Available online at www.HighTechJournal.org

HighTech and Innovation Journal

ISSN: 2723-9535

HighTech and Innovation Journal

Vol. 7, No. 2, June, 2026

Music-Driven Enhanced Dance Performance Generation by Integrating Seq2Seq and Human Pose Recognition

Wen He [1*]

[1] *First College of Arts, Chengdu Sport University, Chengdu, 641418, China.*

Received 06 November 2025; Revised 24 April 2026; Accepted 03 May 2026; Published 01 June 2026

Abstract

To address the accuracy bottleneck in the naturalness and rhythm synchronization of music-driven dance generation, an enhanced dance generation model integrating sequence-to-sequence modeling and human pose recognition was developed to improve the synchronization, naturalness, and structural consistency of generated movements. The model uses multi-scale music features as input, extracts temporal music semantics through a bidirectional long short-term memory network and an attention mechanism, and optimizes motion structure by incorporating skeleton keypoint feedback, thereby achieving joint modeling of music semantics and human motion. Experimental results on the AIST++ and DanceTrack datasets demonstrate that the proposed model achieves a beat alignment error as low as 0.12 s, a joint point error of 11.2 px, and a motion smoothness score of 2.41. In the generation of a 90-second dance sequence, the beat error is reduced by more than 32% compared with mainstream models, and the model achieves a high score of 0.97 in the evaluation of complex dance symmetries such as "arm-lifting rotation." These results indicate that the joint modeling of music semantics and skeletal structure effectively improves movement coordination and rhythm matching in dance generation, enabling the production of natural and coordinated dance movements adaptable to different dance styles.

Keywords: Music-Driven Dance Generation; Seq2Seq Architecture; Human Pose Recognition; Cross-Modal Alignment; Exercise Enhancement; Style Transfer.

1. Introduction

As a dynamic visual expression of music in both spatial and temporal dimensions, the core challenge of automatic dance generation lies in achieving precise, natural, and structurally reasonable cross-modal mapping between music rhythm, semantics, and human movement [1, 2]. In recent years, deep learning-based sequence-to-sequence (Seq2Seq) architectures have provided a fundamental framework for learning temporal transformation relationships from music to dance [3]. Researchers have made preliminary progress in rhythm alignment and short-sequence dance generation using models such as Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and Transformers [4]. In parallel, the advancement of human pose recognition technologies, represented by OpenPose, has provided a reliable data foundation for the precise representation and evaluation of dance movements [5].

However, existing studies still face significant limitations. On the one hand, most approaches focus primarily on temporal modeling while paying insufficient attention to the biomechanical rationality and spatial coordination of generated movements. This often leads to issues such as abnormal joint positions and violations of anatomical constraints, resulting in large joint position errors and low posture symmetry scores. On the other hand, during long-sequence dance generation, existing models have limited ability to dynamically respond to complex rhythms and

* Corresponding author: hewen202577808@163.com

https://doi.org/10.28991/HIJ-2026-07-02-011

semantic variations in music. This can result in beat drift and repetitive movements, making it difficult to maintain long-term synchronization and expressiveness, ultimately causing the accumulation of beat alignment errors over time [6].

The root of these problems lies in the lack of a mechanism within current architectures that enables closed-loop collaborative optimization between high-level music semantics and low-level structural constraints of human motion. To address this research gap, a music-driven Enhanced Dance Generation Model by Fusing Seq2Seq with Human Gesture Recognition (Seq2Seq HGR-EDG) is proposed. The core concept of this research is the construction of an integrated framework that incorporates a "perception–generation–feedback optimization" loop. The theoretical basis of this framework is to formulate dance generation as a Seq2Seq translation problem constrained by multiple factors, where music semantics serve as the driving source and the physiological and kinematic characteristics of the human skeleton act as hard constraints that must be satisfied. Specifically:

- Based on cross-modal representation learning theory, temporal semantic representations of music signals are learned through multi-scale music feature extraction and a Bi-LSTM encoder.
- According to attention mechanism theory, an attention-enhanced Seq2Seq decoder is employed to achieve dynamic and non-uniform mapping from music representations to action sequences, thereby improving beat alignment.
- The key innovation lies in introducing feedback control theory based on discriminative models, where the human pose recognition module functions as an online structural feedback mechanism. This module performs real-time rationality assessment and structural correction of generated actions according to a human kinematic model, thereby forming a closed-loop collaborative optimization process between generation and discrimination.

This theoretical framework aims to move beyond the traditional open-loop generative paradigm by introducing structural feedback constraints to ensure that generated movements satisfy both the naturalness and biomechanical rationality of human motion while maintaining synchronization with music. Unlike recent approaches that optimize structure through latent-space constraints or post-evaluation mechanisms, the innovation of this study lies in constructing a real-time feedback loop based on explicit human kinematic rules. This enables skeletal-level structural correction at each generation step, representing a transition from structure perception to structure control.

The overall structure of the research consists of four parts: the first part summarizes the research results and deficiencies in the field of music-driven dance generation and human gesture recognition at home and abroad. The second part elaborates on the overall architecture of the proposed Seq2Seq-HGR-EDG model and the design of its core modules. The third part includes ablation experiments and comparative experiments to comprehensively verify and analyze the performance and generation effects of the proposed model. The fourth part conducts an in-depth discussion of the significance of the experimental results and the limitations of the research. The fifth part summarizes the full text work and looks forward to future research directions.

2. Related Works

The field of music-driven dance generation continues to explore diverse technical approaches to enhance choreography efficiency and output quality. To streamline music-driven dance composition, Zeng [7] proposed an AI tool based on rhythm and beat analysis. By integrating a multilayer perceptron model with techniques like synthetic minority class oversampling and information gain, it achieves high-accuracy prediction and automatic recommendation of innovative dance steps. To strengthen control over dance movements, Yang et al. [8] proposed a music-driven dance generation method based on keyframe interpolation. By employing normed flow learning and temporal embedding, it achieves robust transitions between key poses, generating high-quality dance movements that balance rhythmic matching and diversity. To address repetitive patterns in dance generation, Kim et al. [9] introduced an improved music-driven multi-dance generation framework. By integrating adversarial learning with style-aware latent representations, they achieved highly synchronized, diverse, and realistic motion generation across multiple dance styles. To generate dance motions with strong rhythmic and aesthetic alignment to music, Au et al. [10] proposed a music-driven dance model based on dance repertoires. Through cross-modal embedding learning and normative flow control, they achieved rhythmically synchronized, stylistically consistent, and highly diverse dance choreography.

Human pose recognition technology continues to innovate in multimodal dance analysis, pushing accuracy boundaries from fundamental feature extraction to culturally specific motion modeling. To enhance dance pose recognition accuracy, Wang et al. [11] proposed a dance motion capture and pose recognition method based on visual sensors and 3D convolutional neural networks. By integrating joint coordinates and velocity features, it achieved higher recognition rates than other methods across multiple datasets. To improve Indian dance motion recognition and analysis, Jhansi Rani et al. [12] proposed a deep learning approach combining data augmentation with quantum technology. Employing a generative adversarial network for Indian classical dance enhanced data augmentation, significantly improving dance motion recognition accuracy. To enhance 3D dance motion recognition precision, Zhou et al. [13] introduced a 3D motion recognition method based on a proprioceptive interaction device and neural networks. By

extracting human joint features through deep images and random decision forests, combined with a hierarchical extreme learning machine network, they achieved high-precision recognition of dance action categories. To generate music with specified emotions, Bao & Sun [14] proposed an emotion-lyric and melody generator. Through automatic emotion recognition based on a bidirectional Transformer encoder pre-trained language model and a Seq2Seq encoder-decoder architecture, they achieved large-scale emotion-driven music generation without manual annotation.

In summary, existing music-driven dance generation research exhibits significant shortcomings in long-term action synchronization, biomechanical plausibility, and cross-style adaptability. To address this, the study innovatively integrates Seq2Seq modeling with human pose recognition technology, constructing a multi-stage collaborative enhancement framework. Its core significance lies in transcending traditional technical boundaries to provide intelligent choreography systems with a cross-modal generation paradigm that combines physiological plausibility and creativity, propelling virtual performing arts toward a higher stage of intelligence.

