BeatDance: Generating Beat-Consistent 3D Dance with Hierarchical Spatial-Temporal Modeling

Xiaojian Shen[a], Dahu Shi[b], Jianrong Zhang[c], Hai Li[d], Hongwei Zhao[d], Dawei Zhang[e], Yunzhi Zhuge[f], Zhiliang Wu[b,*], Guanghui Yue[g], Wei Zhou[h]

[a]*College of Software, Jilin University, Changchun, 130012, China*
[b]*College of Computer Science and Technology, Zhejiang University, Hangzhou, 310007, China*
[c]*ReLER, AAII, University of Technology Sydney, Sydney, Australia*
[d]*College of Computer Science and Technology, Jilin University, Changchun, 130012, China*
[e]*School of Computer Science and Technology, Zhejiang Normal University, Jinhua, 321004, China*
[f]*School of Information and Communication Engineering, Dalian University of Technology, Dalian, 116081, China*
[g]*School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, Shenzhen, 518060, China*
[h]*School of Computer Science and Informatics, Cardiff University, CF10 3AT Cardiff, U.K.*

Abstract

Generating realistic 3D dance from music is a challenging task that requires accurate synchronization with musical rhythms while capturing the spatial complexity of human motion. Although existing methods can generate physically plausible dance motions, they often struggle to achieve precise alignment with music, such as the beat. To address this limitation, we propose a novel diffusion-based framework, BeatDance, with two components: 1) We present a Hierarchical Decoupled Attention (HDA) module, which first disentangles the learning of human pose and temporal dynamics. A hierarchical structure is then employed to capture both short-term and long-term dependencies, thereby enhancing spatial-temporal modeling. 2) We adopt cycle-consistent learning by introducing an auxiliary dance-to-music module. During training, discrepancies between the reconstructed and original music induce a stronger loss signal, effectively encouraging the consistency property between the music and dance motion. Extensive experimental results demonstrate that our proposed approach outperforms recent competitive methods on two benchmark datasets. The project page is available at https://shenxiaojian.github.io/BeatDance/.

*Corresponding author: Zhiliang Wu (E-mail: wu_zhiliang@zju.edu.cn)

Keywords: 3D Dance Generation, Diffusion Models, Spatial-Temporal Modeling

1. Introduction

Music plays an important role in shaping emotions and narratives, which are often embodied through human motion. This natural connection has inspired growing interest in generating 3D dance conditioned on the music. Automated motion synthesis from music reduces reliance on motion-capture systems and costly animation tools, enabling more efficient content creation across various applications such as filmmaking [1, 2], virtual production [3], and video editing [4, 5].

To achieve high-quality dance generation, many works leverage the autoencoder [7, 8] as a core component of the model architecture. Bailando [9] models the upper and lower body motions separately using VQ-VAE, and generates motion code indices auto-regressively. TM2D [10] combines the training of music-to-dance and text-to-motion, which helps produce smoother dance motion. Recently, diffusion models have shown remarkable proficiency in music-driven 3D dance generation [11, 12, 13, 14].

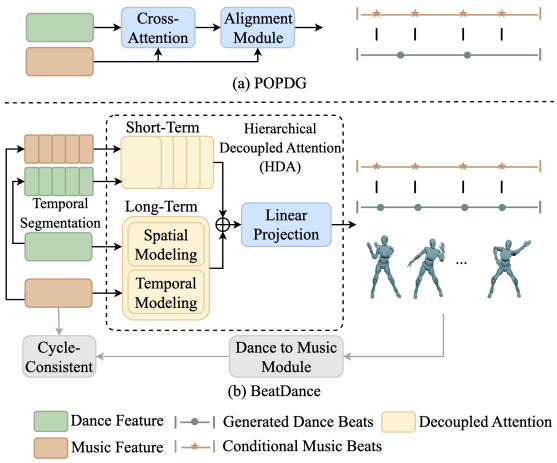

Figure 1: Comparison with the existing state-of-the-art method. (a) POPDG [6] uses a cross-attention and an alignment module to associate music and dance. However, it may face challenges in achieving beat-level synchronization. (b) We propose Hierarchical Decoupled Attention (HDA) and a cycle-consistent learning strategy. HDA hierarchically decouples spatial and temporal information, while cycle-consistent learning leverages the inverse dance-to-music mapping to further enhance temporal alignment.

For example, EDGE [15] proposes to use Jukebox [16] to extract music features, and the model also supports dance editing, dance completion, and joint in-painting. POPDG [6] proposes a new dataset, *i.e.*, PopDanceSet, which comprises a wide range of diverse and complex dance motions. It introduces an additional self-attention and an alignment module to learn the relationships among human body parts and insert music features into the model, respectively.

Despite the encouraging progress, these methods still face challenges in generation precision. First, they only perform a coarse-grained fusion of music and motion features without explicitly modeling the spatial-temporal dependencies simultaneously between music and dance, which lead to beat misalignment (such as POPDG illustrated in Figure 1(a)). Second, these works typically focus on long-term dependency, overlooking the short-term patterns that are crucial for maintaining motion continuity and synchronizing with local-level musical beats.

To address these limitations, we propose BeatDance, a diffusion-based framework for music-driven 3D dance generation. At the core of this framework is our proposed Hierarchical Decoupled Attention (HDA), which consists of two components. First, we present a decoupled attention that explicitly separates the learning of spatial and temporal information. Specifically, the spatial attention divides the body into several anatomical parts, each with an independent linear projection. For the temporal dynamics, we update the temporal feature by applying a Taylor series expansion over a learned joint music-dance attention map. Second, we employ a hierarchical design that organizes these motion features from local to global levels for short-term and long-term dependencies. This design provides the capacity to model fine-grained music-dance correspondences across multiple temporal scales. Meanwhile, to further exploit this capacity, we introduce a dance-to-music branch based on a latent diffusion model. We perform cycle-consistent learning to map dance motion back to its corresponding music, which improves the alignment between the two modalities. The improved music-dance alignment is particularly reflected in more accurate beat synchronization. It also benefits broader aspects of dance quality: the part-specific spatial modeling facilitates the learning of complex dance poses and expressive whole-body movements, while the short-term and long-term temporal modeling promote smooth transitions and

choreographic coherence, respectively.

Taking these components together, our approach is able to generate expressive, smooth, and choreographically coherent dance motions that are well synchronized with the music, particularly aligning closely with the underlying beat (as shown in Figure 1(b)). We conduct extensive experiments on two datasets, PopDanceSet and AIST++. For example, on the PopDanceSet dataset, BeatDance achieves a PBC of 7.91 and BAS of 0.244, outperforming POPDG of 5.95 and 0.233, respectively.

In summary, our contributions include:

- We present BeatDance, a diffusion-based framework for music-driven 3D dance generation that achieves state-of-the-art physical plausibility and beat alignment on both benchmarks, while maintaining competitive motion diversity.
- We propose a Hierarchical Decoupled Attention (HDA), which separately models motion spatial structures and temporal dynamics. It also incorporates a hierarchical architecture to capture both short-term and long-term dependencies between the music and dance.
- We introduce a cycle-consistent learning framework by predicting music features from generated dance motion through a dance-to-music module, which further enhances alignment with the musical beat.

2. Related Works

In this section, we review works closely related to ours in two areas. We first discuss text-driven motion generation, which shares similar architectural foundations with our approach. We then review music-driven dance generation, with a focus on recent diffusion-based methods and their limitations in achieving precise cross-modal alignment.

2.1. Text-Driven Motion Generation

Text-driven motion generation aims to synthesize realistic human motion from a text description. Early works [17, 18] directly deploy a text encoder and a motion decoder to learn the mapping from the two modalities. Recently, some works [19, 20]

employ a Vector Quantized Variational Autoencoder (VQ-VAE) [21] to project human motions into discrete representations. For example, T2M-GPT [22] encodes motion sequences into discrete tokens using VQ-VAE and then learns to map text descriptions to these tokens with a Generative Pretrained Transformer (GPT). Another trend is the use of diffusion models for motion synthesis. MotionDiffuse [23] and MDM [24] are among the first efforts to apply diffusion models [25] to this field. Subsequent works achieve performance improvements by encoding the motion into latent space [26, 27, 28], adopting a retrieval-augmented generation strategy [29], and exploring spatial-temporal decoupled attention [30, 31, 32, 33, 34, 35, 36]. For example, MLD [37] learns low-dimensional latent vectors of a motion sequence with a Variational Autoencoder (VAE). Then, it employs a latent diffusion model to generate latent vectors from text. FineMoGen [38] proposes a spatio-temporal mixture attention to separately model spatial and temporal dependencies. Although it shares a related modeling principle with our method, our work focuses on the distinct challenge of fine-grained music-dance alignment: the generated dance should respond to rapidly changing local rhythmic patterns while maintaining long-term motion coherence. To address this challenge, our HDA explicitly models fine-grained relationships between the two modalities and hierarchically captures short-term rhythmic patterns and long-term dependencies. Furthermore, we propose a cycle-consistent training paradigm that further improves the music-dance alignment.

