Align Your Rhythm: Generating Highly Aligned Dance Poses with Gating-Enhanced Rhythm-Aware Feature Representation

Congyi Fan[1], **Jian Guan**[1]*, **Xuanjia Zhao**[1], **Dongli Xu**[2], **Youtian Lin**[3], **Tong Ye**[1], **Pengming Feng**[4], **Haiwei Pan**[1]

[1]College of Computer Science and Technology, Harbin Engineering University, Harbin, China
[2]Shanghai Academy of Artificial Intelligence for Science, Shanghai, China
[3]School of Intelligence Science and Technology, Nanjing University, Nanjing, China
[4]State Key Laboratory of Space-Ground Integrated Information Technology, Beijing, China

Abstract

Automatically generating natural, diverse, and rhythmic human dance movements driven by music is vital for virtual reality and film industries. However, generating dance that naturally follows music remains a challenge, as existing methods lack proper beat alignment and exhibit unnatural motion dynamics. In this paper, we propose Danceba, a novel framework that leverages gating mechanism to enhance rhythm-aware feature representation for music-driven dance generation, which achieves highly aligned dance poses with enhanced rhythmic sensitivity. Specifically, we introduce Phase-Based Rhythm Extraction (PRE) to precisely extract rhythmic information from musical phase data, capitalizing on the intrinsic periodicity and temporal structures of music. Additionally, we propose Temporal-Gated Causal Attention (TGCA) to focus on global rhythmic features, ensuring that dance movements closely follow the musical rhythm. We also introduce Parallel Mamba Motion Modeling (PMMM) architecture to separately model upper and lower body motions along with musical features, thereby improving the naturalness and diversity of generated dance movements. Extensive experiments confirm that Danceba outperforms state-of-the-art methods, achieving improved rhythmic alignment and motion diversity. Project page: https://danceba.github.io/.

1. Introduction

Music-driven dance generation aims to automatically synthesize realistic and expressive 3D dance movements synchronized precisely with a given musical sequence, enabling natural animation of 3D character models. Due to its broad applications in gaming, virtual reality, film production, and interactive entertainment [14, 34, 38, 43], this task has attracted considerable attention. However, generating dances that accurately reflect musical rhythm while preserving natural and diverse movement dynamics remains a significant challenge, as it requires effectively capturing and utilizing complex temporal relationships between music and motion.

Recently, generative models have demonstrated remarkable success across various tasks, such as text and image generation [9, 29]. Similarly, these generative approaches have also been applied to music-driven dance generation tasks [11, 19, 20, 27, 30, 31, 33, 36, 42, 44]. Among these, Bailando [30] introduces a cross-conditional causal attention mechanism to align musical cues with corresponding dance poses, significantly improving the quality of generated dance movements. However, its attention mechanism primarily captures local correspondence between music and poses, failing to sufficiently utilize global rhythmic characteristics within music, thus limiting the expressiveness and rhythmic coherence of the generated dances.

Follow-up studies [11, 31, 36, 44] have extended Bailando's design to further improve dance generation performance. For instance, Enhancing-Bailando [11] employs a pretrained audio encoder (MERT) [21] to obtain richer musical feature representations. However, this encoder primarily captures semantic information from general audio data rather than explicitly modeling rhythmic characteristics, limiting its effectiveness in precise beat-dance synchronization. Moreover, Bailando and related methods typically concatenate three distinct input modalities, i.e., upper-body pose, lower-body pose, and musical features, and rely heavily on cross-conditional causal attention to align these inputs. Nevertheless, this attention mechanism primarily aligns pose and music at a local feature level, but struggles

*Corresponding author: j.guan@hrbeu.edu.cn

to capture continuous motion dynamics, limiting the naturalness and diversity of generated dance movements.

To further tackle the rhythm alignment issue, Beat-It [42] explicitly introduces beat synchronization via a beat distance predictor. However, such an explicit synchronization approach imposes rigid control constraints and strong regularization during generation, significantly limiting the diversity and expressiveness of generated dances [27]. Thus, a more balanced approach capable of achieving precise rhythmic alignment without sacrificing motion naturalness and diversity is highly desirable.

To this end, we propose a novel framework equipped with a gating-enhanced rhythm-aware feature representation for music-driven dance generation, i.e., Danceba, which achieves highly aligned dance poses with enhanced rhythmic sensitivity. Specifically, Danceba introduces Phase-Based Rhythm Extraction (PRE), a rhythm-aware feature extraction module that explicitly captures rhythmic characteristics from musical phase information. Leveraging the temporal phase information intrinsic to rhythm perception, PRE effectively separates rhythmic from semantic music features, significantly improving the rhythmic alignment between generated dance sequences and musical inputs.

To further enhance beat-dance alignment[1] using the rhythm-aware features extracted by PRE, it is essential to establish a stronger contextual association between musical rhythm and dance movements. To achieve this, we present a Temporal-Gated Causal Attention (TGCA) mechanism, enhanced by a gating strategy to explicitly prioritize and leverage global rhythmic structures within cross-conditional causal attention. By integrating gating mechanism, TGCA effectively strengthens the contextual connection between music rhythm and generated movements, ensuring that generated dance sequences accurately reflect the underlying musical rhythm and exhibit greater coherence and expressive diversity.

Furthermore, to overcome the limitations of traditional attention-based approaches in motion modeling, we introduce Parallel Mamba Motion Modeling (PMMM), inspired by recent advances demonstrating that Mamba architectures achieve state-of-the-art performance in modeling complex human motion sequences [16, 39, 41]. Our parallel Mamba architecture separately models upper and lower-body dance sequences, with input features enhanced by Temporal-Gated Causal Attention (TGCA), which incorporates rhythm-aware representations alongside distinct upper and lower-body pose features. This design effectively captures diverse and nuanced motion characteristics. By leveraging Mamba's superior sequential modeling capability, our method significantly improves the expressiveness, diversity, and temporal coherence of generated dance movements. Finally, an additional TGCA module is applied after PMMM to further reinforce rhythmic alignment, ensuring that the generated dances precisely reflect the musical rhythm.

Extensive experiments show that our method improves FID_k by 48.68%, Div_k by 7.0%, Div_g by 16.3%, and BAS by 12.0% over the second-best results, while maintaining a competitive FID_g of 11.90. This significantly surpasses state-of-the-art methods, producing dance movements that are more natural, diverse, and better synchronized with musical rhythm.

