Publication Date: 30 June 2026
Archs Sci. (2026) Volume 75, Issue 3 Pages 46-57, Paper ID 2025305.

TaylorDance: Music-Driven 3D Dance Generation Based on Spatiotemporally Decoupled Taylor Series Expansion

Dan Wang
School of Music, Jianghan University, Wuhan 430056, China; dwang0868@163.com

Abstract Music-driven 3D dance motion generation is a critical and challenging task in cross-modal content creation, aiming to synthesize physically plausible human dance sequences synchronized with input music in terms of both rhythm and style. Although diffusion-based methods have made significant progress, existing approaches typically model coupled spatiotemporal dimensions and struggle to simultaneously capture the beat structure of music and the complex dynamics of human poses with precision. To address this issue, we propose a novel multiscale spatiotemporal hybrid attention framework. The core innovation lies in decoupling the spatiotemporal complexity of dance generation: for temporal modeling, we design an "implicit keyframe learning with Taylor series expansion" mechanism that adaptively extracts key beat frames via neural networks and employs a third-order Taylor series for higher-order continuous motion synthesis; for spatial modeling, we adopt an anatomy-based partitioning strategy, dividing the human body into seven functional regions for independent and collaborative modeling to enhance complex pose-generation capabilities. Finally, a multiscale architecture integrates spatiotemporal features at different levels of granularity. Extensive experiments on the AIST++ and PopDanceSet datasets demonstrate that our method outperforms state-of-the-art approaches in the physical plausibility of generated motions (PFC and PBC), diversity (Div-k and Div-g), and the critical music–dance beat-alignment score (BAS), thereby validating the effectiveness and advantages of the proposed framework.

Index Terms dance motion generation, diffusion models, spatiotemporal attention, sequence modeling, music–motion synchronization

I. Introduction

With the rapid advancement of artificial intelligence and computer vision technologies, generative models have demonstrated unprecedented potential for content creation. Among these applications, music-driven 3D dance motion generation represents a significant branch of cross-modal generation tasks [1], [2]. It aims to automatically generate rhythmically synchronized, aesthetically pleasing, and coherent 3D human dance sequences based on given musical audio. This technology not only provides efficient, low-cost content-creation tools for digital entertainment, such as virtual idols, game-character animation, and film special effects, but also opens up new possibilities for human–computer interaction, dance-instruction assistance, and physical-rehabilitation training [3]. Its core challenge lies in establishing an intelligent model capable of deeply understanding high-level semantic information, such as musical beats and styles, and mapping this understanding onto expressive, continuous motion sequences that comply with human kinematic and dynamic constraints.

Despite significant progress in this field, generating high-quality, high-fidelity dance movements still presents a series of formidable challenges, primarily manifested in three dimensions. First, dance is a highly complex spatiotemporal art form. In the temporal dimension, movements must strictly adhere to the beat and rhythm of the music, forming coherent sequences with rhythmic variations and dynamic contrasts; in the spatial dimension, the human body—as a multijointed dynamic system—must maintain natural and coordinated postures within anatomical constraints [4], [5]. Traditional methods often blend spatiotemporal features without effectively decoupling these distinct complexities, thereby hindering models from simultaneously capturing long-range rhythmic dependencies and subtle local postural variations.

Second, the generated dance movements must adhere to fundamental physical laws. For instance, foot contact with the ground serves as the foundation for supporting body movement and generating acceleration, while upper-limb and torso motions are also constrained by biomechanical coupling relationships [6]. Movements that violate physical laws severely compromise the realism and usability of the generated results. Effectively embedding implicit physical constraints into generative models therefore remains a critical challenge. Third, an ideal model must strike a balance between "alignment" and "diversity." On the one hand, the key beat points of the generated movements must be precisely synchronized with the musical rhythm. On the other hand, for the same musical segment, the model should generate stylistically diverse, non-

repetitive, and plausible dances to avoid mode collapse. This requires the model both to perceive temporal events in the music accurately and to cover the rich range of modalities within the human motion space [7], [8].

Early studies formalized the problem as a sequence-to-sequence mapping, employing recurrent neural networks (RNNs) or temporal convolutional networks (TCNs) for encoding and decoding, and introduced discriminators from generative adversarial networks (GANs) to enhance the naturalness of the generated movements [9]. However, these methods are prone to error accumulation and motion drift when generating long sequences, and their ability to model complex musical rhythms is limited. To improve motion quality and diversity, studies such as Bailando introduced vector quantization (VQ) techniques. These approaches first discretize the continuous motion space into a motion codebook and then combine it with Transformers for autoregressive prediction. Such approaches enhance controllability through discretization and often incorporate reinforcement learning to optimize beat alignment. However, their two-stage training process is complex, and the discrete representation may result in the loss of subtle continuous motion information. More recently, methods such as EDGE and POPDG have successfully introduced diffusion models into dance generation. The powerful distribution-matching capability of diffusion models enables the generation of highly smooth and diverse motion sequences [10], [11]. EDGE pioneered the integration of diffusion models with the large-scale music model Jukebox, significantly improving generation quality. POPDG further strengthened the modeling of physical connections between joints in the decoder and enhanced the alignment module. Diffusion models have consequently become the current mainstream paradigm, demonstrating considerable potential.

Despite the promising results achieved by existing methods, particularly those based on diffusion models, several limitations remain. First, most models employ coupled spatiotemporal modeling, in which a unified attention mechanism simultaneously handles temporal and interjoint relationships [12]. This may hinder the model's ability to focus on learning distinct aspects of these features. Second, the modeling of temporal dynamics largely relies on the global-attention or local-window-attention mechanisms of standard Transformers, without explicit modeling of the essential temporal structure of dance movements—namely, "keyframes" and "transition frames." This may limit their ability to accurately capture musical beat structures and finely control motion continuity. Finally, spatial modeling typically treats the entire human pose as a single entity or simplifies it into isolated components, failing to leverage anatomical prior knowledge to guide complex pose generation.

To systematically address these challenges and overcome the existing limitations, this paper proposes a novel dance-motion-generation framework called Multiscale Spatiotemporal Hybrid Attention. Our core approach involves structurally decoupling the spatiotemporal complexity of dance generation, modeling each dimension through dedicated modules, and subsequently fusing the resulting features in a collaborative manner.

II. Spatio-Temporal Collaborative Attention Modeling Mechanism

Addressing the dual complexities of temporal continuity and spatial structure in dance movements, this paper introduces a spatio-temporal collaborative attention modeling approach. It divides the motion generation process into two independent yet synergistic branches: temporal dimension modeling and spatial dimension modeling. This dual-branch architecture enhances overall modeling efficiency and generation quality [13], [14].

In the temporal dimension, as illustrated in Figure 1, a deep neural network first encodes key frames within the dance sequence to derive latent representations capturing semantic and rhythmic information (highlighted in red poses in the schematic). This process emphasizes mapping musical beats to corresponding human motion changes, thereby accurately identifying core rhythmic points in the dance. Building upon this foundation, a continuity approximation method based on Taylor series expansion is employed to infer and reconstruct non-key frames (represented by blue poses in the diagram). This ensures smooth transitions and coherent evolution of movements along the temporal axis. Through this sequential modeling strategy, the model not only reduces the computational complexity of directly generating the entire sequence but also significantly enhances the naturalness and fluidity of the dance motion sequence.