2.2. Music-Driven Dance Generation

Music-driven dance generation is challenging due to its demand for complex articulation and precise music-dance alignment. Early methods [39, 40], which rely on similarity-based retrieval and concatenation from existing databases, are inherently limited in generating novel dance motions. With the advent of deep learning, extensive efforts [41, 42, 43, 44] have been made to model the probabilistic mapping between music and dance. Among these advances, VQ-VAEs have emerged as an effective tool for learning discrete motion representations. For example, Bailando [9, 45] encodes motion into discrete codes and uses a GPT-style autoregressive decoder to generate upper- and lower-body movements, with a reinforcement-learning-based scheme

for beat alignment. Similarly, TM2D [10] utilizes a shared VQ-VAE codebook for both music-to-dance and text-to-motion tasks. By unifying the shared latent space, it benefits from large-scale text-motion datasets, resulting in smoother dance motions. More recently, diffusion models have shown remarkable proficiency in music-driven 3D dance generation [11, 14, 46], branching into two distinct research directions. One direction focuses on extremely long dance generation via multi-stage architectures [12, 13]. For example, Lodge [12] employs a coarse-to-fine framework to achieve this through cascaded generation processes. Different from these multi-stage approaches, our HDA captures short-term and long-term dependencies within a unified single-stage module, with a particular focus on improving fine-grained music-dance alignment. The other direction explores single-stage frameworks to streamline the generation pipeline. EDGE [15] is the first to leverage the diffusion model for synthesizing high-fidelity dance in this category. It uses Jukebox [16] to extract music features, which has become the standard setting for subsequent works. POPDG [6] proposes PopDanceSet, a dataset of diverse and complex dance motions. It also introduces spatial-temporal attention blocks and an alignment module that modulates dance features through an affine transformation conditioned on music and motion features. By contrast, our HDA explicitly models fine-grained music-dance correspondences. It first partitions the body into anatomical regions with independent projections, allowing different body parts to separately capture their relationships with the music. It then constructs a joint music-dance attention map and leverages a higher-order Taylor expansion to model complex temporal dynamics. Furthermore, the hierarchical design captures short-term rhythmic patterns and long-term dependencies, enabling the generated motion to respond to local musical beats while maintaining coherent whole-body dynamics. Therefore, while POPDG fuses music and dance features through an affine modulation of dance features, HDA explicitly models how the relationships between music and different body parts evolve across multiple temporal scales.

3. Preliminaries

Denoising Diffusion Probabilistic Models (DDPM) [25] are a class of generative models that learn to synthesize data by reversing a gradual noising process. In this section, we briefly review the key concepts of DDPM that form the foundation of our framework.

Forward process. Given a data sample $\boldsymbol{X}_0 \sim q(\boldsymbol{X}_0)$, the forward process gradually corrupts it by injecting Gaussian noise over T timesteps. By defining the cumulative noise schedule $\alpha_t = \prod_{s=1}^{t}(1-\beta_s)$, where $\{\beta_t\}_{t=1}^{T}$ is a fixed variance schedule, the noisy sample at an arbitrary timestep t can be obtained in closed form:

$$\boldsymbol{X}_t = \sqrt{\alpha_t}\,\boldsymbol{X}_0 + \sqrt{1-\alpha_t}\,\epsilon, \quad \epsilon \sim \mathcal{N}(0,1). \tag{1}$$

When $t = T$, $\alpha_T \approx 0$, so $\boldsymbol{X}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is approximately pure Gaussian noise.

Reverse process. The reverse process learns a denoising network ϵ_θ to iteratively recover the original data from $\boldsymbol{X}_T$. Given a conditioning signal $\boldsymbol{c}$ (*e.g.*, music features in our framework), the network predicts the original data $\boldsymbol{X}_0$ from the noisy input $\boldsymbol{X}_t$ at each step. The training objective is:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\boldsymbol{X}_0, \boldsymbol{c}, \epsilon, t}\Big[\|\boldsymbol{X}_0 - \epsilon_\theta(\boldsymbol{X}_t, t, \boldsymbol{c})\|_2^2\Big]. \tag{2}$$

Following [47], the reverse variance can also be learned jointly via a variational lower-bound loss $\mathcal{L}_{\text{vlb}}$ to better model the reverse transition distribution. At inference time, we adopt DDIM [48], which reformulates the reverse process as a non-Markovian procedure and enables high-quality generation with significantly fewer denoising steps.

4. Method

In this section, we first introduce BeatDance, a diffusion-based framework designed to synthesize realistic and expressive 3D dance sequences that are consistent with the source music. Then we present Hierarchical Decoupled Attention (HDA), which is used to model the intricate spatial-temporal mappings between generated dance and input music, capturing both short-term and long-term dependencies in a hierarchical

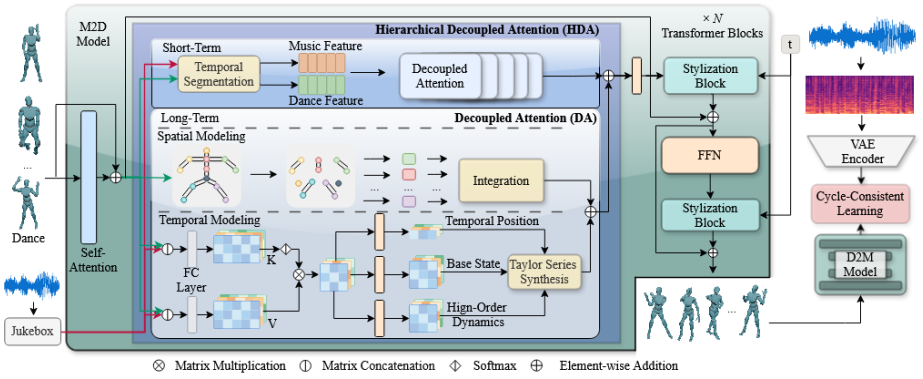

Figure 2: Overview of BeatDance. Hierarchical Decoupled Attention (HDA) employs a hierarchical design to capture both short-term and long-term music-dance dependencies, building upon a Decoupled Attention (DA) that splits the modeling into spatial and temporal branches. Additionally, our cycle-consistent learning mechanism uses an auxiliary Dance-to-Music (D2M) model to reconstruct the source music features from the generated dance, thereby enforcing a tighter music-dance alignment.

fashion. Finally, we propose Cycle-Consistent Learning (CCL) mechanism to reinforce the music-dance consistency by employing an auxiliary dance-to-music model.

4.1. Overview

The overall architecture is shown in Figure 2, which consists of a Music-to-Dance (M2D) model and a Dance-to-Music (D2M) model. We adopt a skeleton-based diffusion architecture for the former, which is trained from scratch with music as the conditioning input. The latter is used to reconstruct the music from the generated dance.

Specifically, in the music-to-dance branch, we follow the standard music-driven 3D dance generation setting to use Jukebox [16] to extract the music embedding from a piece of music. Given a music embedding $\boldsymbol{M} = [\boldsymbol{m}^1, \boldsymbol{m}^2, \cdots, \boldsymbol{m}^T]$ and a dance motion sequence $\boldsymbol{X} = [\boldsymbol{x}^1, \boldsymbol{x}^2, \cdots, \boldsymbol{x}^T]$, where $\boldsymbol{m}^i \in \mathbb{R}^{1\times d_m}$, $\boldsymbol{x}^i \in \mathbb{R}^{1\times d_x}$, T is the length, which is common to both the music feature and the dance sequence. We employ a diffusion-based model to learn the mapping from the music embedding to the corresponding dance motion. In the forward process, noise $\epsilon \sim \mathcal{N}(0, 1)$ is incrementally added to the dance sequence, obtaining $\boldsymbol{X}_t = \sqrt{\alpha_t}\boldsymbol{X}_0 + \sqrt{1-\alpha_t}\epsilon$ at timestep t. Then, we feed $\boldsymbol{X}_t$ into a transformer-based denoising autoencoder ϵ_θ^s to predict the original

dance sequence conditioned on $\boldsymbol{M}$. The model ϵ_θ^s incorporates a Hierarchical Decoupled Attention (HDA) mechanism, which separately models spatial (pose-wise) and temporal dependencies in a hierarchical fashion. This mechanism facilitates the learning of fine-grained correlations between music and dance and enhances the beat alignment (Section 4.2). The optimization of ϵ_θ^s is performed using $\mathcal{L}_{\text{m2d}}$, which can be written as

$$\mathcal{L}_{\text{m2d}} = \mathcal{L}_{\text{diff}} + \lambda_{\text{vlb}}\mathcal{L}_{\text{vlb}} + \lambda_{\text{va}}\mathcal{L}_{\text{va}} + \lambda_{\text{joint}}\mathcal{L}_{\text{joint}} + \lambda_{\text{body}}\mathcal{L}_{\text{body}},$$

where $\mathcal{L}_{\text{diff}} = \mathbb{E}_{\boldsymbol{X},\boldsymbol{M},\epsilon,t}\left[\|\boldsymbol{X} - \epsilon_\theta^s(\boldsymbol{X}_t, t, \boldsymbol{M})\|_2^2\right]$. We incorporate a variational lower-bound loss, denoted as $\mathcal{L}_{\text{vlb}}$, weighted by a coefficient λ_{vlb}, to learn the variance of the diffusion process transitions [47]. Following [15, 6], we also use three Mean Squared Error (MSE) losses, *i.e.*, $\mathcal{L}_{\text{va}}$, $\mathcal{L}_{\text{joint}}$, and $\mathcal{L}_{\text{body}}$. These losses respectively constrain velocity and acceleration, encourage the consistency in joint space, and enhance ground contact for feet, hands, and the neck. λ_{va}, λ_{joint} and λ_{body} are hyperparameters that balance the weight of each loss.