3. Material and Methods

3.1. Design of Music-Driven Enhanced Dance Generation Model

In contemporary society, dance, as an important body language and cultural expression, is widely used in various fields such as artistic performances, sports competitions, social activities, and even psychological therapy. With the development of digital media and artificial intelligence technology, virtual dance generation technology is gradually penetrating into dance education, stage arrangement, and immersive entertainment scenes, becoming an emerging research hotspot for cross-border integration. The natural matching between music semantics and human movements is the core challenge of music driven dance generation. To address this challenge, the Seq2Seq-HGR-EDG model has been proposed. This model aims to achieve the full process generation of dance action sequences from raw music input to high-quality, structurally sound, and highly synchronized with music by jointly modeling the temporal semantic information of music and the structural constraints of the human skeleton. The overall architecture of the Seq2Seq-HGR-EDG model is shown in Figure 1.

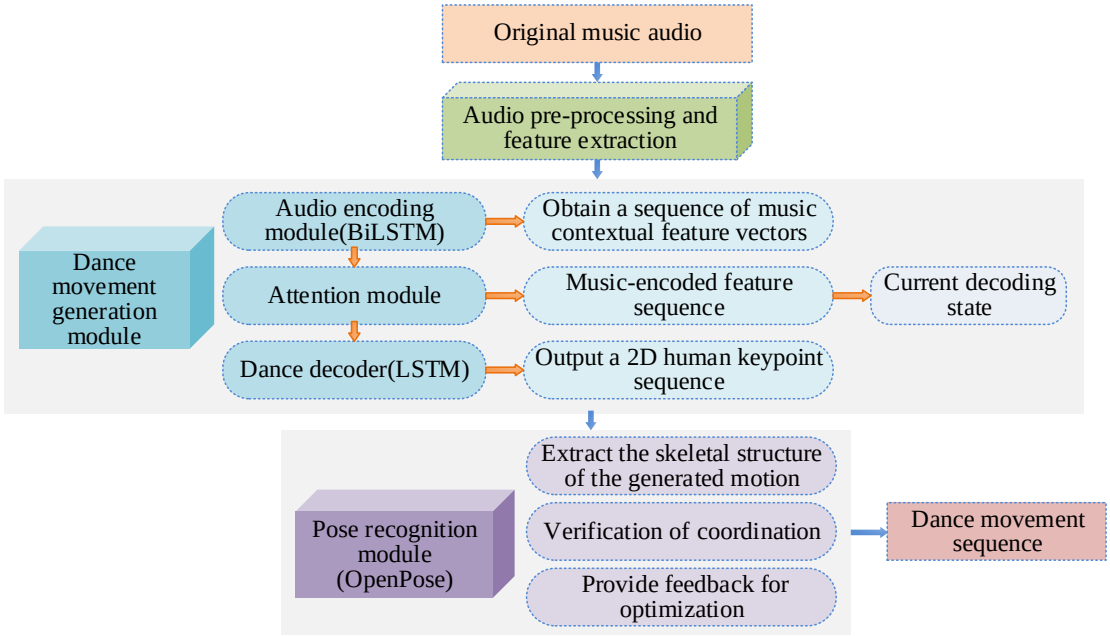


Figure 1. Overall framework diagram of Seq2Seq-HGR-EDG model

As shown in Figure 1, the Seq2Seq-HGR-EDG model first preprocesses the original music signal and extracts multi-dimensional music features. These features are input into a music encoding network based on Bidirectional LSTM (Bi-LSTM) to capture the contextual semantics of the music. Subsequently, the encoded music features are input into the Seq2Seq dance generation module, where the decoder utilizes attention mechanism to dynamically focus on the music segments most relevant to the current generated action, improving the alignment accuracy between music and action. The generated keypoint sequence is further input into the posture recognition and feedback enhancement module, and the skeleton structure analysis is used to determine whether the actions are coordinated and natural. If necessary, structural optimization and repair are carried out, and finally a visualized dance action sequence is output, which is compared and analyzed with real data and evaluated for synchronization. Among them, the multi-scale music feature extraction module is shown in Figure 2 [15, 16].

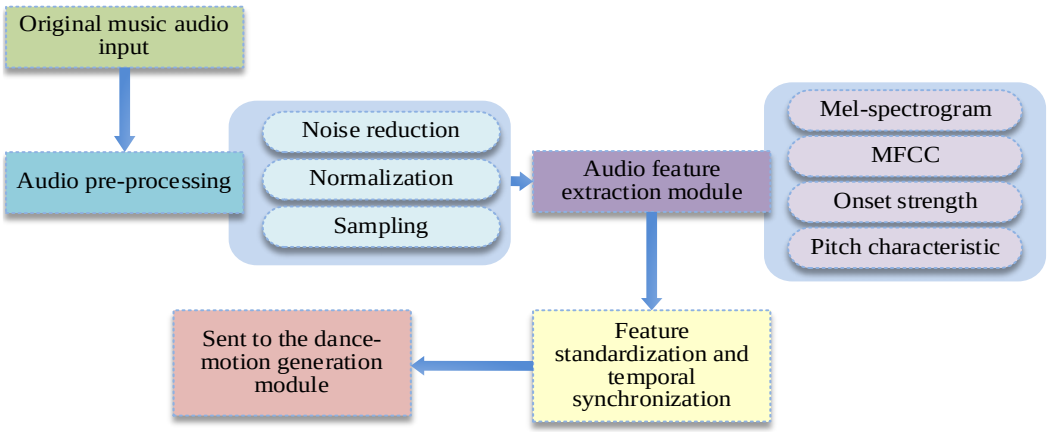


Figure 2. Schematic diagram of multi-scale music feature extraction module

As shown in Figure 2, the music feature encoding module first preprocesses the input raw audio signal, including noise reduction and a uniform sampling rate, to ensure stable audio quality and data consistency. Subsequently, the system employs various methods to extract key music features that reflect frequency, rhythm, and timbre characteristics, including Mel frequency spectrum, Mel frequency cepstral coefficients, beat intensity, and pitch information. To achieve balanced representation of multi-dimensional features, the extracted multi-dimensional features are standardized and aligned with the time scale of the dance action sequence through a time resampling operation. Finally, the processed feature sequence is input into Bi-LSTM for temporal modeling, capturing contextual information of music changes over time and outputting an encoded music feature representation sequence, providing a rich and accurate music semantic foundation for subsequent dance action generation modules.

3.2. Design of Attention Enhanced Seq2Seq Dance Generation Module

In the overall music-driven enhanced dance performance generation model, the quality of dance movements directly determines the expressiveness and naturalness of the system output. To achieve more accurate temporal alignment and semantic mapping between music and action, the Seq2Seq model is studied as the core structure of the dance generation module, and its basic structure is shown in Figure 3 [17].

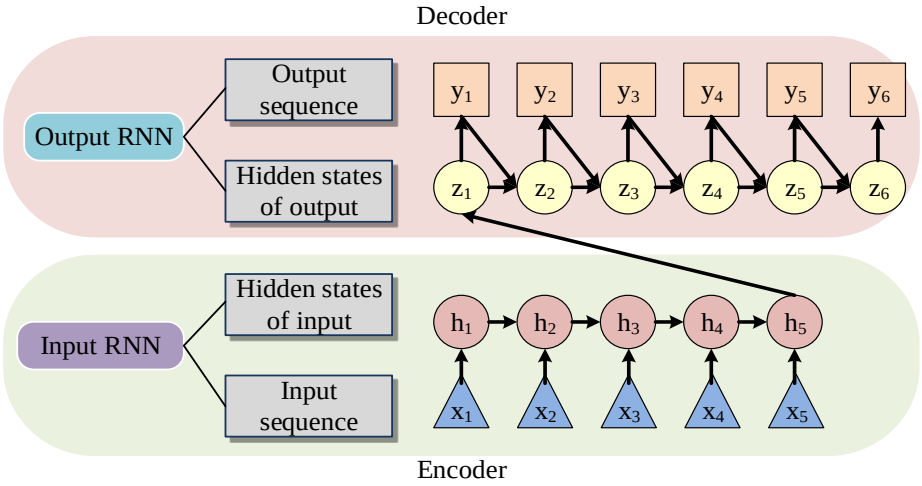


Figure 3. Basic structure of the Seq2Seq model

As shown in Figure 3, the Seq2Seq model consists of an encoder and a decoder. The encoder gradually processes the input sequence through RNN or LSTM and compresses it into a fixed length context vector as the initial input of the decoder. The decoder uses autoregression to generate output sequences, which is suitable for tasks with inconsistent input and output lengths [18]. However, although traditional Seq2Seq models can complete Seq2Seq mapping tasks, they are prone to information loss and insufficient context understanding when dealing with long-term dependencies and complex temporal relationships [19]. To this end, the study introduces an attention mechanism to enhance the model's ability to dynamically focus on the features of different time segments in music. The attention-based Seq2Seq network architecture is shown in Figure 4 [20].