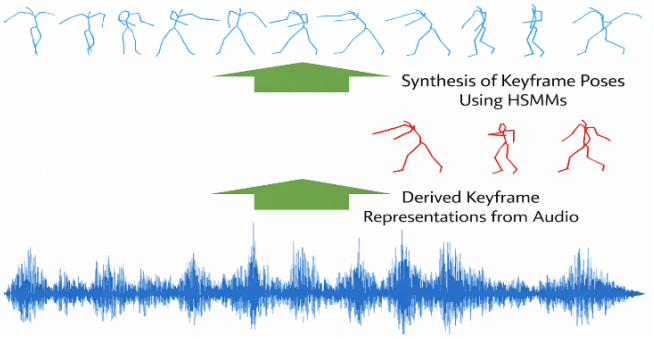

Figure 1: Time Series Modeling Schematic Diagram Using Taylor Series

During the modeling process in spatial dimensions, as shown in Figure 2, this paper adopts a local modeling approach based on human structural prior knowledge to address the high complexity of dance poses in terms of joint distribution and limb coordination. Specifically, the human skeleton is anatomically divided into seven relatively independent and semantically distinct subregions: head, trunk spine, left upper limb, right upper limb, left lower limb, right lower limb, and root node. A spatial feature representation is constructed for each subregion to fully capture the differences in movement patterns and degrees of freedom across body parts.

Building upon this foundation, fine-grained modeling captures spatial relationships within each partition, while cross-partition coordination is unified through a subsequent feature fusion module. This module facilitates information exchange and constraint propagation between body parts while preserving local motion details, thereby generating dance poses that exhibit greater structural consistency alongside richer, more coordinated local expressions.

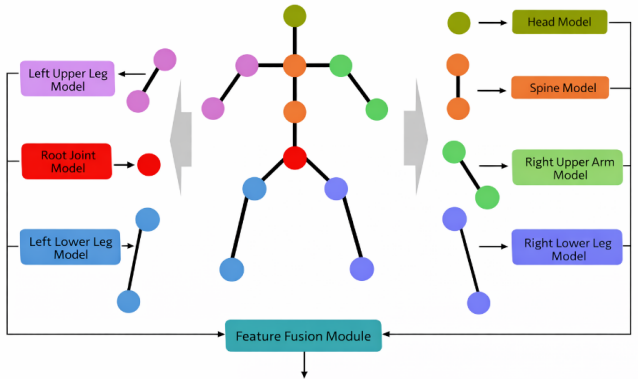

Figure 2: Schematic Diagram of Human Anatomical Segmentation-Based Spatial Modeling

A. Keyframe Representation Methods Based on Latent Space Coding

Let the dance motion sequence features and music sequence features be represented respectively as

$$\mathbf{X}_d \in \mathbb{R}^{T_d \times D_d}, \tag{1}$$

$$\mathbf{X}_m \in \mathbb{R}^{T_m \times D_m}, \tag{2}$$

where, T_d and T_m denote the temporal lengths of the dance sequence and music sequence, respectively, while D_d and D_m represent their corresponding feature dimensions. To characterize the rhythmic correlation between dance and music within a unified latent space, this paper employs implicit keyframe modeling to compress and reconstruct the dance sequence.

First, the dance sequence features are mapped to a latent space to generate a value matrix corresponding to keyframes. This process can be expressed as:

$$\mathbf{V} = \mathbf{X}_d \mathbf{W}_v, \tag{3}$$

where, $\mathbf{W}_v \in \mathbb{R}^{D_d \times D_h}$ represents the learnable parameters of the linear mapping, where D_h denotes the latent feature dimension. $\mathbf{V} \in \mathbb{R}^{T_d \times D_h}$ signifies the value representation of the dance sequence in the latent space.

1) Construction of the Key Matrix

In the implicit keyframe modeling process, the Key Matrix characterizes the rhythmic features of the music sequence in the temporal dimension and provides a matching basis for subsequent attention weight calculations [15]. Considering the high correlation between dance movements and musical beats, this paper maps music features onto a representation space consistent with the dance latent space.

Specifically, the input music sequence feature $\mathbf{X}_m \in \mathbb{R}^{T_m \times D_m}$ undergoes a linear transformation to yield the corresponding key matrix, calculated as follows:

$$\mathbf{K} = \mathbf{X}_m \mathbf{W}_k, \tag{4}$$

where, $\mathbf{W}_k \in \mathbb{R}^{D_m \times D_h}$ denotes the learnable parameter matrix, D_h represents the dimension of the latent feature space, and $\mathbf{K} \in \mathbb{R}^{T_m \times D_h}$ serves as the key representation of the music sequence in the latent space.

2) Generating Implicit Keyframes

After obtaining the value matrix V for the dance sequence and the key matrix K for the music sequence, this paper employs an attention mechanism to compute the correlation between dance and music in the latent space, thereby generating an implicit keyframe representation. This process aims to highlight dance time segments highly synchronized with the musical rhythm, enabling the model to automatically focus on positions with representative motion changes.

First, the similarity score between the key matrix and value matrix is computed, expressed as:

$$\mathbf{A} = \mathrm{Softmax}\left(\frac{\mathbf{K}\mathbf{V}^{\top}}{\sqrt{D_h}}\right), \tag{5}$$

where, $\mathbf{A} \in \mathbb{R}^{T_m \times T_d}$ denotes the attention weight matrix, while $\sqrt{D_h}$ scales the inner product result to mitigate the issue of excessively large values in high-dimensional spaces.

Subsequently, the attention weights are applied to perform a weighted sum over the value matrix, yielding the latent representation of the implicit keyframe:

$$\mathbf{Z} = \mathbf{A}\mathbf{V}, \tag{6}$$

where, $\mathbf{Z} \in \mathbb{R}^{T_m \times D_h}$ represents the implicit keyframe features extracted under musical rhythm guidance. This representation synthesizes the temporal semantics of dance movements with musical beat information, not only compressing redundant content in the original sequence but also providing a more compact and discriminative keyframe description for subsequent motion reconstruction and generation.

B. Continuous Time Series Modeling Method Based on Taylor Series Expansion

For a dance movement sequence comprising T frames, it can be abstracted as a vector function varying over time

$$\mathbf{f}(t) \in \mathbb{R}^{D}, \ t = 1, 2, \ldots, T, \tag{7}$$

where, t denotes the discrete time index, and f(t) represents the D-dimensional action feature vector corresponding to frame t.

In the previous section, the model has obtained an implicit set of key frames.

$$\mathbf{G} \in \mathbb{R}^{M \times K \times D_h}, \tag{8}$$

where, M denotes the number of implicit subspaces, each containing K implicit keyframes, with each keyframe being a D_h-dimensional feature vector. Let the keyframes in the mth subspace be denoted as

$$\{\mathbf{e}_{m,k}\}_{k=1}^{K},\ \mathbf{e}_{m,k} \in \mathbb{R}^{D_h}. \tag{9}$$

The following illustrates the temporal modeling process using a single subspace as an example; the remaining subspaces are processed in parallel using the same strategy.

