin{cases} 1, & \text{If joint "I" is connected to "J"} \\ 0, & \text{Otherwise} \end{cases} \tag{4}$$

Further define the normalized adjacency matrix:

$$\hat{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \tag{5}$$

Where D is the degree matrix. Action sequences can be expressed as three-dimensional tensors:

$$\mathcal{X} \in \mathbb{R}^{T \times N \times C} \tag{6}$$

Where $C = 2$ represents the two-dimensional coordinate channel. In order to enhance the motion expression, the velocity feature is introduced:

$$v_t^{(i)} = x_t^{(i)} - x_{t-1}^{(i)} \tag{7}$$

And acceleration characteristics:

$$a_t^{(i)} = v_t^{(i)} - v_{t-1}^{(i)} \tag{8}$$

The final fusion representation:

$$z_t^{(i)} = \left[x_t^{(i)}, v_t^{(i)}, a_t^{(i)} \right] \in \mathbb{R}^6 \tag{9}$$

Significantly enhances dynamic modeling capability.

In the process of timing alignment and standardization, in order to eliminate the impact of different video lengths, a temporal resampling strategy is used to map all sequences to a fixed length $T' = 128$. The resampling function is defined as:

$$X'_t = X_{\left\lfloor t \cdot \frac{T}{T'} \right\rfloor} \tag{10}$$

At the same time, the spatial coordinates are normalized:

$$\bar{x}_t^{(i)} = \frac{x_t^{(i)} - \mu}{\sigma} \tag{11}$$

Where μ, σ are the global mean and standard deviation respectively. In addition, in order to eliminate the difference between translation and scale, the relative coordinate representation is introduced:

$$x_t^{(i)} = x_t^{(i)} - x_t^{(root)} \tag{12}$$

Where $x_t^{(root)}$ is the root node (torso Center). This processing significantly improves the robustness of the model to different performance scales.

To verify the effectiveness of standardized processing, Table 2 compares the data distribution characteristics before and after standardization.

Table 2. Statistical comparison before and after data standardization

Index	Before standardization	After standardization
Coordinate mean	312.6	0.01
Coordinate variance	14582.3	1.02
Maximum	1024.5	3.21
Minimum value	12.3	-3.08
Different video scales	more	Significantly lower
Model convergence rate (epoch)	48	31
Final error (MPJPE)	42.7	35.2

After standardization, the mean and variance of the data coordinates are optimized from 312.6 and 14582.3 to 0.01 and 1.02, respectively, which effectively eliminates the scale differences between different videos. This improvement directly leads to the double improvement of training efficiency and model accuracy: the number of epochs required for model convergence is reduced from 48 to 31, a decrease of about 35%; The final prediction error (MPJPE) decreased from 42.7 to 35.2, a relative reduction of 17.6%.

In terms of data enhancement, in order to improve the generalization ability of the model to complex movements and style changes, multidimensional enhancement strategies are designed. Time perturbation is achieved by random time scaling:

$$t' = \alpha t, \alpha \sim \mathcal{U}(0.8,1.2) \tag{13}$$

Spatial disturbance is realized by Gaussian noise:

$$\tilde{x}_t^{(i)} = x_t^{(i)} + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{14}$$

Where $\sigma = 0.01$. In addition, style perturbation is introduced to simulate different performance styles through linear transformation:

$$x' = Sx + b \tag{15}$$

Where S is the style transformation matrix and b is the offset vector. This method can simulate the movement differences of different shadow puppet genres (such as Shaanxi and Hebei styles).

In summary, the data processing system constructed in this section not only ensures the data quality, but also enhances the ability of action expression and model robustness, providing a solid data foundation for subsequent Transformer modeling.

4. METHODS: TRANSFORMER MODEL INTEGRATING ATTENTION

In order to fully model the complex time-space dependence and artistic style expression in shadow play action sequence, this paper proposes an improved Transformer model which integrates multi-level attention mechanism. The whole framework takes the structured bone sequence as the input, and realizes the unified modeling of motion dynamics and style features through embedded coding, multi-layer improved Transformer module and task driven output layer [31],[32]. Let the input motion sequence be $\mathcal{X} \in \mathbb{R}^{T\times N\times C}$, where T is the length of time, N is the number of joints, and C is the characteristic dimension (including position, speed, etc.). The model first maps the input code:

$$H_0 = f_{\text{embed}}(\mathcal{X}) \in \mathbb{R}^{T\times N\times d} \tag{15}$$

Where $f_{\text{embed}}(\cdot)$ represents the linear embedding function, and d is the hidden dimension.

In the input encoding stage, to enhance the representation of spatial structure and temporal order information, this paper designs a joint embedding and spatiotemporal position encoding mechanism. Specifically, for the ith joint in frame t, its embedding is expressed as:

$$e_t^{(i)} = W_e z_t^{(i)} + b_e \tag{16}$$

Where $z_t^{(i)} \in \mathbb{R}^C$ is the input feature, $W_e \in \mathbb{R}^{d\times C}$, $b_e \in \mathbb{R}^d$ is the learnable parameter. In order to preserve the spatio-temporal location information, the decomposed location coding is introduced:

$$p_{t,i} = p_t^{(time)} + p_i^{(space)} \tag{17}$$

Where $p_t^{(time)} \in \mathbb{R}^d$ represents temporal position encoding in the form of a sine function:

$$p_t^{(2k)} = \sin\left(\frac{t}{10000^{\frac{2k}{d}}}\right), p_t^{(2k+1)} = \cos\left(\frac{t}{10000^{\frac{2k}{d}}}\right) \tag{18}$$

$p_i^{(space)}$ represents the spatial position coding of joints, which is generated by the skeleton topology index. The final input is expressed as:

$$H_t^{(i)} = e_t^{(i)} + p_{t,i} \tag{19}$$

The design effectively integrates the timing information and structure information of the action.

In the backbone network, the model uses a multi-layer stacked improved Transformer encoder, and each layer contains a multi head attention module and a feedforward network. The standard self-attention calculation is:

$$\text{Attention}(Q,K,V) = \text{Softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{20}$$

Where $Q = HW_Q$, $K = HW_K$, $V = HW_V$, $W_Q, W_K, W_V \in \mathbb{R}^{d\times d_k}$ is the projection matrix. The output passes through the feedforward network:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)\,W_2 + b_2 \tag{21}$$

Where $W_1 \in \mathbb{R}^{d\times d_f}$, $W_2 \in \mathbb{R}^{d_f\times d}$. Finally, the layer output is obtained by residual connection and layer normalization. The output layer is designed as a regression or generation structure according to different tasks, such as in the action prediction task:

$$\hat{Y} = HW_o + b_o \tag{22}$$

Where $\hat{Y}$ is the predicted action sequence, $W_o \in \mathbb{R}^{d\times C}$.

In the design of multi-level attention mechanism, this paper focuses on introducing spatial attention, temporal attention and style perception attention to improve the modeling ability of the model for complex motion expression. Spatial attention aims to capture structural dependencies between joints, and its weight computation incorporates bone adjacency information:

$$\alpha_{ij}^{(t)} = \frac{\exp\left((q_t^{(i)})^{\top}k_t^{(j)} + \lambda A_{ij}\right)}{\sum_j \exp\left((q_t^{(i)})^{\top}k_t^{(j)} + \lambda A_{ij}\right)} \tag{23}$$

Where $q_t^{(i)}, k_t^{(j)}$ are query and key vectors respectively, A_{ij} is the adjacent matrix element, and λ is the structure weight coefficient. This mechanism strengthens the information interaction between physically connected joints.