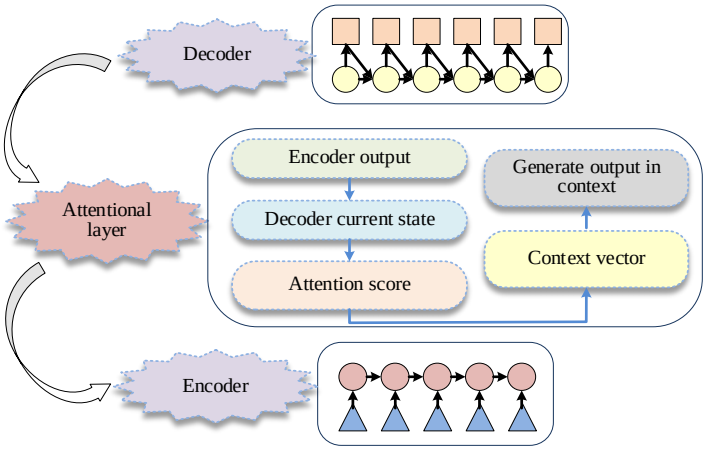


Figure 4. Seq2Seq network architecture based on attention mechanism

As shown in Figure 4, the Seq2Seq network based on attention mechanism consists of an encoder, a decoder, and an attention layer. The attention mechanism enables the decoder to flexibly select key information from the input sequence according to the requirements of different time steps, thereby enhancing the model's expressive power and prediction accuracy [21]. Compared to fixed length context vectors, the attention mechanism achieves dynamic weighted fusion of context, making the generated sequence more accurate and coherent. In addition, this mechanism also enhances the model's ability to capture long-range dependencies, effectively alleviating the gradient vanishing and information forgetting problems that traditional Seq2Seq models encounter in generating long sequences [22]. To further improve the accuracy and expressiveness of dance movement generation, a model for music-driven dance generation tasks is designed based on the attention mechanism-based Seq2Seq architecture, as shown in Figure 5 [23].

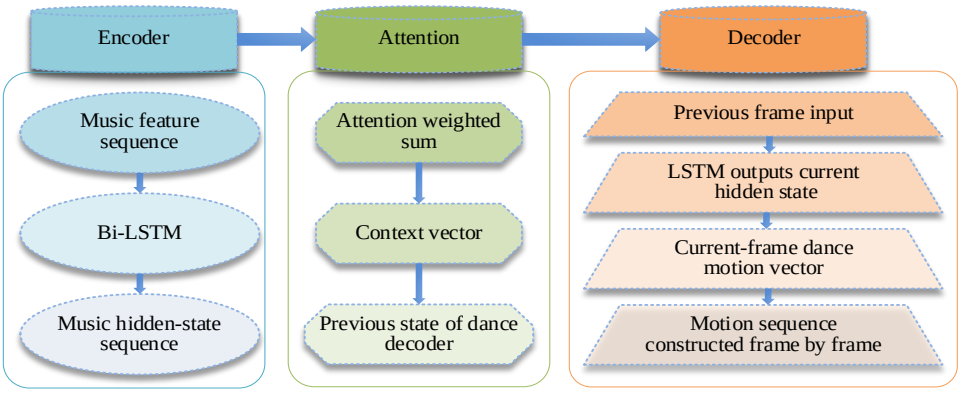


Figure 5. A Seq2Seq dance generation model based on attention mechanism

As shown in Figure 5, the module first processes the input music feature sequence through bidirectional Bi-LSTM by the encoder, extracting the hidden states of each time step, which contain contextual information of the music. Next, the attention mechanism is used to calculate the correlation between the current decoding time and all hidden states of the encoder, generating a weighted context vector that allows the decoder to dynamically focus on the music segment that is most relevant to the current dance action. In the decoder section, the model takes the dance movements from the previous frame or real dance movements as input and recursively generates the current action state through multiple layers of LSTM. Subsequently, the output of the decoder is fused with the context vector and mapped into the final sequence of human keypoint coordinates through a fully connected layer, gradually constructing a complete dance action sequence. The attention weight calculation is shown in Equation 1 [24]:

$$\alpha_{t,i} = \frac{\exp(\text{score}(s_t, h_i))}{\sum_{j=1}^{T} \exp(\text{score}(s_t, h_j))} \quad (1)$$

In Equation 1, $\alpha_{t,i}$ represents the attention weight of encoding time step i during decoding time step t; s_t represents the hidden state of the decoder at time step t; j represents the cyclic index of the encoder time step; $\square_i$ and $\square_j$ represent the hidden states of the encoder at time steps i and j, respectively; T represents the length of the encoding sequence. The calculation of the context vector is shown in Equation 2 [25].

$$c_t = \sum_{T}^{i=1} \alpha_{t,i} h_i \quad (2)$$

In Equation 2, c_t represents the context vector at decoding time step t.

3.3. Design of Human Skeleton Extraction Module Based on Posture Recognition

After constructing a Seq2Seq dance generation model based on attention mechanism, to improve the structural rationality and naturalness of movements, the OpenPose module was introduced to extract two-dimensional human keypoint coordinates from dance videos and supervise and optimize the generated action sequence at the structural level, as shown in Figure 6 [26].

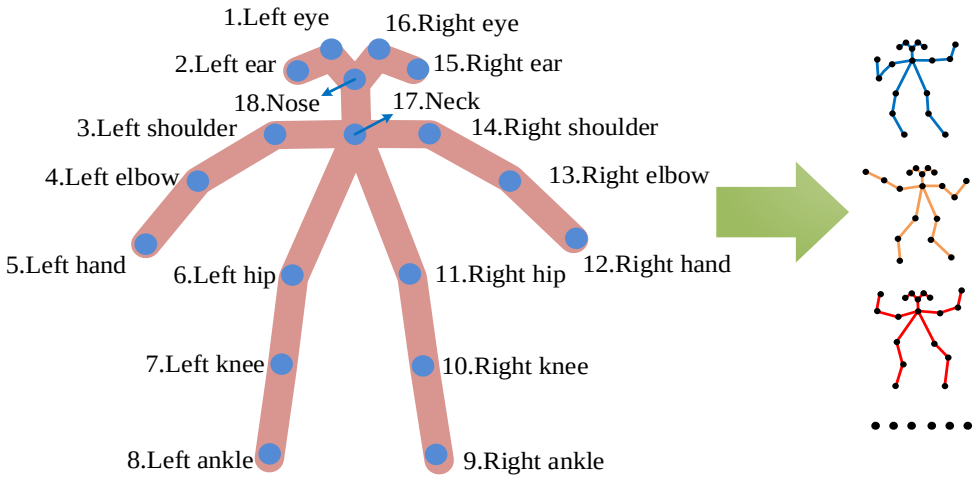


Figure 6. Human posture joints

Figure 6 shows the schematic structure of the key points of human pose, including 18 joint points, covering the main parts such as the head, trunk, upper and lower limbs, and clearly depicting human posture through skeletal connections. Using the OpenPose framework, two-dimensional pose keypoints of the human body can be extracted frame-by-frame from video footage to construct dynamic skeletal motion sequences. Incorporating these extracted pose sequences into the model training process as structural supervision signals helps enhance the anatomical plausibility and limb coordination of generated movements in spatial structure [27]. Specifically, if there is a lack of such structural feedback, generative models that rely solely on temporal alignment often produce several typical errors: 1) spatial asymmetry: for example, in the "arm lift rotation" action, the height or rotation angle of the left and right arms are not consistent, which damages visual coordination; 2) Joint mutation and shaking: Non physiological and intense jumping of joint positions between consecutive frames, resulting in stiff and disjointed movements; 3) Skeleton proportion distortion: The generated limb length undergoes unreasonable expansion and contraction during movement, violating the biomechanical constraints of the human body. These errors will significantly reduce the naturalness and credibility of the generated dance. Based on this foundation, the study constructed a human skeleton extraction module based on pose recognition, as shown in Figure 7.