First, a neural network is employed to predict the temporal position of each implicit keyframe, yielding its corresponding time index within the original motion sequence.

$$\{t_k\}_{k=1}^{K},\ t_k \in [1, T]. \tag{10}$$

Based on this, the motion sequence is treated as a time-varying function that is differentiable near key time points, and a Taylor series expansion of $\mathbf{f}(t)$ is performed at each t_k. For any time t, its approximate expression can be written as

$$\mathbf{f}(t) \approx \mathbf{f}(t_k) + \mathbf{f}'(t_k)(t - t_k) + \frac{1}{2}\mathbf{f}''(t_k)(t - t_k)^2 + \cdots, \tag{11}$$

where, $\mathbf{f}^{'}(t_k)$ and $\mathbf{f}^{''}(t_k)$ denote the first- and second-order derivatives at time point t_k, respectively, characterizing the velocity and acceleration properties of motion changes.

Balancing computational complexity control with practical generation outcomes, this paper employs a third-order Taylor series to approximate the temporal evolution of motion sequences. This order ensures sufficient expressive capability while effectively avoiding computational redundancy and numerical instability issues associated with higher-order terms. Specifically, for any time point t, near the time position t_k corresponding to an implicit keyframe, the motion feature function can be approximated as:

$$\begin{aligned}\mathbf{f}(t) \approx & \mathbf{f}(t_k) + \mathbf{f}'(t_k)(t - t_k) + \frac{1}{2}\mathbf{f}''(t_k)(t - t_k)^2 \\ & + \frac{1}{6}\mathbf{f}'''(t_k)(t - t_k)^3,\end{aligned} \tag{12}$$

where, $\mathbf{f}(t_k)$ represents the motion feature vector at the keyframe, while $\mathbf{f}^{'}(t_k)$, $\mathbf{f}^{''}(t_k)$, and $\mathbf{f}^{'''}(t_k)$ respectively characterize the first-, second-, and third-order temporal trends of the motion at that time point, corresponding to dynamic attributes such as velocity, acceleration, and curvature of change.

Given neural networks' inherent ability to automatically weight and scale features of different orders during parameter learning, the proportional relationships among coefficients corresponding to derivatives of various orders can be modeled autonomously by network weights. Consequently, factorial terms like 1/2 and 1/3 in the Taylor series need not be explicitly introduced but are implicitly incorporated into learnable parameters. Under this assumption, the approximate expression for motion features near time t_k can be further simplified to

$$\begin{aligned}\mathbf{f}(t) \approx & \mathbf{f}(t_k) + \alpha_1 \mathbf{f}'(t_k)(t - t_k) + \alpha_2 \mathbf{f}''(t_k)(t - t_k)^2 \\ & + \alpha_3 \mathbf{f}'''(t_k)(t - t_k)^3,\end{aligned} \tag{13}$$

where, α_1, α_2, and α_3 denote weight coefficients automatically learned by the neural network, which regulate the contribution of different-order derivative terms to the final motion reconstruction.

This expression implicitly encodes the original derivative information of each order and its corresponding scaling coefficient within the network parameters. Based on this, for a set of motion functions approximated by K implicit keyframes at different temporal positions:

$$\left\{\widehat{\mathbf{f}}_k(t)\right\}_{k=1}^{K}. \tag{14}$$

These local approximation results require unified integration to obtain a global action representation at any time t.

This paper employs a time-distance-based Gaussian kernel function to perform weighted fusion of approximation results from different keyframes. The core idea is that implicit keyframes closer to the current time t contribute more significantly to the action at that moment. The corresponding weighting coefficient is defined as:

$$w_k(t) = \frac{\exp\left(-\frac{(t-t_k)^2}{2\sigma^2}\right)}{\sum_{j=1}^{K} \exp\left(-\frac{(t-t_j)^2}{2\sigma^2}\right)}, \tag{15}$$

where, t_k denotes the temporal position corresponding to the kth implicit keyframe, while σ represents the bandwidth parameter of the Gaussian kernel, which controls the smoothness of the temporal neighborhood.

Ultimately, the action feature at time t is obtained by the weighted sum of the local approximation results:

$$\mathbf{f}(t) = \sum_{k=1}^{K} w_k(t)\widehat{\mathbf{f}}_k(t). \tag{16}$$

By introducing a Gaussian weighting fusion mechanism based on temporal distance, the model achieves smooth transitions between local modeling results of different implicit keyframes, thereby effectively enhancing the continuity and consistency of action sequences across the entire temporal dimension.

C. Partitioned Spatial Modeling Approach Based on Human Anatomy Priors

Addressing the high nonlinearity and regional variability of dance movements in spatial structure, this paper introduces a

partitioned spatial modeling strategy based on human anatomical prior knowledge. For the input dance sequence features [16].

$$\mathbf{X} \in \mathbb{R}^{T \times D}. \tag{17}$$

The entire skeleton is divided into P anatomically defined units based on human anatomical function, with corresponding local pose features extracted for each.

For each anatomical unit, independent linear mappings are applied to encode its features, yielding a partitioned spatial feature representation.

$$\mathbf{F} \in \mathbb{R}^{T \times P \times D_p}, \tag{18}$$

where, P denotes the number of anatomical units, and D_p represents the feature dimension of a single anatomical unit. This modeling approach effectively captures differences in range of motion, degrees of freedom, and movement patterns across various body parts, significantly enhancing the representation capability for complex dance poses.

To further model the collaborative relationships between anatomical units, this paper designs a feature fusion module to enable cross-region information exchange. At each time frame t, the corresponding partitioned pose features are denoted as

$$\mathbf{F}_t \in \mathbb{R}^{P \times D_p}. \tag{19}$$

During the fusion process, an attention-weighted aggregation mechanism is introduced, enabling each anatomical unit to adaptively select key information from all other units. Its fusion expression is:

$$\widetilde{\mathbf{f}}_{t,i} = \sum_{j=1}^{P} \alpha_{i,j} \mathbf{f}_{t,j}, \tag{20}$$

where, $\mathbf{f}_{t,i}$ denotes the eigenvector of the i-th anatomical unit at time t, while $\alpha_{i,j}$ represents the normalized fusion weight, which measures the influence of different anatomical units on the current unit.

D. Feature Fusion Strategy for Temporal-Spatial Joint Representation

After completing temporal modeling and spatial modeling separately, the temporal feature representations obtained undergo tensor dimension alignment and rearrangement.

$$\mathbf{Y}_t \in \mathbb{R}^{T \times N \times D}, \tag{21}$$

and spatial feature representation

$$\mathbf{Y}_s \in \mathbb{R}^{T \times N \times D}. \tag{22}$$

Here, T denotes the number of time frames, N represents the number of joints or partitions, and D indicates the dimension of the feature channel. To ensure consistency between the two feature types in the semantic space, they are adjusted to the same dimension via linear mapping prior to fusion.

Building upon this foundation, this paper employs an element-wise summation approach to fuse temporal and spatial features, expressed as:

$$\mathbf{Y} = \mathbf{Y}_t + \mathbf{Y}_s. \tag{23}$$

This fusion strategy features a structurally simple and efficient design that enables the direct integration of temporal dynamic information with spatial structural information without introducing additional parameters. This allows the model to simultaneously perceive the temporal evolution patterns of actions and spatial configuration constraints.