2. Related Work

2.1. Music-Driven Dance Generation

Music-driven dance generation focuses on automatically synthesizing 3D dance movements from musical input, creating temporally coherent pose sequences for animating 3D character models. It has attracted increasing attention in both the computer vision and virtual avatar communities [1, 13] due to its applications in entertainment, gaming, and virtual reality. Early approaches, including [10, 12, 15, 17, 18, 32, 45], investigated diverse methodologies for this task. The study in [17] utilized a standard Transformer [35] with a late-fusion module to predict subsequent dance steps. The FACT framework [18] introduced the AIST++ dataset and employed a cross-modal Transformer with full attention mechanisms for music-to-dance generation. However, these initial methods neglect the rhythm features of music and struggle to effectively integrate musical and motion features, leading to suboptimal generation quality.

Bailando [30] reformulated dance generation as a next-step pose prediction problem using cross-conditional causal attention, achieving high-quality music-driven dance synthesis. Subsequent works [11, 31, 44] extended this framework: Bailando++ [31] enhanced pattern recognition with musical context encoding, while Enhancing-Bailando [11] incorporated a pre-trained MERT audio encoder [21] for improved feature extraction. However, these methods often overlook rhythmic structures or rely on generic audio features. Beat-It [42] introduced explicit beat alignment via a conditional diffusion model [9, 25], ensuring precise synchronization. Yet, its reliance on rigid beat control signals and strong regularization may constrain pose diversity [27]. In contrast, we emphasize rhythm-aware feature representation, achieving superior rhythmic alignment without sacrificing movement diversity.

2.2. Mamba in Human Motion Generation

The integration of Mamba [7] and Mamba2 [2] into human motion generation represents a pivotal advancement, leveraging their efficient sequence modeling capabilities. Within human motion generation, Mamba surpasses con-

[1] Note that, the terms "beat" and "rhythm" are used interchangeably in music processing following DiffDance [27] and Bailando++ [31].

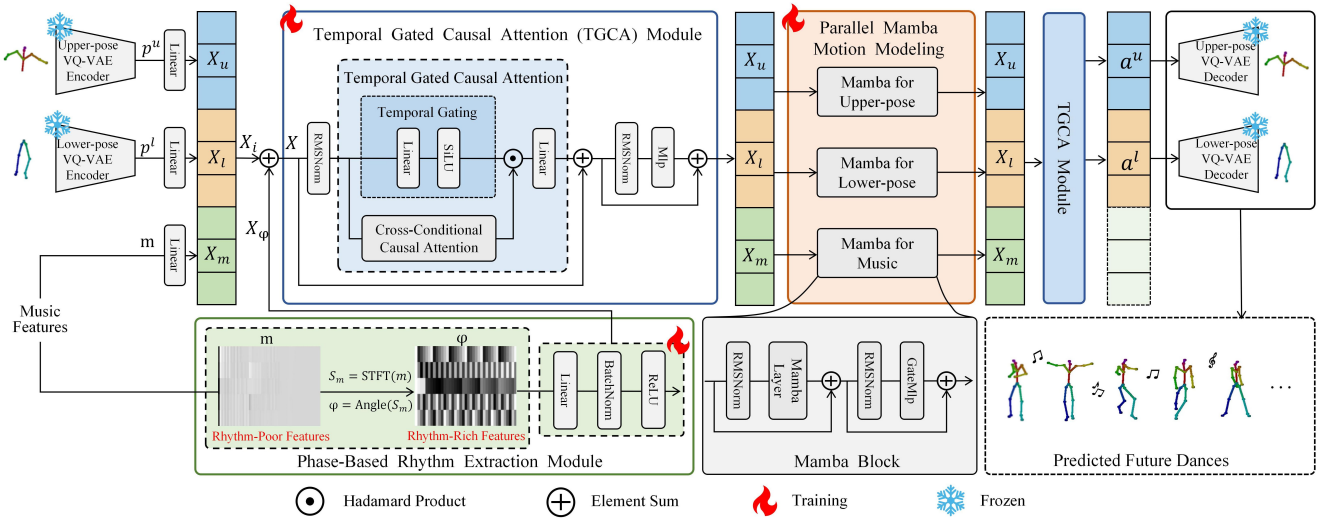

Figure 1. The overall framework of Danceba consists of three core modules: Phase-based Rhythm Extraction (PRE), Temporal-Gated Causal Attention (TGCA), and Parallel Mamba Motion Modeling (PMMM). PRE precisely extracts rhythm-aware features from musical phase information, providing accurate rhythmic signals. These rhythm-enhanced music features are fused with pose embeddings (upper-body and lower-body poses) and processed through the TGCA module, which utilizes gating mechanism to reinforce rhythmic sensitivity and align dance movements accurately with music beats. Parallel Mamba Motion Modeling separately models upper and lower body motion sequences, effectively capturing distinct dance dynamics aligned with rhythm-aware music features, thus significantly enhancing the naturalness, diversity, and temporal coherence of the generated dance movements.

ventional Transformers in performance, as demonstrated by Motion Mamba [41], which highlights its efficacy for extended motion sequence synthesis. Furthermore, InfiniMotion [40] utilizes Mamba's hierarchical and bidirectional temporal processing to generate continuous, smooth motion sequences of indefinite length. Recent approaches including SMCD [28], Light-T2M [39], FTMoMamba [16], and MMD [4] have further incorporated Mamba architectures, advancing the field through enhanced efficiency and high-fidelity motion modeling. Building upon recent advancements, Mamba has achieved state-of-the-art performance in human motion modeling [16, 39, 41], motivating our selection of Mamba over self-attention mechanisms for information exchange within dance sequences, which significantly enhances the diversity of generated dance motions.

3. Proposed Method

This section presents our proposed Danceba framework for music-driven 3D dance generation. By leveraging gating-enhanced rhythm-aware features, Danceba ensures precise alignment between dance poses and music. The overall framework is illustrated in Figure 1, and its components are detailed as follows: Section 3.1 outlines the preliminary framework and key notations. Specifically, we introduce phase-based rhythm feature extraction (Section 3.2), global beat-dance alignment via temporal-gated causal attention (Section 3.3), and Mamba-based parallel motion modeling (Section 3.4).

