InfiniteDance: Scalable Dance Generation Towards in-the-wild Generalization

Ronghui Li[1,⋆], Zhongyuan Hu[1,⋆], Li Siyao[2], Youliang Zhang[1], Haozhe Xie[2], Mingyuan Zhang[2], Jie Guo[3], Xiu Li[1,†], and Ziwei Liu[2]

[1] Tsinghua University
[2] S-Lab, Nanyang Technological University
[3] Peng Cheng Laboratory
Project Page: https://infinitedance.github.io/

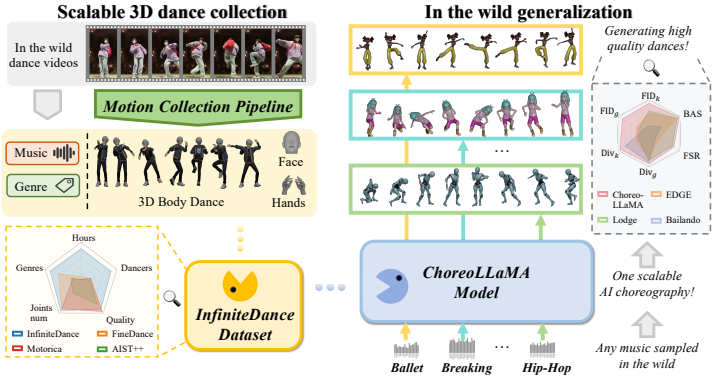

Fig. 1: We propose a fully automated **Motion Collection Pipeline** that extracts high-fidelity and physically plausible 3D dance motions from in-the-wild videos. Based on this pipeline, we construct the **InfiniteDance Dataset**, a large scale music–dance paired dataset. We further propose **ChoreoLLaMA**, trained on InfiniteDance, which generates high-quality 3D dances matching the tempo and style of in-the-wild music.

Abstract. Although existing 3D dance generation methods perform well in controlled scenarios, they often struggle to generalize in the wild. When conditioned on unseen music, existing methods often produce unstructured or physically implausible dance, largely due to limited music-to-dance data and restricted model capacity. This work aims to push the frontier of generalizable 3D dance generation by scaling up both data and model design. **1)** On the data side, we develop a fully automated pipeline that reconstructs high-fidelity 3D dance motions from monocular videos. To eliminate the physical artifacts prevalent in existing reconstruction methods, we introduce a Foot Restoration Diffusion Model (FRDM) guided by foot-contact and geometric constraints that enforce physical plausibility while preserving kinematic smoothness and expressiveness, resulting in a diverse, high-quality multimodal 3D dance

⋆ Equal contribution. † Corresponding author.

dataset totaling 100.69 hours. **2)** On model design, we propose Choreographic LLaMA (ChoreoLLaMA), a scalable LLaMA-based architecture. To enhance robustness under unfamiliar music conditions, we integrate a retrieval-augmented generation (RAG) module that injects reference dance as a prompt. Additionally, we design a slow/fast-cadence Mixture-of-Experts (MoE) module that enables ChoreoLLaMA to smoothly adapt motion rhythms across varying music tempos. Extensive experiments across diverse dance genres show that our approach surpasses existing methods in both qualitative and quantitative evaluations, marking a step toward scalable, real-world 3D dance generation.

1 Introduction

Generating high-quality 3D dance is essential for a wide range of applications, including 3D animation, filmmaking, digital performance, and interactive entertainment. In recent years, there has been increasing interest in deep learning-based methods for automatic dance generation, with the goal of developing an *AI choreographer* that can significantly reduce the costly and time-consuming manual efforts involved in traditional 3D choreography pipelines, thereby enabling scalable content creation.