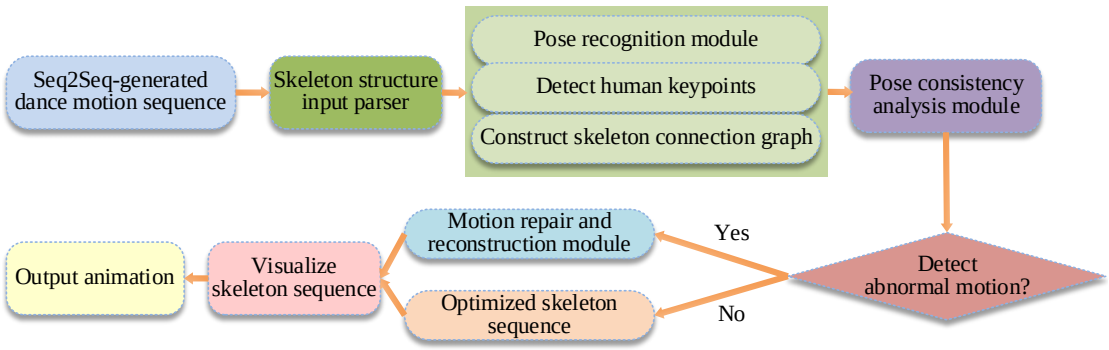


Figure 7. Flowchart of human skeleton extraction module for face recognition

As shown in Figure 7, this module takes a sequence of dance action frames as input and automatically identifies the main joint positions of the human body in each frame through a pose estimation network. It extracts a standardized set of two-dimensional keypoints and constructs the topology structure of the human skeleton. Unlike traditional image

recognition, this module focuses on the geometric connections of skeletons in action sequences and the laws of biological motion. By analyzing the spatial distance, activity amplitude, and temporal coherence between joints, it determines whether there are structural abnormalities or inconsistencies in the action. At the same time, the module introduces a posture symmetry analysis mechanism and bone length stability constraints to identify and repair issues such as left and right limb imbalance, sudden movements, or skeletal distortions. The calculation of the symmetry scoring mechanism is shown in Equation 3, which can directly quantify and correct the first type of asymmetry error.

$$S_{\text{sym}} = 1 - \frac{1}{N}\sum_{k=1}^{N}\frac{\|p_k^L - p_k^R\|}{\|c_k\|} \tag{3}$$

In Equation 3, S_{sym} represents the symmetry score, and the closer the value is to 1, the more symmetrical it is; N represents the total number of symmetric keypoint pairs; c_k represents the reference point corresponding to the kth keypoint; p_k^L and p_k^R are the coordinates of the kth left keypoint and the right keypoint, respectively. The formula for the stability constraint of bone length deviation is shown in Equation 4, which is specifically used to suppress the third type of proportional distortion problem.

$$D_{\text{bone}} = \frac{1}{M}\sum_{(a,b)\in B}\left|\frac{L_{a,b} - \bar{L}_{a,b}}{\bar{L}_{a,b}}\right| \tag{4}$$

In Equation 4, D_{bone} represents the bone length deviation score; M represents the total number of bone segments; $(a,b) \in B$ represents the set of bone segment connection pairs, while B represents the standard bone segment connection relationship in the human body; $L_{a,b}$ represents the actual length of the a -b th bone segment in the current frame; $\bar{L}_{a,b}$ represents the standard length of the bone segment in the reference human body model. After detecting abnormal structures, the system will trigger a local repair algorithm based on skeleton smooth interpolation to finely adjust the trajectory of key points, making the action more in line with the natural motion logic of the human body. This algorithm can effectively smooth out the second type of joint mutation and ensure temporal alignment with the music rhythm during the repair process. As a key feedback link in the dance generation system, this module effectively enhances the structural integrity and motion realism of generated movements, providing a solid guarantee for high-quality music driven dance performance generation.

4. Results

4.1. Performance Analysis of Seq2Seq-HGR-EDG Model

To validate the effectiveness of the Seq2Seq-HGR-EDG model, a unified experimental platform was constructed for training and testing. The initial learning rate was 0.0002, the batch size was set to 32, and the training cycle was 200 rounds. The model parameters were jointly optimized through cross entropy and attitude consistency loss. The specific experimental configuration is shown in Table 1.

Table 1. List of experimental equipment configuration

Category	Component type	Model / Version	Category	Component type	Model / Version
Hardware	Processor	Intel Core i9-13900K	Software	Programming Language	Python 3.9
	GPU	NVIDIA GeForce RTX 3090		Deep Learning Framework	PyTorch 2.0.1
	Memory	64GB DDR4		Audio Processing Tool	NumPy 1.24, Pandas 1.5
	Operating System	Ubuntu 20.04		Video Processing Tool	Matplotlib 3.7, Seaborn 0.12
	-	-		Pose Estimation Framework	Scikit-learn 1.3, XGBoost 1.7

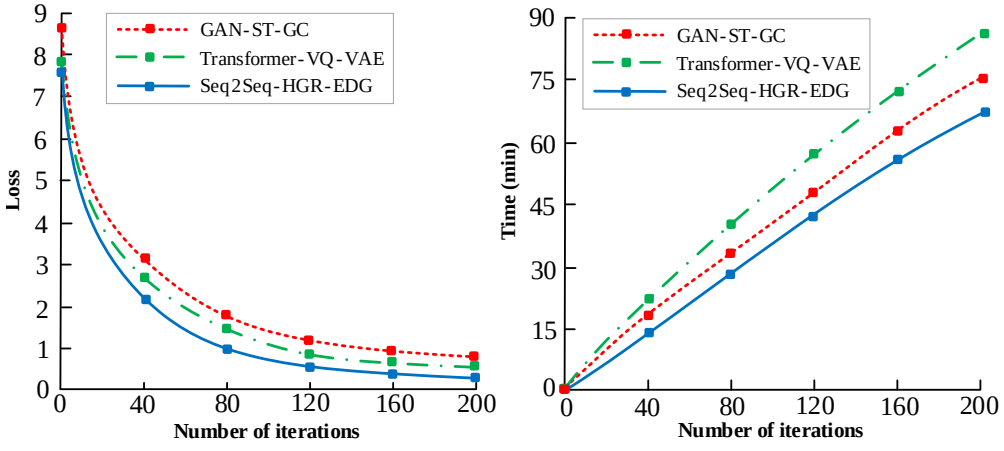
In the experimental environment shown in Table 1, the AIST++ dataset and DanceTrack dataset were selected for model training and evaluation. The AIST++ dataset contains high-quality synchronized dance videos and 3D keypoint annotations from various music styles, making it suitable for modeling music dance collaborative tasks. The DanceTrack dataset provides multi-agent dance tracking sequences and keypoint information, which helps enhance the model's ability to generate actions in complex scenes. At the same time, multiple indicators were used for quantitative evaluation: the beat alignment error was calculated as the average time difference between the significant motion points of the generated action and the music beat points. The joint position error measured the average Euclidean distance between the generated and real joint point coordinates in pixel space. The smoothness of motion was quantified by calculating the amplitude of joint acceleration. The Fréchet starting distance evaluated the overall generation quality by comparing the distribution of the generated sequence and the real sequence in the feature space. To verify the role of each key module in the dance generation process of the Seq2Seq-HGR-EDG model, a ablation experiment was designed. Remove or replace specific components in the model, including the complete Seq2Seq-HGR-EDG model, removal of pose recognition and feedback

module (w/o HGR), Seq2Seq without attention mechanism (w/o Attention), and replacement of multi-scale music features with Mel-Frequency Cepstral Coefficients (MFCC) only (w/o Multi-Scale). The experimental results are shown in Table 2.