As for the dance-to-music branch, we aim to reconstruct the music from the generated dance motion with a latent diffusion model. Building upon this model, we introduce a cycle-consistent learning strategy to further encourage semantic consistency between the generated 3D dance and the source music. Briefly, we first train the dance-to-music model independently. Once trained, we freeze its parameters and integrate it with the music-to-dance model to form a music-to-dance-to-music cycle. For more details, please refer to Section 4.3.1.

Overall optimization goal. Taking these components together, our training objective can be formulated as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{m2d}} + \lambda_{\text{cycle}}\mathcal{L}_{\text{cycle}}, \tag{3}$$

where $\mathcal{L}_{\text{cycle}}$ denotes the cycle-consistent loss (Section 4.3.2), weighted by a coefficient λ_{cycle}.

4.2. Hierarchical Decoupled Attention

Different from using cross-attention and alignment modules that jointly model spatial and temporal information, we propose a Hierarchical Decoupled Attention (HDA),

which explicitly disentangles spatial and temporal modeling (Decoupled Attention) and captures both short-term and long-term dependencies (Hierarchical Mechanism).

4.2.1. Spatial Modeling

To effectively learn complex dance poses, we employ an anatomical partitioning strategy. We divide each frame's raw representation into $N_s = 8$ parts, comprising 7 functionally relevant regions that incorporate the head, spine, left arm, right arm, left leg, right leg, root joint, and an additional global part representing the full body. Each part is fed into an independent linear projection layer (N_s in total), and the outputs are concatenated to obtain $\boldsymbol{Z} \in \mathbb{R}^{T\times N_s\times d_s}$, where d_s represents the dimension of each human body part. Then, a Feed-Forward Network (FFN) is applied to each body part independently. To enable interaction across different parts, we employ a learnable parameter matrix $\boldsymbol{W}^s \in \mathbb{R}^{N_s\times N_s}$ for the feature integration. Specifically, at the t-th time step, the spatial feature is computed as

$$\boldsymbol{Z}^{\text{out}} = \boldsymbol{W}^s \cdot \text{FFN}(\boldsymbol{Z}), \tag{4}$$

where $\cdot$ denotes the matrix multiplication. Finally, the spatial feature is reshaped into $\boldsymbol{Z}^{\text{spatial}} \in \mathbb{R}^{T\times d'_s}$ with $d'_s = N_s \times d_s$, which serves as the final output.

4.2.2. Temporal Modeling

We first map the dance feature $\boldsymbol{X}$ and music embedding $\boldsymbol{M}$ to keys ($\boldsymbol{K}^d$, $\boldsymbol{K}^m$) and values ($\boldsymbol{V}^d$, $\boldsymbol{V}^m$) via linear projection and concatenate them to obtain $\boldsymbol{K} = [\boldsymbol{K}^d; \boldsymbol{K}^m] \in \mathbb{R}^{2T\times d_k}$ and $\boldsymbol{V} = [\boldsymbol{V}^d; \boldsymbol{V}^m] \in \mathbb{R}^{2T\times d_v}$, where d_k and d_v are the dimensions of key and value, respectively. Note that $d_v = d'_s$. Then, the attention map $\boldsymbol{A} \in \mathbb{R}^{d_k\times d_v}$ is computed as $\boldsymbol{A} = \text{softmax}(\boldsymbol{K}^\top)\boldsymbol{V}$, which effectively encodes the music and dance information into d_k distinct latent bins. Each bin is characterized by a d_v-dimensional feature vector, which serves as the basis for our temporal dynamics modeling.

Specifically, $\boldsymbol{A}$ is processed by 5 distinct FFNs to produce a set of dynamic components: a temporal position matrix $\hat{\boldsymbol{T}} \in \mathbb{R}^{d_k\times 1}$, a base state matrix $\boldsymbol{F}_0 \in \mathbb{R}^{d_k\times d_v}$, and three derivative matrices $\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_3 \in \mathbb{R}^{d_k\times d_v}$, which encode the scaled first-, second-, and third-order motion dynamics, respectively. Following [38], we update the attention

map using a Taylor expansion-based mechanism, which allows the model to incorporate higher-order temporal variations in a structured and interpretable manner:

$$\boldsymbol{Z}^{\mathrm{TS}} = \boldsymbol{F}_0 + \boldsymbol{F}_1 \odot (\hat{\boldsymbol{t}} - \hat{\boldsymbol{T}}) + \boldsymbol{F}_2 \odot (\hat{\boldsymbol{t}} - \hat{\boldsymbol{T}})^2 + \boldsymbol{F}_3 \odot (\hat{\boldsymbol{t}} - \hat{\boldsymbol{T}})^3, \tag{5}$$

where $\odot$ denotes element-wise multiplication, and the $\hat{\boldsymbol{t}} = [\hat{t}^1, \hat{t}^2, \cdots, \hat{t}^T] \in \mathbb{R}^T$ with $\hat{t}^i \in \mathbb{R}$ is the time sequence vector. Please find more details about the broadcasting process in Algorithm 1.

To integrate d_k latent bins in the $\boldsymbol{Z}^{\mathrm{TS}} \in \mathbb{R}^{T \times d_k \times d_v}$ into a unified representation, we employ a dynamic weighting strategy. This approach assigns a relevance score to each bin based on its temporal proximity to the current frame, allowing the model to selectively emphasize temporally coherent features. Formally, for each timestep t, the temporal feature $\boldsymbol{Z}_t^{\mathrm{temporal}} \in \mathbb{R}^{1 \times d_v}$ is computed via a weighted sum over d_k bins:

$$\boldsymbol{Z}_t^{\mathrm{temporal}} = \sum_{k=1}^{d_k} \gamma_{t,k} \cdot \boldsymbol{Z}_{t,k}^{\mathrm{TS}}, \tag{6}$$

where $\gamma_{t,k}$ and $\boldsymbol{Z}_{t,k}^{\mathrm{TS}}$ denote the (t,k)-th entry of the weight matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{T \times d_k}$ and $\boldsymbol{Z}^{\mathrm{TS}}$, respectively. Then, $\gamma_{t,k}$ is processed by a softmax-normalized Gaussian kernel:

$$\gamma_{t,k} = \frac{\exp(-(t - \hat{\boldsymbol{T}}_k)^2 / \sigma^2)}{\sum_{i=1}^{d_k} \exp(-(t - \hat{\boldsymbol{T}}_i)^2 / \sigma^2)}, \tag{7}$$

where $\sum_{k=1}^{d_k} \gamma_{t,k} = 1$, $\hat{\boldsymbol{T}}_k$ denotes the temporal position associated with the k-th bin, and σ^2 is a variance hyperparameter that controls the smoothness of the weighting distribution. In this way, the model can focus more on features that are temporally aligned with the current context.

Finally, the full temporal representation $\boldsymbol{Z}^{\mathrm{temporal}} \in \mathbb{R}^{T \times d_v}$ is obtained by concatenating the frame-wise features $\boldsymbol{Z}_t^{\mathrm{temporal}}$ along the temporal dimension.

4.2.3. Spatio-Temporal Feature Fusion

With $\boldsymbol{Z}^{\mathrm{spatial}}$ and $\boldsymbol{Z}^{\mathrm{temporal}}$ available, we integrate them via an additive operation,

$$\boldsymbol{Z}^{\mathrm{global}} = \boldsymbol{Z}^{\mathrm{spatial}} + \boldsymbol{Z}^{\mathrm{temporal}}. \tag{8}$$

Such a fusion allows the model to jointly leverage spatial-temporal configuration, resulting in a more informative global-level representation.

4.2.4. Hierarchical Mechanism

We also propose a hierarchical mechanism to capture long-term and short-term music-dance dependencies. The aforementioned global feature $\boldsymbol{Z}^{\text{global}}$ is used to represent long-term dependencies, while short-term dynamics are modeled within local-level temporal segments. Specifically, we first partition the music and dance features into K non-overlapping short-term windows along the temporal dimension, obtaining $\hat{\boldsymbol{M}} = [\boldsymbol{M}_l^1, \boldsymbol{M}_l^2, \cdots, \boldsymbol{M}_l^K]$ and $\hat{\boldsymbol{X}} = [\boldsymbol{X}_l^1, \boldsymbol{X}_l^2, \cdots, \boldsymbol{X}_l^K]$. We independently apply the decoupled attention mechanism to each window. Subsequently, the features extracted from all windows are concatenated along the temporal axis to form the local-level representation $\boldsymbol{Z}^{\text{local}}$. Finally, to incorporate hierarchical information, global and local representations can be combined as follows:

$$\boldsymbol{Z}^{\text{h}} = \text{FFN}(\boldsymbol{Z}^{\text{global}} + \boldsymbol{Z}^{\text{local}}), \tag{9}$$

where $\boldsymbol{Z}^{\text{h}}$ denotes the output of our Hierarchical Decoupled Attention (HDA) module, which effectively captures both global dynamics and local details.