Despite notable progress in recent years, state-of-the-art 3D dance generation methods are still not ready to be deployed in real-world applications. This is mainly due to two key limitations: **(1) Dance motion quality remains suboptimal.** Current methods still often produce basic artifacts such as foot skating and mesh penetration. **(2) Limited generalization across in-the-wild music conditions.** Although many existing approaches can excel in controlled settings, they often collapse into *unstructured motions* under diverse musical conditions. On one hand, this limitation arises from the scarcity and imbalance of existing music-to-dance datasets, which fail to provide sufficient scale and diversity for models to learn the broad genres of musical and choreographic patterns needed for robust generalization. On the other hand, existing methods often depend on handcrafted and biased music-conditioning designs, yielding limited adaptability to diverse musical styles and tempos. For example, Lodge [19] performs reliably on fast-tempo tracks but struggles with slower rhythms, as its manually crafted rules are tailored for high-energy dance.

Inspired by recent breakthroughs in large-scale models, particularly in LLMs and video generation, we explore whether scaling up data and model capacity can benefit 3D dance generation. In this work, we propose *InfiniteDance*, a scalable 3D dance generation framework by jointly scaling both *data* and *model* capacity, with the goal of advancing deep learning-based methods toward more practical and deployable dance synthesis.

First, on the data side, we introduce an automatic pipeline that converts monocular videos into 3D dance motion and use it to construct a large-scale, high-quality dataset. Although MoCap-based datasets provide high-fidelity motion, their scale is limited to only a few hours. In contrast, monocular reconstruction is more scalable but frequently suffers from artifacts, including penetration,

jittering, floating, and foot skating. Recent motion imitation methods [25, 46], built on physical simulators [7, 28], improve physical plausibility but often suffer from foot jittering due to inaccurate estimation of diverse in-the-wild ground friction. To address these issues, we propose a Foot Restoration Diffusion Model (FRDM), which repairs foot-related artifacts in a self-supervised manner using velocity, position, and rotation cues, while preserving fidelity to the original motion. With geometric guidance during inference, FRDM significantly improves motion realism and brings reconstruction quality close to professional MoCap. Based on this pipeline, we build *InfiniteDance*, a dataset containing 100.69hours of high-quality 3D dance–music pairs covering 30 dance genres, along with RGB videos, 2D keypoints, and other annotations.

Second, on the modeling side, we propose *ChoreoLLaMA*, a scalable architecture that maps music and motion conditions into learnable tokens instead of relying on handcrafted priors. We use a pretrained LLaMA [1, 39, 40] backbone and feed continuous token embeds extracted by MuQ (for music) and an RVQ-VAE (for dance). To mitigate performance degradation under unseen musical conditions, we adopt a cross-modal Retrieval-Augmented Generation (RAG) strategy that selects reference dance motions paired with the music to guide generation. Furthermore, we introduce Cadence-MoE, a Mixture of Cadence Experts designed to learn choreography behaviors across different rhythmic patterns. It jointly models music, genre, and dance tokens under varying tempos, while an adaptive gating network dynamically selects expert modules, enhancing style alignment and improving tempo consistency.

Powered by large-scale training on *InfiniteDance*, our model produces stable and expressive dance motions, outperforming existing methods both qualitatively and quantitatively and pushing 3D dance generation closer to real-world applicability. In conclusion, our key contributions are as follows:

1. We propose a novel 3D dance collection pipeline that captures high-quality, physically plausible, and expressive motion from monocular videos. At its core is an efficient Foot Restoration Diffusion Model (FRDM) that effectively resolves foot-ground contact artifacts while preserving the geometric fidelity of the original motion.

2. We construct a large-scale, high-quality 3D dance dataset, *InfiniteDance*, comprising **100.69** hours of motion across 30 genres, paired with rich annotations including RGB videos, 2D keypoints, music, and genre labels.

3. We design a scalable LLaMA-based choreography framework that leverages retrieved reference dances to improve generalization to in-the-wild music and employs a Cadence-MoE to mitigate generation bias caused by dataset imbalance, thereby enhancing music–dance style consistency.