3.1. Preliminary

Given a dance sequence $\mathbf{P} \in \mathbb{R}^{T\times(J\times 3)}$, where T is the time length and J is joint amount. Using two pretrained Pose VQ-VAEs [30] $\mathcal{F}^u_{VAE}$ and $\mathcal{F}^l_{VAE}$ with codebook $\mathcal{Z}^u$ and $\mathcal{Z}^l$ for the upper and lower half bodies, it separately encodes upper and lower body movements into compositional code pairs $\mathbf{p} = [\mathbf{p}^u, \mathbf{p}^l]$:

$$\mathbf{p}^u = \mathcal{F}^u_{VAE}(\mathbf{P}), \mathbf{p}^l = \mathcal{F}^l_{VAE}(\mathbf{P}), \tag{1}$$

where $\mathbf{p}^u \in (\mathcal{Z}^u)^{T'}$, $\mathbf{p}^l \in (\mathcal{Z}^l)^{T'}$. Here $T' = T/\lambda$, λ is the temporal down-sampling rate. The input music feature $\mathbf{m} \in \mathbb{R}^{T'\times D_m}$, where D_m is the channel dimension of the input music features. Then we embed music features $\mathbf{m}$, upper p^u and lower p^l pose codes to learnable features with three separate linear layers: $\mathbf{X}_\mathrm{m} = \mathrm{Linear}(\mathbf{m}), \mathbf{X}_\mathrm{u} = \mathrm{Linear}(\mathbf{p}^u)$ and $\mathbf{X}_\mathrm{l} = \mathrm{Linear}(\mathbf{p}^l)$, where $\mathbf{X}_\mathrm{m} \in \mathbb{R}^{T'\times D}$, $\mathbf{X}_\mathrm{u} \in \mathbb{R}^{T'\times D}$, $\mathbf{X}_\mathrm{l} \in \mathbb{R}^{T'\times D}$ and D is the channel dimension of the features.

The dance generation task is reframed as selecting the most probable future pose code from the codebook $\mathcal{Z}$, conditioned on the music and prior movements. Since the upper and lower bodies are modeled separately, maintaining coherence and avoiding asynchrony (e.g., opposing directions) requires cross-conditioned prediction, leveraging mutual information between existing movements:

$$\hat{p}^u_t = \arg\max_k \mathbb{P}\left(\mathbf{z}^u_k \mid \mathbf{m}_{1\ldots t}, p^u_{0\ldots t-1}, p^l_{0\ldots t-1}\right), \tag{2}$$

$$\hat{p}^l_t = \arg\max_k \mathbb{P}\left(\mathbf{z}^l_k \mid \mathbf{m}_{1\ldots t}, p^u_{0\ldots t-1}, p^l_{0\ldots t-1}\right). \tag{3}$$

At each time step t, MotionGPT [30] estimates the probabilities of pose codes $z_i \in \mathbb{Z}$ and selects the one with the most probable pose codes as the predicted upper pose $\hat{p}_t^u$ and predicted lower pose $\hat{p}_t^l$.

Specifically, the input feature $\mathbf{X}_i \in \mathbb{R}^{(3\times T')\times D}$ are concatenated by upper body motion feature, lower body motion feature and music feature as follows: $\mathbf{X}_i = \mathrm{Concat}(\mathbf{X}_\mathrm{m}, \mathbf{X}_\mathrm{u}, \mathbf{X}_\mathrm{l})$. Subsequently, $\mathbf{X}_i$ is passed through successive Autoregressive Transformer layers $\mathcal{F}_{AR}$, followed by a linear transformation and a softmax layer, generates the probability distributions for predicting body pose collectively:

$$\mathbf{a}^h = \mathrm{Softmax}(\mathrm{Linear}(\mathcal{F}_{AR}(\mathbf{X}_i)), \tag{4}$$

where $\mathbf{a}^h = [\mathbf{a}^u, \mathbf{a}^l]$, $\mathbf{a}^u$ denotes that of upper body and $\mathbf{a}^l$ denotes that of lower body.

Finally, the model is optimized via supervised training with the cross-entropy loss on action probability a_t^h at each time step t and for both upper body and lower body:

$$\mathcal{L}_{CE} = \frac{1}{T'} \sum_{t=0}^{T'-1} \sum_{h=u,l} \mathrm{CrossEntropy}\left(\mathrm{a}_t^h, p_{t+1}^h\right). \tag{5}$$

In inference, the MotionGPT predicts pose codes according to the initial pose code and the entire music, followed by Pose VQ-VAE decoders generating a new dance.

3.2. Phase-Based Rhythm Feature Extraction

To enhance rhythmic feature extraction in music-driven dance generation, we propose Phase-Based Rhythm Extraction (PRE), which decouples rhythm from musical semantics by leveraging the inherent periodicity and temporal positioning of phase representations. The architecture of PRE is shown in the green part of Figure 1. Unlike existing methods that entangle rhythm with melody and harmony, PRE enables independent and precise rhythm modeling as shown in Figure 2, forming a solid foundation for gating-enhanced rhythm-aware mechanism.

Specifically, we adopt the Short-Time Fourier Transform (STFT) instead of the conventional Fourier Transform (FT), since the non-stationary nature of music and the sliding-window approach of STFT effectively captures sudden tempo shifts and complex rhythmic patterns [37]. Specifically, we first perform STFT on the input music feature to extract phase angles:

$$\begin{aligned} \mathbf{S}_m &= \mathrm{STFT}\left(\mathbf{m}\right), \\ \boldsymbol{\varphi} &= \mathrm{Angle}\left(\mathbf{S}_m\right), \end{aligned} \tag{6}$$

where $\mathbf{S}_m$ represents the complex spectrogram obtained via STFT, $\mathrm{Angle}(\cdot)$ computes the phase angle of a complex-valued input and $\boldsymbol{\varphi}'$ denotes the phase angles extracted from $\mathbf{S}_m$. By taking the angle of the STFT output, we obtain the

