Harmonious Music-driven Group Choreography with Trajectory-Controllable Diffusion

Yuqin Dai[1], Wanlu Zhu[1], Ronghui Li[2], Zeping Ren[2], Xiangzheng Zhou[1], Jixuan Ying[2], Jun Li[1*], Jian Yang[1*]

[1]PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
[2]Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
{daiy, wanluzhu, xzhou, junli, csjyang}@njust.edu.cn {lrh22, rzp22, yingjx23}@mails.tsinghua.edu.cn

Abstract

Creating group choreography from music is crucial in cultural entertainment and virtual reality, with a focus on generating harmonious movements. Despite growing interest, recent approaches often struggle with two major challenges: ***multi-dancer collisions*** and ***single-dancer foot sliding***. To address these challenges, we propose a Trajectory-Controllable Diffusion (TCDiff) framework, which leverages non-overlapping trajectories to ensure coherent and aesthetically pleasing dance movements. To mitigate collisions, we introduce a Dance-Trajectory Navigator that generates collision-free trajectories for multiple dancers, utilizing a distance-consistency loss to maintain optimal spacing. Furthermore, to reduce foot sliding, we present a footwork adaptor that adjusts trajectory displacement between frames, supported by a relative forward-kinematic loss to further reinforce the correlation between movements and trajectories. Experiments demonstrate our method's superiority.

Project Page — https://wanluzhu.github.io/TCDiffusion/

Introduction

Dance, one of the most expressive art forms, has a profound impact on cultural, cinematic, and academic domains (Artemyeva and Moshenska 2018; Lee et al. 2021; Yao et al. 2023; Xue et al. 2024). The process of choreography creation has traditionally been labor-intensive, spurring the development of automated learning models for dance generation. Consequently, music-driven choreography, initially focused on solo dancers (Li et al. 2021; Tseng, Castellon, and Liu 2023; Li et al. 2023), has garnered considerable attention. As the demand for more immersive and interactive experiences grows, the need for multi-person choreography has become increasingly prominent (Yao et al. 2023), leading to a greater emphasis on music-driven group choreography that prioritizes both cohesion and diversity in group movements (Schwartz 1998). However, despite initial recognition and exploration (Yao et al. 2023; Le et al. 2023b,a; Siyao et al. 2023; Yang et al. 2024), these approaches continue to face two significant challenges:

*Corresponding authors

Figure 1: Visualizations of two key issues in baseline models: multi-dancer collisions (Le et al. 2023a) (in the black box) and single-dancer foot sliding (Yang et al. 2024) (in the red box). In contrast, our approach eliminates these issues, delivering superior visual aesthetics.

Multi-dancer collision. In many group choreography frameworks (Le et al. 2023b,a; Yao et al. 2023; Tseng, Castellon, and Liu 2023), model inputs are typically constructed by concatenating the movements and coordinates of each dancer. However, this strategy results in a significant imbalance, as movements often encompass over 100 dimensions, whereas coordinates are constrained to just three. In group choreography, dancers' coordinates can vary substantially, yet their movements frequently display notable similarities. For example, in the AIOZ-GDance dataset (Le et al. 2023b), over 80% of the movements are similar. This similarity leads to *dancer ambiguity*, complicating the model's ability to distinguish between individual dancers, and often leading to collisions, as illustrated in Figure 1.

Single-dancer foot sliding. Foot sliding occurs when a dancer's feet appear to unnaturally glide or shift across the ground, despite accurate movement of the rest of the body, as depicted in the middle of Figure 1. This issue often arises from difficulties in accurately modeling the correlation between the global trajectory and the local rotations of other body parts (Yang, Yang, and Wang 2023). In multi-person choreography, the dancer ambiguity issue further complicates the modeling of this correlation, making it even harder to align footwork with displacement.

To address these challenges, we propose a two-stage method, **Trajectory-Controllable Diffusion** (TCDiff), which first predicts dancers' coordinates and subsequently generates their movements accordingly. ***To mitigate multi-dancer collisions,*** we introduce the Dance-Trajectory Navigator (DTN), designed to resolve dancer ambiguity

from representation imbalance by focusing on critical positional coordinates. This approach centers on a distance-consistency loss that regulates spatial distances, effectively preventing collisions. Furthermore, we present a simple yet effective fusion projection plug-in that significantly reduces dancer ambiguity while requiring minimal memory. ***For single-dancer foot sliding***, We introduce a footwork adaptor that derives foot movements by analyzing displacement between consecutive trajectory frames. In addition, we propose a Relative Forward-Kinematic (RFK) loss to strengthen the root-motion relationship by enhancing the connection between the root node and the other joints of a dancer. In summary, our main contributions are:

- We propose a Dance-Trajectory Navigator that can generate distinct dancer trajectories by exploring a distance-consistency loss to avoid dancer collision.
- We introduce a Footwork Adaptor that utilizes trajectory shifts between adjacent frames to generate precise footwork. It incorporates a relative forward-kinematic loss to strengthen the correlation between root node and dance motion, effectively reducing single-dancer foot sliding.
- Leveraging these components, we develop a novel two-stage multi-dancer generation framework, Trajectory-Controllable Diffusion, to produce high-quality dance movements. Experimental results demonstrate the superiority of our approach over existing methods.

Related Work

Music-driven Single-dancer Generation

The single-dancer generation is the most relevant area to group dance generation, yet it remains a significant challenge (Joshi and Chakrabarty 2021). Early approaches based on motion retrieval paradigms (Kovar and Gleicher 2002; Fan, Xu, and Geng 2011; Ofli et al. 2011; Lee, Lee, and Park 2013) often result in deformed actions. Recent advancements leverage large datasets (Lee et al. 2019; Huang et al. 2020; Valle-Pérez et al. 2021; Li et al. 2021, 2023; Han et al. 2023) to synthesize motions using deep learning techniques, including auto-regressive models (Alemi, Françoise, and Pasquier 2017; Yalta et al. 2019; Ahn et al. 2020; Siyao et al. 2022; Xu et al. 2024) and generative models (Kim et al. 2022; Tseng, Castellon, and Liu 2023; Li et al. 2024c,b). In recent years, diffusion-based models (Ho, Jain, and Abbeel 2020; Sohl-Dickstein et al. 2015; Tseng, Castellon, and Liu 2023; Ren, Huang, and Li 2025) have emerged, achieving state-of-the-art performance with high diversity and fidelity. Solo dance generation prioritizes realism, making artifacts like foot sliding unacceptable (Yang, Yang, and Wang 2023). Existing methods address this by imposing physical constraints via constraint losses and foot contact labels (Zhang et al. 2021, 2023; Tseng, Castellon, and Liu 2023; Li et al. 2024c). In addition to generating natural dance movements, enhancing dance action controllability remains an important yet underexplored area. Current methods offer temporal and spatial controls, such as genre (Kim et al. 2022), text (Gong et al. 2023; Li et al. 2024a), and joint control (Tseng, Castellon, and Liu 2023; Raab et al. 2023). However, achieving consistent and plausible group dynamics with single-dancer models is challenging, as multi-person dance requires modeling inter-dancer correlations. Additionally, single-dancer models often face issues with dancer ambiguity.