Temporal attention is used to model long-sequence dependencies. Its core lies in information aggregation across time steps:

$$\beta_{tt'}^{(i)} = \frac{\exp\left(\left(q_t^{(i)}\right)^{\top}k_{t'}^{(i)}\right)}{\sum_{t'}\exp\left(\left(q_t^{(i)}\right)^{\top}k_{t'}^{(i)}\right)} \tag{24}$$

Where $\beta_{tt'}^{(i)}$ represents the dependent weight of joint i at time t and t'. This mechanism can capture long-distance movement patterns, such as continuous dancing or repetitive movements.

To further model the unique artistic style of shadow play, this paper proposes a style-aware attention mechanism. First, define the global style vector:

$$s = \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N} H_t^{(i)} \tag{25}$$

Then it is introduced into attention Computing:

$$\gamma_{ij}^{(t)} = \frac{\exp\left((q_t^{(i)})^{\top}k_t^{(j)} + (W_s s)^{\top}k_t^{(j)}\right)}{\sum_j \exp(\cdot)} \tag{26}$$

Where $W_s \in \mathbb{R}^{d\times d}$ is the style mapping matrix. This mechanism enables the model to perceive the overall action style and adjust the local action expression, so as to better restore

the intangible cultural heritage performance characteristics.

In terms of architecture improvement, this paper further proposes a hierarchical Transformer structure to achieve the collaboration of local and global modeling. First, the sequence is divided into several subsequences:

$$\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_M\} \tag{27}$$

The length of each subsequence is $L = T/M$. The local Transformer models each subsequence separately:

$$H_m = \text{Transformer}_{local}(\mathcal{X}_m) \tag{28}$$

Then integrate through global Transformer:

$$H_{global} = \text{Transformer}_{global}([H_1, H_2, \dots, H_M]) \tag{29}$$

This structure effectively reduces the difficulty of long sequence modeling.

In order to reduce the computational complexity, the sparse attention mechanism is introduced, and only the attention between the key time steps is calculated:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK_\Omega^\top}{\sqrt{d_k}}\right)V_\Omega \tag{30}$$

Where Ω represents the selected set of key indices (such as key frames), whose size is much smaller than T, thereby reducing complexity from $O(T^2)$ to $O(T\log T)$ or $O(T)$.

In addition, in order to make full use of the bone topology, this paper designs a graph structure enhancement module. Features are enhanced by graph convolution:

$$H' = \sigma\left(\hat{A}HW_g\right) \tag{31}$$

Where $\hat{A}$ is the normalized adjacency matrix, $W_g \in \mathbb{R}^{d\times d}$ is the weight matrix, and $\sigma(\cdot)$ is the activation function. This module is alternately stacked with the Transformer, endowing the model with both graph structure modeling and global dependency modeling capabilities.

To sum up, the integrated attention Transformer model proposed in this paper effectively improves the modeling ability of the complex space-time relationship and artistic style in the shadow play action sequence through multi-level attention mechanism and structural innovation, and provides a solid foundation for subsequent experimental verification.

5. LOSS FUNCTION AND TRAINING STRATEGY

To ensure optimal model performance across multiple dimensions—motion reconstruction accuracy, dynamic continuity, structural rationality, and artistic style expression—this paper constructs a multi-objective joint optimization loss function system, and carries out parameter learning combined with stable and efficient training strategies. The overall optimization objective is defined as:

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_{rec} + \lambda_2\mathcal{L}_{pred} + \lambda_3\mathcal{L}_{smooth} + \lambda_4\mathcal{L}_{bone} + \lambda_5\mathcal{L}_{style} \tag{32}$$

Where $\lambda_1 \sim \lambda_5$ is the weight coefficient of each loss term, which is used to balance the importance of different optimization objectives.

For basic reconstruction and prediction, this paper adopts mean squared error (MSE) as the core metric. For the input sequence $\mathcal{X} = \{x_t^{(i)}\}$ and the model output $\hat{\mathcal{X}} = \{\hat{x}_t^{(i)}\}$, the reconstruction loss is defined as:

$$\mathcal{L}_{rec} = \frac{1}{TN} \sum_{t=1}^{T} \sum_{i=1}^{N} \| x_t^{(i)} - \hat{x}_t^{(i)} \|_2^2 \tag{33}$$

Where T is the number of time steps, N is the number of joints, $x_t^{(i)} \in \mathbb{R}^C$ is the real key point position, and $\hat{x}_t^{(i)}$ is the predicted value. For the future action prediction task, the prediction loss is introduced:

$$\mathcal{L}_{pred} = \frac{1}{T'N} \sum_{t=T+1}^{T+T'} \sum_{i=1}^{N} \| x_t^{(i)} - \hat{x}_t^{(i)} \|_2^2 \tag{34}$$

Where T' is the length of the prediction time window. The loss strengthens the learning ability of the model to the future dynamic change trend.

In order to ensure the continuity and naturalness of the action in the time dimension, this paper introduces the action smoothing constraint. By constraining the speed change between adjacent time steps, the smoothing loss is defined as:

$$\mathcal{L}_{smooth} = \frac{1}{(T-1)N} \sum_{t=2}^{T} \sum_{i=1}^{N} \| (x_t^{(i)} - x_{t-1}^{(i)}) - (\hat{x}_t^{(i)} - \hat{x}_{t-1}^{(i)}) \|_2^2 \tag{35}$$

This constraint encourages the generated motion to be consistent with real data in terms of velocity changes, thereby avoiding jitter or discontinuity.

In terms of structural consistency, considering the strict bone connection relationship of shadow puppet movements, this paper designs the bone length constraint loss. Let there be a connection between joint I and j, and the bone length is defined as:

$$d_t^{(i,j)} = \| x_t^{(i)} - x_t^{(j)} \|_2 \tag{36}$$

Then the bone constraint loss is:

$$\mathcal{L}_{bone} = \frac{1}{T \mid E \mid} \sum_{t=1}^{T} \sum_{(i,j) \in E} \mid d_t^{(i,j)} - \hat{d}_t^{(i,j)} \mid \tag{37}$$

Where E is the set of bone edges, $\mid E \mid$ is the number of edges, and $\hat{d}_t^{(i,j)}$ is the predicted bone length. This loss effectively prevents unreasonable limb deformation.

To address the unique artistic forms of intangible cultural heritage shadow play, this paper introduces a style consistency loss as a core innovation. Firstly, the style of action sequence is defined as the global feature statistics, where $H_t^{(i)} \in \mathbb{R}^d$ is the middle layer feature representation of the model. The corresponding prediction sequence style is expressed as $\hat{s}$. The loss of style consistency is defined as:

$$\mathcal{L}_{style} = \| s - \hat{s} \|_2^2 \tag{38}$$

In addition, to further capture the statistical characteristics of the style, the covariance

matrix constraint is introduced:

$$\Sigma = \frac{1}{TN}\sum(H_t^{(i)} - s)(H_t^{(i)} - s)^\top \tag{39}$$

The final style loss is extended to:

$$\mathcal{L}_{style} = \| s - \hat{s} \|_2^2 + \| \Sigma - \hat{\Sigma} \|_F^2 \tag{40}$$

Where $\|\cdot\|_F$ is the Frobenius norm. This design enables the model not only to learn movement trajectories but also to maintain consistency in style distribution.