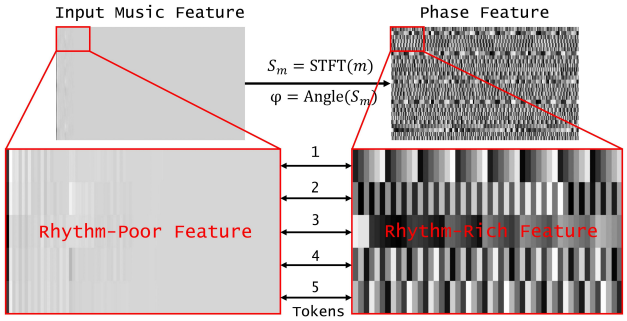

Figure 2. Visualizing the input music features and phase features. We notice that only a small portion of the input music features contains meaningful musical information, i.e., rhythm-poor features. And after a short-time Fourier transform (STFT), each corresponding token has a rich periodic Phase Feature, i.e., rhythm-rich features.

music phase data $\boldsymbol{\varphi}' \in \mathbb{C}^{T_\varphi \times D_\varphi}$, where T_φ represents the number of time frames obtained after the STFT transformation, and D_φ is the channel dimension of the $\boldsymbol{\varphi}'$.

The extracted phase angles $\boldsymbol{\varphi}'$ inherently reflect temporal shifts and periodic structures of musical rhythms: low-frequency phase variations correspond to broader rhythmic patterns, while high-frequency variations capture finer rhythmic details. We also apply a center-cropping strategy, selectively focusing on the most informative phase regions:

$$\boldsymbol{\varphi} = \mathrm{CenterCrop}(\boldsymbol{\varphi}'), \tag{7}$$

where $\boldsymbol{\varphi} \in \mathbb{C}^{T'\times D_\varphi}$ has the same temporal length T', addressing the potential temporal misalignment and preserving critical rhythmic cues.

Subsequently, the phase features through a linear transformation, enhancing phase features representing musical rhythms. The process is shown in the following equation:

$$\mathbf{X}_\varphi = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Linear}(\varphi)\right)\right), \tag{8}$$

where the embedded phase features $\mathbf{X}_\varphi \in \mathbb{C}^{T'\times D}$ encode rich rhythmic information, providing a clearer and more structured representation of rhythmic and musical elements. Finally, we fuse the input feature $\mathbf{X}_i$ with the rhythmic features $\mathbf{X}_\varphi$ by concatenating them: $\mathbf{X}_\gamma = \mathrm{Concat}(\mathbf{X}_\varphi, \mathbf{X}_\varphi, \mathbf{X}_\varphi)$. Then, we sum the rhythm-enhanced features with the original music, upper body, and lower body motion features: $\mathbf{X} = \mathbf{X}_f + \mathbf{X}_\gamma$, where $\mathbf{X}$ represents the rhythmically enhanced features.

3.3. Global Beat Attention via Temporal Gating

Leveraging these rhythm-rich features, we next investigate how to integrate them into the alignment process between music and dance. However, prior methods [11, 30, 31, 36, 44] relying on cross-conditional causal attention mechanism (denoted as $\mathrm{C}^3\mathrm{Attention}$) exhibit a critical misalignment issue. As shown in Figure 3 (a), we visualize the

cross-conditional causal attention mechanism and observe that the next predicted token lacks direct control signals from historical tokens, failing to establish a stable global beat attention. This misalignment leads to error accumulation, which we identify as a key factor behind the unnatural dance generation in prior work.

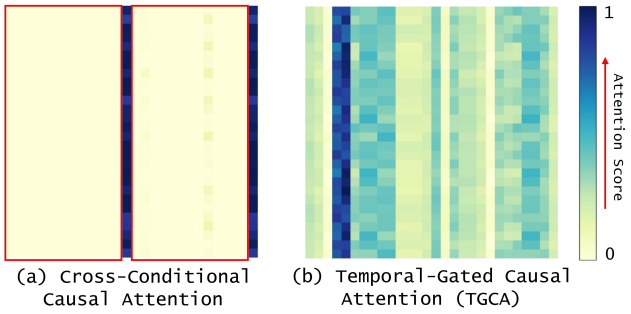

Figure 3. Heat map of attention. The two displayed images show the results learned by Cross-Conditional Causal Attention (a) and Temporal-Gated Causal Attention (b) under the same piece of music. The brighter the color, the higher the attention; the darker the color, the lower the attention. TGCA the next token has a clear control signal, preventing overly random dance movements and achieving enhanced rhythm-aware feature representation.

Building on this, we propose Temporal Gated Causal Attention (TGCA) to enhance global rhythm feature attention, improving music-dance alignment and ensuring the next token is guided by explicit historical information, preventing random action generation. As illustrated in the blue section of Figure 1, the rhythmically enhanced features $\mathbf{X}$ are passed through TGCA module, which can be expressed as the following:

$$\text{Gating}(\mathbf{X}) = \text{SiLU}(\text{Linear}(\mathbf{X})). \tag{9}$$

The final output is obtained by the element-wise multiplication of cross-conditional causal attention $\text{C}^3\text{Attention}$ and Gating, as shown in the following equation:

$$\begin{aligned} \mathbf{X}_{attn} &= \text{TGCA}(\mathbf{X}) \\ &= \text{C}^3\text{Attention}(\mathbf{X}) \odot \text{Gating}(\mathbf{X}). \end{aligned} \tag{10}$$

As shown in Figure 3 (b), our proposed gated causal attention mechanism allows historical tokens to better predict the next token, implying an enhanced influence of global rhythmic patterns on dance generation.

3.4. Mamba-Based Parallel Motion Modeling

After TGCA enhances dance movements in $\mathbf{X}_{attn}$ with global rhythmic perception, we looking into dance modeling for generation. Dance generation faces two key challenges: (1) Dance is a complex spatial-temporal sequence encoding trajectory, posture, and velocity over time, requiring a model capable of capturing both local transitions and global rhythmic variations. (2) The upper and lower body exhibit distinct movement patterns and speeds, necessitating separate processing to maintain coherence and diversity. To address these issues, we propose a parallel Mamba motion modeling architecture that separately models music, upper-body, and lower-body movements while leveraging state space modeling to ensure fluid and diverse motion generation.