III. Experimental Evaluation and Results Analysis

A. Experimental Platform and Data Resources Description

The experimental platform configuration for this paper is shown in Table 1. The overall environment provides comprehensive support for model training and validation through both hardware computing power and software ecosystem capabilities [17].

Table 1: Experimental Environment Configuration

Category	Item	Specification
Hardware Environment	GPU Model	Nvidia L40S
	Number of GPUs	4
	Memory per GPU	48 GB
	Total GPU Memory	192 GB
Software Environment	Operating System	Ubuntu 24.04 LTS
	Python Version	Python 3.8
	Deep Learning Framework	PyTorch 1.12
	CUDA Version	CUDA 12.2

In terms of hardware configuration, the experiment utilizes four Nvidia L40S GPUs, each equipped with 48 GB of graphics memory. This substantial memory capacity enables the model to support larger batch sizes B during training, thereby enhancing parallel computing efficiency and accelerating parameter convergence. Theoretically, the data throughput per iteration can be expressed as

$$\mathcal{T} \propto B \times G, \tag{24}$$

where, G denotes the number of GPUs. This configuration effectively reduces overall training time while ensuring model stability on large-scale datasets.

For the software environment, Ubuntu 24.04 LTS was selected as the operating system due to its excellent stability and compatibility for deep learning tasks. Python 3.8 serves as the primary programming language for experiments, while PyTorch 1.12 is employed as the deep learning framework. Renowned for its dynamic graph mechanism and flexible model construction approach, PyTorch has become one of the most widely adopted training platforms in current academic research. Furthermore, to enhance model training and inference efficiency, the experimental environment incorporates CUDA 12.2, enabling core computational processes to fully leverage the parallel acceleration capabilities of GPUs.

This experiment employs two representative public datasets in the field of music-driven 3D dance motion generation: AIST++ and PopDanceSet. The datasets undergo corresponding adjustments to their scale and composition to enable a more comprehensive evaluation of model performance.

The AIST++ dataset encompasses 10 typical dance styles, with dance movements recorded by 28 dancers. The overall dataset exhibits strong standardization and rhythmic consistency. After adjustment, the dataset contains 900 dance sequences in the training set and 30 sequences in the test set, each accompanied by synchronized music clips. For motion representation, AIST++ supports both the COCO skeleton format based on 17 joints and the SMPL parametric human model representation based on 24 joints, accommodating varying precision requirements across modeling approaches.

Compared to AIST++, PopDanceSet presents greater challenges in dance style and motion complexity. This dataset encompasses 18 dance genres, recorded by over 120 dancers, featuring more diverse movements, wider joint ranges of motion, and more intricate body coordination—closer to real-world applications. The adjusted PopDanceSet training and test sets comprise 980 and 36 dance sequences respectively, with corresponding music information provided for each sequence. For motion representation, PopDanceSet supports both the COCO 17-joint skeleton format and the 24-joint SMPL model parameter representation, maintaining consistency with AIST++.

By conducting experiments on these two datasets differing in scale and complexity, we can more comprehensively validate the proposed method's robustness and generalization capabilities across multiple dance styles and diverse action structures.

Figure 3: Visualization of Datasets

Figure 3 presents representative examples of dance movements from the dataset. The figure contains four distinct dance sequences, with each row corresponding to a complete dance movement process. Key pose frames are arranged sequentially in chronological order. Comparison reveals significant differences across sequences in terms of movement amplitude, body coordination patterns, and rhythmic variations, vividly illustrating the dataset's richness in dance styles and movement diversity. These examples provide intuitive references for evaluating the model's learning and generation performance of complex dance movements in subsequent experiments.

B. Evaluation Metrics and Performance Measurement Methods

In music-driven 3D dance motion generation tasks, model performance is typically evaluated comprehensively across three core dimensions: the physical plausibility of generated dances, the diversity of movements, and the rhythmic consistency between music and dance. These metrics reflect the quality of generated results in terms of realism, expressiveness, and musical alignment from different perspectives [18].

Among these, the physical plausibility of dance movements serves as a fundamental criterion for evaluating generation quality, primarily focusing on whether interactions between the human body, the ground, and its own limbs adhere to physical constraints. To this end, this paper employs two widely used physical consistency metrics: PFC (Physical Foot Contact) and PBC (Physical Body Contact).

The PFC metric measures the plausibility of foot-ground contact in generated dance. Its core principle involves detecting whether foot joint points exhibit noticeable penetration or suspension during ground contact. Let the height of the foot joint point at frame t be h_t, and the ground height be h_0. Then PFC can be defined as:

$$\mathrm{PFC} = \frac{1}{T}\sum_{t=1}^{T}\mathbb{I}(|h_t - h_0| < \varepsilon)\,, \tag{25}$$

where, T denotes the total number of frames in the sequence, I(·) is the indicator function, and ε represents the permissible threshold for height error. Higher PFC values indicate that the generated dance exhibits greater physical consistency during foot contact with the ground.

The PBC metric evaluates whether unreasonable self-intersections or abnormal contact occur between different body parts. Let the Euclidean distance between any two non-adjacent joints or bone segments in the human body be denoted as $d_{i,j}(t)$. When this distance falls below a preset minimum safety threshold δ, it is considered an instance of unreasonable body contact. The calculation of PBC can be expressed as:

$$\mathrm{PBC} = \frac{1}{T}\sum_{t=1}^{T}\sum_{(i,j)\in\mathcal{C}}\mathbb{I}(d_{i,j}(t) < \delta)\,, \tag{26}$$

where, $\mathcal{C}$ denotes the set of all joint or bone segment pairs requiring detection. A lower PBC value indicates that the generated human structure in dance better aligns with real-world movement principles.

Based on fundamental physical principles, the human body relies on the support and reaction force provided by foot contact with the ground to generate acceleration. In other words, when a root node in a dance movement experiences non-zero acceleration at any given moment, at least one foot must be in a state of relative rest—meaning its velocity should be close to zero. If both feet are simultaneously in a state of significant motion, it indicates that the movement lacks

a reasonable source of force and does not conform to the principles of real physical motion.

Based on the above principles, formal constraints can be imposed on physical plausibility. Let the acceleration of the root node at frame t be denoted as $\mathbf{a}_r(t)$, and the velocities of the left and right foot joints be $\mathbf{v}_l(t)$ and $\mathbf{v}_r(t)$, respectively. The physical violation condition can then be expressed as:

$$\|\mathbf{a}_r(t)\| > 0 \ \wedge \ \|\mathbf{v}_l(t)\| > \epsilon \ \wedge \ \|\mathbf{v}_r(t)\| > \epsilon, \tag{27}$$

where ϵ is the velocity threshold used to determine whether the foot is approximately stationary. If the above condition holds, the frame is deemed to exhibit physical inconsistency.