In terms of optimization strategy, this paper uses the multi task joint training framework to optimize the model parameters through the above multiple losses. AdamW optimizer is used during training, and its parameter update rule is:

$$\theta_{t+1} = \theta_t - \eta\left(\frac{m_t}{\sqrt{v_t} + \epsilon} + \lambda\theta_t\right) \tag{41}$$

Where θ_t is the model parameter, η is the learning rate, m_t, v_t are the first-order and second-order moment estimates, respectively, and λ is the weight attenuation coefficient. Learning rate adopts cosine annealing strategy:

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta_{max} - \eta_{min})\left(1 + \cos\frac{t\pi}{T_{max}}\right) \tag{42}$$

To improve training stability and convergence speed.

In addition, to avoid the inconsistency of gradient scales of different loss terms, a dynamic weight adjustment mechanism is introduced:

$$\lambda_k = \frac{1}{\sigma_k^2} \tag{43}$$

Where σ_k is the uncertainty parameter of the kth task, which is obtained through learning. This strategy can adaptively balance the multi task training process.

On the whole, the optimization system proposed in this section not only ensures the movement accuracy, but also takes into account the structural rationality and artistic style expression. Through multi task joint training and dynamic weight adjustment, the unified improvement of model performance and stability is achieved.

6. EXPERIMENTAL SETUP

In order to comprehensively verify the effectiveness and superiority of the proposed Transformer model with attention mechanism in the modeling of the action sequence of intangible cultural heritage shadow play, this paper constructs a systematic experimental setup, including the computing environment, model parameters, comparison method selection, evaluation index system and multi task experimental configuration. All experiments are conducted in a unified environment to ensure reproducibility and fairness of results. Specifically, the experimental platform is based on NVIDIA RTX4060 GPU and is implemented by PyTorch deep learning framework. CUDA version is 11.8. The batch size $B = 32$ is used for model training, the initial learning rate is set as $\eta_0 = 1 \times 10^{-4}$, and the weight attenuation coefficient is $\lambda = 1 \times 10^{-5}$. The hidden dimension of the model is set as $d = 256$, the number of multiple attention heads is $h = 8$, and the number of Transformer layers is $L = 6$. The input sequence length is unified as $T = 128$. In the training process, a gradient clipping strategy is used to limit the gradient norm:

$$\| \nabla\theta \|_2 \leq \tau \tag{44}$$

Where θ is the model parameter, and the threshold $\tau = 1.0$ is used to prevent the gradient explosion problem. The convergence criterion of the model is based on the loss of verification set. When there is no improvement in 10 epochs, the training is stopped in advance.

For comparison, this paper selects three representative mainstream models: LSTM model based on cyclic structure, ST-GCN model based on graph convolution and standard Transformer model. The LSTM model models the time dependency recursively, and its hidden state is updated as:

$$h_t = \text{LSTM}(x_t, h_{t-1}) \tag{45}$$

Where h_t is the current hidden state and x_t is the input feature. ST-GCN model uses graph convolution to process bone structure, and its core calculation is:

$$H' = \sum_k \hat{A}_k H W_k \tag{46}$$

Where $\hat{A}_k$ is the k-th normalized adjacency matrix, and W_k is the corresponding weight matrix. The standard Transformer uses the global self attention mechanism, and its computational complexity is $O(T^2 d)$. Where T is the sequence length and d is the feature dimension. Through the above comparison model, the advantages of this method can be evaluated from different modeling paradigms (temporal recursion, graph structure modeling, global attention).

For evaluation metrics, this paper constructs a multi-dimensional quantitative system to evaluate from three aspects: accuracy, dynamic quality and computational efficiency. First, the average joint position error (MPJPE) is used to measure the prediction accuracy, which is defined as:

$$\text{MPJPE} = \frac{1}{TN} \sum_{t=1}^{T} \sum_{i=1}^{N} \| x_t^{(i)} - \hat{x}_t^{(i)} \|_2 \tag{47}$$

Where $x_t^{(i)}$ is the real joint position and $\hat{x}_t^{(i)}$ is the predicted value. Secondly, in order to evaluate the fluency of movement, the speed consistency index is introduced:

$$\text{Smooth} = \frac{1}{(T-1)N} \sum_{t=2}^{T} \sum_{i=1}^{N} \| v_t^{(i)} - \hat{v}_t^{(i)} \|_2 \tag{48}$$

Where $v_t^{(i)} = x_t^{(i)} - x_{t-1}^{(i)}$, represents the real speed, and $\hat{v}_t^{(i)}$ is the predicted speed. The lower the indicator, the smoother the action. In addition, in terms of efficiency evaluation, the number of model parameters and floating-point operations (flops) are used as indicators, and the model complexity can be expressed as:

$$\text{FLOPs} \approx 2T^2 d + 4Td^2 \tag{49}$$

It is used to measure the difference between different methods in calculating resource consumption.

In order to more intuitively show the configuration differences of different models, **Table 3** gives the main parameter settings of each comparison method.

Table 3. Comparison of parameter configurations for different models

Model	Number of layers	Hide dimensions	Parameter quantity (m)	Flops (g)	Enter length
LSTM	3	256	3.2	12.5	128
ST-GCN	9	256	3.8	15.2	128
Transformer	6	256	5.6	21.3	128
Proposed	6	256	4.9	16.8	128
Sparse Transformer	6	256	4.9	9.7	128
Hierarchical model	6	256	5.1	11.3	128

Under the condition of similar parameters (about 3.2m to 5.6m), the parameter of the complete model proposed in this paper is 4.9M, which is lower than 5.6m of the standard Transformer, but higher than 3.2m of LSTM. More importantly, by introducing sparse attention and hierarchical modeling strategy, the derived model of this method has significant advantages in computational complexity (flops). The flops of sparse Transformer is only 9.7g, which is 54.5% lower than that of standard Transformer (21.3g); The hierarchical model also reduced by 46.9%, which reflects the effectiveness of the architecture in efficiency optimization.

In terms of experimental task setting, this paper evaluates the performance of the model from three perspectives: motion reconstruction, motion prediction and motion generation. In the reconstruction task, the model inputs the complete sequence and reconstructs the original action, with the goal of minimizing the reconstruction error; In the prediction task, input the previous T_{obs} frame and predict the future T_{pred} frame, which is defined as:

$$\hat{X}_{T_{obs}+1:T_{obs}+T_{pred}} = f\left(X_{1:T_{obs}}\right) \tag{50}$$

Where $f(\cdot)$ represents the model mapping function. In the generation task, the action sequence is generated by randomly initializing a latent vector $z \sim \mathcal{N}(0, I)$:

$$\hat{X} = G(z) \tag{51}$$

Where $G(\cdot)$ is the generation module. This task is mainly used to evaluate the ability of the model in style expression and diversity.

7. RESULTS AND ANALYSIS

Under the unified experimental setup, the proposed model is systematically evaluated on multiple tasks and indicators. Firstly, the performance of different methods in motion prediction task is compared from the perspective of overall performance. The average error reduction rate is defined as:

$$\Delta_{imp} = \frac{E_{baseline} - E_{model}}{E_{baseline}} \times 100\% \tag{52}$$

Where, $E_{baseline}$ represents the baseline model error, and E_{model} represents the method error in this paper. This indicator is used to quantify the performance improvement. **Figure 1** shows the change trend of MPJPE when different methods predict the increase of time step.