Table 2. Comparison of ablation results

Model Configuration	Beat Alignment Error (s)	Joints Position Error (px)	Motion Smoothness	Fréchet Inception Distance (FID)	Accuracy (%)	Inference Time (ms/frame)
Full Model	0.12	11.20	2.41	0.89	92.41	12.33
w/o Multi-Scale	0.16	13.52	3.15	0.82	87.53	11.70
w/o Attention	0.19	15.13	3.97	0.8	84.64	11.41
w/o HGR	0.17	14.81	3.82	0.79	85.82	11.92

According to Table 2, the complete Seq2Seq HGR-EDG model performed the best in all indicators, with a beat alignment error of only 0.12s, joint position error of 11.20px, motion smoothness of 2.41, FID value of 0.89, accuracy of 92.4%, and inference time controlled at 12.3ms/frame, reflecting a good balance between generation accuracy, visual quality, and computational efficiency of the model. When removing multi-scale music features, the joint position error significantly increased to 13.52px, and the smoothness of motion deteriorated. This indicates that using only single features such as MFCC is difficult to fully characterize the rhythm, harmony, and timbre information of music, resulting in a lack of sufficient semantic guidance for the decoder to generate complex body coordination actions, thereby affecting the spatial accuracy and coherence of the actions. Removing the attention mechanism increased the beat alignment error to 0.19 seconds and reduced the motion smoothness to 3.97, with the most prominent negative impact. This confirms the crucial role of attention mechanisms in modeling music dance long temporal, non-uniform alignment relationships. Removing the posture recognition and feedback module mainly led to an increase in joint point errors and FID values, resulting in a decrease in accuracy. This directly proves the effectiveness of the online structural feedback mechanism introduced in the study. The HGR module provided biomechanical supervision signals through symmetry scoring and bone length constraints, which can correct structural distortions that violate human kinematic knowledge in generated actions in real time, significantly improving the visual rationality and structural accuracy of generated actions. The ablation experiment shows that multi-scale features, attention mechanisms, and pose feedback form the backbone of the model's high performance from three levels: input representation, temporal alignment, and structural constraints. On this basis, two mainstream models were introduced into the study: Generative Adversarial Network-Spatio-Temporal Graph Convolutional Network (GAN-ST-GCN), and a dance generation model based on Transformer and Vector Quantized-Variational Auto Encoder (Transformer-VQ-VAE) for comparative experiments. The results are shown in Figure 8.



(a) Comparison of loss value (b) Comparison of calculation time

Figure 8. Comparison of iterative convergence performance and computation time

Figure 8 (a) shows the curve of training loss as a function of iteration times. From the figure, the Seq2Seq HGR-EDG model had the fastest loss reduction in the first 200 iterations, rapidly decreasing from the initial 7.63 to 0.13, which is significantly better than GAN-ST-GCN and Transformer VQ-VAE. This fast initial convergence is mainly attributed to the effective combination of Bi-LSTM encoder and attention mechanism: Bi-LSTM can efficiently capture the forward and backward context of music, providing rich initialization semantic representation for the decoder; The attention mechanism helps the decoder quickly locate effective music context, reducing the randomness of search alignment relationships in the early stages of training, thereby accelerating loss reduction. In contrast, the adversarial training mechanism of GAN-ST-GCN may be unstable in the early stages, while the discrete encoding process of

Transformer VQ-VAE may introduce additional optimization difficulties. As shown in Figure 8 (b), during 200 iterations, the total time consumption of Seq2Seq-HGR-EDG was 68.42 minutes, which was an average reduction of 12.87% compared to GAN-ST-GCN and Transformer-VQ-VAE, and the computation time growth was slow. This indicates that the model has better computational efficiency while ensuring high performance. This result is attributed to the compactness of the model structure and the suppression of invalid learning paths by the attitude feedback mechanism, thereby reducing the cost of invalid gradient propagation and repetitive operations.

4.2. Analysis of Music-Driven Dance Generation Effects

Based on the two open-source datasets AIST++ and DanceTrack, a comprehensive evaluation of the model's dance generation performance was conducted. The study first employed beat alignment error as the evaluation metric, with results shown in Figure 9.

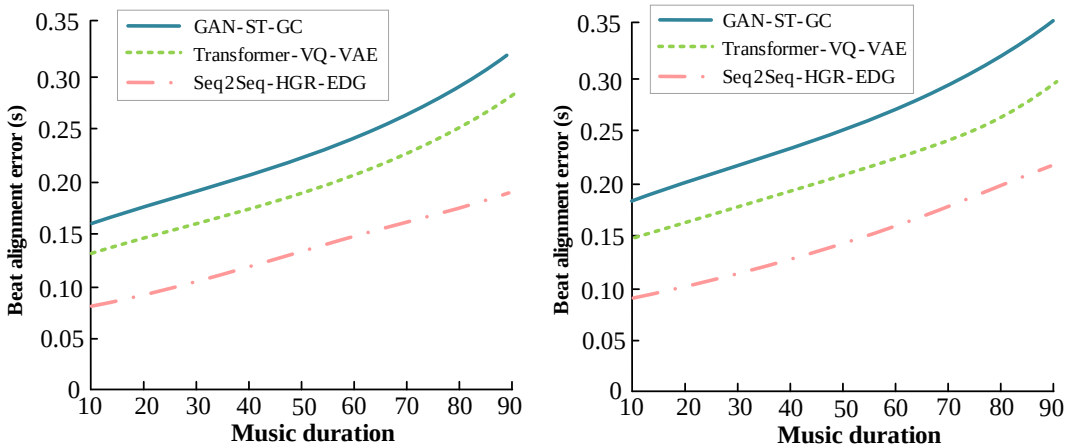


Figure 9. Comparison of beat alignment errors between the three models on different data sets

Figure 9 (a) shows a comparison of beat alignment errors for three models on the AIST++ dataset. When the music duration reached 90 seconds, the beat alignment error of Seq2Seq-HGR-EDG was 0.19 seconds, which was an average reduction of 35.59% compared to the GAN-ST-GCN model and Transformer-VQ-VAE model. This is due to the dynamic focusing ability of the attention mechanism on long-term music segments, as well as the posture recognition module maintaining motion beat synchronization by repairing joint drift. Based on the comparison of beat alignment errors between the three models in Figure 9 (b) on the DanceTrack dataset, when the music duration reached 90 seconds, the beat alignment error of the Seq2Seq-HGR-EDG model was 0.22 seconds, which was an average decrease of 32.31% compared to the other two models. The two datasets show the same trend of change, mainly due to the enhanced cross style generalization of multi-scale music features, and the synergistic effect of temporal modeling and structural optimization, which results in a much lower slope of beat alignment error growth compared to the baseline model. Meanwhile, the study compared the motion smoothness of three models, and the results are shown in Figure 10.

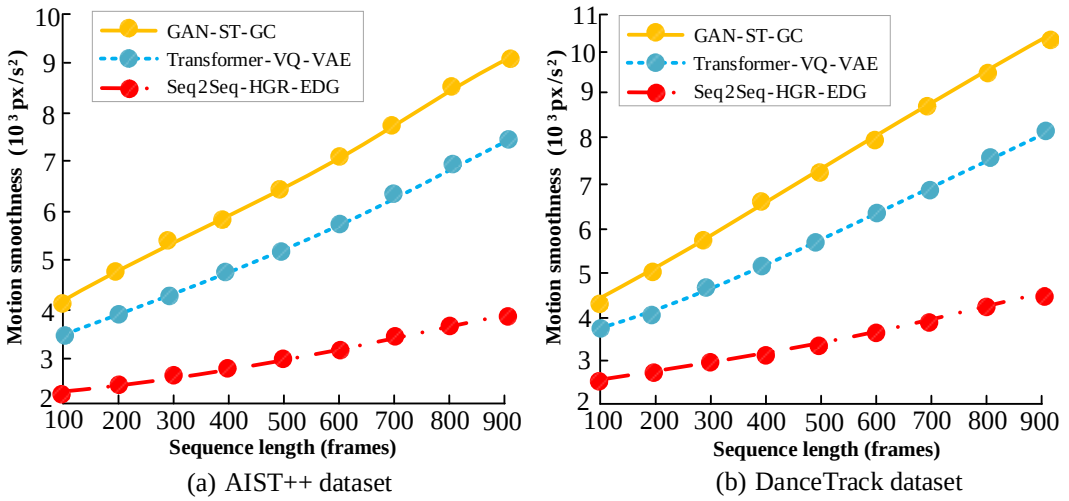
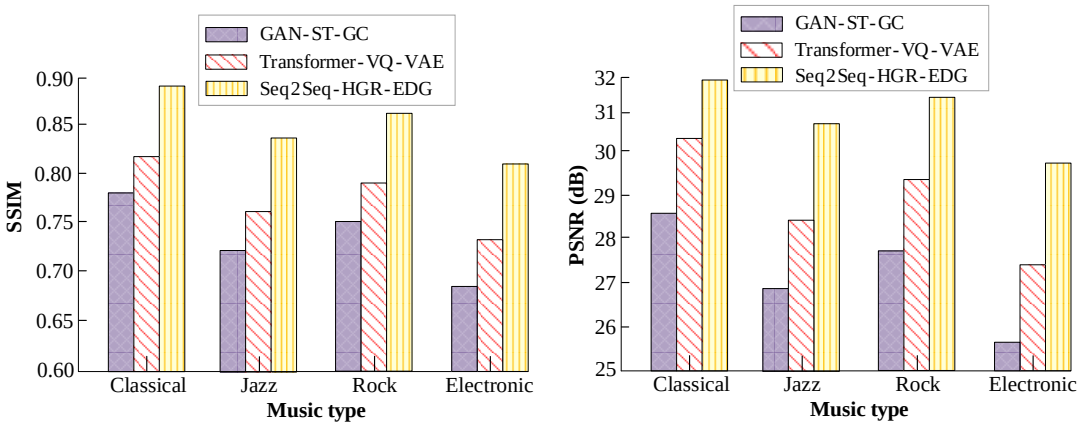