Specifically, to ensure each action is primarily influenced by relevant preceding movements, we introduce a GateMlp in the Mamba block, as shown in Figure 1. This mechanism selectively retains essential motion features while filtering out less relevant information, thereby enhancing the coherence and expressiveness of the generated dance sequences. The process is formulated as follows:

$$\mathbf{X}'_{mb} = \text{Mamba}(\text{RMSNorm}(\mathbf{X}_{attn})) + \mathbf{X}_{attn}, \tag{11}$$

$$\mathbf{X}_{mb} = \text{GateMlp}(\text{RMSNorm}(\mathbf{X}'_{mb})) + \mathbf{X}'_{mb}. \tag{12}$$

The gating mechanism dynamically regulates the frequency, speed, and amplitude of movements in response to musical beats, ensuring that dance sequences remain synchronized with the music. Finally, three parallel Mamba architectures model upper- and lower-body dance sequences separately, therefore ensure the generated dance movements are expressive, diverse, and temporally coherent.

4. Experiments

4.1. Experimental Setup

Dataset: We follow prior works [27, 30, 36] and perform all experiments on the AIST++ dataset [18], the most widely used benchmark for music-driven dance generation that avoids observable biases (e.g., offensive postures). AIST++ consists of 991 high-quality 3D pose sequences recorded at 60 FPS in skinned multi-person linear (SMPL) format [22], with 951 sequences designated for training and 40 reserved for evaluation. Following [27, 30, 36], we generate 40 pieces of dance sequences on the test set of AIST++ dataset, and sample the generated dance sequence with length of 20 seconds for performance evaluation.

Evaluation Metrics: We follow previous studies [18, 30] to quantitatively evaluate the generated samples in three key aspects: quality, diversity, and beat-dance alignment. Prior to evaluation, it is necessary to extract kinetic features [26] and geometric features [24] from both the ground truth and the generated samples via a specific toolbox [6]. For dance quality, we compute two types of the Fréchet Inception Distance (FID) [8], including FID_k based on kinetic features and FID_g based on geometric features. Lower FID_k indicates that the generated dance has more similar kinetic features to the original dance, meaning the generated dances are more natural and smooth. Lower FID_g indicates that

Table 1. Comparison with state-of-the-art methods on the AIST++ dataset. Underlining indicates the best performance among existing methods. Blue indicates results that surpass the best existing method. ↓ indicates that lower values are better, while ↑ indicates that higher values are better. '$*$' represents the dances generated by "Li *et al.*" that exhibit significant jitteriness, resulting in extremely high velocity variation, as also noted in [18]. '†' indicates the reproduced results obtained using Bailando++ official code and published checkpoints. '‡' denotes the results that are averaged over five independent training runs.

Method	Venue	Motion Quality		Motion Diversity		Beat Align Score↑
		FID_k ↓	FID_g ↓	Div_k ↑	Div_g ↑	
Ground Truth	–	17.10	10.60	8.19	7.45	0.2374
Li *et al.* [17]	Arxiv 2020	86.43	43.46	*6.85	3.32	0.1607
DanceRevolution [12]	ICLR 2021	73.42	25.92	3.52	4.87	0.1950
DanceNet [45]	TOMM 2022	69.18	25.49	2.86	2.85	0.1430
FACT [18]	ICCV 2021	35.35	22.11	5.94	6.18	0.2209
Bailando [30]	CVPR 2022	28.16	9.62	7.83	6.34	0.2332
EDGE [33]	CVPR 2023	42.16	22.12	3.96	4.61	0.2334
DiffDance [27]	ACMMM 2023	24.09	20.68	6.02	2.89	0.2418
†Bailando++ [31]	TPAMI 2023	22.74	11.58	7.94	6.34	0.2263
E3D2 [36]	AAAI 2024	26.25	8.94	7.96	6.49	0.2232
Lodge [20]	CVPR 2024	37.09	18.79	5.58	4.85	0.2423
Ours (Best)		**11.67**(-11.07)	11.90	**8.52**(+0.58)	**7.55**(+1.21)	**0.2714**(+2.91%)
‡Ours (Average)		**15.48**(-7.26)	13.57	7.85	**7.49**(+1.15)	**0.2779**(+3.56%)

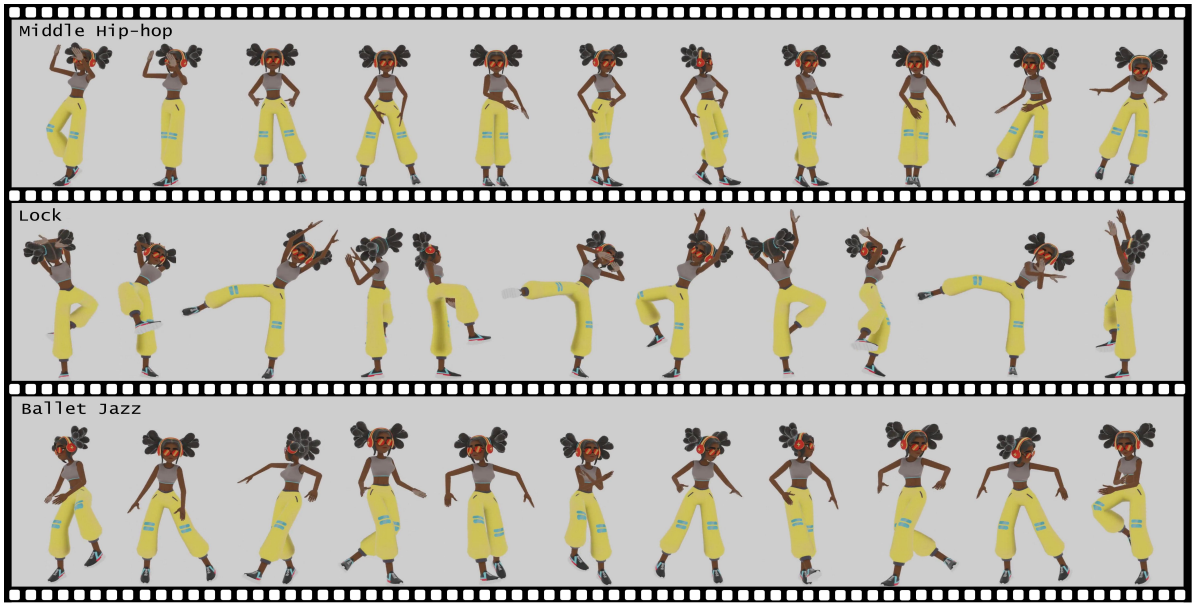

Figure 4. Dance examples generated by our proposed method on various types of music. The demo videos can be found in the supplementary material. The character is sourced from Mixamo [23].

the generated dance has more similar geometric features, reflecting the quality of choreography. For dance diversity, we employ metrics Div_k and Div_g [18], which denote the distance of kinetic and distance of geometric features, respectively. Higher Div_k and Div_g signify the greater amplitude and diversity of motion, respectively. Finally, to evaluate beat-dance alignment, we use the Beat Align Score (BAS) [18], which quantifies the average distance between a music beat and its nearest dance beat. Higher BAS indicates greater degree of beat-dance alignment.