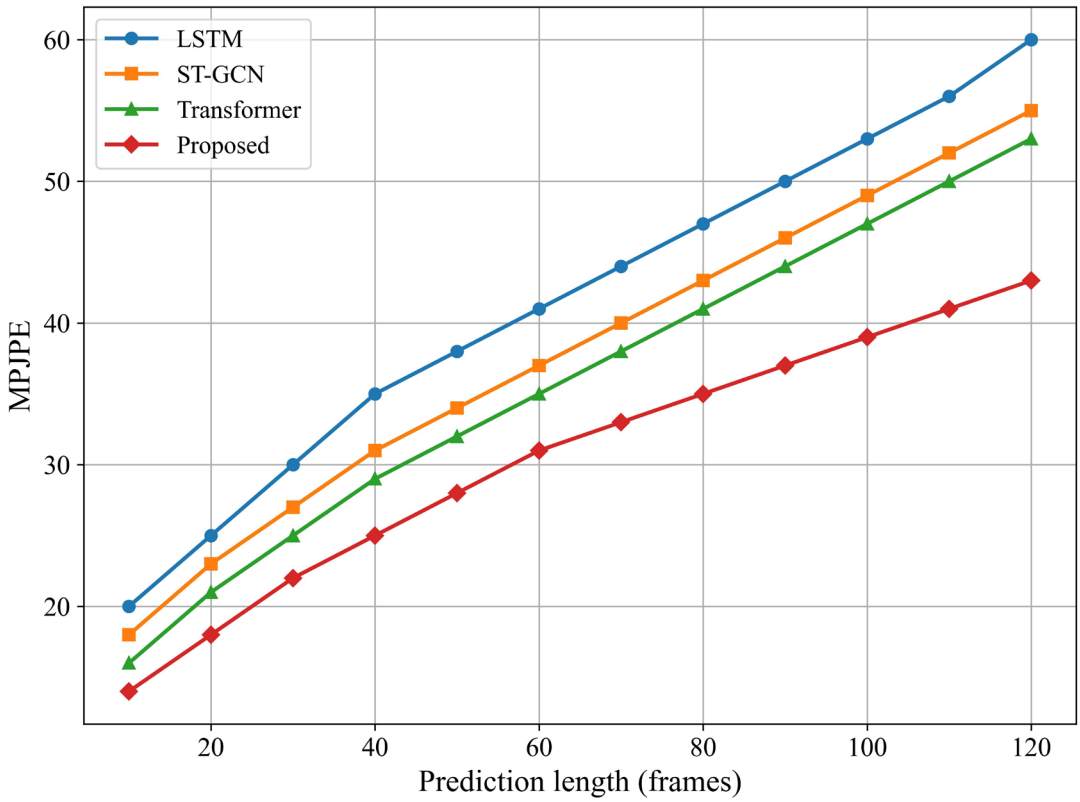

Figure 1. MPJPE curves of different models with prediction step size: the horizontal axis is the number of prediction frames, and the vertical axis is MPJPE

As the number of predicted frames increases from 10 to 120, the errors of all models exhibit an upward trend, but there are significant differences in the growth rate. LSTM model error rose steepest, from about 20 to nearly 60, indicating that it is prone to error accumulation in long series modeling. The performance of ST-GCN and standard Transformer was relatively stable, but still increased significantly in the late stage. However, the overall curve of this method is always at the bottom, and the growth slows down in the middle and rear segments. The error at 120 frames is about 43, which is approximately 10 units lower than that of the Transformer, a reduction of nearly 20%. This shows that this model has better stability and error control ability in long time series prediction.

The main experimental results are further quantified by Table 4.

Table 4. Performance comparison of main experiment (action prediction task)

Model	MPJPE↓	Smooth↓	Style loss↓	Promotion rate (%)
LSTM	41.8	12.5	0.082	-
ST-GCN	38.2	10.7	0.075	8.6
Transformer	35.6	9.8	0.069	14.8
Proposed	31.4	8.2	0.052	24.9
Hierarchical model	30.8	7.9	0.050	26.3

In terms of the core index MPJPE, this method reached 31.4, which was 24.9%, 17.8%

and 11.8% lower than LSTM (41.8), ST-GCN (38.2) and standard Transformer (35.6), respectively. In terms of motion smoothness and style loss, this method is also the best, which are 8.2 and 0.052, respectively. In particular, the style loss is reduced by 24.6% compared with the Transformer, which proves the effectiveness of multi-level attention mechanism in capturing intangible cultural heritage art style. The column of "improvement rate" in the table further quantifies the cumulative improvement relative to the LSTM baseline. The proposed reaches 24.9%, and the hierarchical model reaches 26.3%.

In the ablation experiment, to evaluate the contribution of each module, the module contribution rate is defined:

$$C_m = \frac{E_{base} - E_{\overline{om}}^{w}}{E_{base}} \tag{53}$$

Where $E_{w/om}$ represents the error after removing module m. In the ablation experiment, Figure 2 visually shows the change of model performance after removing different modules through the histogram.

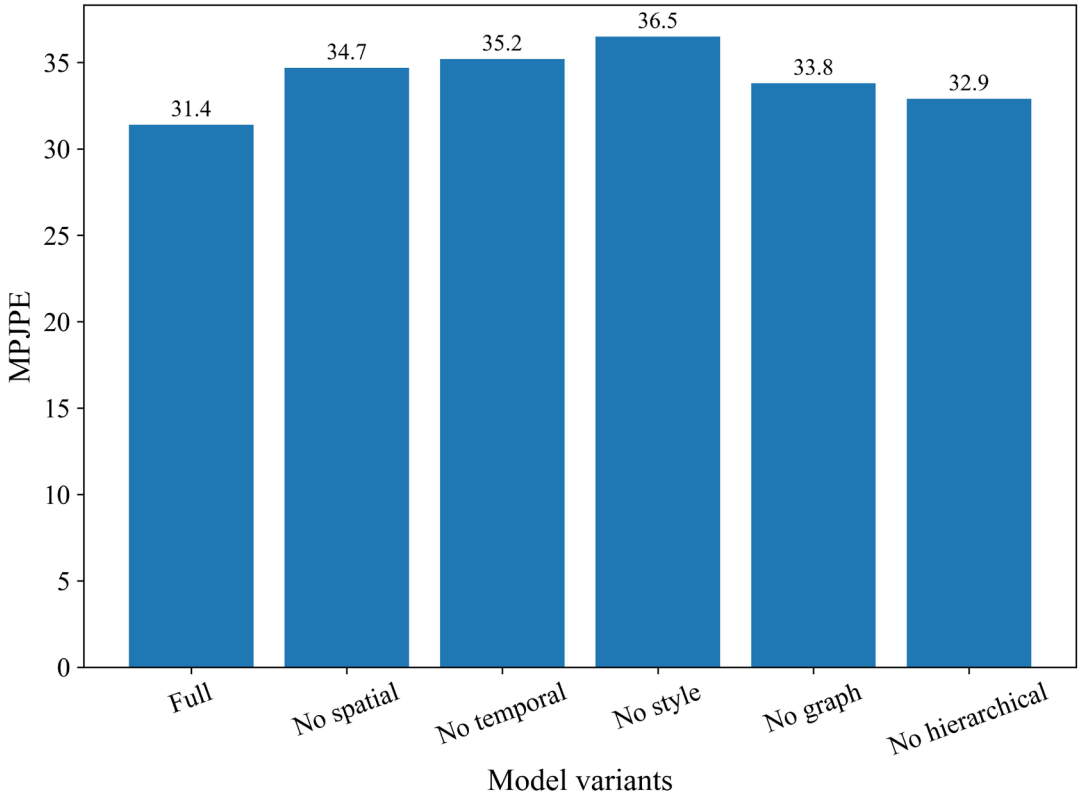

Figure 2. Histogram of impact of module ablation on MPJPE

The complete model achieves the lowest MPJPE (about 31.4), while removing style attention leads to the most significant error increase, to about 36.5, with an increase of more than 5 units, which is the most influential of all modules. After removing temporal attention and spatial attention, the errors increased to about 35.2 and 34.7 respectively, indicating that these two parts are equally important for basic motion modeling. In contrast, the performance degradation caused by the removal of graph structure module and hierarchical structure is small, but it still shows a certain degree of deterioration.