2 Related Works

2.1 Choreography Dataset

With the rise of data-driven generative models, 3D dance datasets have become essential for choreography research. Current approaches for collecting dance mo-

tion data include marker-based or inertial motion capture, multi-view camera setups, manual keyframing, and monocular video-based pose estimation.

AIST++ [21] marked a milestone by providing 5.2 hours of music-dance paired data; They leverage multi-view camera systems and SMPLify [23] to reconstruct SMPL-format motion. MotoricaDance [2] employed professional motion capture equipment, offering 6.2 hours of high-quality data. FineDance [20] further collected 14.6 hours of dance data by a marker-based MoCap system and professional dancers with fine-grained hands. Although using MoCap equipment can ensure motion quality, the setup is expensive and restricts the capture environments.

Danceformer [15] introduced another category of 3D dance datasets created via manual keyframing by professional animators, and DanceCamera3D [43] further collected MMD (MikuMikuDance) motions keyframed by enthusiasts. However, such motions remain animator-edited and often lack natural dynamics.

Recently, PoPDanceSet [27] used monocular video-based motion capture to collect in-the-wild dance data, estimating 3D poses from online clips with HybrIK [16, 17] to obtain 3.56 hours of motion. However, monocular estimation lacks physical modeling and often introduces artifacts such as jitter, mesh penetration, floating, and foot skating.

In summary, existing datasets struggle to achieve both scalability and high quality, and none of them include facial expressions. Our dataset addresses these gaps by providing 100.69 hours of high-quality 3D dance motion data across 30 diverse genres, with detailed hand movements and expressive facial annotations.

2.2 Music Driven Dance Generation

Generating music-synchronized dance has been studied extensively. Traditional motion-graph-based approaches [3–5, 29] retrieve candidate dance clips based on musical features and stitch them together using hand-crafted choreography rules. However, these methods struggle to generalize across dance genres due to the complexity and diversity of choreographic patterns.

With the rise of deep learning, some early works [13, 21, 24, 47, 50] treat music2dance as a seq2seq task and use LSTM or Transformer to generate frame-by-frame. However, these methods suffer from motion-freezing issues because of error accumulation. Diffusion-based methods [6, 8, 19, 20, 37, 41, 48] model motion features in continuous space through iterative denoising. EDGE [41] and FineDance [20] produce high-quality short clips, and Lodge [19] introduces a coarse-to-fine framework for long-term choreography, producing impressive results for fast-paced dances. However, its handcrafted dance primitives and rule-based choreography augmentation do not generalize well to diverse dance genres.

Token-based autoregressive methods [21, 24, 35] use discrete motion tokens for compact and high-quality motion representation and then train a sequence model to learn music-dance dependencies. Based on this, Bailando [35] enhances rhythmicity via reinforcement learning. However, it only takes low-level music features as input, making it challenging for the sequence model to model the high-level musical structure. As a result, the generated dances often lack coherent

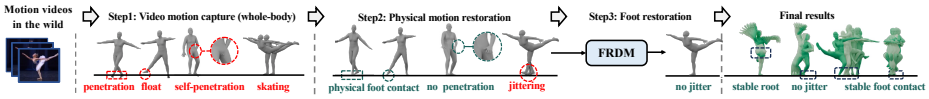

Fig. 2: Overview of our motion collection pipeline. **Step 1:** We estimate whole-body motion from monocular videos, which contain noticeable artifacts. **Step 2:** We refine these motions via motion imitation in a physics simulator for physical plausibility, though this often introduces foot jittering. **Step 3:** We apply our Foot Restoration Diffusion Model (FRDM) to correct foot motions. The **final results** show stable root and foot contacts without jittering or penetration.

choreographic structure and may exhibit repetitive or meaningless motions, such as random hand waving without semantic intent.

Existing methods are limited to certain dance genres, resulting in poor generalization to diverse genres and unfamiliar music. Our approach supports multiple genres and improves choreography quality and generalization through a Retrieval Augmented Generation (RAG) mechanism and Cadence-MoE.