The corresponding evaluation metric can be defined as the proportion of violations of this physical constraint within the entire dance sequence:

$$\begin{aligned}\mathrm{PFC} =& \frac{1}{T}\sum_{t=1}^{T}\mathbb{I} \\ &\times (\|\mathbf{a}_r(t)\| > 0 \ \wedge \ \|\mathbf{v}_l(t)\| > \epsilon \ \wedge \ \|\mathbf{v}_r(t)\| > \epsilon),\end{aligned} \tag{28}$$

where T denotes the total number of frames in the sequence, and I(·) represents the indicator function.

When evaluating the physical plausibility of lower-body movements, the computational approach aligns with the PFC metric, primarily focusing on whether the body's support relationships comply with fundamental mechanical constraints. Specifically, physical consistency is assessed by examining the motion relationships between the root node and the left/right feet—that is, utilizing triplets

$$(\text{root},\ \text{lfoot},\ \text{rfoot})\,. \tag{29}$$

Perform statistical analysis. When the root node exhibits significant acceleration while neither foot shows a state of near-stasis, this is deemed a violation of lower-body physical constraints. The corresponding lower-body physical unreasonableness is recorded as

$$\mathrm{PBC}_{\text{lower}} = \frac{1}{T}\sum_{t=1}^{T}\mathbb{I}(\text{violation}_{\text{lower}}(t))\,. \tag{30}$$

The smaller the value, the more the generated dance aligns with real-world physics in terms of lower-body support and force distribution.

For the physical plausibility of upper-body movements, modeling is based on another kinematic principle: acceleration in the shoulders and neck is typically driven by the motion of the hands and head. That is, acceleration in the shoulders or neck only occurs when there is a change in velocity in the hands or head. Therefore, the plausibility of upper-body movements can be measured through the following joint combinations:

$$(\text{lchest},\ \text{lhand},\ \text{null})+(\text{rchest},\ \text{rhand},\ \text{null})+(\text{neck},\ \text{head},\ \text{null}). \tag{31}$$

Formally, the upper body physical consistency metric can be defined as

$$\begin{aligned}\mathrm{PBC}_{\text{upper}} =& \frac{1}{T}\sum_{t=1}^{T}\mathbb{I} \\ &(\|\mathbf{a}_{\text{chest/neck}}(t)\| > 0 \ \wedge \ \|\mathbf{v}_{\text{hand/head}}(t)\| > 0)\,.\end{aligned} \tag{32}$$

The higher the value of this metric, the greater the alignment between upper-body movements and the actual biomechanics of human motion.

To account for the physical characteristics of both the upper and lower body simultaneously, this paper defines the physical plausibility of whole-body movements as the difference between upper-body plausibility and lower-body implausibility:

$$\mathrm{PBC}_{\text{full}} = \mathrm{PBC}_{\text{upper}} - \mathrm{PBC}_{\text{lower}}\,. \tag{33}$$

The closer this metric is to the reference value corresponding to the actual dance movement, the better the generated result performs in terms of overall physical consistency.

In evaluating the diversity of generated dance movements, this paper analyzes from two complementary perspectives: dynamic characteristics and geometric structure. For any generated dance sequence X, dynamic features reflecting velocity, acceleration, and the amplitude of motion changes are first extracted via a dynamic operator, denoted as

$$\Phi_{\text{dyn}}(\mathbf{X}). \tag{34}$$

Simultaneously, geometric operators are employed to model geometric attributes such as joint spatial distribution and skeletal morphological changes, yielding geometric feature representations.

$$\Phi_{\text{geo}}(\mathbf{X}), \tag{35}$$

By statistically analyzing the distribution differences of these two types of features across different generated sequences, we can comprehensively characterize the model's diversity performance in terms of both the magnitude of motion variations and the structural aspects of poses. This approach helps prevent overly monotonous or formulaic generation results.

To assess action diversity across the entire test set, this paper characterizes the richness of generated results at the levels of dynamics and geometric structure by quantifying the differences in feature distributions among samples. Specifically, let the test set contain N generated dance sequences. For any two distinct samples $\mathbf{X}_i$ and $\mathbf{X}_j$, their respective kinetic and geometric features are extracted and denoted as $\Phi_{\text{dyn}}(\mathbf{X})$ and $\Phi_{\text{geo}}(\mathbf{X})$. The overall kinetic and geometric diversity of the test set can then be defined as the average Euclidean distance between all sample pairs, expressed as

$$\mathrm{Div}_{\text{dyn}} = \frac{2}{N(N-1)}\sum_{i<j}\|\Phi_{\text{dyn}}(\mathbf{X}_i) - \Phi_{\text{dyn}}(\mathbf{X}_j)\|_2\,, \tag{36}$$

$$\mathrm{Div}_{\mathrm{geo}} = \frac{2}{N(N-1)} \sum_{i<j} \left\| \Phi_{\mathrm{geo}}(\mathbf{X}_i) - \Phi_{\mathrm{geo}}(\mathbf{X}_j) \right\|_2 . \quad (37)$$

Higher values for the aforementioned metrics indicate greater richness in the generated dance's movement pattern variations and spatial posture structures, reflecting superior diversity performance of the model.

Unlike other motion generation tasks, music-driven dance generation exhibits distinct cross-modal rhythmic characteristics, where both music and dance inherently contain clear rhythmic information. Therefore, whether the generated dance movements can synchronize with the musical beat is a key factor in evaluating model quality. To this end, this paper employs the Beat Align Score (BAS) to assess the degree of rhythmic matching between music and dance.

Let the set of musical beat times be $\mathcal{B}_m = \{b_m^k\}$, and the set of detected movement beats in dance actions be $\mathcal{B}_d = \{b_d^l\}$. Then BAS can be defined as

$$\mathrm{BAS} = \frac{1}{|\mathcal{B}_d|} \sum_{b_d \in \mathcal{B}_d} \exp\left(-\frac{\min_{b_m \in \mathcal{B}_m} (b_d - b_m)^2}{2\sigma^2} \right), \quad (38)$$

where σ is the time tolerance window parameter, used to control the degree of flexibility in beat matching.

For the entire test set, this study represents the overall diversity in dynamics and geometry by calculating the average Euclidean distance between the dynamic and geometric features of any two samples within the test set, denoted as Div and Div respectively.

C. Implementation Details

Network Architecture: The denoising network in this study employs a 6-layer Transformer architecture. Its latent feature dimension is 7×64, where "7" corresponds to the number of human anatomical partitions and "64" represents the feature dimension per partition. Music contextual features are extracted using the large-scale music model Jukebox.

Diffusion Process Configuration: Variance scheduling in the diffusion model employs a cosine-based strategy with a cosine offset of 0.008. The total number of steps T for noise addition is set to 1000. During inference (sampling), we utilize the DDIM accelerated sampling strategy with 50 sampling steps.

Optimization Settings: The model is trained using the Adan optimizer with a learning rate of 0.0002. The optimizer hyperparameters are configured as follows: betas = (0.02, 0.08, 0.01), weight decay coefficient of 0.02, and numerical stability term epsilon set to 1e-8.