Music-driven Multi-dancer Generation

Multi-dancer generation is an emerging area currently in its nascent stage. To the best of our knowledge, only a handful of studies (Yao et al. 2023; Le et al. 2023b,a; Yang et al. 2024) have focused on generating scenarios for more than two dancers. Among these, GDanceR (Le et al. 2023b) and GCD (Le et al. 2023a) utilize no special structures to avoid the imbalance issue in motion representation, resulting in a tendency for dancer ambiguity. CoDancers (Yang et al. 2024) splits group motions into single-person motions, preventing the occurrence of dancer ambiguity. However, completely isolating individual features results in incomplete group information, which leads to disharmony generation results. Therefore, a more optimal feature separation strategy is necessary. In this work, we introduce TCDiff, a method that first generates dancers' trajectories and then produces logical movements. Together with our proposed effective plugin, Fusion Projection, this approach significantly reduces dancer ambiguity.

Multi-agent Trajectory Prediction

To accurately model dancer trajectories, we leverage insights from the trajectory prediction field, which closely aligns with our task by focusing on understanding agent movement. Past methods heavily rely on hand-crafted rules for describing motions and interactions, including the Gaussian Process (Kalman 1960), and Markov Models (Kim, Lee, and Essa 2011). However, these methods struggle with complex real-world scenarios. Recent deep learning approaches for temporal modeling are LSTM (Alahi et al. 2016a) and their variants. To model complex interactions attention-based methods and graph-based approaches have been developed, such as SGCN (Shi et al. 2021) and STAR (Yu et al. 2020). Predicting trajectories is challenging due to the multi-modality issue (Gu et al. 2022a; Gupta et al. 2018; Alahi et al. 2016b), where the same input can lead to different outcomes. Consequently, current methods aim to learn a distribution rather than a single trajectory, employing generative models like GANs (Gupta et al. 2018) and CVAEs (Ivanovic et al. 2020) and DDPMs (Gu et al. 2022b). However, dancer trajectories are complicated by reliance on music, adding a new factor and even stationary dancers show intricate positional variations. In this work, TCDiff utilizes our proposed Dance-Trajectory Navigator, an auto-regressive model that can generate smooth, continuous, and non-overlapping trajectories for dancers.

Background

Music-driven Group Choreography

Given a music sequence $\mathcal{M} = \{\boldsymbol{m}_i\}_{i=1}^{L}$, group choreography is to generate a corresponding group dance movement sequence $\boldsymbol{x} = \{\boldsymbol{x}^i\}_{i=1}^{L}$, where $\boldsymbol{x}^i = \{\boldsymbol{x}^{i,c}\}_{c=1}^{C}$, L, C is the

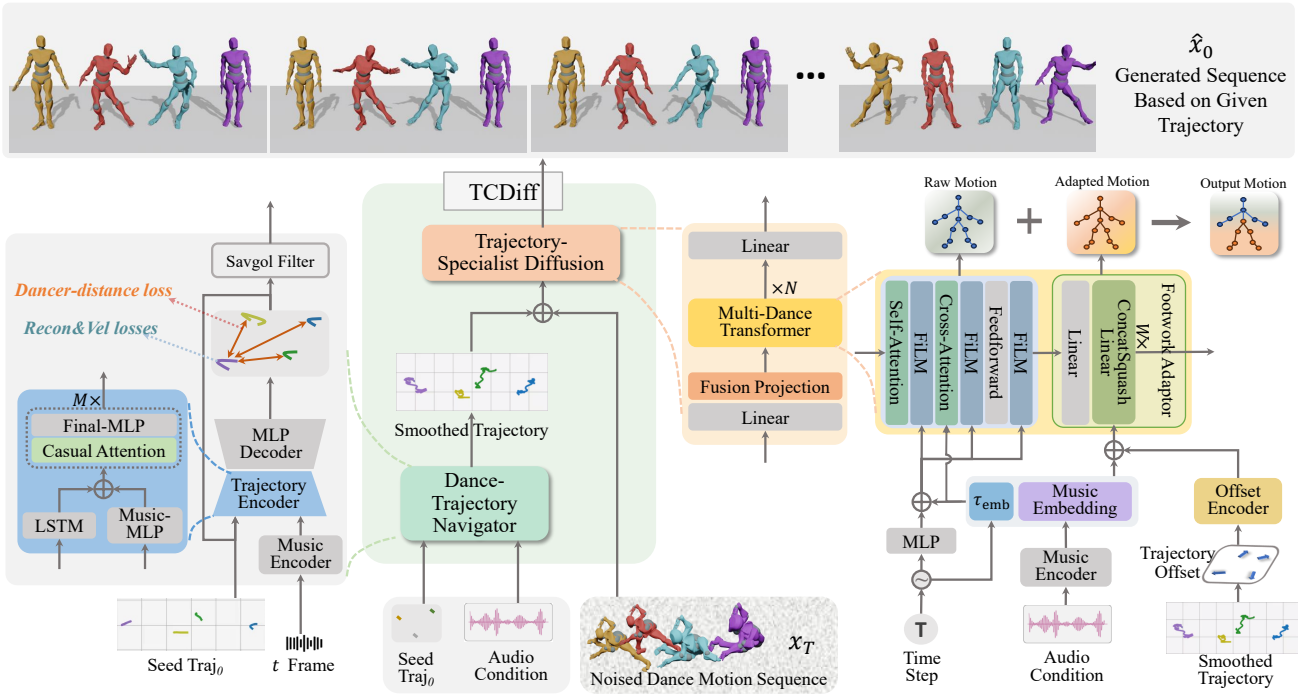

Figure 2: Our TCDiff framework consists of two components: Dance-Trajectory Navigator (DTN) and Trajectory-Specialist Diffusion (TSDiff). Initially, DTN is designed to extract disjoint trajectories (dancer positions) for mitigating dancer ambiguity, as dancers' coordinates exhibit distinct differences and are less prone to confusion. Subsequently, TSDiff utilizes the trajectories for conditional diffusion to generate corresponding dance movements. During this process, a Fusion Projection enhances group information before inputting it into the multi-dance transformer, while a footwork adaptor adjusts the final footwork.

music clip length and the number of dancers, respectively. For simplicity, we use $\boldsymbol{x}_t$ to represent $\{\boldsymbol{x}_t^{i,c}\}_{i=1,c=1}^{L,C}$ at t step.