3 *InfiniteDance* Dataset

In this section, we introduce a scalable automated pipeline for 3D motion acquisition, based on which we build a large-scale 3D dance dataset.

3.1 High-quality 3D Motion Collection Pipeline

We propose a novel pipeline that extracts high-quality, physically plausible 3D motions from monocular videos. As shown in Fig. 2, our pipeline is as follows:

The first step is to extract high-quality whole-body motions from monocular videos using video-based motion estimation methods. We first preprocess the videos by using YOLOv8 [42] to extract single-person video sequences. Given its strong generalization ability and gravity-aware modeling, we employ GVHMR [32] to estimate body motion. We use SMPLest-X [44] to obtain SMPL-X expression and hand parameters, as it captures visible features and estimates occluded faces and hands accurately.

The motions estimated in the previous step are used as references for motion *imitation* [26] within a physical simulation environment, which helps correct non-physical artifacts by enforcing physical constraints. This step effectively eliminates common artifacts such as body interpenetration, floating, and foot skating. However, because the physics-based simulation cannot accurately model the diverse ground–surface frictions involved in different dance movements, it often converts foot-skating artifacts into noticeable *foot jittering*.

Foot Restoration: To further address the foot-jittering issue commonly observed in motion imitation, we introduce a Foot Restoration Diffusion Model (FRDM), which will be described in detail later.

3.2 Dataset Statistics

Table 1: Comparisons of 3D Dance Datasets. ✋ denotes whether containing hand (finger) motion. 😊 represents facial expression. $\bar{T}$(sec) denotes the average seconds per sequence. Acquisition methods: MoCap (captured with professional motion capture equipment), Pseudo (previous video-based motion estimation), Animator (manually keyframed by animators).

Dataset	Acquisition	Joints num	✋	😊	Genres	Representation	Dancers	T(h)	$\bar{T}$(s)
Dance w/. Melody [38]	MoCap	21	✗	✗	4	3D joints	-	1.6	92.5
Music2Dance [51]	MoCap	55	✓	✗	2	3D joints	2	0.96	-
EA-MUD [36]	Pseudo	24	✗	✗	4	3D joints	-	0.35	73.8
PhantomDance [15]	Animator	24	✗	✗	13	SMPL	100+	9.6	133.3
AIST++ [21]	Pseudo	17/24	✗	✗	10	SMPL	30	5.2	13.3
MMD [4]	MoCap	52	✓	✗	4	FBX	-	9.9	-
FineDance [20]	MoCap	52	✓	✗	22	SMPL-X	27	14.6	152.3
POPDG [27]	Pseudo	24	✗	✗	19	SMPL	-	3.56	-
Motorica [2]	MoCap	52	✗	✗	8	BVH	-	6.22	-
AIOZ [43]	Pseudo	24	✗	✗	7	SMPL	4000+	16.7	37.5
DCM [14]	Animator	24	✗	✗	4	MMD	-	3.2	106.67
DD100 [34]	MoCap	52	✓	✗	10	SMPL-X	10	1.92	69.3
InterDance [18]	MoCap	52	✓	✗	15	SMPL-X	-	3.93	142.7
InfiniteDance (Ours)	Our Pipeline	**55**	✓	✓	**30**	**SMPL-X**	1000+	**100.69**	30.01

Our 100.69h dataset is curated from platforms like YouTube and TikTok, prioritizing stable cameras and fully visible dance videos. It spans 6 major genres and 33 fine-grained subcategories (*e.g.*, Ballet, Jazz, K-pop), with taxonomy verified by professionals. Unlike existing datasets [20, 21], we preserve detailed hand and facial movements. Furthermore, the clips (6s to 4 mins) capture not only common routines but also technically demanding choreography like pirouettes, floorwork, and single-leg spins. Full statistics are in the supplementary material.