4.3. Cycle-Consistent Learning Mechanism

The Cycle-Consistent Learning (CCL) mechanism incorporates a dance-to-music model to enhance temporal consistency between the generated dance sequences and the source music. The motivation is that a high-quality music-driven dance should preserve the information in the source music that is visually reflected by motion, such as rhythmic accents, temporal structure, and motion-related musical semantics. If the generated dance loses these music-related cues, the auxiliary D2M model cannot reconstruct the corresponding music features well. Therefore, reconstructing music features from generated motion provides a direct cross-modal supervision signal, encouraging the dance to preserve music-related information and remain aligned with the input music. Notably, we reconstruct music features rather than raw audio, because music features provide a compact and motion-relevant target while avoiding acoustic details that are weakly reflected by body motion. To achieve this, this mechanism employs an auxiliary task that first generates dance motion from input music, then reconstructs the music features from the generated dance with the dance-to-music model.

Algorithm 1 Broadcasting Details in Taylor Expansion

Input:

Time sequence vector $\hat{\boldsymbol{t}} \in \mathbb{R}^T$;

Temporal position matrix $\hat{\boldsymbol{T}} \in \mathbb{R}^{d_k \times 1}$;

Base state matrix $\boldsymbol{F}_0 \in \mathbb{R}^{d_k \times d_v}$;

Dynamics matrices $\{\boldsymbol{F}_i\}_{i=1}^3 \in \mathbb{R}^{d_k \times d_v}$.

Output: Final output matrix $\boldsymbol{Z}^{\mathrm{TS}} \in \mathbb{R}^{T \times d_k \times d_v}$.

// Reshape inputs to enable broadcasting

$\hat{\boldsymbol{t}}' \leftarrow \mathrm{reshape}(\hat{\boldsymbol{t}}, (T, 1, 1))$

$\hat{\boldsymbol{T}}' \leftarrow \mathrm{reshape}(\hat{\boldsymbol{T}}, (1, d_k, 1))$

$\boldsymbol{F}'_0 \leftarrow \mathrm{reshape}(\boldsymbol{F}_0, (1, d_k, d_v))$

$\boldsymbol{F}'_1 \leftarrow \mathrm{reshape}(\boldsymbol{F}_1, (1, d_k, d_v))$

$\boldsymbol{F}'_2 \leftarrow \mathrm{reshape}(\boldsymbol{F}_2, (1, d_k, d_v))$

$\boldsymbol{F}'_3 \leftarrow \mathrm{reshape}(\boldsymbol{F}_3, (1, d_k, d_v))$

// Compute the time difference via broadcasting

$\Delta \leftarrow \hat{\boldsymbol{t}}' - \hat{\boldsymbol{T}}'$ {Resulting shape is $(T, d_k, 1)$}

// Compute all terms and sum them up using broadcasting

$\boldsymbol{Z}^{\mathrm{TS}} \leftarrow \boldsymbol{F}'_0 + \boldsymbol{F}'_1 \odot \Delta + \boldsymbol{F}'_2 \odot \Delta^2 + \boldsymbol{F}'_3 \odot \Delta^3$

return $\boldsymbol{Z}^{\mathrm{TS}}$

4.3.1. Dance-to-Music Generation

We propose a Dance-to-Music (D2M) model to reconstruct music from the predicted dance motion $\hat{\boldsymbol{X}} = [\hat{\boldsymbol{x}}^1, \hat{\boldsymbol{x}}^2, \cdots, \hat{\boldsymbol{x}}^T]$. Specifically, the model consists of a Variational Autoencoder (VAE) and a Latent Diffusion Model (LDM). To obtain high-quality music latents $\hat{\boldsymbol{M}} = [\hat{\boldsymbol{m}}^1, \hat{\boldsymbol{m}}^2, \cdots, \hat{\boldsymbol{m}}^{T'}]$ and achieve high-fidelity reconstruction, we adopt a pretrained VAE, which is trained on 20,000 mel-spectrograms, each representing a 5-second music clip randomly sampled from a Spotify playlist. Subsequently, a latent diffusion model leverages an autoencoder ϵ_θ^l to predict the noise, conditioned on the dance motion via a cross-attention module. Similar to the music-to-dance model, we use the MSE loss to optimize ϵ_θ^l, which can be computed as:

$$\mathcal{L}_{\mathrm{d2m}} = \mathbb{E}_{\hat{\boldsymbol{M}}, \hat{\boldsymbol{X}}, \epsilon, t}\Big[\|\epsilon - \epsilon_\theta^l(\hat{\boldsymbol{M}}_t, t, \hat{\boldsymbol{X}})\|_2^2\Big]. \tag{10}$$

4.3.2. Cycle-Consistent Learning

We leverage the D2M model to establish a cycle-consistent mapping. Specifically, we first pre-train the D2M model and freeze its parameters to improve M2D training efficiency and stability. During the M2D training phase, the predicted dance sequences $\hat{\boldsymbol{X}}$ (or the denoised output $\epsilon_\theta^s(\boldsymbol{X}_t, t, \boldsymbol{M})$) are fed into the frozen D2M model. Specifically, conditioned on the predicted dance sequence, the frozen D2M model performs one denoising step in the latent space to estimate the music latent representation $\boldsymbol{M}_R$. Rather than performing a complete iterative reverse diffusion process, which would introduce substantial computational overhead and a long back-propagation path, we employ this single-step inference strategy during M2D training. We introduce a cycle consistency loss that minimizes the L2 distance between the reconstructed music latents $\boldsymbol{M}_R$ and the original music latents $\boldsymbol{M}_O$:

$$\mathcal{L}_{\text{cycle}} = \|\boldsymbol{M}_R - \boldsymbol{M}_O\|_2^2. \tag{11}$$

The gradients from the $\mathcal{L}_{\text{cycle}}$ are propagated back through the M2D model, effectively regularizing its training. This additional supervision further helps the M2D model generate dance motions that better align with the semantic structure, *e.g.*, beats, of the music.

5. Experiments

In this section, we present a comprehensive evaluation of BeatDance. We first describe the experimental settings, including the datasets, implementation details, and evaluation metrics used throughout our evaluation. We then compare our method against state-of-the-art approaches through both quantitative and qualitative analysis. Finally, we conduct ablation studies to analyze the contribution of each proposed component.

5.1. Experimental Settings

In this subsection, we describe the two benchmark datasets used for evaluation, specify the implementation details of our model, and introduce the evaluation metrics adopted to assess the quality of generated dances.

5.1.1. Datasets

AIST++ [49] is a large-scale 3D dance motion dataset reconstructed from multi-view dance videos. It contains 1,408 sequences with a total duration of 5.2 hours, covering 10 genres of street dance performed by 30 subjects. The genres span both old-school styles (Break, Pop, Lock, and Waack) and new-school styles (Middle Hip-hop, LA-style Hip-hop, House, Krump, Street Jazz, and Ballet Jazz). Each genre contains both basic (85%) and advanced (15%) choreographies. The accompanying music spans tempos from 80 to 135 BPM. Per-frame annotations include SMPL pose parameters for 24 joints with global translation, and 17 COCO-format joint locations in both 2D and 3D. The dataset is carefully split to ensure no overlap of music or choreography between training and test sets. Following [15], we adopt the same split and preprocessing, adjusting all training samples to 5 seconds at 30 fps.

PopDanceSet [6] is a more recent dataset designed to reflect the aesthetic preferences of contemporary audiences. It is constructed by filtering dance videos from a popular online platform using a popularity function. The dataset totals 12,819 seconds across 19 music genres — from CPOP and KPOP to house dance and rock — performed by 132 subjects. Per-frame annotations include 24 SMPL pose parameters and 17 COCO-format 3D joint locations, extracted using a monocular pose estimation model. Compared to AIST++, PopDanceSet spans a broader range of dance styles and music genres, and features more intricate choreographic movements, posing greater challenges for music-dance alignment. For fair comparison, we follow its official data split and process all samples to 5 seconds at 30 fps.

Together, these two datasets enable a comprehensive evaluation of BeatDance: AIST++ serves as a well-established benchmark, while PopDanceSet provides a more choreographically complex scenario.

5.1.2. Implementation Details

All experiments are conducted on 4 NVIDIA A6000 GPUs. Our framework consists of two branches: the music-to-dance (M2D) generation branch and the auxiliary dance-to-music (D2M) branch used for cycle-consistent learning. We describe the implementation details of each branch below.

Dance Representation. The dance feature vector comprises 156 dimensions: a 3-dimensional root joint translation, 6-DoF rotations for 24 SMPL body joints (144 dimensions in total), and 9 binary contact indicators for the feet, hands, and neck. For the Hierarchical Decoupled Attention module, the 24 body joints are partitioned into N_s=8 groups: 7 functionally relevant regions incorporating the head, spine, left arm, right arm, left leg, right leg, and root joint, along with an additional global full-body part. Each group is processed by an independent linear projection to preserve part-specific motion characteristics.