Motion and Music Representations

We represent motion features of a dancer utilizing a SMPL (Loper et al. 2023) pose $\boldsymbol{d} \in \mathbb{R}^{144}$ extracted from a 24-joint SMPL model in 6D rotation format (Zhou et al. 2019), along with 4-dimensional foot contact labels $\boldsymbol{f} \in \mathbb{R}^4$ and a 3-dimensional root node $\boldsymbol{p} \in \mathbb{R}^3$ for the positions of the dancer. This results in a motion representation $\boldsymbol{x} = [\boldsymbol{f}, \boldsymbol{p}, \boldsymbol{d}] \in \mathbb{R}^{151}$. Note that compared to 3D keypoint representations (Zhang, Black, and Tang 2021; Zanfir et al. 2021; Ma et al. 2023), the use of rotation format tends to yield better motion consistency, as observed in (Siyao et al. 2022). For the music feature, we follow prior works (Kim et al. 2022; Li et al. 2024c) to utilize Librosa (McFee et al. 2015) to extract a representation $\mathcal{M} \in \mathbb{R}^{35}$, comprising a 1-dimensional envelope, 20-dimensional MFCC, 12-dimensional chroma, along with 1-dimensional one-hot peaks and 1-dimensional one-hot beats.

Diffusion Model

We generate dance movements via a diffusion-based method (Ho, Jain, and Abbeel 2020), which establishes a Markov noising process that gradually contaminates clean data $\boldsymbol{x}_0$ into standard Gaussian noise $\boldsymbol{x}_T \sim \mathcal{N}(0, \boldsymbol{I})$ through T pollution steps. The corruption process is defined as:

$$q\left(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}\right) = \mathcal{N}\left(\sqrt{\alpha_t}\boldsymbol{x}_{t-1}, \left(1-\alpha_t\right)\boldsymbol{I}\right), \quad (1)$$

where $\alpha_t \in (0,1)$ are pre-defined hyper-parameters, and $\boldsymbol{I}$ is the identity matrix. Since the music sequence $\mathcal{M}$ is frequently integrated as a conditioning factor (Tseng, Castellon, and Liu 2023), the motion creation involves reversing the forward diffusion process by estimating $\widehat{\boldsymbol{x}}_\theta\left(\boldsymbol{x}_t, t, \mathcal{M}\right) \approx \boldsymbol{x}$ with parameters θ for all t. Thus, the basic objective function (Ho, Jain, and Abbeel 2020) is defined as:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{\boldsymbol{x},t}\left[\left\|\boldsymbol{x} - \widehat{\boldsymbol{x}}_\theta\left(\boldsymbol{x}_t, t, \mathcal{M}\right)\right\|_2^2\right]. \quad (2)$$

We extend the diffusion model by incorporating dancers' trajectories to produce more realistic dance movements.

Methodology

In this section, we introduce our **T**rajectory-**C**ontrollable **Diff**usion (**TCDiff**) framework, which can generate synchronized group dance movements from a music clip. Our pipeline consists of a Dance-Trajectory Navigator **(DTN)** and a Trajectory-Specialist Diffusion **(TSDiff)** in Figure 2.

Dance-Trajectory Navigator

To mitigate dancer ambiguity resulting from similar movements among dancers, our Dance-Trajectory Navigator (DTN) module aims to suppress the interference of similar movements and prioritize the modeling of coordinates,

as shown in the left of Figure 2. Starting with the music sequence $\mathcal{M} = \{\boldsymbol{m}_i\}_{i=1}^{L}$ and a seed trajectory $\boldsymbol{p}^0 = \{\boldsymbol{p}^{0,1}, \ldots, \boldsymbol{p}^{0,C}\}$, the DTN module takes the coordinates $\boldsymbol{p}^i = \{\boldsymbol{p}^{i,1}, \ldots, \boldsymbol{p}^{i,C}\}$ and the music frame $\boldsymbol{m}_i$ at sequence step i as inputs to a music encoder (Kim et al. 2022) and a trajectory encoder, along with an MLP decoder, to recursively generate the next dancer coordinates $\widehat{\boldsymbol{p}}^{i+1} = \{\widehat{\boldsymbol{p}}^{(i+1),1}, \ldots, \widehat{\boldsymbol{p}}^{(i+1),C}\}$. The final trajectory $\widehat{\boldsymbol{p}} = \{\widehat{\boldsymbol{p}}^1, \ldots, \widehat{\boldsymbol{p}}^L\}$ is obtained by applying a Savgol filter (Press and Teukolsky 1990) to smooth the dancer coordinates. The MLP decoder is implemented as a simple multilayer perceptron. Next, we describe the trajectory encoder.

Trajectory Encoder is tasked with extracting detailed features from both the input music and coordinate sequences, which are then fed into the MLP decoder to generate trajectory predictions. For feature pre-processing, the music sequence undergoes processing by a music-MLP, while the coordinate sequence is inputted into a sequence model (Hochreiter and Schmidhuber 1997) to extract temporal features. Additionally, instead of relying on absolute positional encoding (Dosovitskiy et al. 2020; Vaswani et al. 2017) or naive positional encoding (Hu et al. 2019; Liu et al. 2021), we utilize identity encoding (IE) and temporal positional encoding (TPE) (Peng, Mao, and Wu 2023) to capture temporal information. Furthermore, we introduce a trajectory attention module comprising Casual Attention (Vaswani et al. 2017; Radford et al. 2018), effectively directing the model's focus to past information through masking, in conjunction with an MLP network as:

$$\text{Attn} = \text{Softmax}\left(\boldsymbol{M}\boldsymbol{P}^T/\sqrt{d} + \boldsymbol{B}\right)\boldsymbol{P} \times mask, \quad (3)$$

where $\boldsymbol{M}$, $\boldsymbol{P}$ are the processed music and position features using both the music-MLP and the LSTM, respectively. The $mask$ is the causal mask with $mask_{i,j} = -\infty \times 1(i > j) + 1(i \leq j)$, where $1(\cdot)$ is the indicator function. $\boldsymbol{B}$ is the bias, and d is a scaling factor to ensure the stability of the model's training process. We replicate the trajectory attention module M times in this context.