Further quantify the ablation results, as shown in Table 5.

Table 5. Ablation results

Model configuration	MPJPE↓	Smooth↓
Complete model	31.4	8.2
No spatial attention	34.7	9.6
No time for attention	35.2	9.9
Without style attention	36.5	10.3
No graph structure module	33.8	9.1
No hierarchical structure	32.9	8.7

Removing the "style attention" module has the greatest impact on the performance of the model, resulting in the MPJPE rising from 31.4 to 36.5 (16.2% worse), and the smoothness index rising from 8.2 to 10.3, confirming the special importance of the style perception mechanism for intangible cultural heritage movement modeling. Removing "temporal attention" and "spatial attention" also brought significant performance degradation. MPJPE rose to 35.2 and 34.7 respectively, indicating that both are the core of modeling basic action dependency. Removing "graph structure module" and "hierarchical structure" also leads to performance degradation, but the range is relatively small, indicating that they are an effective supplement to performance improvement.

In terms of generalization ability, this paper designs a cross play test to decouple the distribution of training set and test set. The generalization error is defined as:

$$E_{gen} = \frac{1}{TN} \sum \| x_{test} - \hat{x}_{train} \| \tag{54}$$

In the generalization capability analysis, Figure 3 shows the error distribution of different models in the cross play test in the form of box diagram.

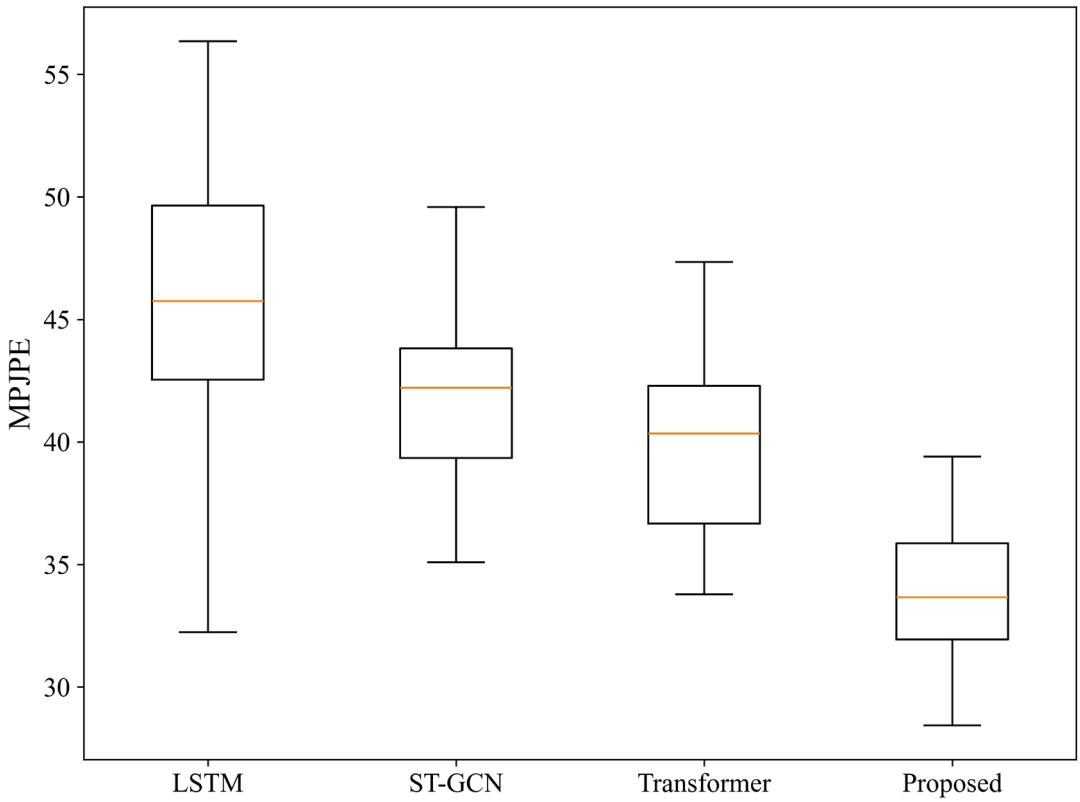

Figure 3. Cross scene generalization error box diagram of different models

The proposed method exhibits the smallest interquartile range and the lowest median error, which is roughly concentrated at about 34, and the upper and lower parts must be shorter, indicating that the error distribution is more concentrated and stable. In contrast, LSTM has the widest range of distribution, with a median of nearly 45, which fluctuates greatly, indicating that it is unstable under different data distributions. Although Transformer and ST-GCN have improved, there is still a certain degree of discreteness.

The corresponding numerical results are shown in Table 6.

Table 6. Comparison of generalization performance

Model	Cross-repertoire MPJPE↓	Variance↓
LSTM	45.2	6.8
ST-GCN	41.6	5.3
Transformer	38.9	4.7
Proposed	34.5	3.2
Hierarchical model	33.8	3.0

In the cross play test, the MPJPE of this method is 34.5, and the error variance is 3.2, both of which are significantly better than the comparison method. Compared with LSTM (45.2 ± 6.8), the average error of this method is reduced by 23.7%, and the stability (variance) is improved by 52.9%; Compared with the standard Transformer (38.9 ± 4.7), the error is reduced by 11.3%, and the stability is improved by 31.9%. This fully proves that the motion

representation learned in this model is more generalized than over fitting the surface features of a specific drama.

In the robustness analysis, Gaussian noise and missing key points are introduced to simulate the real environment interference. The noise model is defined as:

$$\tilde{x} = x + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{55}$$

The proportion of missing data is set as $p \in [0,0.5]$. Define the robustness index:

$$R = \frac{E_{noise}}{E_{clean}} \tag{56}$$

Where E_{noise} is the noise input error. Table 7 shows the performance changes under different interference conditions.

Table 7. Robustness test results

Noise intensity σ	LSTM	Transformer	Proposed
0.01	1.12	1.08	1.05
0.05	1.35	1.21	1.12
0.1	1.62	1.38	1.25
10% missing	1.28	1.19	1.10
30% missing	1.55	1.34	1.18
50% missing	1.83	1.52	1.31

Under all interference conditions, the proposed method consistently achieves the lowest performance degradation ratio. For example, under high noise intensity (σ=0.1), the error ratios of LSTM, Transformer and the proposed method are 1.62, 1.38 and 1.25, respectively. The degradation degree of the proposed method is 22.8% lower than that of LSTM and 9.4% lower than that of Transformer. Under extreme missing data conditions (50% missing), the error ratio of this method is 1.31, while LSTM and Transformer are 1.83 and 1.52, respectively, with more obvious advantages. This is due to the confidence weighting, interpolation repair and other preprocessing strategies in this method and the structural robustness of the model itself.

In terms of computational efficiency, the unit time efficiency is defined as:

$$\eta = \frac{T}{t_{infer}} \tag{57}$$

Where t_{infer} is the reasoning time. Combined with the complexity analysis, Table 8 shows the efficiency of different models.