Music Representation. We use the pretrained Jukebox [16] model to extract music features, which yields 4,800-dimensional embeddings per frame. These embeddings encode rich semantic and rhythmic information from the raw audio, providing a strong conditioning signal for the M2D generation branch. All music clips are processed at 30 fps to match the dance frame rate, yielding 150 frames per 5-second clip.

Music-to-Dance Branch. The M2D diffusion model adopts a cosine-based variance schedule with a cosine offset of 0.008 and sets the total denoising steps T to 1000. The denoising network is a 6-layer Transformer with latent dimensions of 8×64, where 8 corresponds to the number of body parts and 64 is the per-part feature dimensionality. The local hierarchical window size τ is set to 30 frames, and the Taylor series expansion order for the temporal attention map is set to 3, both determined through ablation studies (Table 4 and Table 5). During training, we use the Adan [50] optimizer with a learning rate of 2×10^{-4}, a weight decay of 0.02, and a batch size of 128. During inference, we employ DDIM sampling with 50 steps, which reduces inference time by 20× relative to standard DDPM sampling while maintaining comparable generation quality.

Dance-to-Music Branch. The D2M branch synthesizes audio represented as mel-spectrograms with 256 Mel-frequency bands. The VAE encoder and decoder each consist of four convolutional blocks for progressive downsampling and upsampling, respectively, compressing the mel-spectrogram into a compact latent representation. The latent diffusion model adopts a U-Net architecture with a four-layer convolutional encoder and a symmetric four-layer decoder connected via skip connections, enabling multi-scale feature

reuse. For training, we employ the AdamW optimizer [51] with a learning rate of 1×10^{-4}, a weight decay of 10^{-6}, and a batch size of 32.

Training Strategy. The M2D model is the primary focus of this work and is trained with the support of the auxiliary D2M branch via a two-stage procedure. In the first stage, the D2M branch is pre-trained independently to ensure it can provide reliable cycle-consistency supervision. In the second stage, its parameters are frozen, and it is used exclusively to provide cycle-consistency supervision for M2D: dance sequences generated by M2D are passed through the frozen D2M branch to reconstruct the source music, and the resulting $\mathcal{L}_{\text{cycle}}$ is back-propagated to refine the M2D model. Throughout training, only the M2D parameters are updated.

5.1.3. Evaluation Metrics

We follow previous methods [15, 6] to evaluate generated dances from three dimensions: physical plausibility (PFC and PBC), diversity (Div_k and Div_g), and beat alignment with music (BAS). These metrics characterize complementary aspects of visual dance quality: PFC and PBC reflect whether the motion appears physically plausible, Div_k and Div_g reflect the variety of generated movements, and BAS reflects whether visible motion accents follow the musical rhythm. Since no single automatic metric fully captures the perceptual quality of a dance, we interpret them jointly.

Physical Foot Contact (PFC). PFC evaluates the physical plausibility of lower-body dance motion. It is computed as:

$$\text{PFC} = \frac{1}{N \cdot \max_{1\le j\le N} \|\overline{\boldsymbol{a}}^j_{\text{COM}}\|} \sum_{i=1}^{N} s^i, \tag{12}$$

where N is the number of frames and s^i is the per-frame plausibility score:

$$s^i = \|\overline{\boldsymbol{a}}^i_{\text{COM}}\| \cdot \|\boldsymbol{v}^i_{\text{LF}}\| \cdot \|\boldsymbol{v}^i_{\text{RF}}\|. \tag{13}$$

Here, $\boldsymbol{v}^i_{\text{LF}}$ and $\boldsymbol{v}^i_{\text{RF}}$ denote the velocities of the left and right foot at frame i, and $\overline{\boldsymbol{a}}^i_{\text{COM}}$ is the center-of-mass (COM) acceleration with its vertical component clamped to be

non-negative:

$$\overline{\boldsymbol{a}}^i_{\mathrm{COM}} = \begin{pmatrix} a^i_{\mathrm{COM},x} \\ a^i_{\mathrm{COM},y} \\ \max(a^i_{\mathrm{COM},z},\ 0) \end{pmatrix}. \tag{14}$$

PFC is designed to expose foot-sliding artifacts. Intuitively, when the body accelerates, at least one foot should provide stable ground contact; a frame is therefore penalized when the COM accelerates while both feet move simultaneously. A lower PFC indicates fewer visually implausible foot-ground artifacts.

Physical Body Contact (PBC). PBC extends PFC to assess the physical plausibility of full-body dance motion. An auxiliary function f is first defined as:

$$f(x, y, z) = \frac{\sum_{i=1}^{N} \|\overline{\boldsymbol{a}}^i_x\| \cdot \|\boldsymbol{v}^i_y\| \cdot \|\boldsymbol{v}^i_z\|}{\max_{1 \le j \le N} \|\overline{\boldsymbol{a}}^j_x\|}, \tag{15}$$

where $\overline{\boldsymbol{a}}_x$ and $\boldsymbol{v}_y$ denote the acceleration and velocity of the joints specified in each argument of f. PBC is then defined as:

$$\begin{aligned} \mathrm{PBC} = \frac{1}{N}[&- f(\mathrm{root}, \mathrm{lfoot}, \mathrm{rfoot}) \\ &+ f(\mathrm{lchest}, \mathrm{lhand}, \mathrm{null}) \\ &+ f(\mathrm{rchest}, \mathrm{rhand}, \mathrm{null}) \\ &+ f(\mathrm{neck}, \mathrm{head}, \mathrm{null})]. \end{aligned} \tag{16}$$

For example, in $f(\mathrm{root}, \mathrm{lfoot}, \mathrm{rfoot})$, $\overline{\boldsymbol{a}}_x$ is the acceleration of the root joint, and $\boldsymbol{v}_y$, $\boldsymbol{v}_z$ are the velocities of the left and right foot, respectively. PBC extends the foot-contact assessment to the full body by combining root-foot, chest-hand, and neck-head motion relationships. It reflects whether the generated whole-body movements are physically coordinated. Intuitively, a dance with a realistic PBC exhibits fewer foot-sliding artifacts and more natural movements in which the hands and head move coherently with the torso. Following POPDG [6], values closer to the ground-truth PBC are considered better.

Diversity (Div_k and Div_g). We evaluate diversity from two complementary perspectives: kinematic (Div_k) and geometric (Div_g). Each score is computed as the average

pairwise Euclidean distance between the feature vectors of all generated sequences in the test set, calculated independently on kinematic and geometric features. Div_k measures variation in motion dynamics, such as differences in velocity and acceleration patterns, whereas Div_g measures variation in geometric pose and choreographic configurations. Visually, they indicate whether a model generates a broad range of movement styles and body shapes instead of repeatedly producing similar sequences. However, abnormally jittery or distorted motions can also inflate pairwise distances. Therefore, following EDGE [15], we regard values close to the ground-truth diversity as preferable to blindly maximizing either score.

Beat Alignment Score (BAS). BAS measures the rhythmic synchronization between the generated dance and the input music. It is defined as:

$$\text{BAS} = \frac{1}{N_p} \sum_{i=1}^{N_p} \exp\left(-\frac{\min_{b_j^m \in B^m} \|b_i^p - b_j^m\|^2}{2\sigma^2} \right), \tag{17}$$

where $B^p = \{b_i^p\}$ and $B^m = \{b_j^m\}$ are the sets of kinematic and music beats, N_p is the number of kinematic beats, and σ is a normalization hyperparameter. The kinematic beats are extracted as local minima of the kinetic velocity curve, and the music beats are detected using the librosa library. For each kinematic beat, BAS assigns a larger contribution when a nearby music beat exists; consequently, a higher BAS indicates tighter temporal synchronization. In visual terms, high BAS corresponds to salient motion accents, such as brief pauses, direction changes, or transitions, occurring near musical beats.

5.2. Comparison with State-of-the-Art Methods

In this subsection, we compare BeatDance against competitive baselines on two benchmark datasets. We first present quantitative results across all evaluation metrics, followed by qualitative visualizations to illustrate the superiority of our method in generating expressive and beat-consistent dance motions. For a fair comparison, all methods use the same experimental setting: 4,800-dimensional Jukebox music embeddings per frame, the same 156-dimensional motion representation, and 5-second music and motion sequences sampled at 30 fps. We evaluate all generated motions using the same protocol and metrics described in Section 5.1.3.

Table 1: Comparison with the state-of-the-art methods on PopDanceSet and AIST++ test sets. Metrics are marked as: ↑ higher is better, ↓ lower is better, → closer to GT is better. **Bold** (underlined) indicates the best (second-best) results. † denotes diffusion-based methods.