DTN Loss. The objectives of the DTN module are determined by some observations that the ground truth typically exhibits characteristics of continuity and non-overlap. However, it is insufficient to guarantee this if relying solely on the reconstruction loss $\mathcal{L}_{\text{recon}}$ (Zhang et al. 2019). Therefore, we apply a velocity loss $\mathcal{L}_{\text{v}}$ (Tseng, Castellon, and Liu 2023) to restrict position variation for ensuring the continuity. To further approximate the ground truth for reducing the overlapping prediction, simultaneously, we introduce a distance-consistency loss $\mathcal{L}_{DC}$,

$$\Delta\boldsymbol{p}^{w,ij} = \left(\boldsymbol{p}^{w,i} - \boldsymbol{p}^{w,j}\right) - \left(\widehat{\boldsymbol{p}}^{w,i} - \widehat{\boldsymbol{p}}^{w,j}\right), \quad (4)$$

$$\mathcal{L}_{DC} = \frac{1}{C-1}\sum_{w=1}^{L}\binom{C}{2}_{ij}\left\|\Delta\boldsymbol{p}^{w,ij}\right\|_2^2, \quad (5)$$

which ensures that the spacing among dancers is within an appropriate range. The overall loss $\mathcal{L}_{dtn}$ for DTN is $\mathcal{L}_{dtn} = \mathcal{L}_{recon} + \lambda_v\mathcal{L}_v + \lambda_{DC}\mathcal{L}_{DC}$, where λ_v and λ_{DC} are the balanced hyper-parameters.

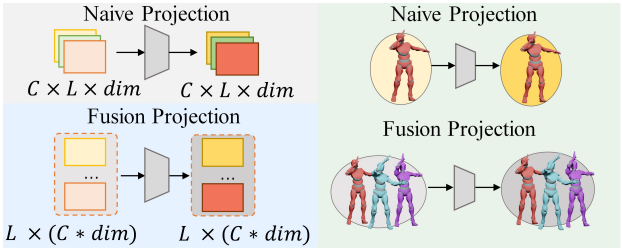

Figure 3: Fusion Projection Module.

Trajectory-Specialist Diffusion

After extracting disjoint trajectories from the DTN module, we introduce a Trajectory-Specialist Diffusion (TSDiff) to generate dancer movements characterized by non-overlapping steps and enhanced grip, in accordance with the provided trajectories. TSDiff consists of a Fusion Projection, a multi-dancer transformer, and simple linear layers for input and output. To ensure non-overlapping movements, we leverage the disjoint trajectories for conditional denoising. To mitigate foot-slide, we present a footwork adaptor within the multi-dancer transformer to adjust foot movements based on trajectory information, thereby reducing foot-slide occurrences. To fully exploit the known trajectory feature, we refrain from introducing noise to the provided positional information $\widehat{\boldsymbol{p}}$, opting instead for conditional motion denoising. Since the linear layers are simple single linear networks, the Fusion Projection and Multi-Dancer Transformer modules are as follows:

Fusion Projection (FP). We propose a simple yet effective solution that tackles dancer ambiguity challenges via its unique feature processing, as shown in Figure 3. The primary rationale behind the FP module is that high-dimensional inputs can capture more distinctive features. By increasing the dimensionality of input features through stacking, we construct high-dimensional input nodes for an expanded MLP, effectively reducing ambiguity among dancers. Relevant evidence is provided in the ablation study.

Multi-Dancer Transformer employs an initial decoder block $D(\cdot)$ (Tseng, Castellon, and Liu 2023) to generate raw motion, which is subsequently refined by our Footwork Adaptor to achieve realistic foot placement. To be more specific, given frame-level music information $\mathcal{M} = \{\boldsymbol{m}_i\}_i^L$, the diffusion final time step T, and the offset of the given trajectory $\mathcal{V} = \{\boldsymbol{v}_i\}_i^L$ (e.g., velocity) as conditions, we first utilize $D(\cdot)$ for feature processing to derive raw motion $\widehat{\boldsymbol{r}}$:

$$\widehat{\boldsymbol{r}} = D(\boldsymbol{x}_T, T, \mathcal{M}). \quad (6)$$

Subsequently, we propose ***Footwork Adaptor*** module $FA(\cdot)$ for correction to obtain the adapted motion $\widehat{\boldsymbol{a}}$.

$$\widehat{\boldsymbol{a}} = FA(\widehat{\boldsymbol{r}}, T, \mathcal{M}, \mathcal{V}). \quad (7)$$

Given that positional changes are predominantly driven by footwork, our approach focuses exclusively on modulating the dancers' lower body movements. Therefore, we split $\widehat{\boldsymbol{r}}$ and $\widehat{\boldsymbol{a}}$ into upper and lower body $\widehat{\boldsymbol{r}} = \{\widehat{\boldsymbol{r}}_{upper}, \widehat{\boldsymbol{r}}_{lower}\}$, and $\widehat{\boldsymbol{a}} = \{\widehat{\boldsymbol{a}}_{upper}, \widehat{\boldsymbol{a}}_{lower}\}$. The down part that is closely related

to footwork is picked as the final generated result:

$$\widehat{x}_0 = \widehat{\boldsymbol{r}}_{upper} \oplus \widehat{\boldsymbol{a}}_{lower}. \tag{8}$$

The Footwork Adaptor consists of a linear layer and a ConcatSquashLinear layer, as shown in Figure 2, which has been proven to be effective in various coordinate prediction domains (Gu et al. 2022a; Luo and Hu 2021).

Conditional Motion Denoising. To leverage the provided trajectory information, instead of adding noise to all information like the original diffusion does, we exclude adding noise to trajectory information. For denoising, we do not denoise $\widehat{\boldsymbol{p}}$ in $\boldsymbol{x} = [\boldsymbol{d}, \boldsymbol{f}, \widehat{\boldsymbol{p}}]$, but treat it as a condition for motion denoising. At each step t in forward processing, we concatenate the noised dance pose information $\boldsymbol{d_t}$, the noised contact label $\boldsymbol{f}_t$ and the given position $\widehat{\boldsymbol{p}}$ into one vector:

$$\boldsymbol{x}_t = \boldsymbol{d_t} \oplus \boldsymbol{f_t} \oplus \widehat{\boldsymbol{p}}. \tag{9}$$

Because the above data format is consistent with the original, we can still use simple loss in Eq. 2 for optimization. At this point, motion $\boldsymbol{d}$ and contact label $\boldsymbol{f}$ are denoised while coordinates $\widehat{\boldsymbol{p}}$ are reconstructed. This enhances the model's learning and memory capabilities for coordinates while improving its ability to extract features from trajectories.