Table 8. Comparison of calculation efficiency

Model	Flops (g)	Reasoning time (ms)	Throughput (frames/s)
LSTM	12.5	18.6	688
Transformer	21.3	27.5	465
Proposed	16.8	21.7	589
Sparse model	9.7	14.2	901
Hierarchical model	11.3	16.8	762

Although the standard Transformer has the highest flops (21.3g) and the slowest reasoning time (27.5ms), its throughput is only 465 frames per second. Through sparse and hierarchical design, the complete method proposed in this paper reduces flops to 16.8g, reduces reasoning time to 21.7ms, and improves throughput to 589 frames per second while maintaining high accuracy (MPJPE=31.4). Its variant model performs better: the sparse model achieves the highest throughput of 901 frames/second with 9.7 GFLOPs; The layered model also achieves 762 frames/second throughput with 11.3g flops, which provides the possibility for the digital application of real-time intangible cultural heritage movement.

To sum up, the experimental results verify the effectiveness and advancement of this method from the perspectives of accuracy, structural contribution, generalization ability, robustness and efficiency, and fully prove its application value in the modeling of intangible cultural heritage movement sequence.

8. DISCUSSION

The Transformer model combined with multi-level attention mechanism proposed in this paper has achieved remarkable results in the modeling task of shadow play action sequence, but after in-depth analysis of the experimental results and model behavior, there are still some problems worth discussing. The first is the effectiveness boundary of the attention mechanism of style perception. Ablation experiments show that removing the style attention module has the greatest damage to the performance of the model, which verifies the core role of the module in intangible cultural heritage movement modeling. However, further analysis found that the contribution of style attention is closely related to the style consistency of the training data. When the performance style of a given play type in the training dataset is highly consistent, the style vectors can be effectively clustered, and the style representation learned by the model is also clear; However, when the data contains mixed performances of different genres under the same play, the discrimination of style vector decreases significantly, and the sensitivity of the model to style differences also decreases. This phenomenon suggests that style perception attention essentially depends on the statistical rules of style in the training data. When the style boundary is fuzzy or there are continuous style gradients, the simple global style vector may not be enough to capture the subtle style differences. In the future, we can consider the introduction of multi-scale style representation or style decoupling strategy based on comparative learning, so that the model can distinguish different levels of style attributes, such as separating the macro genre style from the micro movement rhythm style.

Secondly, the performance differences of the model in different movement types are worth in-depth analysis. Although the **proposed** is better than the comparison method in MPJPE index on the whole, it is found that the performance improvement is not evenly distributed after stratified statistics by action category. For actions with obvious periodic regularity, such as repetitive dancing or walking gaits in shadow play, the temporal attention mechanism of the

proposed method can effectively capture their periodic patterns, and the prediction error is reduced by more than 30% compared with the baseline model. However, for the action clips with strong randomness and weak regularity, such as improvised fighting or emotional violent jitter, the advantages of the proposed are significantly reduced, and even equal to the standard Transformer in some extreme cases. This shows that the current model still tends to learn statistical laws from data, and its generalization ability is limited for unconventional actions that lack repetitive patterns. From a cognitive perspective, human performers still follow some implicit rules in improvisation, such as speed change patterns or strength control habits under specific emotions. These deeper action generation mechanisms have not been effectively captured by the current model. Enabling the model to learn implicit action generation rules from limited samples—rather than merely memorizing motion patterns—is key to further improving performance.

The scaling law between the computational efficiency of the model and the sequence length is also worth discussing. Although the sparse attention and hierarchical modeling strategy proposed in this paper reduces the theoretical complexity from the square level to the approximate linear level, the actual acceleration effect during deployment is influenced by hardware architecture and sequence characteristics. Specifically, when the sequence length is short, the cost of index calculation and memory access introduced by the sparse strategy may exceed the amount of calculation saved, resulting in an increase in the actual reasoning time instead of a decrease. Only when the sequence length exceeds 200 frames, the advantage of sparse attention really appears. This means that for the medium length sequences commonly seen in shadow plays, it is necessary to dynamically decide whether to enable the sparse strategy based on the specific sequence length, rather than using the same architecture across the board. In addition, the selection of local window size in the hierarchical model has a significant impact on performance. Too small a window will cause the local Transformer to fail to capture meaningful action fragments, and too large a window will degenerate into a global Transformer, losing the original intention of hierarchical design. In the experiment, we found that the optimal window size is about 16 frames, which corresponds to the average length of a basic action unit in the shadow play, implying that the optimal structure of the model is intrinsically related to the timing granularity of the data itself.

Regarding the interpretability of the model, Visual analysis of attention weights reveals some interesting phenomena. The spatial attention weight is not evenly distributed on all bone connections, but focuses on the connection points between the trunk and limbs and the key joints controlled by the bamboo stick, which is highly consistent with the actual control principle of shadow play - the performer mainly drives the movement of the whole shadow play character by controlling several key points. The temporal attention weight shows an obvious phased pattern. The model assigns higher weights to the start frame, transition frame, and end frame of an action, while less attention is paid to the intermediate frame in the process of uniform motion. This attention allocation strategy is similar to the visual attention mechanism of human observation. However, the mechanism of style attention is not completely clear. The visualization results show that the influence of style vector on different time steps and joints is different, but the semantic interpretation of this difference still needs to be further studied. For example, whether the model deconstructs style as a combination of rhythm features, amplitude features and posture preference features cannot be directly interpreted from the attention weight.

Finally, the application prospect and potential limitations of this method in intangible cultural heritage digital protection need to be examined objectively. On the positive side, the model can learn the action style of a specific genre from a small number of samples, and generate new action sequences that conform to the style, which provides a technical possibility for action archiving and virtual recovery of endangered genres. At the same time, the sparse variant of the model has the advantage of reasoning efficiency, so it is expected to be embedded in the interactive display system to realize the real-time interaction between the audience and the virtual shadow puppet characters. However, it must be acknowledged that digital modeling is an approximation and simplification of real performance after all. The artistic charm of

shadow play lies not only in the movement track itself, but also in the subtle interaction between the performer and the audience when improvising, the coordination of singing and movement, and the visual beauty brought by shadow play carving technology. These factors lie beyond the scope of pure motion modeling. Therefore, the proposed should be positioned as a component in the intangible cultural heritage protection toolbox, rather than a complete solution. In the future work, it is worth exploring to integrate the motion model with audio drive, interactive control and other technologies to build a more complete digital shadow puppet performance system.

9. CONCLUSION

Aiming at the key scientific problems in the digital modeling of shadow puppet movements of intangible cultural heritage, such as the difficulty in capturing long time sequence dependence, the lack of artistic style expression, and the contradiction between computational efficiency and modeling ability, this paper proposes an improved Transformer model integrating multi-level attention mechanism. At the data level, this paper constructs a structured data processing pipeline for representing intangible cultural heritage actions, including multi view action collection, skeleton key point extraction and confidence weighting, dynamic feature enhancement of velocity and acceleration, and time sequence alignment and standardization based on time resampling and relative coordinate transformation. Through multidimensional data enhancement strategies, including time scaling, spatial noise injection and style change matrix simulation, the generalization ability of the model to complex movements and style changes is significantly improved. The data set constructed covers martial arts, literary drama, mythology, history, folklore and other drama categories, with a total number of frames exceeding one million and an effective key point rate of more than 90%, providing a high-quality data base for subsequent modeling research.