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

To validate the superiority of our method, we conduct comparative experiments with the state-of-the-art methods [12, 17, 18, 20, 27, 30, 31, 33, 36, 45], as well as the ground truth. Following [27, 30, 36], we generate 40 pieces of dance sequences on the AIST++ test set, and sample the generated dance sequence with the length of 20 seconds to compute the metrics mentioned above. Among the compared methods, "Li *et al.*" [17],

DanceNet [45] and FACT [18] use the results from the AIST++ benchmark [18], while the results of DanceRevolution [12] are derived from its reproduction as presented in Bailando [30]. EDGE [33] utilizes the reproduced results from the Lodge [20]. The Bailando, DiffDance [27] and Lodge methods use the experimental results from their respective publications.

The comparative results are presented in Table 1. Our method achieves state-of-the-art performance, significantly outperforming existing methods in terms of FID_k, Div_g, Div_k, and Beat Align Score (BAS), achieving the best results. Additionally, our method ranks third on the FID_g metric. The generated dance examples of our proposed Danceba on various types of music are shown in Figure 4, and the demo videos are provided in the supplementary files.

Motion Quality Metrics: As illustrated in Table 1, our method outperforms existing methods in overall quality performance. To be specific, our method improves on FID_k by 48.68% (11.67 vs. 22.74), while exhibits competitive edge of FID_g. Further analysis of the employed quality metrics reveals distinct roles for the two metrics types. The metric FID_k, defined based on velocity and energy, primarily captures the kinetic characteristics of the generated dance, while FID_g, derived from multiple manually designed motion templates, emphasizes the geometric structural properties of the dance. Thus, our method demonstrates superior performance in modeling the kinetic aspects of sequential motions, as evidenced by its exceptional FID_k scores. Although it exhibits a modest gap in FID_g (11.90 vs. 11.58) compared to the state-of-the-art Bailando++ [31], it achieves a 19.0% higher Div_g (7.55 vs. 6.34), indicating improved motion diversity with comparable generative quality. Despite this minor trade-off, our method outperforms existing approaches overall, achieving a robust balance of efficiency and quality across a broader range of metrics.

Motion Diversity Metrics: Regarding diversity evaluation, Danceba achieves the best diversity performance, which improves Div_k by 7.0% (8.52 vs. 7.96) and even slightly outperforms the Ground Truth (8.52 vs. 8.19). Meanwhile, our method improves Div_g by 16.3% (7.55 vs. 6.49) and also slightly surpasses the Ground Truth (7.55 vs. 7.45), as illustrate in Table 1. For a visual comparison, we also present a comparison with the state-of-the-art method Bailando, as illustrated in Figure 5. It can be observed that Bailando exhibits relatively small movement amplitudes, with some dance movements missing and the actions appearing stiff. In contrast, our method shows larger movement amplitudes and movement diversity.

Beat-dance Alignment Metrics: Danceba achieves the best performance in beat-dance alignment, as quantified by the BAS metric. Specifically, it improves BAS by 12.0% (0.2714 vs. 0.2423) compared to the previous best result and surpasses the ground truth by 14.3% (0.2714 vs. 0.2374), as shown in Table 1. Figure 6 further visualizes the alignment between music beats and motion beats of the generated dance compared to Bailando++ [31]. The results show that Danceba achieves a significantly smaller beat distance, demonstrating its ability to generate dance sequences that more accurately follow musical rhythms.

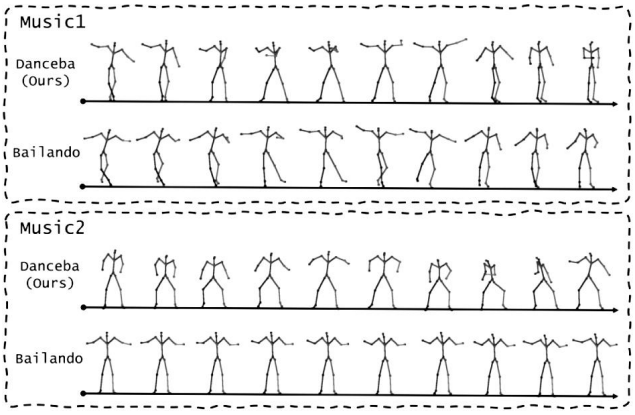

Figure 5. Comparison with the state-of-the-art method Bailando [30]. Visual comparisons with Bailando can be found in the supplementary video.

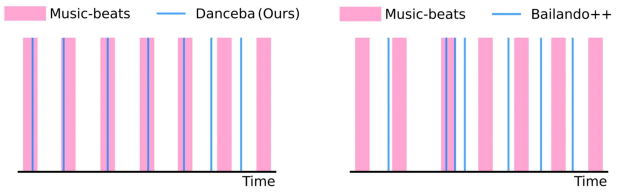

Figure 6. Beats alignment visualization, where the horizontal axis shows frame indices of beat events. Comparing Danceba with Bailando++ [31], we can find that the distance between motion beats and music beats generated by our method is smaller. This indicating that Danceba performs better in terms of rhythmic alignment.

4.3. Ablation Studies

We conduct ablation studies to evaluate the contributions of the key modules (i.e., PRE, TGCA, and PMMM) in Danceba, where "w/o" denotes the removal of a specific module. "w/o PMMM" refers to the setting where PMMM is replaced with TGCA layers to ensure fairness. The results are presented in Table 2. Additionally, we assess our method using a single Mamba structure instead of the parallel Mamba structure for motion modeling, referred to as Danceba-Single, with results shown in Table 3.