TSDiff Loss. To enable trajectory-conditional generation, we introduce the Relative Forward-Kinematic (RFK) loss $\mathcal{L}_{\text{RFK}}$ to enhance root-motion correlation. The $\mathcal{L}_{\text{RFK}}$ adjusts the individual dancers' relative distance between root nodes and other body joints and can be formulated as:

$$\mathcal{L}_{\text{RFK}} = \frac{1}{L}\sum_{i=1}^{L}\left\|(\text{FK}(\boldsymbol{d}) - \text{FK}(\boldsymbol{p})) - (\text{FK}(\widehat{\boldsymbol{d}}) - \text{FK}(\widehat{\boldsymbol{p}}))\right\|_2^2. \tag{10}$$

Here, $\text{FK}(\cdot)$ is the forward kinematic function that calculates the positions of joints given the 6D rotation motion. We adopt joint velocity loss $\mathcal{L}_{\text{vel}}$ and the foot contact loss $\mathcal{L}_{\text{contact}}$ from (Tseng, Castellon, and Liu 2023). The overall objective of our proposed TSDiff is $\mathcal{L}_{\text{TSDiff}} = \mathcal{L}_{\text{simple}} + \lambda_{\text{RFK}}\mathcal{L}_{\text{RFK}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}} + \lambda_{\text{contact}}\mathcal{L}_{\text{contact}}$, where λ_{RFK}, λ_{vel} and λ_{contact} are the balanced hyper-parameters.

Experiments

Experimental Settings

Implementation Details. For our Dance-Trajectory Navigator, the $\lambda_v = \lambda_{DC} = 2$, and the hidden size of all module layers is set to 64. The Trajectory transformer, which is stacked with $M = 6$ transformer layers, is equipped with 8 heads of attention. The $\lambda_{\text{RFK}} = 0.6$, $\lambda_{vel} = 3$, the $\lambda_{joint} = 0.6$, and the $\lambda_{contact} = 10$. Both the LSTM model and the Music-MLP consist of 3 layers each. The Final-MLP processes the information passed to it through 4 layers, utilizing LeakyReLU non-linearity as the activation function. The sequence length $L = 120$, the hidden dimension is 512, with $N = 8$ layers and 8 heads of attention. We apply a 3-layer MLP as a Fusion Projection, followed by ReLU activation at each layer. Additionally, we stack $W = 3$ Concat Squash Linear with a hidden size of $d_{csl} = 128$ and $d_{ctx} = 512$. The entire framework was trained on 4 Nvidia 4090 GPUs for 3 days. We use a single 4090 GPU to train the Dance-Trajectory Navigator for 26 hours, utilizing batch sizes of 750, 400, 256, and 170 for 2, 3, 4, and 5 dancers, respectively. Similarly, the TSDiff model was trained on 4 NVIDIA 4090 GPUs for 2 days, employing batch sizes of 60, 53, 32, and 20, in that order.

Figure 4: Generated results with different dancer counts.

Dataset. AIOZ-GDance dataset (Le et al. 2023b) is an extensive repository of group dance performances comprising 16.7 hours of synchronized music and 3D multi-dancer motion data. This dataset encompasses a diverse array of over 4000 dancers, spanning 7 distinct dance styles and 16 music genres. Following the partition setting (Le et al. 2023b). we randomly sample all videos into train, validation and test sets with 80%, 10% and 10% of total videos, respectively.

Compared methods. We compare our proposed MotionDiffuse with three baseline models: GDanceR (Le et al. 2023b), GCD (Le et al. 2023a), and CoDancers(Yang et al. 2024). To the best of our knowledge, these represent all the available group dance generation models capable of producing choreography for two or more dancers. Additionally, we incorporate EDGE (Tseng, Castellon, and Liu 2023), the prevailing single-dancer model, to underscore our approach by training it on the AIOZ-GDance dataset.

Metrics. We employ metrics for both multi-dancer and single-dancer evaluations to assess our model. For multi-dancer assessment (Le et al. 2023b). Group Motion Realism (GMR) measures feature similarity via Frechet Inception Distance, Group Motion Correlation (GMC) evaluates coherence through cross-correlation between generated dancers, and Trajectory Intersection Frequency (TIF) assesses the frequency of collisions among dancers. For single-dancer evaluation, Frechet Inception Distance (FID) (Li et al. 2021; Heusel et al. 2017) quantifies the similarity between individual dances and ground-truth dances. Generation Diversity (Div) (Li et al. 2021; Huang et al. 2020) appraises the variety of dance movements using kinetic features. Motion-Music Consistency (MMC) (Li et al. 2021) evaluates how well generated dances synchronize with the music beat. Physical Foot Contact score (PFC) (Tseng, Castellon, and Liu 2023) indicates the physical plausibility of footwork by considering the correlation between the cen-

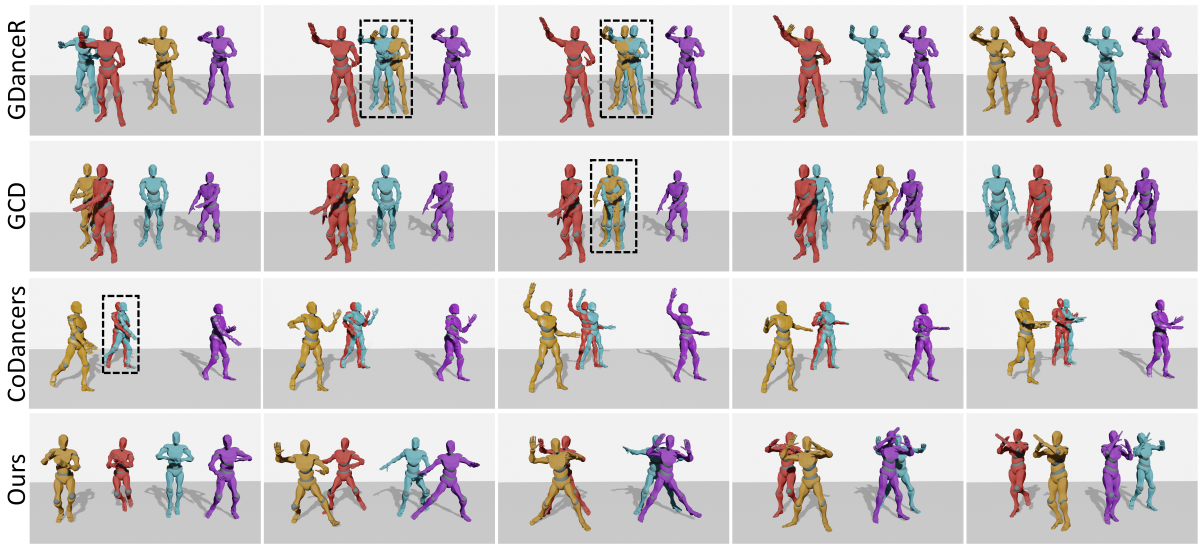

Figure 5: Visual comparison with Baselines. These methods often result in collisions (in black box) or lack of foot movements during exchanges. In contrast, our model minimizes dancer overlaps and generates more natural footwork.