Figure 10. Comparison of motion smoothness between the three models on different data sets

As shown in Figure 10 (a), when the sequence length increased to 900 frames, the motion smoothness of Seq2Seq-HGR-EDG was 3.89×10^3px/s^2, which is only 42.6% of GAN-ST-GCN. At 500 frames, the motion smoothness value of the key nodes was 2.98×10^3px/s^2, which is lower than the benchmark level of Transformer-VQ-VAE at 100 frames. Indicating that the acceleration changes of eq2Seq HGR EDG generated actions are smoother and more in line with the inertial characteristics of human motion. This is mainly due to the dual smoothing constraints within the model: one is the encoding decoding process, where the memory units of Bi-LSTM help maintain the temporal continuity of action states, and the attention mechanism aggregates contextual information to generate more natural transitions for action interpolation. Secondly, the explicit physical constraints imposed by the attitude feedback module directly suppress unreasonable joint shaking and limb extension, ensuring the smoothness of the motion trajectory from a physical perspective. Based on Figure 10 (b), in the 900 frame complex scene of the DanceTrack dataset, the motion smoothness of Seq2Seq-HGR-EDG was 4.35×10^3px/s^2, and the fluctuation difference across datasets was 0.46, which is 57% smaller than the baseline model. Further explanation shows that posture symmetry analysis coordinates the motion relationships between multiple individuals, reduces abrupt changes caused by interaction prediction conflicts, and achieves low fluctuation smooth generation across datasets. On this basis, the study compared the quality of dance images generated under four different music genres: classical music, jazz music, rock music, and electronic music, using the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) indicators. The results are shown in Figure 11.



(a) Comparison of SSIM for different music genres (b) Comparison of PSNR for different music genres

Figure 11. SSIM and PSNR comparison of dance images generated by different music types

Figure 11 (a) shows a comparison of SSIM metrics for dance scenes generated by three models under different music genres. The Seq2Seq-HGR-EDG model performed the best in all music genres, especially in classical music scenes, reaching 0.89, which is much higher than the 0.82 of Transformer-VQ-VAE and the 0.78 of GAN-ST-GCN. This indicates that the generated dance posture is more similar to the real posture in terms of structure, brightness, and contrast, which is directly attributed to the strict maintenance of the human skeleton topology by the posture feedback module, resulting in more accurate joint connection relationships and more realistic limb shapes. According to Figure 11 (b), Seq2Seq HGR-EDG achieved high PSNR values of 31.92dB, 30.75dB, 31.74dB, and 29.63dB in classical, jazz, rock, and electronic music genres, respectively, indicating that the generated action sequence has less noise and clearer texture details. Based on Figure 11, the Seq2Seq-HGR-EDG model can better preserve the texture details and key structures of dance images, and has strong robustness and wide adaptability in various music styles. At the same time, the symmetry optimization effects of three models on different dance movements were studied and tested, and the results are shown in Table 3.

Table 3. Comparison of symmetry optimization effect of dance posture

Dance Stance	Seq2Seq-HGR-EDG	GAN-ST-GCN	Transformer-VQ-VAE
Lift and rotate	0.97	0.89	0.91
Straddle forward	0.94	0.85	0.88
Alternate arms and legs	0.95	0.87	0.9
Single leg support	0.92	0.81	0.86
Extended transverse spread	0.96	0.88	0.91
Diagonal arm sweep	0.93	0.84	0.89
Torso twist step	0.94	0.82	0.87

From Table 3, the Seq2Seq-HGR-EDG model exhibited the best symmetry score in all five typical dance postures, especially in the "arm lift rotation" and "extended lateral extension" postures, where the scores reached 0.97 and 0.96, respectively. This highlighted the core role of the symmetry scoring mechanism in the posture feedback module. During the generation process, the decoder was not only driven by music features, but also guided by gradients optimized towards higher symmetry scores, dynamically adjusting the motion trajectories of the left and right limbs to achieve precise mirroring or collaborative symmetry. In contrast, GAN-ST-GCN scored only 0.81 in the "single leg support" posture, exposing its insufficient modeling ability for dynamic center of gravity offset actions. Although the Transformer VQ-VAE model was overall superior to GAN-ST-GCN, it still had a certain gap with a score of only 0.87 in the "trunk twisting step" involving complex coordination between the upper limbs and trunk. The results indicate that using explicit, kinematic-based structural indicators as one of the optimization objectives can significantly improve the quality of generated actions in advanced aesthetic and coordination dimensions. Finally, visual analysis was conducted on the dance action sequences generated by the three models, and the results are shown in Figure 12.

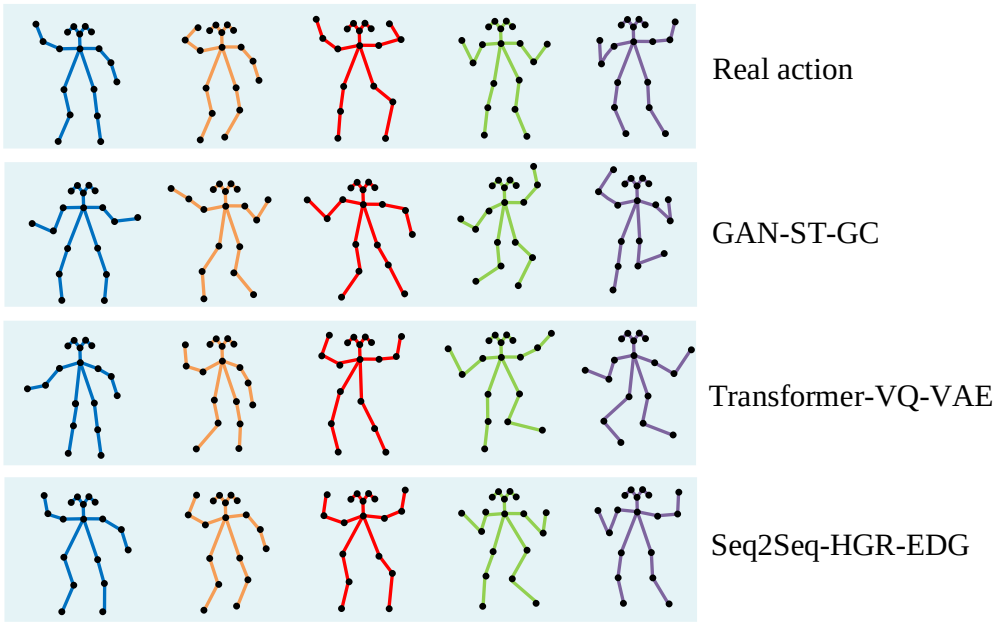


Figure 12. Visualize the comparison results of the motion sequences generated by the dance

From Figure 12, the actions generated by GAN-ST-GCN could roughly present limb dynamics, but some postures had joint misalignment and stiff movements, which may be due to the limited ability of its graph structure modeling to express temporal dependencies. The actions generated by Transformer-VQ-VAE have improved coherence, but there were still a few joint positions in the posture that are not precise enough. The Seq2Seq HGR-EDG model proposed in the study generated actions that are closest to real actions in terms of joint connectivity naturalness, limb motion amplitude, and overall posture richness. For example, the generated arm swing trajectory was smooth and conformed to the motion arc, and the torso twist was coordinated and unified with the lower limb gait. This high degree of naturalness and rationality is due to its integration of Bi-LSTM's temporal memory ability, dynamic focusing mechanism of attention mechanism, and posture recognition feedback optimization strategy. The three work together to enhance the model's dual constraint ability on music rhythm and human skeletal structure.