In terms of model architecture, the core innovation of this paper is embodied in the organic integration of three levels. First, a multi-level attention fusion mechanism is designed, in which spatial attention introduces bone adjacency priors to strengthen the physical connection constraints between joints, temporal attention captures long-distance action dependence through information aggregation across time steps, and style aware attention uses global feature statistics to adjust local attention calculation, so that the model can learn and maintain the unique performance styles of different shadow puppet genres. Secondly, a graph structure enhanced Transformer architecture is proposed. By alternately stacking graph convolution operation and self-attention mechanism, the model has the flexible modeling ability of inductive bias of bone topology and global space-time dependence. Thirdly, aiming at the challenge of long sequence modeling, the sparse attention mechanism is introduced to only calculate the interaction between key time steps, and the hierarchical modeling strategy is combined to realize the collaborative modeling of local details and global semantics, which reduces the computational complexity from the square of sequence length to the approximate linear level. On the optimization goal, this paper constructs a multi task joint loss function, which integrates the speed smoothing constraint, bone length consistency constraint and style consistency constraint based on feature statistics on the basis of motion reconstruction accuracy, and realizes the adaptive balance of multi-objective through the dynamic weight adjustment mechanism.

Through systematic experimental verification, this paper draws the following main conclusions. In motion prediction tasks, the proposed method achieves an MPJPE of 31.4, which was reduced by 24.9%, 17.8% and 11.8% respectively compared with LSTM, ST-GCN and standard Transformer, and the loss of style was reduced by more than 24%, indicating that the model has significant advantages in motion accuracy and style fidelity. Ablation experiments confirmed that style perception attention was the module that contributed the most to the performance of the model. After removal, MPJPE rose to 36.5%, with a deterioration rate of 16.2%, highlighting the core position of style modeling in the digitization of intangible cultural heritage movements. In the cross play generalization test, the error variance of this

method is only 3.2, which is 52.9% lower than 6.8 of LSTM, which proves that the motion representation learned by the model has strong generalization ability rather than over fitting the surface features of a specific play. The robustness test shows that under the extreme condition of 50% missing key points, the error ratio of this method is 1.31, which is significantly better than 1.83 of LSTM and 1.52 of Transformer, reflecting the strong robustness of the model to noise and missing data. In terms of computational efficiency, through sparse and hierarchical design, the variant of the proposed model achieves a throughput of 901 frames per second with 9.7 GFLOPs, which is approximately 19.4 times that of the standard Transformer, providing technical possibilities for real-time interactive applications.

In general, the integrated attention Transformer model proposed in this paper has achieved the collaborative improvement of accuracy, style fidelity, generalization ability and computational efficiency in the task of modeling the action sequence of intangible cultural heritage, providing an effective technical scheme for the digital protection and intelligent inheritance of intangible cultural heritage. It should be pointed out that there are still some aspects to be improved in this method, such as the strong dependence of style-aware attention on style consistency in the training data, the improvement of the modeling ability of unconventional actions with strong randomness and weak regularity, and the acceleration effect of sparse strategy on short sequences is not obvious. Future work can be explored in the following directions: introducing unsupervised or semi supervised style decoupling learning paradigm to reduce the dependence of the model on style annotation data; Explore the fusion architecture of diffusion model and Transformer to further improve the modeling ability of complex action distribution; A more interactive virtual shadow puppet performance system is constructed by combining the motion model with audio driven, natural language command control and other technologies; And extend this method to other types of performance intangible cultural heritage to verify its versatility and migration ability.

Abbreviations

LSTM, Long Short-Term Memory;
GRU, Gated Recurrent Unit;
ST-GCN, Spatial Temporal Graph Convolutional Network;
CNN, Convolutional Neural Network;
HRNet, High-Resolution Network;
MPJPE, Mean Per Joint Position Error;
MSE, Mean Squared Error;
FLOPs, Floating Point Operations;
GPU, Graphics Processing Unit;
CUDA, Compute Unified Device Architecture;
AdamW, Adaptive Moment Estimation with Weight Decay;
ReLU, Rectified Linear Unit;
ELU, Exponential Linear Unit;
GAN, Generative Adversarial Network;
OpenPose, Open Pose Estimation;
ICP, Iterative Closest Point;
Frobenius, Frobenius Norm;
ICH, Intangible Cultural Heritage.

Supplementary Material

Not applicable.

Appendix

Not applicable.

Ethics approval and consent to participate.

This study did not involve human participants, animal subjects, or any data requiring ethical approval. Therefore, ethics approval and consent to participate are not applicable.

Acknowledgements

The authors would like to thank the editors of this journal and all the anonymous reviewers who provided valuable comments on this work.

Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article. However, **M.L.** is an editorial board member of this journal. To ensure transparency, the authors confirm that this manuscript was handled by an independent editor and that **M.L.** was not involved in the peer review or decision-making process.

Author contributions

All authors have read and agreed to the published version of the manuscript. The authors' contributions are specified as follows: **Y.L.**: contributed to Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing – Original draft, Writing – Review & Editing, Visualization, Supervision, and Project administration. **S.F.**: contributed to Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – Original draft, Writing – Review & Editing, Visualization, and Supervision. **M.L.**: contributed to Investigation, Data Curation, Writing – Review & Editing, Visualization, Software, and Formal analysis.

Funding information

This work was supported by the **China Adult Education Association** (Grant No.: **2025-0588Y**).

Data availability

The data that support the findings of this study are available upon request from the corresponding authors, **M.L.**

Disclaimer

The views and opinions expressed in this article are those of the authors and are the product of professional research. It does not necessarily reflect the official policy or position of any affiliated institution, funder, agency, or that of the publisher. The authors are responsible for this article's results, findings, and content.

Declaration of AI and AI-assisted Technologies in the Writing Process

During the writing of this article, the author used ChatGPT for spelling and grammar checking. After using this tool, the author reviewed and edited the content as needed and assumes full responsibility for the final published content.

REFERENCES

[1] Lin, C., Xia, G., Nickpour, F., & Chen, Y. (2025, June). Bridging Tradition and Innovation: Using Artistic Genes to Assess Cultural Authenticity in Digital Shadow Play. In *International Conference on Human-Computer Interaction* (pp. 213-232). Cham: Springer Nature Switzerland. DOI: **https://doi.org/10.1007/978-3-032-13164-5_14**

[2] Li, T., & Cao, W. (2021). Research on a method of creating digital shadow puppets based on parameterized templates. *Multimedia Tools and Applications*, *80*(13), 20403-20422. DOI: **https://doi.org/10.1007/s11042-021-10726-1**

[3] Hou, Y., Kenderdine, S., Picca, D., Egloff, M., & Adamou, A. (2022). Digitizing intangible cultural heritage embodied: State of the art. *Journal on Computing and Cultural Heritage (JOCCH)*, *15*(3), 1-20. DOI: **https://doi.org/10.1145/3494837**

[4] Rallis, I., Voulodimos, A., Bakalos, N., Protopapadakis, E., Doulamis, N., & Doulamis, A. (2020). Machine learning for intangible cultural heritage: a review of techniques on dance analysis. *Visual Computing for Cultural Heritage*, 103-119. DOI: **https://doi.org/10.1007/978-3-030-37191-3_6**

[5] Zhou, Y., Wang, R., Li, H., & Kung, S. Y. (2020). Temporal action localization using long short-term dependency. *IEEE Transactions on Multimedia*, *23*, 4363-4375. DOI: **https://doi.org/10.1109/TMM.2020.3042077**

[6] Jin, Y., Long, Y., Chen, C., Zhao, Z., Dou, Q., & Heng, P. A. (2021). Temporal memory relation network for workflow recognition from surgical video. *IEEE Transactions on Medical Imaging*, *40*(7), 1911-1923. DOI: **https://doi.org/10.1109/TMI.2021.3069471**

[7] Hermans, C. (2025). Of rhythm and movement: physical play and dance as (participatory) sense-making practices. *Research in Dance Education*, *26*(3), 313-328. DOI: **https://doi.org/10.1080/14647893.2023.2211524**