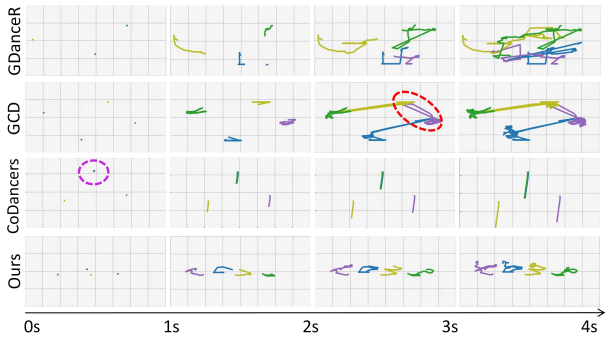

Figure 6: Top-down comparisons of dancer trajectories. GDanceR produces overlapping paths, while GCD causes an unnatural shift, returning to the starting point and ultimately overlapping with a blue dancer. CoDancers generates unreasonable initial positions for the dancers, leading to significant overlap. In contrast, our model effectively minimizes overlaps, showcasing the superior performance of our DTN.

Method	Group-dance Metric			Single-dance Metric			
	GMR↓	GMC↑	TIF↓	FID↓	Div↑	MMC↑	PFC↓
EDGE	63.35	61.72	0.36	31.40	9.57	0.26	2.63
GDanceR	51.27	79.01	0.22	43.90	9.23	0.25	3.05
GCD	31.47	80.97	0.17	31.16	10.87	**0.26**	2.53
CoDancers	26.10	74.05	**0.10**	**23.98**	9.48	0.25	3.26
Ours	**13.86**	**81.98**	0.13	37.47	**15.10**	0.25	**0.51**

Table 1: Quantitative comparison with the baselines.

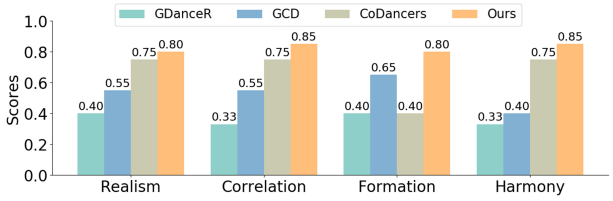

Figure 7: User study based on four criteria: motion realism, music-motion correlation, formation aesthetics, and harmony of dancers. Our model has garnered greater user favor, showcasing our superiority in aesthetic appeal.

ter of mass and foot velocity.

Comparison to the State of the Art

Qualitative Visual Comparison. The performance of our model is illustrated in Figures 4 and 5, highlighting its ability to generate aesthetically pleasing results across various group sizes. Top-view dancer trajectories in Figure 6 further demonstrate our model's superiority in minimizing overlaps. In contrast, GDanceR (Le et al. 2023b) and GCD (Le et al. 2023a) overly prioritize movement similarity, neglecting positional differences, which leads to dancer ambiguity and confusion in positioning. CoDancers (Yang et al. 2024) produces unreasonable initial positions due to incomplete group information, resulting in severe overlaps. Additionally, baseline methods suffer from dancer ambiguity, limiting their ability to align footwork actions with positional changes, thereby hindering the generation of accurate footwork.

Quantitative Results. Tables 1 and 2 compare our model's performance with baseline methods. Our model consistently outperforms in group-dance metrics and excels in Div and PFC for single-dance Metrics. TCDiff effectively captures inter-dancer correlations (high GMC) with a slight trade-off in individual fidelity (FID). Leveraging enriched group dance information, it generates realistic and

Method	Dancers#	GMR↓	GMC↑	TIF↓	FID↓	Div↑	MMC↑
GDanceR	2	53.83	75.44	0.286	48.82	9.36	0.248
	3	55.85	74.07	0.204	44.47	9.36	0.245
	4	58.79	77.71	0.162	47.32	9.24	0.248
	5	55.05	78.72	0.218	44.19	8.99	0.249
GCD	2	34.09	80.26	0.167	32.62	10.41	0.266
	3	36.25	79.93	0.184	33.94	10.02	0.266
	4	36.28	81.82	0.125	35.89	9.87	0.251
	5	38.43	81.44	0.168	35.08	9.92	0.264
CoDancers	2	24.53	72.88	0.080	26.31	9.01	0.251
	3	27.23	74.34	0.084	24.85	9.15	0.254
	4	26.44	75.34	0.097	25.76	9.43	0.258
	5	26.34	74.22	0.113	25.45	9.77	0.253
Ours	2	15.77	81.92	0.121	41.26	16.20	0.263
	3	10.97	81.51	0.123	48.00	19.28	0.253
	4	13.44	81.70	0.149	23.32	10.89	0.253
	5	15.36	82.77	0.109	37.31	14.01	0.236

Table 2: Comprehensive comparison with SOTA methods across varying numbers of dancers.

Method	Group-dance Metric			Single-dance Metric			
	GMR↓	GMC↑	TIF↓	FID↓	Div↑	MMC↑	PFC↓
w/o CMD	33.25	80.61	0.148	43.77	15.22	0.23	0.95
w/o FA and FP	25.57	74.80	0.149	27.18	**17.86**	0.19	4.29
w/o FA	21.39	80.13	0.149	29.21	12.79	0.22	3.25
w/o FP	24.25	**82.63**	0.148	**23.77**	16.00	0.21	0.72
Full	**13.86**	81.98	**0.126**	37.47	15.10	**0.25**	**0.51**

Table 3: Ablation study of Conditional Motion Denoising (CMD), Footwork Adaptor (FA) and Fusion Projection (FP).

harmonious multi-person sequences (low GMR), boosting diversity and quality. In contrast, single-dancer models like EDGE (Tseng, Castellon, and Liu 2023) perform well on solo metrics but struggle in multi-person scenarios, showing high TIF due to dancer ambiguity. GDanceR (Le et al. 2023b) produces low-quality movements (high TIF), reflecting poor ambiguity handling. Similarly, GCD (Le et al. 2023a) suffers from unnatural transitions, leading to high TIF and limited coherence. CoDancers (Yang et al. 2024) reduces ambiguity (low TIF) but compromises inter-dancer correlations (low GMC) and formation integrity, resulting in discordant group formations, as shown in our user study (Figure 7). By decoupling dancer coordinates and movements into two stages, our method mitigates ambiguity, improves coordination, and enhances formation quality.

Ablation Study

Effectiveness of Conditional Motion Denoising. Table 3 shows that our Conditional Motion Denoising (CMD) improves GMR, GMC, FID, MMC, and PFC metrics. CMD effectively converts trajectory denoising loss into reconstruction loss, enhancing generation quality and maximizing the use of trajectory features. Additionally, clean trajectory data enables more accurate computation of the RFK loss and the FA, further boosting the model's trajectory-based generation performance.