Dataset	Method	Motion Quality		Motion Diversity		Alignment
		PFC ↓	PBC →	Div_k →	Div_g →	BAS ↑
PopDanceSet	Ground Truth	1.5824	8.7365	9.0219	7.2931	0.174
	Bailando [9]	3.9751	4.8863	5.1835	5.4342	0.230
	EDGE† [15]	3.8366	4.0348	<u>6.1709</u>	5.7568	0.224
	POPDG† [6]	<u>1.8253</u>	5.9492	**7.1342**	<u>5.8314</u>	0.233
	DanceEditor† [46]	1.8972	<u>7.8429</u>	5.6338	5.2459	<u>0.238</u>
	Ours†	**1.7332**	**7.9080**	5.7236	**5.8507**	**0.244**
AIST++	Ground Truth	1.1806	8.0353	8.5190	8.9384	0.453
	Bailando [9]	1.0633	<u>6.3632</u>	14.8835	**6.6585**	0.474
	EDGE† [15]	1.0792	5.1172	2.5946	2.4477	0.466
	POPDG† [6]	<u>0.9952</u>	4.7050	2.4775	2.7832	0.468
	DanceEditor† [46]	1.6540	4.6520	<u>2.7542</u>	<u>2.7951</u>	<u>0.475</u>
	Ours†	**0.9877**	**7.1591**	**2.7965**	2.7441	**0.476**

5.2.1. Quantitative Results

Table 1 presents quantitative evaluation results on the PopDanceSet and AIST++ datasets. Our method achieves the best results on four of the five metrics on each dataset, including PFC, PBC, and BAS on both datasets. The remaining differences in diversity metrics are discussed below.

Motion quality. BeatDance achieves state-of-the-art performance in both PFC and PBC on PopDanceSet and AIST++. Physical plausibility is a foundational criterion in dance generation: motions with foot-sliding artifacts or implausible physical dynamics are immediately perceptible to viewers, rendering them unsuitable regardless of their beat alignment or diversity. This improvement is driven by HDA and CCL in a complementary fashion. HDA models the fine-grained spatial-temporal structure of human motion. Its spatial branch learns region-specific physical dynamics through independent per-body-part projections. Its hierarchical temporal structure further ensures

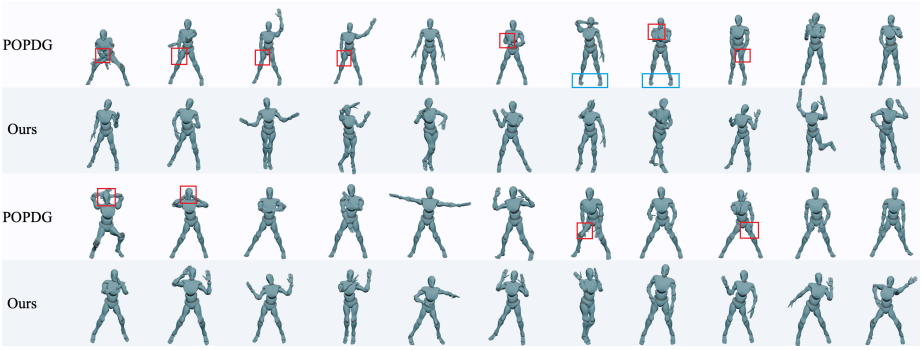

Figure 3: Visual results on the PopDanceSet. We visually compare BeatDance with the state-of-the-art method POPDG [6]. Penetration artifacts (red) and Distorted motions (blue) are highlighted.

coherent full-body coordination across multiple scales. CCL additionally discourages physically implausible configurations by constraining the generated motion to remain semantically aligned with the source music. Together, they produce the state-of-the-art PFC and PBC scores across both benchmarks.

Beat alignment. BeatDance also achieves state-of-the-art BAS on both PopDanceSet and AIST++. As the defining objective of music-driven dance generation, beat alignment is indispensable: a dance that fails to respond to the rhythmic structure of the music defeats the fundamental purpose of the task, regardless of its diversity. This advantage stems from the synergy between HDA and CCL. HDA equips the model with the capacity to learn fine-grained correspondences between musical rhythm and dance motion across multiple temporal scales, which is a prerequisite for precise beat synchronization. CCL then provides a direct optimization signal for this alignment. The music→dance→music round-trip can only close with low reconstruction error when the generated dance genuinely tracks the rhythmic structure of the source music. Together, the two components enforce beat-level synchronization.

Motion diversity. BeatDance achieves the best geometric diversity (Div_g) on PopDanceSet and the best kinematic diversity (Div_k) on AIST++. Diversity is *at best an auxiliary indicator*, as pairwise distance can be inflated by physically implausible outlier poses. A model generating degraded motions may therefore score artificially

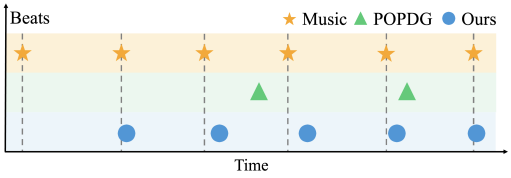

Figure 4: Visual comparison of beat consistency.

high on this metric, a phenomenon also noted in prior work [6, 9]. For the two metrics where we do not rank first, the gaps are readily explained. On PopDanceSet, POPDG achieves a higher Div_k than ours (7.13 vs. 5.72 for BeatDance), but this comes at the cost of substantially lower motion quality (PBC 5.95 for POPDG vs. 7.91 for BeatDance) and beat alignment (BAS 0.233 for POPDG vs. 0.244 for BeatDance). As shown in Figure 3, POPDG's generated motions exhibit clear quality issues, including penetration artifacts and distorted poses. Such implausible poses may partially account for its higher Div_k, as outlier motions tend to inflate pairwise distance measures. On AIST++, our Div_g (2.74) is competitive with the leading diffusion-based result (2.80). The only method that clearly surpasses ours is Bailando, which achieves a Div_g of 6.66, but this advantage does not generalize. Bailando achieves the second-lowest Div_g scores on PopDanceSet, suggesting that its high AIST++ diversity may reflect dataset-specific overfitting of its discrete choreographic memory rather than genuine expressive capability.

5.2.2. Qualitative Results

Figure 3 shows the qualitative comparison on the PopDanceSet dataset. Compared to POPDG [6], our method generates more expressive dance motions. We highlight penetration artifacts and distorted motions using red and blue rectangles, respectively. Furthermore, we provide a beat synchronization analysis in Figure 4. Our model (blue circles) exhibits significantly higher rhythmic consistency than POPDG (green triangles). Specifically, BeatDance captures 5 of 6 musical beats, whereas POPDG only aligns with 2. These results further validate the effectiveness of HDA and CCL, which enhance the dance motion quality and the music-dance consistency, particularly in terms of beat-level alignment. It is worth noting that Figure 4 is provided only as an illustrative example for an intuitive understanding of beat alignment. The BAS results

in Table 1 are computed over the complete test sets, where BeatDance achieves the best performance on both PopDanceSet and AIST++. Regarding diversity, we observe that even though POPDG scores higher on Div_k, the qualitative results in Figure 3 reveal that a portion of its apparent kinematic variety arises from physically implausible configurations (*e.g.*, penetration artifacts and distorted limb poses highlighted in red and blue), rather than genuine choreographic diversity. BeatDance, by contrast, maintains consistent limb topology across generated sequences while still producing varied whole-body postures adapted to the input music, suggesting that its lower Div_k reflects tighter physical constraints rather than a lack of expressive range.

5.2.3. User Study

To complement the objective metrics, we conducted a pairwise user study with 20 participants using samples from the PopDanceSet test set. BeatDance was compared with POPDG and DanceEditor across multiple music clips, yielding 10 comparisons per participant and 200 pairwise judgments in total. Each pair consisted of a BeatDance result and a result from one competing method generated from the same music clip. For each pair, participants selected the dance with better overall quality while considering motion naturalness, expressiveness, and music-motion compatibility. Method identities were hidden, and both the comparison order and the left-right placement of the videos were randomized. As shown in Table 2, BeatDance was preferred over both competing methods, demonstrating its perceptual advantage in overall dance quality.

Table 2: Pairwise user-study results on the PopDanceSet test set.

Comparison	BeatDance Win Rate
BeatDance vs. POPDG	91%
BeatDance vs. DanceEditor	94%

5.3. Ablation Studies

To ensure experimental efficiency while maintaining a fair comparison, we adopt a scaled-down model configuration (half dimension and layers) for all ablation experiments, following the protocol in POPDG [6]. We empirically observed that the relative

Table 3: Ablation study of different components. DA, HM, and CCL denote the Decoupled Attention, Hierarchical Mechanism, and Cycle Consistency Learning, respectively.

Method	PFC	PBC	Div_k	Div_g	BAS
w/o DA	15.5267	6.3755	**8.9975**	6.1423	0.218
w/o HM	1.6046	5.7520	6.9480	**6.3608**	0.240
w/o CCL	1.5594	5.8300	6.6467	5.5107	0.252
Ours	**1.5390**	**7.7493**	6.7770	6.1801	**0.273**

performance trends remain consistent between the scaled-down and full models.