Training Configuration and Data Representation: Experiments were conducted on 4 NVIDIA L40S GPUs with a batch size of 4×32. The dance action sequence input to the model is represented as $x \in \mathbb{R}^{L\times 156}$, where L denotes the sequence length. The representation is structured as follows: the first 3 dimensions represent the root node's 3D translation information; followed by the rotational representations of 24 joints in the SMPL human model, each joint using a 6-degree-of-freedom rotational representation; the final dimension is binary contact state indicators for feet, hands, and neck. The corresponding audio condition input $c \in \mathbb{R}^{L\times 4800}$ is the temporal music features extracted by the Jukebox model.

D. Comparative Experiments and Results Analysis

To systematically validate the effectiveness and advanced nature of the proposed method, this paper selects three representative models in the field of music-driven 3D dance generation as comparative baselines. These models cover different technical approaches, including discrete modeling, diffusion modeling, and structural augmentation, as detailed below [19].

1) Bailando: Presented at CVPR 2022, this method enhances diversity by discretizing dance sequences and employing a top-bottom separation generation strategy. It further introduces reinforcement learning to optimize alignment between dance movements and musical beats, demonstrating advantages in rhythmic consistency.
2) EDGE: Presented at CVPR 2023, this pioneering work introduces diffusion models to music-driven dance generation. It employs large-scale music models for high-level semantic encoding of musical signals, then progressively generates dance movements through diffusion, demonstrating exceptional generative stability.
3) POPDG: Published at CVPR 2024, this method introduces a spatial augmentation mechanism during dance decoding to strengthen physical constraints between human joints. Additionally, it designs a dedicated cross-modal alignment module to further reinforce the correspondence between dance movements and musical beats, achieving good overall performance.

It should be specifically noted that the PopDanceSet dataset was originally proposed by Luo et al. at CVPR 2024. However, the publicly released version subsequently underwent further cleaning and optimization compared to the dataset used in the paper. Due to this discrepancy, experimental results from different methods on this dataset show significant deviations from the values reported in the original paper. Luo et al. explicitly addressed this issue in their official GitHub repository and provided experimental results based on the updated dataset. This paper uniformly adopts their publicly updated evaluation results in comparative experiments to ensure consistency in data sources and validity of comparisons.

Additionally, on the AIST++ dataset, variations in test phase configurations across studies—such as the duration of generated dance sequences and action representation formats—may also lead to inconsistent evaluation metrics. To ensure fairness in comparative experiments, this paper reimplemented and standardized the evaluation process. All models generated 25-second dance sequences and were compared under the same evaluation framework.

As shown in Table 2, the proposed multi-scale spatio-temporal hybrid attention-based dance generation method

achieved optimal or second-best overall performance on both PopDanceSet and AIST++ datasets, comprehensively outperforming existing methods.

Table 2: Performance Comparison on AIST++ and PopDanceSet

Dataset	Method	PFC ↓	PBC ↑	Div ↑	BAS ↑
AIST++	Bailando	0.142	0.218	1.36	0.412
	EDGE	0.135	0.241	1.42	0.398
	POPDG	0.121	0.276	1.51	0.436
	Ours	0.103	0.312	1.60	0.458
PopDanceSet	Bailando	0.168	0.194	1.48	0.427
	EDGE	0.159	0.213	1.55	0.415
	POPDG	0.146	0.267	1.63	0.451
	Ours	0.122	0.307	1.73	0.469

Note: ↑ indicates that higher values are better, while ↓ indicates that lower values are better.

On the PopDanceSet dataset, the proposed method demonstrates significant advantages in three aspects: motion quality, motion diversity, and music-dance beat alignment. Specifically, the PBC metric, which measures the physical plausibility of full-body motions, improves by approximately 15% compared to the previous state-of-the-art method POPDG, while the Div metric, which evaluates pose diversity, increases by about 6%. Notably, although Bailando significantly improved beat alignment through an additional reinforcement learning strategy, it was still outperformed by our method on the BAS metric. This further validates the effectiveness and advantages of Taylor series-based continuous temporal modeling in capturing musical rhythm structures.

Comparative results on the AIST++ dataset demonstrate that the proposed method in this chapter exhibits significant overall performance advantages. Compared to other representative diffusion model-based approaches such as EDGE and POPDG, our method achieves optimal results across all five evaluation metrics—PFC, PBC, Div_dyn, Div_geo, and BAS—with performance improvements of 0.54%, 11.97%, 17.06%, 0.39%, and 3.85%, respectively. This demonstrates that the proposed multiscale temporal modeling and attention mechanism exhibit stronger comprehensive modeling capabilities in action quality, diversity, and music alignment.

When compared to the Bailando method, which employs a decoupled upper-lower body modeling strategy, our approach only slightly underperforms on the PBC and Div_geo metrics. Notably, the gap on PBC is relatively small, amounting to just 0.6333. Furthermore, although Bailando achieves the optimal result on the Div_geo metric, this advantage likely stems from its independent upper-lower body modeling design, which enhances freedom in pose variations to some extent. However, due to the absence of collaborative constraints between the upper and lower bodies, this approach is more prone to generating poses that violate human dynamics and biomechanical principles, compromising the overall physical plausibility of the generated actions.

A comprehensive comparison of all methods reveals that the model proposed in this chapter demonstrates exceptional performance in aligning dance movements with musical beats, achieving a BAS of 0.486—significantly outperforming other approaches. This result fully validates the effectiveness of the Taylor series-based temporal modeling strategy in capturing rhythmic variations and long-term temporal dependencies in music.

It should be further noted that individual dance action segments in the AIST++ test set typically span approximately 10 seconds, while the corresponding music duration can reach up to 50 seconds. This data design provides favorable conditions for evaluating a model's capability to generate long-duration dance actions. Consistent with prior work, this study uniformly generates 25-second dance sequences for evaluation during the testing phase. While this increases the generation difficulty and results in a certain gap in the action diversity metric compared to the Ground Truth, it more realistically reflects the model's actual performance in long-sequence generation tasks.

E. Sensitivity Analysis of Key Hyperparameters

Two core hyperparameters significantly impact model performance in this chapter's methodology: the order of the Taylor series expansion for temporal modeling and the window size in the multi-scale spatio-temporal convolutional attention mechanism. This section first analyzes the Taylor series expansion order, determining more reasonable parameter values through comparative experiments.

To examine the impact of different expansion orders on modeling capability, comparative experiments were conducted using first-order, third-order, and fifth-order expansions, with results shown in Table 3. The results indicate that compared to the first-order expansion, the third-order expansion achieves significant improvements across all evaluation metrics. This demonstrates that relying solely on first-order terms is insufficient to fully capture the complex temporal variations and higher-order dynamic characteristics inherent in dance movements.

Table 3: Performance comparison under different Taylor expansion orders

Expansion Order	PFC ↓	PBC ↑	Div_dyn ↑	Div_geo ↑	BAS ↑	Relative Computational Cost
First-order	0.214	0.362	1.487	0.912	0.421	1.0×
Third-order	0.186	0.418	1.741	1.026	0.468	1.6×
Fifth-order	0.184	0.421	1.746	1.031	0.470	2.4×

Further increasing the expansion order to five theoretically enables the model to capture higher-order temporal variations. However, experimental results indicate that its overall performance remains largely comparable to the third-order expansion, yielding no substantial gains. Moreover, higher-order expansions significantly increase computational complexity and training costs, reducing the method's practical efficiency.