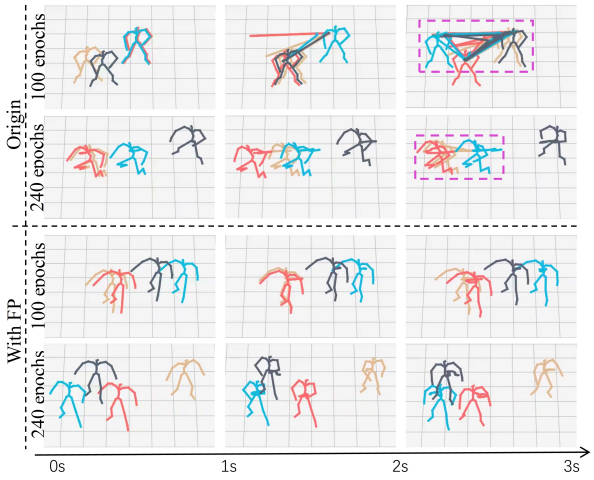

Figure 8: Comparison of EDGE results with and without our FP module. The figure displays various dancers and their movement trajectories. Incorporating the FP module significantly reduces overlap and irregular swapping phenomena.

Effectiveness of Fusion Projection. As shown in Table 3, FP noticeably improves metrics such as GMR, GMC, MMC, and PFC, leading to more lifelike group dance sequences with better synchronization to music beats. The TIF value remains consistent across models, as it primarily depends on the dancer positions generated by the Dance-Trajectory Navigator. FP enhances the extraction of group-level information while slightly compromising individual fidelity. This trade-off significantly improves visual performance, as illustrated in Figure 8, which compares EDGE's generation results with and without our module. Although EDGE was initially designed for single-dancer generation and often encounters dancer ambiguity, the integration of FP effectively mitigates this issue, underscoring the module's impact.

Effectiveness of Footwork Adaptor. We highlight the importance of natural footwork in enhancing visual aesthetics (Tseng, Castellon, and Liu 2023). As shown in Table 3, our Footwork Adaptor markedly enhances the model's PFC value. Moreover, when utilized independently, it boosts the model's GMC, underscoring the FA's dual benefits: preventing foot slipping and augmenting dance coherence.

Conclusion

This paper introduces TCDiff, a novel framework for high-quality multi-dancer movement generation, along with the concept of dancer ambiguity to guide future research. Our lightweight Fusion Projection module effectively mitigates dancer ambiguity with minimal computational cost. Experiments confirm our model's superior performance.

Acknowledgments

This work was supported by the National Science Fund of China under Grant Nos. U24A20330, 62361166670 and 62072242.

References

Ahn, H.; Kim, J.; Kim, K.; and Oh, S. 2020. Generative autoregressive networks for 3d dancing move synthesis from music. *IEEE Robotics and Automation Letters*, 5(2): 3501–3508.

Alahi, A.; Goel, K.; Ramanathan, V.; Robicquet, A.; Fei-Fei, L.; and Savarese, S. 2016a. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 961–971.

Alahi, A.; Goel, K.; Ramanathan, V.; Robicquet, A.; Fei-Fei, L.; and Savarese, S. 2016b. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 961–971.

Alemi, O.; Françoise, J.; and Pasquier, P. 2017. GrooveNet: Real-time music-driven dance movement generation using artificial neural networks. *networks*, 8(17): 26.

Artemyeva, G.; and Moshenska, T. 2018. Role and importance of choreography in gymnastic and dance sports. *Slobozhanskyi herald of science and sport*, (4 (66)): 27–30.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fan, R.; Xu, S.; and Geng, W. 2011. Example-based automatic music-driven conventional dance motion synthesis. *IEEE transactions on visualization and computer graphics*, 18(3): 501–515.

Gong, K.; Lian, D.; Chang, H.; Guo, C.; Jiang, Z.; Zuo, X.; Mi, M. B.; and Wang, X. 2023. Tm2d: Bimodality driven 3d dance generation via music-text integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9942–9952.

Gu, T.; Chen, G.; Li, J.; Lin, C.; Rao, Y.; Zhou, J.; and Lu, J. 2022a. Stochastic trajectory prediction via motion indeterminacy diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17113–17122.

Gu, T.; Chen, G.; Li, J.; Lin, C.; Rao, Y.; Zhou, J.; and Lu, J. 2022b. Stochastic trajectory prediction via motion indeterminacy diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17113–17122.

Gupta, A.; Johnson, J.; Fei-Fei, L.; Savarese, S.; and Alahi, A. 2018. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2255–2264.

Han, B.; Ren, Y.; Peng, H.; Zhang, T.; Ling, Z.; Yin, X.; and Han, F. 2023. EnchantDance: Unveiling the Potential of Music-Driven Dance Movement. *arXiv preprint arXiv:2312.15946*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Hu, H.; Zhang, Z.; Xie, Z.; and Lin, S. 2019. Local relation networks for image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3464–3473.

Huang, R.; Hu, H.; Wu, W.; Sawada, K.; Zhang, M.; and Jiang, D. 2020. Dance revolution: Long-term dance generation with music via curriculum learning. *arXiv preprint arXiv:2006.06119*.

Ivanovic, B.; Leung, K.; Schmerling, E.; and Pavone, M. 2020. Multimodal deep generative models for trajectory prediction: A conditional variational autoencoder approach. *IEEE Robotics and Automation Letters*, 6(2): 295–302.

Joshi, M.; and Chakrabarty, S. 2021. An extensive review of computational dance automation techniques and applications. *Proceedings of the Royal Society A*, 477(2251): 20210071.

Kalman, R. E. 1960. A new approach to linear filtering and prediction problems.

Kim, J.; Oh, H.; Kim, S.; Tong, H.; and Lee, S. 2022. A brand new dance partner: Music-conditioned pluralistic dancing controlled by multiple dance genres. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3490–3500.

Kim, K.; Lee, D.; and Essa, I. 2011. Gaussian process regression flow for analysis of motion trajectories. In *2011 International Conference on Computer Vision*, 1164–1171. IEEE.

Kovar, L.; and Gleicher, M. 2002. Pighin Frédéric. *Motion graphs. ACM T Graphic SIGGRAPH*, 2002: 473–482.

Le, N.; Do, T.; Do, K.; Nguyen, H.; Tjiputra, E.; Tran, Q. D.; and Nguyen, A. 2023a. Controllable Group Choreography Using Contrastive Diffusion. *ACM Transactions on Graphics (TOG)*, 42(6): 1–14.

Le, N.; Pham, T.; Do, T.; Tjiputra, E.; Tran, Q. D.; and Nguyen, A. 2023b. Music-Driven Group Choreography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8673–8682.

Lee, H.-Y.; Yang, X.; Liu, M.-Y.; Wang, T.-C.; Lu, Y.-D.; Yang, M.-H.; and Kautz, J. 2019. Dancing to music. *Advances in neural information processing systems*, 32.