5. Discussion

5.1. The Significance of the Results

The study comprehensively validated the performance of the proposed Seq2Seq-HGR-EDG model in music-driven dance generation tasks through the design of ablation experiments and multi-model comparative experiments. The results showed that the complete model outperformed variants and other mainstream models that remove key modules in maintaining the naturalness, temporal rationality, and pose symmetry of dance movements. Especially in dance pose generation involving complex limb coordination and high symmetry requirements, Seq2Seq-HGR-EDG exhibited significant advantages. This achievement not only enriches the theoretical system of cross modal generation of music and dance, but also provides a new technological path for virtual dance generation, intelligent entertainment, and artistic creation.

5.2. Comparison with Existing Research

Current mainstream approaches in the field of music-driven dance generation mainly rely on RNNs or conditional generative adversarial networks, with research generally focusing on capturing temporal features and maintaining action continuity [28]. For example, Cheng et al. [29] proposed a hierarchical generation framework that decouples dance synthesis into two modules: "structure" and "style." Their work emphasized action style modeling and user controllability; however, the "structure" module still relied primarily on implicit data-driven learning and did not incorporate explicit human pose feedback mechanisms. As a result, its structural control capability remained limited when handling complex dance movements that require strong spatial symmetry and biomechanical rationality.

In contrast, the present study innovatively employs a pose recognition module as a real-time structural guidance mechanism, directly improving the spatial consistency and biomechanical rationality of generated actions through explicit optimization objectives such as symmetry scoring and bone-length constraints. From another perspective, existing studies generally concentrate on two core challenges: "temporal alignment" and "spatial rationality." For instance, the coherent dance action generation model proposed by Jiang & Yan [30] achieved coherent motion generation through latent-space learning, emphasizing the physical rationality of movements and amplitude regulation, but it did not deeply model the collaborative relationship between music and dance.

Although the works of Kim et al. [9] and Au et al. [10] sought to improve movement diversity and music matching, their methods were largely based on adversarial learning or normalizing flows, resulting in limited transparency and controllability of the generation process. Moreover, the stability of long-sequence alignment under complex rhythmic conditions still requires improvement. The present study achieves finer-grained responses to musical rhythm and semantics through multi-scale music feature extraction and attention-enhanced temporal modeling, thereby demonstrating superior performance on key synchronization metrics such as beat alignment error.

At the same time, the proposed enhancement mechanism effectively integrates multimodal features, improving the model's perception of rhythm and emotion. This overcomes the limitations of traditional methods in terms of generation quality and expressive richness. Furthermore, the use of finer-grained pose data as feedback improves movement fluency and the naturalness of motion transitions, highlighting the technological advancement and innovation of the proposed approach.

5.3. Limitations of the Research

Although the experimental results validate the effectiveness of the proposed model, several limitations still remain. First, the training and testing processes mainly rely on the AIST++ dataset, which covers a limited range of dance styles. As a result, the model's generalization capability across the diverse dance styles encountered in real-world scenarios requires further validation. Second, there are substantial differences in dance style, annotation format, and data quality between the two open-source datasets used in this study, namely AIST++ and DanceTrack. AIST++ contains high-quality street dance sequences that are accurately aligned with music and provides 3D keypoint annotations. In contrast, DanceTrack focuses on 2D pose sequences in multi-target tracking scenarios, and the correspondence between dance movements and music is not as strictly aligned as in AIST++. These differences lead to slightly lower model performance on the DanceTrack dataset, particularly in terms of beat alignment error and motion smoothness, compared to the performance achieved on AIST++.

This observation reflects the model's dependence on high-quality and strongly aligned data. Moreover, its cross-dataset generalization capability, especially its adaptability to weak music–dance associations or differing annotation standards, still requires improvement. To address this issue, future work should develop more robust data preprocessing strategies, such as unified 2D/3D pose projection, cross-dataset rhythm feature normalization, and domain adaptation techniques to reduce dataset discrepancies. Such improvements would enable a more realistic and comprehensive evaluation of the model's generalization ability.

Furthermore, although the proposed framework introduces human pose feedback, the feedback mechanism is currently based on offline recognition results and does not achieve real-time coupling with the generation process. This limitation restricts the applicability of the model in interactive scenarios. In addition, the current evaluation metrics mainly focus on symmetry and naturalness and do not fully capture more subjective dimensions such as creativity and artistic expression in generated dance movements.

5.4. Supplementing and Challenging Existing Theories

Starting from multimodal input and a posture feedback mechanism, this study proposes a more tightly coupled generation framework between music and movement, extending previous research that relied primarily on audio rhythm modeling or graph convolutional networks. The results indicate the existence of a learnable deep alignment relationship between multi-scale audio features and human posture structures. This finding provides a new paradigm for cross-modal dance generation, as well as a theoretical foundation and modeling framework for multimodal sequence generation tasks, with strong generalization capability and expansion potential.

6. Conclusion

To address the core problems of poor rhythm synchronization, unnatural movements, and structural distortion in music-driven dance generation, an enhanced generation model, Seq2Seq HGR-EDG, was proposed by integrating a Seq2Seq network with human pose recognition technology. The core of the proposed method lies in capturing music semantics through multi-scale music feature extraction and attention-enhanced temporal modeling, while innovatively incorporating an OpenPose-based skeleton extraction and feedback mechanism to optimize structure and calibrate generated actions in real time. This forms a closed-loop framework for the collaborative enhancement of music understanding and action generation.

Experimental results demonstrate that the proposed model significantly outperforms mainstream comparison models on the AIST++ and DanceTrack datasets. The model achieves outstanding performance indicators, including a beat alignment error as low as 0.12 s, a joint position error of 11.2 px, and a motion smoothness score of 2.41, while maintaining lower error accumulation during long-sequence generation. In addition, the model performs exceptionally well in terms of structural rationality and visual quality of generated dance postures. The symmetry score for movements such as "arm lifting and rotation" reaches as high as 0.97, while the generated images achieve SSIM and PSNR values of 0.89 and above 31.9 dB, respectively. These results confirm the effectiveness of the proposed framework in improving movement naturalness, synchronization accuracy, and cross-style robustness.

This study provides a new paradigm for cross-modal dance generation that balances temporal alignment and spatial structural constraints. Future work will explore the integration of graph attention networks to further enhance action semantic relationship modeling, as well as the incorporation of 3D pose prediction technologies to support the development of more immersive virtual reality and interactive application scenarios.

7. Nomenclature

Symbol	Meaning and explanation
$\alpha_{t,i}$	The attention weight of encoding time step i during decoding time step t
t	Decoding time step
i	Coding time step
s_t	The hidden state of the decoder at time step t
j	Represents the cyclic index of the encoder time step
□$_i$	The hidden states of the encoder at time step i
□$_j$	The hidden states of the encoder at time step j
T	The length of the encoding sequence
c_t	The context vector at decoding time step t
S_{sym}	The symmetry score, and the closer the value is to 1, the more symmetrical it is
N	The total number of symmetric keypoint pairs
c_k	The reference point corresponds to the kth keypoint;
p_k^L	The coordinate of the kth left keypoint
p_k^R	The coordinate of the kth right keypoint
D_{bone}	The bone length deviation score
M	The total number of bone segments
$(a, b) \in B$	The set of bone segment connection pairs, B represents the standard bone segment connection relationship in the human body
$L_{a,b}$	The actual length of the a -b th bone segment in the current frame
$\bar{L}_{a,b}$	The standard length of the bone segment in the reference human body model

8. Declarations

8.1. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

8.2. Funding

The author received no financial support for the research, authorship, and/or publication of this article.