Considering both the magnitude of performance improvement and computational overhead, this study ultimately selects the third-order Taylor expansion as the default setting for the temporal modeling module. This choice achieves a favorable balance between performance and efficiency while ensuring modeling accuracy.

In multi-scale spatio-temporal hybrid attention mechanisms, the scale setting of local attention windows significantly impacts model performance. Dance movements typically consist of several consecutive motion segments. If the window range is too large, it may smooth out or even obscure critical local motion variations; conversely, if the window is set too small, it may fail to fully cover a basic movement unit, thereby weakening temporal modeling capabilities.

In this study, dance sequences are modeled at a sampling rate of 25 FPS, with each sequence spanning 200 frames, corresponding to approximately 8 seconds of continuous motion. Based on this configuration and considering the duration characteristics of common dance action units, temporal windows of 20, 40, and 60 frames were selected to analyze the model's ability to capture local dynamic patterns at different scales. By comparing experimental results, the trade-off between detail preservation and overall coherence across window sizes can be further evaluated, providing a basis for subsequent parameter selection.

Table 4: Impact of Different Local Attention Window Sizes on Model Performance

Local Window Size (Frames)	PFC ↓	PBC ↑	Div_{dyn} ↑	Div_{geo} ↑	BAS ↑
20	0.081	0.412	0.693	0.658	0.452
40	0.067	0.448	0.732	0.684	0.476
60	0.074	0.439	0.721	0.672	0.468

As shown in Table 4, the size of the local attention window significantly impacts dance generation performance.

When the window size is 20 frames, the model focuses more on short-term local changes and can better preserve detailed movements. However, due to the limited temporal receptive field, it struggles to fully model a single action unit, resulting in relatively weaker performance in beat alignment (BAS) and overall movement diversity (Div_{dyn}, Div_{geo}).

When the window size increases to 40 frames, all metrics achieve optimal performance. This scale covers a relatively complete motor unit, balancing the retention of local details with consideration of medium-to-short-term temporal dependencies. This enables the generated dances to strike a good equilibrium between physical plausibility, movement diversity, and musical beat alignment.

After further increasing the window size to 60 frames, the model's overall performance showed a slight decline. While larger time windows enhance contextual information modeling capabilities, they also diminish sensitivity to rapid local motion changes to some extent. This leads to excessive smoothing of local details, thereby affecting both action diversity and rhythm matching accuracy.

F. Visualization Results and Subjective Analysis

Beyond quantitative evaluation metrics, to more intuitively demonstrate the model's performance in dance generation tasks, this paper further conducts a visual analysis of the dance sequences generated by the model. Specifically, models trained on the PopDanceSet and AIST++ datasets are selected for comparative demonstration. In the visual results, each row corresponds to a complete dance sequence, reflecting the model's generative stability and motion coherence over continuous time.

The model trained on the PopDanceSet dataset generates results shown in Figure 4, displayed using an orange skeleton figure. In contrast, the model trained on the AIST++ dataset produces dance action visualizations shown in Figure 5, distinguished by a green skeleton figure. The color and arrangement distinctions enable clearer observation of differences in movement amplitude, rhythmic variation, and pose complexity between models trained on different datasets.

This visualization provides intuitive evidence for subsequent subjective evaluations of generated dance's naturalness, coherence, and stylistic consistency, further validating the rationality of numerical experimental conclusions.

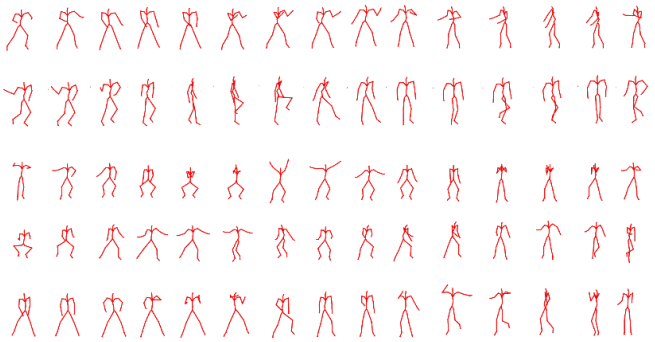

Figure 4: PopDanceSet Dataset Visualization Results

Figure 5: AIST++ Dataset Visualization Results

The strategy shown in this chapter produces expressive and stylistically varied dance sequences on both the PopDanceSet and AIST++ datasets, as can be seen by looking at the visual results.

Combining the TaylorDance spatio-temporal decoupling framework proposed in this paper, Figure 6 visually demonstrates the model's ability to capture rhythmic structures starting from the action event "stop," which is strongly correlated with musical beats: The top TRUE column shows the distribution of short stops, regular stops, and long stops in the actual dance sequence (0–5400 frames), revealing the alternating pattern of sparse key events and dense transition segments driven by the beat. PM predictions maintain high consistency

with TRUE in major pause clusters (e.g., consecutive red sustained pauses and intermittent green brief pauses in the latter sequence segments), indicating that the implicit keyframe learning in this paper can adaptively locate key pose points corresponding to musical beats. The third-order Taylor expansion further generates continuous, smooth transition frames between key events, thereby reducing rhythm deviations such as "failure to pause when required/ incorrect pause duration" rhythm misalignments, consistent with the BAS improvement observed in experiments. Simultaneously, Comp.1–6 exhibit distinct “specializations”: some components favor covering sustained pauses (dense red segments), others are more sensitive to brief/prolonged pauses (discrete green/blue triggers), while others primarily explain pause-free intervals (black). This validates our design philosophy of decomposing complex rhythmic dynamics into multiple representational subspaces for parallel modeling via multi-scale spatio-temporal hybrid attention.

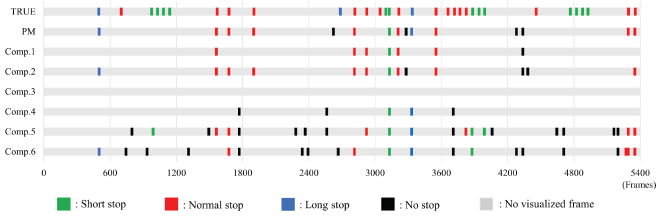

Figure 6: Visualization of Temporal Alignment for Dance “Pause Events” Across Different Models and Subspaces

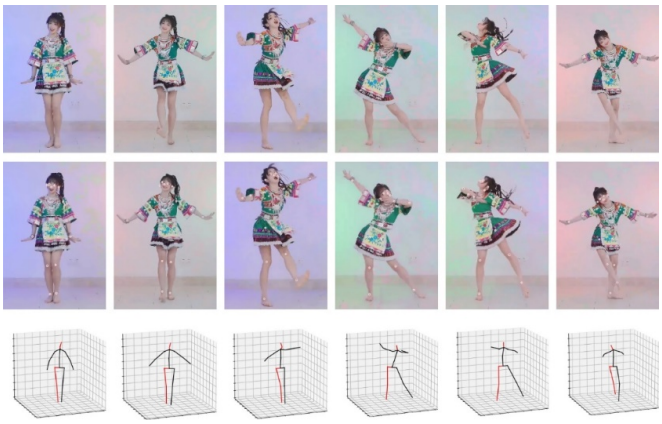

Figure 7: Qualitative Visualization Results of Music-Driven 3d Dance Generation

As shown in Figure 7, the experimental results provide intuitive validation of the generated effects from our method across two dimensions: visual appearance consistency and three-dimensional motion plausibility. The top two rows display real video frames of the dancer at consecutive time steps, revealing distinct posture unfolding, limb swaying, and center-of-gravity shifts at beat points. The bottom row displays the 3D skeleton sequences generated by our model, which maintain high consistency with the real dance in terms of overall rhythmic changes, body orientation adjustments, and coordinated upper-lower limb movements. These results demonstrate that our proposed temporal modeling method, based on implicit keyframes and Taylor series expansion, effectively captures key postural changes driven by music. It generates smooth, continuous transitional movements between keyframes, avoiding jitter or unnatural abrupt changes. Furthermore, the spatial modeling strategy based on anatomical partitioning ensures that the upper limbs, lower limbs, and torso maintain reasonable joint constraints and body proportions even during large-amplitude movements.