5.3.1. Investigating Each Component

Effect of Decoupled Attention. We replace DA with vanilla cross-attention to assess its contribution. As shown in Table 3, removing DA causes a dramatic drop in PFC (1.54 → 15.53), PBC (7.75 → 6.38), and BAS (0.273 → 0.218). In DA, each anatomical group is assigned an independent linear projection, preserving part-specific dynamics and allowing each body region to independently align with musical rhythm. Replacing it with shared attention entangles the representations of all body parts, disrupting both local physical constraints and fine-grained music-motion correspondence. Although *w/o DA* achieves the highest Div_k (8.9975), this is consistent with the pattern observed in Section 5.2.1: physically implausible motions tend to inflate pairwise kinematic distance. The near-unchanged Div_g (6.14 vs. 6.18) further supports this interpretation, as geometric diversity is less sensitive to such artifacts.

Effect of Hierarchical Mechanism. We remove the short-term branch of HM, retaining only long-term attention. As shown in Table 3, the primary degradations appear in PBC (7.75 → 5.75) and BAS (0.273 → 0.240), while PFC changes only marginally (1.54 → 1.60). This pattern aligns with the design motivation of HM: musical beats are sub-second local events that global attention, spread over the entire sequence, lacks the temporal resolution to capture precisely. The local windows (size τ=30 frames, *i.e.*, 1 second) provide short-range receptive fields that improve beat-level alignment and full-body coordination, as reflected in the BAS and PBC gains. PFC, which primarily depends on part-specific spatial modeling already provided by DA, is less sensitive to

Table 4: Ablation study of the order of the Taylor series expansion.

Order	PFC	PBC	Div_k	Div_g	BAS
1	1.2401	7.2673	6.6585	5.8027	0.229
2	1.3925	6.9268	6.6322	5.4058	0.242
3	1.5390	**7.7493**	**6.7770**	**6.1801**	**0.273**
4	1.7447	7.3459	6.5358	5.6094	0.255
5	**1.1339**	6.4334	6.2685	5.1131	0.256

the removal of local temporal context. The slight increases in Div_k and Div_g (6.78 $\rightarrow$ 6.95 and 6.18 $\rightarrow$ 6.36) suggest that removing local constraints allows the model to explore a slightly broader motion distribution, at the cost of physical plausibility and rhythmic consistency.

Effect of Cycle-Consistent Learning. We remove CCL and train the M2D model without cycle-consistency supervision. As shown in Table 3, removing CCL leads to consistent degradation across nearly all metrics: PBC drops from 7.75 to 5.83, BAS from 0.273 to 0.252, Div_k from 6.78 to 6.65, and Div_g from 6.18 to 5.51, with only PFC remaining largely unchanged (1.54 $\rightarrow$ 1.56). This broad pattern suggests that CCL fosters holistic cross-modal alignment between music and full-body motion, especially beat alignment.

5.3.2. Effect of Taylor Series Expansion Order

We investigate the effect of the Taylor series expansion order used to approximate the joint music-dance attention map. As shown in Table 4, order 3 achieves the best performance across most metrics, including PBC (7.75), Div_k (6.78), Div_g (6.18), and BAS (0.273). Lower orders (1 and 2) yield consistently lower BAS (0.229 and 0.242), suggesting that a low-order approximation lacks the capacity to capture the complex non-linear dependencies between music and dance, resulting in weaker beat alignment. As the order increases beyond 3, performance degrades across most metrics: PBC drops from 7.75 to 6.43 and BAS from 0.273 to 0.256 at order 5, indicating that excessively high-order approximations introduce optimization difficulties that impair music-

Table 5: Ablation study of the local window size τ.

τ	PFC	PBC	Div_k	Div_g	BAS
5	2.6125	**8.5605**	7.5657	5.8857	0.232
15	1.6489	6.9349	6.6073	5.5570	0.249
30	1.5390	7.7493	6.7770	**6.1801**	**0.273**
50	**1.4113**	6.7600	**8.7246**	6.0828	0.218
75	1.8742	5.3648	7.2374	5.5517	0.234

dance correspondence. Although order 5 achieves the best PFC (1.13), this comes at the cost of significant degradation in all other metrics and therefore does not represent an overall improvement. We thus adopt order 3 as the default setting.

5.3.3. Effect of Local Window Size

We analyze the effect of the local window size τ on model performance. As shown in Table 5, a small window (τ=5) achieves the highest PBC but the worst PFC, as the limited temporal context is insufficient to maintain global motion coherence. Conversely, a large window (τ=50) yields the best PFC but the worst BAS (0.218), as the coarse temporal resolution makes it difficult to resolve sub-second beat events. Our chosen window size of 30 achieves the best BAS (0.273) and Div_g (6.18) while maintaining competitive performance across other metrics, achieving a favorable balance between local rhythmic precision and long-range motion coherence. Given that tempo varies across clips, τ=30 should be understood as an approximate temporal context rather than a fixed musical unit such as a beat, bar, or phrase. This one-second context can cover local beat-level patterns and short motion transitions while avoiding overly coarse temporal context that weakens sensitivity to beat events. We use τ=30 by default.

5.3.4. Effect of Generative Architecture

To choose the most appropriate architecture for the dance-to-music branch, we compare two widely used architectures: Generative Adversarial Network (GAN) [52] and Diffusion models [25]. Each model is evaluated under two settings: operating on

Table 6: Ablation study of different generative architectures for the dance-to-music generation task.

Method	BCS ↑	BHS ↑	FAD ↓	BAS ↑
Ground Truth	100	100	0	0.174
GAN	86.5	52.8	7.63	0.191
VAE+GAN	100.2	54.6	7.52	0.193
Diffusion	110.8	62.1	7.60	0.193
VAE+Diffusion	**113.8**	**64.8**	**6.86**	**0.195**

raw data and within a VAE [53] latent space. To evaluate rhythmic consistency and content similarity between the generated music and the ground truth, we adopt four metrics: BAS, BCS (Beats Coverage Score) [54], BHS (Beats Hit Score) [54], and FAD (Fréchet Audio Distance) [55]. BCS and BHS evaluate rhythmic fidelity, where BCS measures the difference in total beat count between the synthesized and reference music, and BHS examines their temporal beat alignment. FAD is used to quantify content similarity and overall audio quality. As shown in Table 6, the VAE + Diffusion configuration achieves the best performance across all metrics, indicating its superior capability in modeling both rhythm and audio content.

6. Conclusion

In this paper, we propose a diffusion-based framework for 3D music-driven dance generation, which comprises a Hierarchical Decoupled Attention (HDA) and a cycle-consistent learning (CCL) mechanism. We employ HDA to capture both long-term and short-term dependencies between the source music and the generated dance by explicitly disentangling spatial and temporal modeling. Additionally, CCL effectively enhances the music-dance consistency by an auxiliary music-to-dance-to-music task. Consequently, our proposed approach outperforms recent competitive methods on two benchmark datasets, particularly in physical plausibility and beat alignment. Extensive experiments also show its superior performance in maintaining beat consistency. Despite these promising results, BeatDance still has several limitations. First, the effectiveness of CCL depends on the reliability of the auxiliary D2M model. Although

the D2M model is independently pre-trained and the ablation study validates the practical benefit of CCL, inaccurate D2M estimation under challenging music patterns may introduce noisy or biased supervision. Second, BeatDance may show weaker beat synchronization when the input contains unseen music genres with highly complex rhythmic patterns. Future work will explore more robust D2M supervision, explicit style control, and stronger generalization to complex musical rhythms.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 62502447 and the Earth System Big Data Platform of the School of Earth Sciences, Zhejiang University.

References

[1] C. Gu, J. Yu, C. Zhang, Learning disentangled representations for controllable human motion prediction, Pattern Recognition 146 (2024) 109998.

[2] Z. Wu, K. Chen, K. Li, H. Fan, Y. Yang, Bvinet: Unlocking blind video inpainting with zero annotations, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2025, pp. 14017–14027.

[3] H. Jia, N. Zhao, Y. Xu, L. Zhu, Y. Yang, Gas: Geometry-appearance synergy for consistent video customization, in: Proceedings of the International Conference on Multimedia Modeling (ICMM), 2026, pp. 648–662.

[4] Z. Wu, C. Sun, H. Xuan, G. Liu, Y. Yan, Waveformer: wavelet transformer for noise-robust video inpainting, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Vol. 38, 2024, pp. 6180–6188.

[5] J. Wang, Z. Wu, H. Xuan, Y. Yan, Text-video completion networks with motion compensation and attention aggregation, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 2990–2994.

[6] Z. Luo, M. Ren, X. Hu, Y. Huang, L. Yao, Popdg: Popular 3d dance generation with popdanceset, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 26984–26993.

[7] L. Siyao, T. Gu, Z. Yang, Z. Lin, Z. Liu, H. Ding, L. Yang, C. C. Loy, Duolando: Follower gpt with off-policy reinforcement learning for dance accompaniment, in: International Conference on Learning Representations (ICLR), 2024.

[8] Y. Xu, L. Zhu, Y. Yang, Gg-editor: Locally editing 3d avatars with multimodal large language model guidance, in: Proceedings of the ACM International Conference on Multimedia (ACMMM), 2024, pp. 10910–10919.