8.3. Institutional Review Board Statement

Not applicable.

8.4. Informed Consent Statement

Not applicable.

8.5. Declaration of Competing Interest

The author declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

9. References

[1] Xinlei, S. (2023). Folk dance and music art of the new generation: China's experience. Voprosy Istorii, 3(1), 170-177. doi:10.31166/voprosyistorii202303statyi31.

[2] Han, B., Li, Y., Shen, Y., Ren, Y., & Han, F. (2024). Dance2MIDI: Dance-driven multi-instrument music generation. Computational Visual Media, 10(4), 791–802. doi:10.1007/s41095-024-0417-1.

[3] He, D. (2025). Seq2Seq Text Recognition Method for Large-Scale Corpus Linguistics Knowledge Based on Transformer. International Journal of High Speed Electronics and Systems, 34(01), 2540069. doi:10.1142/S0129156425400695.

[4] Li, K., & Santos, E. (2024). Artificial Intelligence Choreography: 3D Dance Generation Based on Deep Generative Adversarial Networks. Journal of Network Intelligence, 9(3), 1725–1741. doi:10.6025/jni/2024/9/3/1725-1741.

[5] Kim, W., Sung, J., Saakes, D., Huang, C., & Xiong, S. (2021). Ergonomic postural assessment using a new open-source human pose estimation technology (OpenPose). International Journal of Industrial Ergonomics, 84, 103164. doi:10.1016/j.ergon.2021.103164.

[6] Zhou, Z., Huo, Y., Huang, G., Zeng, A., Chen, X., Huang, L., & Li, Z. (2025). QEAN: quaternion-enhanced attention network for visual dance generation. Visual Computer, 41(2), 961–973. doi:10.1007/s00371-024-03376-5.

[7] Zeng, D. (2025). AI-Powered Choreography Using a Multilayer Perceptron Model for Music-Driven Dance Generation. Informatica (Slovenia), 49(20), 137–148. doi:10.31449/inf.v49i20.8103.

[8] Yang, Z., Wen, Y. H., Chen, S. Y., Liu, X., Gao, Y., Liu, Y. J., Gao, L., & Fu, H. (2024). Keyframe Control of Music-Driven 3D Dance Generation. IEEE Transactions on Visualization and Computer Graphics, 30(7), 3474–3486. doi:10.1109/TVCG.2023.3235538.

[9] Kim, J., Kwon, B., Kim, J., & Lee, S. (2023). MNET++: Music-Driven Pluralistic Dancing Toward Multiple Dance Genre Synthesis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(12), 15036–15050. doi:10.1109/TPAMI.2023.3312092.

[10] Au, H. Y., Chen, J., Jiang, J., & Guo, Y. (2024). ReChoreoNet: Repertoire-based Dance Re-choreography with Music-conditioned Temporal and Style Clues. Machine Intelligence Research, 21(4), 771–781. doi:10.1007/s11633-023-1478-9.

[11] Wang, Q., Tong, G., & Zhou, S. (2023). A Study of Dance Movement Capture and Posture Recognition Method Based on Vision Sensors. HighTech and Innovation Journal, 4(2), 283–293. doi:10.28991/HIJ-2023-04-02-03.

[12] Jhansi Rani, C., & Devarakonda, N. (2023). Generative adversarial network based data augmentation and quantum based convolution neural network for the classification of Indian classical dance forms. Journal of Intelligent and Fuzzy Systems, 45(4), 6107–6125. doi:10.3233/JIFS-231183.

[13] Zhou, Q., Jiang, D. L., & Wang, G. (2024). 3D Dance Movement Recognition Based on Somatic Interaction Devices and Neural Networks. Journal of Network Intelligence, 9(4), 2290–2303.

[14] Bao, C., & Sun, Q. (2023). Generating Music with Emotions. IEEE Transactions on Multimedia, 25, 3602–3614. doi:10.1109/TMM.2022.3163543.

[15] Liang, X., Li, W., Huang, L., & Gao, C. (2024). DanceComposer: Dance-to-Music Generation Using a Progressive Conditional Music Generator. IEEE Transactions on Multimedia, 26(6), 10237–10250. doi:10.1109/TMM.2024.3405734.

[16] Cai, X., Wang, T., Lu, R., Jia, S., & Sun, H. (2023). Automatic generation of Labanotation based on human pose estimation in folk dance videos. Neural Computing and Applications, 35(35), 24755–24771. doi:10.1007/s00521-023-08206-8.

[17] Li, W., Wu, L., Wen, X., Feng, Q., Zhou, T., Yang, L., & Yin, Z. (2024). Runoff simulation study based on LSTM-Seq2seq model optimized by attention mechanism. Journal of Glaciology and Geocryology, 46(3), 980–992. doi:10.7522/j.issn.1000-0240.2024.0078.

[18] Li, W., Li, K., Yue, Y., Wang, J., Xu, H., & Luo, Y. (2024). ISAR Range Alignment Based on a Spatiotemporal Attention-Seq2Seq Network. Journal of Signal Processing, 40(9), 1659–1673. doi:10.12466/xhcl.2024.09.008.

[19] Yang, L., Wei, C., Yang, J., Ma, J., Guo, H., Cheng, L., & Li, Z. (2024). Seq2Seq-AFL: Fuzzing via sequence-to-sequence model. International Journal of Machine Learning and Cybernetics, 15(10), 4403–4421. doi:10.1007/s13042-024-02153-z.

[20] Tingting, L., Bo, L., & Chunzhu, L. (2024). Aircraft trajectory prediction within terminal area based on Seq2Seq-attention model. Science Technology and Engineering, 24(9), 3882-3895.

[21] Huang, J., Huang, X., Yang, L., & Tao, Z. (2024). Dance-conditioned artistic music generation by creative-GAN. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 107(5), 836-844. doi:10.1587/transfun.2023EAP1059.

[22] Piekut, B. (2024). Sound against Music. TDR - The Drama Review - A Journal of Performance Studies, 68(2), 35–54. doi:10.1017/S1054204324000066.

[23] Zhang, C., Zhang, H., Pu, T., & Pan, J. (2025). Supply Chain Demand Forecasting Based on Data Mining Algorithm and Seq2Seq. International Journal of Control, Automation and Systems, 23(1), 89–104. doi:10.1007/s12555-024-0141-8.

[24] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv Preprint, arXiv:1409.0473. doi:10.48550/arXiv.1409.0473.

[25] Shi, Y., & Han, S. (2025). Multimedia interactive creative dance choreography integrating intelligent chaotic art algorithms. Journal of Computational Methods in Sciences and Engineering, 25(4), 2976–2991. doi:10.1177/14727978251318055.

[26] Zhou, Q., Li, M., Zeng, Q., Aristidou, A., Zhang, X., Chen, L., & Tu, C. (2023). Let's all dance: Enhancing amateur dance motions. Computational Visual Media, 9(3), 531–550. doi:10.1007/s41095-022-0292-6.

[27] Siyao, L., Yu, W., Gu, T., Lin, C., Wang, Q., Qian, C., Loy, C. C., & Liu, Z. (2023). Bailando++: 3D Dance GPT with Choreographic Memory. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(12), 14192–14207. doi:10.1109/TPAMI.2023.3319435.

[28] Hasanvand, M., Nooshyar, M., Moharamkhani, E., & Selyari, A. (2023). Machine Learning Methodology for Identifying Vehicles Using Image Processing. Artificial Intelligence and Applications, 1(3), 154–162. doi:10.47852/bonviewAIA3202833.

[29] Cheng, Y., Jiang, Y., & Wang, Y. (2024). Music-stylized hierarchical dance synthesis with user control. Virtual Reality and Intelligent Hardware, 6(5), 339–357. doi:10.1016/j.vrih.2024.06.004.

[30] Jiang, H., & Yan, Y. (2024). Sensor based Dance Coherent Action Generation Model using Deep Learning Framework. Scalable Computing: Practice and Experience, 25(2), 1073–1090. doi:10.12694/scpe.v25i2.2648.