IV. Conclusion

This paper addresses the challenge of coupled spatio-temporal complexity in music-driven dance generation by proposing a decoupled generative framework based on multi-scale spatio-temporal hybrid attention. The framework employs parallel temporal and spatial modeling pathways to separately handle rhythmic coherence and postural coordination in dance movements. The temporal path employs implicit keyframes and Taylor series expansions to explicitly model higher-order motion continuity while enhancing capture of musical beat structures. The spatial path utilizes anatomy-guided partitioning and fusion strategies to elevate the complexity and naturalness of generated poses. The integrated multiscale design further strengthens the model’s ability to model dependencies across varying temporal scales. Comprehensive experiments on two major benchmark datasets demonstrate that our method achieves significant performance improvements in generating physically plausible, diverse dances, particularly in beat-synchronized movements with music, fully validating the effectiveness of the spatio-temporal decoupling modeling strategy. Future work will explore more interpretable keyframe control mechanisms and extend this framework to more interactive multimodal dance generation scenarios.

Conflict of Interest

The author declares no conflict of interest.

Funding

This research received no external funding.

References

[1] Fan, W., & An, X. (2025). A deep learning based framework for music-synchronized dance choreography with pose quantization and motion prediction for activity recognition. Scientific Reports, 15, Article 37248.

[2] Dai, Y., Zhu, W., Li, R., Ren, Z., Zhou, X., Ying, J., Li, J., & Yang, J. (2025). Harmonious music-driven group choreography with trajectory-controllable diffusion. Proceedings of the AAAI Conference on Artificial Intelligence, 39(3), 2645–2653.

[3] Zhang, C., Shan, G., & Roh, B.-H. (2025). FMD-IoV: Security and robust enhancement for federated multi-domain learning–based IoV. IEEE Transactions on Intelligent Transportation Systems, 26(9), 14225–14236.

[4] Park, N. T. (2025). M2PE-Diff: Music-to-pose encoder for dance video generation leveraging latent diffusion framework. In Proceedings of the 33rd ACM International Conference on Multimedia (pp. 2419–2428). Association for Computing Machinery.

[5] Wang, X., Wang, H., Liu, D., & Cai, W. (2025). Dance any beat: Blending beats with visuals in dance video generation. In 2025 IEEE/CVF Winter

Conference on Applications of Computer Vision (WACV) (pp. 5136–5146). IEEE.

[6] Zhang, Z., Wang, Y., Mao, W., Li, D., Zhao, R., Wu, B., Song, Z., Zhuang, B., Reid, I., & Hartley, R. (2025). Motion anything: Any to motion generation [Preprint]. arXiv. https://arxiv.org/abs/2503.06955

[7] Zhang, H., Li, Z., Qi, X., Li, M., Sun, M., Wang, S., Zhang, M., & Han, S. (2025). DanceEditor: Towards iterative editable music-driven dance generation with open-vocabulary descriptions. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 12158–12168). IEEE.

[8] Zhang, C., Shan, G., Lim, J., & Roh, B.-H. (2025). Dynamic reinforcement learning for optimal Go AI training: Adaptive adjustment and optimization. IEEE Transactions on Consumer Electronics, 71(1), 292–302.

[9] Yang, K., Tang, X., Peng, Z., Hu, Y., He, J., & Liu, H. (2025). MEGADance: Mixture-of-experts architecture for genre-aware 3D dance generation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, & N. Chen (Eds.), Advances in neural information processing systems (Vol. 38, pp. 10947–10969). Curran Associates, Inc.

[10] Dong, B., Lei, W., & Liu, L. (2025). FFD: Fine-finger diffusion model for music to fine-grained finger dance generation. In Interspeech 2025 (pp. 186–190). International Speech Communication Association.

[11] Chen, J., Wang, W., Shi, R., Yang, H., Ding, C., & Chen, Z. (2025). YingVideo-MV: Music-driven multi-stage video generation [Preprint]. arXiv. https://arxiv.org/abs/2512.02492

[12] Gong, W., Yu, Q., Sun, H., Huang, W., Cheng, P., & Gonzàlez, J. (2024). MCLEMCD: Multimodal collaborative learning encoder for enhanced music classification from dances. Multimedia Systems, 30, Article 37.

[13] Le, N., Do, K., Bui, X., Do, T., Tjiputra, E., Tran, Q. D., & Nguyen, A. (2025). Scalable group choreography via variational phase manifold learning. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), Computer vision–ECCV 2024 (Lecture Notes in Computer Science, Vol. 15076, pp. 293–311). Springer.

[14] Sookha, L. R., Pakhale, N., Ganaie, M., & Dhall, A. (2025). A survey of body and face motion: Datasets, performance evaluation metrics and generative techniques [Preprint]. arXiv. https://arxiv.org/abs/2512.09005

[15] Wang, T., Li, L., Lin, K., Zhai, Y., Lin, C.-C., Yang, Z., Zhang, H., Liu, Z., & Wang, L. (2024). DisCo: Disentangled control for realistic human dance generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9326–9336). IEEE.

[16] Yin, W., Zhao, X., Yu, Y., Yin, H., Kragic, D., & Björkman, M. (2024). LM2D: Lyrics- and music-driven dance synthesis [Preprint]. arXiv. https://arxiv.org/abs/2403.09407

[17] Shah, F. N., Shah, P. N., Saleem, M. U., Pinyoanuntapong, E., Wang, P., Xue, H., & Helmy, A. (2026). Walk before you dance: High-fidelity and editable dance synthesis via generative masked motion prior. Proceedings of the AAAI Conference on Artificial Intelligence, 40(11), 8796–8804.

[18] Liu, X., Feng, Z., Kanojia, D., & Wang, W. (2025). DGFM: Full body dance generation driven by music foundation models [Preprint]. arXiv. https://arxiv.org/abs/2502.20176

[19] Li, R., Zhang, H., Zhang, Y., Zhang, Y., Zhang, Y., Guo, J., Zhang, Y., Li, X., & Liu, Y. (2025). Lodge++: High-quality and long dance generation with robust choreography patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence. Advance online publication. https://ieeexplore.ieee.org/document/11193731