Note that, all ablation results are averaged over five independent training runs to ensure the credibility and robustness of our results. The motion quality score (FID) is computed as the average of FID_k and FID_g, while the motion diversity score (Div) is the average of Div_k and Div_g.

Phase-Based Rhythm Extraction: To validate the effec-

Table 2. Ablation study of three key modules (i.e., PRE, TGCA, and PMMM), and "w/o" indicates Danceba without specified module. Visual comparisons with Bailando can be found in the supplementary video.

METHOD	FID ↓	DIV ↑	BAS ↑
GROUND TRUTH	13.85	7.82	0.2374
W/O TGCA	33.22	6.74	–
W/O PMMM	25.24	5.78	–
W/O PRE	20.62	6.53	0.2474
DANCEBA	14.53	7.67	0.2779

tiveness of PRE, we replace the rhythm feature extraction module with a learnable linear layer as used in previous works [11, 30, 31, 36, 44]. Consequently, the motion quality (i.e., FID), motion diversity (i.e., Div), and beat alignment score (i.e., BAS) decrease by 6.09 (41.91%), 1.14 (14.86%), and 0.0305 (10.96%), respectively, as illustrated in Table 2. Specifically, this performance drop highlights the contribution of PPE, which effectively captures the critical rhythm features of music, a crucial aspect that previous methods have largely overlooked.

Temporal-Gated Causal Attention: As shown in Table 2, we validate the effectiveness of the proposed TGCA by training a Danceba variant that replaces TGCA with conventional causal cross-conditional attention. The results demonstrate that the full Danceba model outperforms this variant in both FID and Div, highlighting TGCA's ability to effectively leverage global rhythmic features extracted by PRE. This significantly enhances the contextual association between music and dance movements, resulting in dance sequences that not only accurately reflect musical rhythm but also exhibit greater coherence and expressive diversity.

Parallel Mamba Motion Modeling: To evaluate the effectiveness of Parallel Mamba Motion Modeling, we train a model without this module. As shown in Table 2, Danceba outperforms its variant without using PMMM in terms of both FID and Div, demonstrating the importance of parallel Mamba modeling. These results validate Mamba's superior sequential modeling capability, which is particularly well-suited for dance generation as it requires both fine-grained inter-modal interactions and long-sequence modeling.

To further assess the advantage of using a parallel Mamba architecture instead of a single Mamba architecture for motion modeling, we compare Danceba with Danceba-Single. Unlike Danceba-Single, which treats the body as a unified entity and disregards the distinction between upper and lower-body movements, PMMM explicitly models their distinct motion dynamics. Additionally, PMMM integrates rhythm-aware representation alongside these distinct pose features, enhancing beat-dance alignment with the musical rhythm. As shown in Table 3, Danceba significantly improves expressiveness and temporal coherence, with FID_k and FID_g scores improving by 67.02 and 60.40, respectively.

Table 3. Ablation study on parallel Mamba motion modeling, where Danceba-Single denotes our method using single Mamba architecture.

METHOD	FID_k ↓	FID_g ↑
GROUND TRUTH	17.10	10.60
DANCEBA-SINGLE	82.50	83.97
DANCEBA	15.48(**-67.02**)	13.57(**-60.40**)

5. Limitations

Although Danceba achieves superior dance generation quality compared to state-of-the-art methods, it has two key limitations: (1) Music feature encoding: Danceba uses a simple linear layer instead of a pre-trained audio model (e.g., Jukebox [3], MERT [21], or CLAP [5]). While computationally efficient, this design may fail to capture nuanced musical structures, affecting rhythm-motion synchronization. Future work could explore pre-trained music encoders to enhance feature extraction. (2) 3D pose quantization: Danceba adopts the Pose VQ-VAE framework from Bailando [30] for fair comparison with prior SOTA methods [30, 31, 36]. However, it has a limited encoding space, which may limit motion diversity. Such a trade-off between quantization efficiency and expressiveness could hinder the preservation of fine-grained motion details. Exploring hierarchical quantization or adaptive codebook learning could further improve motion quality.

6. Conclusion

This paper has presented a novel framework Danceba for music-driven dance generation that enhances rhythm-aware feature representation and motion modeling. By introducing phase-based rhythm extraction and temporal-gated causal attention, our approach ensures precise beat-dance synchronization and effectively integrates rhythmic structures into dance movements. Additionally, parallel Mamba motion modeling separately captures the distinct motion dynamics of the upper and lower body while incorporating rhythm-aware representations, improving the naturalness and diversity of the generated dance. Extensive experiments demonstrate that Danceba significantly outperforms state-of-the-art methods, generating dance movements that are more rhythmic, natural, and diverse while maintaining strong alignment with musical beats. Future work could integrate pre-trained audio encoders for richer music representation, explore advanced quantization to enhance motion expressiveness and diversity, or extend to style-conditioned motion synthesis for broader applicability.

Acknowledgments

This work was partly supported by the Fundamental Research Funds for the Central Universities (Grant No. 3072025YY0601), and the Beijing Nova Program (Grant No. 20230484261).

References

[1] Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Trans. Graph. (TOG)*, 2023. 2

[2] Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proc. Int. Conf. Mach. Learn. (ICML)*, 2024. 2

[3] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020. 8

[4] Kun Dong, Jian Xue, Zehai Niu, Xing Lan, Ke Lu, Qingyuan Liu, and Xiaoyu Qin. Realistic full-body motion generation from sparse tracking with state space model. In *Proc. ACM Int. Conf. Multimedia (ACMMM)*, 2024. 3

[5] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP learning audio concepts from natural language supervision. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2023. 8

[6] Deepak Gopinath and Jungdam Won. Fairmotion - tools to load, process and visualize motion capture data. Github, 2020. 5

[7] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2

[8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017. 5

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020. 1, 2

[10] Daniel Holden, Jun Saito, and Taku Komura. A deep learning framework for character motion synthesis and editing. *ACM Trans. Graph. (TOG)*, 2016. 2

[11] Qiaochu Huang, Xu He, Boshi Tang, Haolin Zhuang, Liyang Chen, Shuochen Gao, Zhiyong Wu, Haozhi Huang, and Helen Meng. Enhancing expressiveness in dance generation via integrating frequency and music style information. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2024. 1, 2, 4, 8