Lee, L.-H.; Lin, Z.; Hu, R.; Gong, Z.; Kumar, A.; Li, T.; Li, S.; and Hui, P. 2021. When creators meet the metaverse: A survey on computational arts. *arXiv preprint arXiv:2111.13486*.

Lee, M.; Lee, K.; and Park, J. 2013. Music similarity-based approach to generating dance motion sequence. *Multimedia tools and applications*, 62: 895–912.

Li, R.; Dai, Y.; Zhang, Y.; Li, J.; Yang, J.; Guo, J.; and Li, X. 2024a. Exploring Multi-Modal Control in Music-Driven Dance Generation. *arXiv preprint arXiv:2401.01382*.

Li, R.; Yang, S.; Ross, D. A.; and Kanazawa, A. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13401–13412.

Li, R.; Zhang, H.; Zhang, Y.; Zhang, Y.; Zhang, Y.; Guo, J.; Zhang, Y.; Li, X.; and Liu, Y. 2024b. Lodge++: High-quality and Long Dance Generation with Vivid Choreography Patterns. *arXiv preprint arXiv:2410.20389*.

Li, R.; Zhang, Y.; Zhang, Y.; Zhang, H.; Guo, J.; Zhang, Y.; Liu, Y.; and Li, X. 2024c. Lodge: A coarse to fine diffusion network for long dance generation guided by the characteristic dance primitives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1524–1534.

Li, R.; Zhao, J.; Zhang, Y.; Su, M.; Ren, Z.; Zhang, H.; Tang, Y.; and Li, X. 2023. FineDance: A Fine-grained Choreography Dataset for 3D Full Body Dance Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10234–10243.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 851–866.

Luo, S.; and Hu, W. 2021. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2837–2845.

Ma, X.; Su, J.; Wang, C.; Zhu, W.; and Wang, Y. 2023. 3D Human Mesh Estimation from Virtual Markers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 534–543.

McFee, B.; Raffel, C.; Liang, D.; Ellis, D. P.; McVicar, M.; Battenberg, E.; and Nieto, O. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, 18–25.

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

Peng, X.; Mao, S.; and Wu, Z. 2023. Trajectory-Aware Body Interaction Transformer for Multi-Person Pose Forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17121–17130.

Press, W. H.; and Teukolsky, S. A. 1990. Savitzky-Golay smoothing filters. *Computers in Physics*, 4(6): 669–672.

Raab, S.; Leibovitch, I.; Tevet, G.; Arar, M.; Bermano, A. H.; and Cohen-Or, D. 2023. Single Motion Diffusion. *arXiv preprint arXiv:2302.05905*.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training.

Ren, Z.; Huang, S.; and Li, X. 2025. Realistic Human Motion Generation with Cross-Diffusion Models. In *European Conference on Computer Vision*, 345–362. Springer.

Schwartz, J. L. 1998. The Passacaille in Lully's" Armide": Phrase Structure in the Choreography and the Music. *Early Music*, 26(2): 301–320.

Shi, L.; Wang, L.; Long, C.; Zhou, S.; Zhou, M.; Niu, Z.; and Hua, G. 2021. SGCN: Sparse graph convolution network for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8994–9003.

Siyao, L.; Gu, T.; Yang, Z.; Lin, Z.; Liu, Z.; Ding, H.; Yang, L.; and Loy, C. C. 2023. Duolando: Follower GPT with Off-Policy Reinforcement Learning for Dance Accompaniment. In *The Twelfth International Conference on Learning Representations*.

Siyao, L.; Yu, W.; Gu, T.; Lin, C.; Wang, Q.; Qian, C.; Loy, C. C.; and Liu, Z. 2022. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11050–11059.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Tseng, J.; Castellon, R.; and Liu, K. 2023. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 448–458.

Valle-Pérez, G.; Henter, G. E.; Beskow, J.; Holzapfel, A.; Oudeyer, P.-Y.; and Alexanderson, S. 2021. Transflower: probabilistic autoregressive dance generation with multimodal attention. *ACM Transactions on Graphics (TOG)*, 40(6): 1–14.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xu, Z.; Lin, Y.; Han, H.; Yang, S.; Li, R.; Zhang, Y.; and Li, X. 2024. Mambatalk: Efficient holistic gesture synthesis with selective state space models. *arXiv preprint arXiv:2403.09471*.

Xue, H.; Luo, X.; Hu, Z.; Zhang, X.; Xiang, X.; Dai, Y.; Liu, J.; Zhang, Z.; Li, M.; Yang, J.; et al. 2024. Human motion video generation: A survey. *Authorea Preprints*.

Yalta, N.; Watanabe, S.; Nakadai, K.; and Ogata, T. 2019. Weakly-supervised deep recurrent neural networks for basic dance step generation. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 1–8. IEEE.

Yang, K.; Tang, X.; Diao, R.; Liu, H.; He, J.; and Fan, Z. 2024. CoDancers: Music-Driven Coherent Group Dance Generation with Choreographic Unit. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, ICMR '24, 675–683. New York, NY, USA: Association for Computing Machinery. ISBN 9798400706196.

Yang, S.; Yang, Z.; and Wang, Z. 2023. LongDanceDiff: Long-term Dance Generation with Conditional Diffusion Model. *arXiv preprint arXiv:2308.11945*.

Yao, S.; Sun, M.; Li, B.; Yang, F.; Wang, J.; and Zhang, R. 2023. Dance with You: The Diversity Controllable Dancer Generation via Diffusion Models. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8504–8514.

Yu, C.; Ma, X.; Ren, J.; Zhao, H.; and Yi, S. 2020. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, 507–523. Springer.

Zanfir, M.; Zanfir, A.; Bazavan, E. G.; Freeman, W. T.; Sukthankar, R.; and Sminchisescu, C. 2021. Thundr: Transformer-based 3d human reconstruction with markers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12971–12980.

Zhang, P.; Ouyang, W.; Zhang, P.; Xue, J.; and Zheng, N. 2019. Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12085–12094.

Zhang, S.; Zhang, Y.; Bogo, F.; Pollefeys, M.; and Tang, S. 2021. Learning motion priors for 4d human body capture in 3d scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11343–11353.

Zhang, Y.; Black, M. J.; and Tang, S. 2021. We are more than our joints: Predicting how 3d bodies move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3372–3382.

Zhang, Y.; Zhang, H.; Hu, L.; Yi, H.; Zhang, S.; and Liu, Y. 2023. Real-time Monocular Full-body Capture in World Space via Sequential Proxy-to-Motion Learning. *arXiv preprint arXiv:2307.01200*.

Zhou, Y.; Barnes, C.; Lu, J.; Yang, J.; and Li, H. 2019. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5745–5753.