[9] L. Siyao, W. Yu, T. Gu, C. Lin, Q. Wang, C. Qian, C. C. Loy, Z. Liu, Bailando: 3d dance generation by actor-critic gpt with choreographic memory, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11050–11059.

[10] K. Gong, D. Lian, H. Chang, C. Guo, Z. Jiang, X. Zuo, M. B. Mi, X. Wang, Tm2d: Bimodality driven 3d dance generation via music-text integration, in: Proceedings of the International Conference on Computer Vision (ICCV), 2023, pp. 9942–9952.

[11] C. Zhang, Y. Tang, N. Zhang, R.-S. Lin, M. Han, J. Xiao, S. Wang, Bidirectional autoregressive diffusion model for dance generation, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 687–696.

[12] R. Li, Y. Zhang, Y. Zhang, H. Zhang, J. Guo, Y. Zhang, Y. Liu, X. Li, Lodge: A coarse to fine diffusion network for long dance generation guided by the characteristic dance primitives, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1524–1534.

[13] R. Li, H. Zhang, Y. Zhang, Y. Zhang, Y. Zhang, J. Guo, Y. Zhang, X. Li, Y. Liu, Lodge++: High-quality and long dance generation with vivid choreography patterns, arXiv preprint arXiv:2410.20389 (2024).

[14] Z. Wang, H. Zhuang, L. Li, Y. Zhang, J. Zhong, J. Chen, Y. Yang, B. Tang, Z. Wu, Explore 3d dance generation via reward model from automatically-ranked demonstrations, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2024, pp. 301–309.

[15] J. Tseng, R. Castellon, K. Liu, Edge: Editable dance generation from music, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 448–458.

[16] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, I. Sutskever, Jukebox: A generative model for music, arXiv preprint arXiv:2005.00341 (2020).

[17] M. Petrovich, M. J. Black, G. Varol, Temos: Generating diverse human motions from textual descriptions, in: Proceedings of the European Conference on Computer Vision (ECCV), 2022, pp. 480–497.

[18] C. Guo, X. Zuo, S. Wang, L. Cheng, Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts, in: Proceedings of the European Conference on Computer Vision (ECCV), 2022, pp. 580–597.

[19] S. Lu, L.-H. Chen, A. Zeng, J. Lin, R. Zhang, L. Zhang, H.-Y. Shum, Humantomato: Text-aligned whole-body motion generation, in: International Conference on Machine Learning (ICML), 2024, pp. 32939–32977.

[20] E. Pinyoanuntapong, P. Wang, M. Lee, C. Chen, Mmm: Generative masked motion model, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1546–1555.

[21] A. Van Den Oord, O. Vinyals, et al., Neural discrete representation learning, Advances in Neural Information Processing Systems (NeurIPS) 30 (2017) 6309–6318.

[22] J. Zhang, Y. Zhang, X. Cun, S. Huang, Y. Zhang, H. Zhao, H. Lu, X. Shen, T2m-gpt: Generating human motion from textual descriptions with discrete representations, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14730–14740.

[23] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, Z. Liu, Motiondiffuse: Text-driven human motion generation with diffusion model, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 46 (6) (2024) 4115–4128.

[24] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, A. H. Bermano, Human motion diffusion model, in: International Conference on Learning Representations (ICLR), 2023.

[25] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, Advances in Neural Information Processing Systems (NeurIPS) 33 (2020) 6840–6851.

[26] Z. Zhang, A. Liu, I. Reid, R. Hartley, B. Zhuang, H. Tang, Motion mamba: Efficient and long sequence motion generation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2024, pp. 265–282.

[27] G. Yue, W. Li, C. Zhao, et al., Text-guided semantic alignment network with spatial-frequency interaction for infrared-visible image fusion under extreme illumination, IEEE Transactions on Image Processing (TIP) 34 (2025) 7943–7958.

[28] A. Khamis, et al., Instruction-driven 3d facial expression generation and transition, IEEE Transactions on Multimedia (TMM) 27 (2025) 6140–6153.

[29] M. Zhang, X. Guo, L. Pan, Z. Cai, F. Hong, H. Li, L. Yang, Z. Liu, Remodiffuse: Retrieval-augmented motion diffusion model, in: Proceedings of the International Conference on Computer Vision (ICCV), 2023, pp. 364–373.

[30] H. Chen, Z. Zhao, R. Wang, H. Yu, C. He, X. Zhang, Smrnet: Stacked motion residual learning with spatiotemporal modeling in diffusion models for human motion prediction, Pattern Recognition (2026) 113281.

[31] G. Yue, S. Li, R. Cong, T. Zhou, B. Lei, T. Wang, Attention-guided pyramid context network for polyp segmentation in colonoscopy images, IEEE Transactions on Instrumentation and Measurement (TIM) 72 (2023) 1–13.

[32] Y. Ma, F. W. Li, X. Liang, Uncertainty-aware calibrated 3d human motion forecasting with latent conformal prediction, Pattern Recognition (2026) 113144.

[33] Z. Wu, K. Li, Y. Xu, H. Fan, Y. Yang, Dlvinet: Advancing dual-lens video inpainting beyond parallax constraints, in: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Vol. 40, 2026, pp. 10888–10896.

[34] G. Yue, H. Xiao, H. Xie, T. Zhou, W. Zhou, W. Yan, B. Zhao, T. Wang, Q. Jiang, Dual-constraint coarse-to-fine network for camouflaged object detection, IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) 34 (5) (2023) 3286–3298.

[35] Z. Wu, C. Sun, H. Xuan, Y. Yan, Deep stereo video inpainting, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), 2023, pp. 5693–5702.

[36] R. Ding, K. Qu, J. Tang, Ksof: Leveraging kinematics and spatio-temporal optimal fusion for human motion prediction, Pattern Recognition 161 (2025) 111206.

[37] X. Chen, B. Jiang, W. Liu, Z. Huang, B. Fu, T. Chen, G. Yu, Executing your commands via motion diffusion in latent space, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18000–18010.

[38] M. Zhang, H. Li, Z. Cai, J. Ren, L. Yang, Z. Liu, Finemogen: Fine-grained spatio-temporal motion generation and editing, Advances in Neural Information Processing Systems (NeurIPS) 36 (2023) 13981–13992.

[39] M. Lee, K. Lee, J. Park, Music similarity-based approach to generating dance motion sequence, Multimedia Tools and Applications 62 (3) (2013) 895–912.

[40] F. Ofli, E. Erzin, Y. Yemez, A. M. Tekalp, Learn2dance: Learning statistical music-to-dance mappings for choreography synthesis, IEEE Transactions on Multimedia (TMM) 14 (3) (2011) 747–759.

[41] S. Wu, S. Lu, L. Cheng, Music-to-dance generation with optimal transport, arXiv preprint arXiv:2112.01806 (2021).

[42] Y. Huang, J. Zhang, S. Liu, Q. Bao, D. Zeng, Z. Chen, W. Liu, Genre-conditioned long-term 3d dance generation driven by music, in: ICASSP, 2022, pp. 4858–4862.

[43] W. Li, B. Ren, H. Xu, S. Cao, Y. Xie, Autodance: Music driven dance generation, in: International Symposium on Artificial Intelligence and its Application on Media (ISAIAM), 2021, pp. 55–59.

[44] D.-G. Lee, S.-W. Lee, Human interaction recognition framework based on interacting body part attention, Pattern Recognition 128 (2022) 108664.

[45] L. Siyao, W. Yu, T. Gu, C. Lin, Q. Wang, C. Qian, C. C. Loy, Z. Liu, Bailando++: 3d dance gpt with choreographic memory, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 45 (12) (2023) 14192–14207.

[46] H. Zhang, Z. Li, X. Qi, M. Li, M. Sun, S. Wang, M. Zhang, S. Han, Danceeditor: Towards iterative editable music-driven dance generation with open-vocabulary descriptions, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 12158–12168.

[47] A. Q. Nichol, P. Dhariwal, Improved denoising diffusion probabilistic models, in: International Conference on Machine Learning (ICML), 2021, pp. 8162–8171.

[48] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, in: International Conference on Learning Representations (ICLR), 2021.

[49] R. Li, S. Yang, D. A. Ross, A. Kanazawa, Ai choreographer: Music conditioned 3d dance generation with aist++, in: Proceedings of the International Conference on Computer Vision (ICCV), 2021, pp. 13401–13412.

[50] X. Xie, P. Zhou, H. Li, Z. Lin, S. Yan, Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 46 (12) (2024) 9508–9520.

[51] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations (ICLR), 2019.

[52] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Communications of the ACM 63 (11) (2020) 139–144.

[53] D. P. Kingma, M. Welling, Auto-encoding variational bayes, in: International Conference on Learning Representations (ICLR), 2014.

[54] H.-Y. Lee, X. Yang, M.-Y. Liu, T.-C. Wang, Y.-D. Lu, M.-H. Yang, J. Kautz, Dancing to music, Advances in Neural Information Processing Systems (NeurIPS) 32 (2019) 3586–3596.

[55] K. Kilgour, M. Zuluaga, D. Roblek, M. Sharifi, Fréchet audio distance: A metric for evaluating music enhancement algorithms, arXiv preprint arXiv:1812.08466 (2018).