[12] Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Daxin Jiang. Dance revolution: Long-term dance generation with music via curriculum learning. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021. 2, 6, 7

[13] Jinwoo Kim, Heeseok Oh, Seongjean Kim, Hoseok Tong, and Sanghoon Lee. A brand new dance partner: Music-conditioned pluralistic dancing controlled by multiple dance genres. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022. 2

[14] Matthew Kyan, Guoyu Sun, Haiyan Li, Ling Zhong, Paisarn Muneesawang, Nan Dong, Bruce Elder, and Ling Guan. An approach to ballet dance training through MS Kinect and visualization in a CAVE virtual reality environment. *ACM Trans. Intell. Syst. Technol. (TIST)*, 2015. 1

[15] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2019. 2

[16] Chengjian Li, Xiangbo Shu, Qiongjie Cui, Yazhou Yao, and Jinhui Tang. FTMoMamba: Motion generation with frequency and text state space models. *arXiv preprint arXiv:2411.17532*, 2024. 2, 3

[17] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. Learning to generate diverse dance motions with Transformer. *arXiv preprint arXiv:2008.08171*, 2020. 2, 6

[18] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3D dance generation with AIST++. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021. 2, 5, 6, 7

[19] Ronghui Li, Junfan Zhao, Yachao Zhang, Mingyang Su, Zeping Ren, Han Zhang, Yansong Tang, and Xiu Li. Finedance: A fine-grained choreography dataset for 3D full body dance generation. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023. 1

[20] Ronghui Li, YuXiang Zhang, Yachao Zhang, Hongwen Zhang, Jie Guo, Yan Zhang, Yebin Liu, and Xiu Li. Lodge: A coarse to fine diffusion network for long dance generation guided by the characteristic dance primitives. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024. 1, 6, 7

[21] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, et al. MERT: Acoustic music understanding model with large-scale self-supervised training. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024. 1, 2, 8

[22] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM Trans. Graph. (TOG)*, 2015. 5

[23] Mixamo. https://www.mixamo.com/. 6

[24] Meinard Müller, Tido Röder, and Michael Clausen. Efficient content-based retrieval of motion capture data. In *Proc. ACM Conf. SIGGRAPH*, 2005. 5

[25] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[26] Kensuke Onuma, Christos Faloutsos, and Jessica K. Hodgins. Fmdistance: A fast and effective distance function for motion capture data. In *Eurographics*, 2008. 5

[27] Qiaosong Qi, Le Zhuo, Aixi Zhang, Yue Liao, Fei Fang, Si Liu, and Shuicheng Yan. DiffDance: Cascaded human motion diffusion model for dance generation. In *Proc. ACM Int. Conf. Multimedia (ACMMM)*, 2023. 1, 2, 5, 6, 7

[28] Ziyun Qian, Zeyu Xiao, Zhenyi Wu, Dingkang Yang, Mingcheng Li, Shunli Wang, Shuaibing Wang, Dongliang Kou, and Lihua Zhang. SMCD: High realism motion style transfer via Mamba-based diffusion. *arXiv preprint arXiv:2405.02844*, 2024. 3

[29] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 2019. 1

[30] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3D dance generation by actor-critic GPT with choreographic memory. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[31] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando++: 3D dance GPT with choreographic memory. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 2023. 1, 2, 4, 6, 7, 8

[32] Guofei Sun, Yongkang Wong, Zhiyong Cheng, Mohan S Kankanhalli, Weidong Geng, and Xiangdong Li. Deepdance: music-to-dance motion choreography with adversarial learning. *IEEE Trans. Multimedia (TMM)*, 2020. 2

[33] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2023. 1, 6, 7

[34] Shuhei Tsuchida. Dance information processing: Computational approaches for assisting dance composition. *New Gener. Comput.*, 2024. 1

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017. 2

[36] Zilin Wang, Haolin Zhuang, Lu Li, Yinmin Zhang, Junjie Zhong, Jun Chen, Yu Yang, Boshi Tang, and Zhiyong Wu. Explore 3D dance generation via reward model from automatically-ranked demonstrations. In *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2024. 1, 4, 5, 6, 8

[37] Simon Welker, Julius Richter, and Timo Gerkmann. Speech enhancement with score-based generative models in the complex STFT domain. *arXiv preprint arXiv:2203.17004*, 2022. 4

[38] Shuo Yan, Gangyi Ding, Zheng Guan, Ningxiao Sun, Hongsong Li, and Longfei Zhang. OutsideMe: Augmenting dancer's external self-image by using a mixed reality system. In *Proc. Annu. ACM Conf. Extended Abstr. Hum. Factors Comput. Syst. (CHI)*, 2015. 1

[39] Ling-An Zeng, Guohong Huang, Gaojie Wu, and Wei-Shi Zheng. Light-t2m: A lightweight and fast model for text-to-motion generation. *arXiv preprint arXiv:2412.11193*, 2024. 2, 3

[40] Zeyu Zhang, Akide Liu, Qi Chen, Feng chen, Reid Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. InfiniMotion: Mamba boosts memory in Transformer for arbitrary long motion generation. *arXiv preprint arXiv:2407.10061*, 2024. 3

[41] Zeyu Zhang, Akide Liu, Ian Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. Motion Mamba: Efficient and long sequence motion generation. In *Proc. Springer Eur. Conf. Comput. Vis. (ECCV)*, 2025. 2, 3

[42] Chenxi Zheng, Jing Qin, and Shengfeng He. Beat-It: Beat-synchronized multi-condition 3D dance generation. In *Proc. Springer Eur. Conf. Comput. Vis. (ECCV)*, 2024. 1, 2

[43] Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)*, 2023. 1

[44] Haolin Zhuang, Shun Lei, Long Xiao, Weiqin Li, Liyang Chen, Sicheng Yang, Zhiyong Wu, Shiyin Kang, and Helen Meng. GTN-Bailando: Genre consistent long-term 3D dance generation based on pre-trained genre token network. In *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2023. 1, 2, 4, 8

[45] Wenlin Zhuang, Congyi Wang, Jinxiang Chai, Yangang Wang, Ming Shao, and Siyu Xia. Music2Dance: DanceNet for music-driven dance generation. *ACM Trans. Multimedia Comput. Commun. Appl. (TOMM)*, 2022. 2, 6, 7