[8] Romat, H., Fender, A., Meier, M., & Holz, C. (2021, March). Flashpen: A high-fidelity and high-precision multi-surface pen for virtual reality. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)* (pp. 306-315). IEEE. DOI: **https://doi.org/10.1109/VR50410.2021.00053**

[9] Sun, J., Zheng, C., Xie, E., Liu, Z., Chu, R., Qiu, J., ... & Li, Z. (2025). A survey of reasoning with foundation models: Concepts, methodologies, and outlook. *ACM Computing Surveys*, *57*(11), 1-43. DOI: **https://doi.org/10.1145/3729218**

[10] Luo, Q., Zeng, W., Chen, M., Peng, G., Yuan, X., & Yin, Q. (2023, July). Self-attention and transformers: Driving the evolution of large language models. In *2023 IEEE 6th International conference on electronic information and communication technology (ICEICT)* (pp. 401-405). IEEE. DOI: **https://doi.org/10.1109/ICEICT57916.2023.10245906**

[11] Hassija, V., Palanisamy, B., Chatterjee, A., Mandal, A., Chakraborty, D., Pandey, A., ... & Kumar, D. (2025). Transformers for vision: A survey on innovative methods for computer vision. *Ieee Access*. DOI: **https://doi.org/10.1109/ACCESS.2025.3571735**

[12] Mazzia, V., Angarano, S., Salvetti, F., Angelini, F., & Chiaberge, M. (2022). Action transformer: A self-attention model for short-time pose-based human action recognition. *Pattern Recognition*, *124*, 108487. DOI: **https://doi.org/10.1016/j.patcog.2021.108487**

[13] Zhang, E. Y., Cheok, A. D., Pan, Z., Cai, J., & Yan, Y. (2023). From turing to transformers: A comprehensive review and tutorial on the evolution and applications of generative transformer models. *Sci*, *5*(4), 46. DOI: **https://doi.org/10.3390/sci5040046**

[14] Moutik, O., Sekkat, H., Tigani, S., Chehri, A., Saadane, R., Tchakoucht, T. A., & Paul, A. (2023). Convolutional neural networks or vision transformers: Who will win the race for action recognitions in visual data?. *Sensors*, *23*(2), 734. DOI: **https://doi.org/10.3390/s23020734**

[15] Ren, Q., Li, M., Li, H., & Shen, Y. (2021). A novel deep learning prediction model for concrete dam displacements using interpretable mixed attention mechanism. *Advanced Engineering Informatics*, *50*, 101407. DOI: **https://doi.org/10.1016/j.aei.2021.101407**

Tutek, M., & Šnajder, J. (2022). Toward practical usage of the attention mechanism as a tool for interpretability. *IEEE access*, *10*, 47011-47030. DOI: **https://doi.org/10.1109/ACCESS.2022.3169772**

[16] Yang, Z. B., Zhang, J. P., Zhao, Z. B., Zhai, Z., & Chen, X. F. (2020). Interpreting network knowledge with attention mechanism for bearing fault diagnosis. *Applied Soft Computing*, *97*, 106829. DOI: **https://doi.org/10.1016/j.asoc.2020.106829**

[17] Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent neural networks: A comprehensive review of architectures, variants, and applications. *Information*, *15*(9), 517. DOI: **https://doi.org/10.3390/info15090517**

[18] Ahmad, T., Wu, J., Alwageed, H. S., Khan, F., Khan, J., & Lee, Y. (2023). Human activity recognition based on deep-temporal learning using convolution neural networks features and bidirectional gated recurrent unit with features selection. *IEEE access*, *11*, 33148-33159. DOI: **https://doi.org/10.1109/ACCESS.2023.3263155**

[19] Zan, T., Jia, X., Guo, X., Wang, M., Gao, X., & Gao, P. (2025). Research on variable-length control chart pattern recognition based on sliding window method and SECNN-BiLSTM. *Scientific Reports*, *15*(1), 5921. DOI: **https://doi.org/10.1038/s41598-025-86849-4**

[20] Bouktif, S., Fiaz, A., Ouni, A., & Serhani, M. A. (2020). Multi-sequence LSTM-RNN deep learning and metaheuristics for electric load forecasting. *Energies*, *13*(2), 391. DOI: **https://doi.org/10.3390/en13020391**

[21] Ahmad, T., Jin, L., Zhang, X., Lai, S., Tang, G., & Lin, L. (2021). Graph convolutional neural network for human action recognition: A comprehensive survey. *IEEE Transactions on Artificial Intelligence*, *2*(2), 128-145. DOI: **https://doi.org/10.1109/TAI.2021.3076974**

[22] Yang, X., Li, S., Niu, S., & Yue, X. (2026). Graph network learning for human skeleton modeling: a survey. *Artificial Intelligence Review*, *59*(1), 31. DOI: **https://doi.org/10.1007/s10462-025-11442-0**

[23] Feng, L., Zhao, Y., Zhao, W., & Tang, J. (2022). A comparative review of graph convolutional networks for human skeleton-based action recognition. *Artificial Intelligence Review*, *55*(5), 4275-4305. DOI: **https://doi.org/10.1007/s10462-021-10107-y**

[24] Yu, H., Fan, X., Hou, Y., Pei, W., Ge, H., Yang, X., ... & Zhang, M. (2023). Toward realistic 3d human motion prediction with a spatio-temporal cross-transformer approach. *IEEE*

Transactions on Circuits and Systems for Video Technology, *33*(10), 5707-5720. DOI: **https://doi.org/10.1109/TCSVT.2023.3255186**

[25] Jiao, L., Zhang, X., Liu, X., Liu, F., Yang, S., Ma, W., ... & Zhang, J. (2023). Transformer meets remote sensing video detection and tracking: A comprehensive survey. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *16*, 1-45. DOI: **https://doi.org/10.1109/JSTARS.2023.3289293**

[26] Soydaner, D. (2022). Attention mechanism in neural networks: where it comes and where it goes. *Neural Computing and Applications*, *34*(16), 13371-13385. DOI: **https://doi.org/10.1007/s00521-022-07366-3**

[27] Yuan, C., Liu, J., Wang, H., & Yang, Q. (2025). Object Detection in Complex Traffic Scenes Based on Environmental Perception Attention and Three-Scale Feature Fusion. *Applied Sciences*, *15*(6), 3163. DOI: **https://doi.org/10.3390/app15063163**

[28] Hou, Y., Kenderdine, S., Picca, D., Egloff, M., & Adamou, A. (2022). Digitizing intangible cultural heritage embodied: State of the art. *Journal on Computing and Cultural Heritage (JOCCH)*, *15*(3), 1-20. DOI: **https://doi.org/10.1145/3494837**

[29] Kim, M., Hwang, T., & So, J. (2025). Real-Time Live Streaming Framework for Cultural Heritage Using Multi-Camera 3D Motion Capture and Virtual Avatars. *Applied Sciences*, *15*(22), 12208. DOI: **https://doi.org/10.3390/app152212208**

[30] Shen, J., Chen, L., He, X., Zuo, C., Li, X., & Dong, L. (2025). An Interactive Human-in-the-Loop Framework for Skeleton-Based Posture Recognition in Model Education. *Biomimetics*, *10*(7), 431. DOI: **https://doi.org/10.3390/biomimetics10070431**

[31] Wen, B. (2025). A multimodal transformer framework with biomechanical constraints for injury prediction and human motion analysis. *Journal of Computational Methods in Sciences and Engineering*, 14727978251348632. DOI: **https://doi.org/10.1177/14727978251348632**