3.3 Foot Restoration Diffusion Model

Given a flawed motion sequence $\boldsymbol{x}$ with foot-ground contact artifacts, our goal is to produce a corrected motion $\tilde{\boldsymbol{x}}$ that exhibits more stable foot contacts while preserving the original full-body geometric consistency. Since these artifacts mainly arise from instability in the root, knees, and feet, we restrict our correction to these joints and keep the upper body unchanged.

Optimizing for foot-skating artifacts directly is difficult, as SMPL pose parameters tend to amplify small errors near the root. We therefore operate in the linear joint position space. We adopt the motion representation similar to HumanML3D [10]. For a motion sequence $\boldsymbol{x} \in \mathbb{R}^{L\times 259}$, where L is the motion frame number, each frame $\boldsymbol{x}$ can be represented as:

$$\boldsymbol{x} = [\boldsymbol{r}, \boldsymbol{j}^v, \boldsymbol{j}^p, \boldsymbol{j}^r] \tag{1}$$

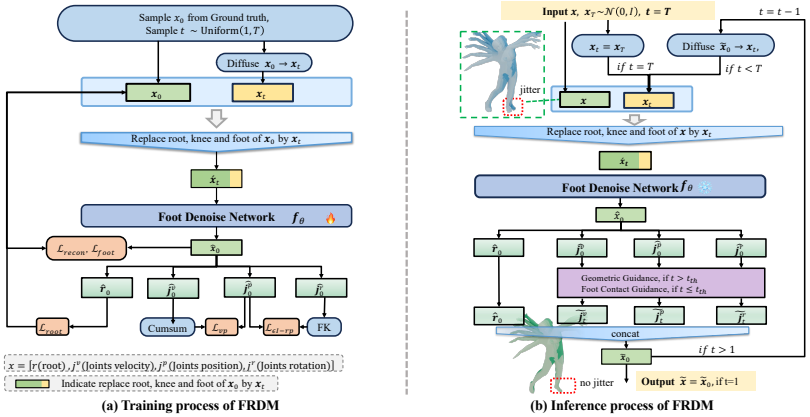

Fig. 3: (a) The Foot Restoration Diffusion Model (FRDM) can be trained in a self-supervised manner. We sample $\boldsymbol{x}_0$ from ground-truth motions and obtain $\boldsymbol{x}_t$ by adding noise. To repair only the artifacts in the root, knees, and feet, we replace these parts in $\boldsymbol{x}_0$ with the corresponding components from $\boldsymbol{x}_t$ to obtain $\acute{\boldsymbol{x}}_t$. We then train a foot denoising network $\boldsymbol{f_\theta}.\hat{\boldsymbol{x}}_0 = \boldsymbol{f_\theta}(\acute{\boldsymbol{x}}_t, t)$, $\boldsymbol{j}_0^p = \text{Cumsum}(\boldsymbol{j}_0^v), \boldsymbol{j}_0^p = \text{FK}(\boldsymbol{j}_0^r)$. **(b)** Given the motion $\boldsymbol{x}$ with foot artifacts, we first sample $\boldsymbol{x}_T \sim \mathcal{N}(0, I)$, and get $\acute{\boldsymbol{x}}_t$ by replacing the root, knee and foot region of $\boldsymbol{x}$ by those of $\boldsymbol{x}_t$. In the early denoising steps $t > t_{th}$, where t_{th} is a threshold, we apply geometric guidance to keep the restored motion geometrically consistent with the original input. In the last steps $t \leq t_{th}$, we use foot-contact guidance to explicitly improve foot stability.

where $\boldsymbol{r} = [\dot{\boldsymbol{r}}^a, \dot{\boldsymbol{r}}^x, \dot{\boldsymbol{r}}^z, \boldsymbol{r}^y]$ is the root data, $\dot{\boldsymbol{r}}^a$ is the angular velocity along the Y-axis (yaw angle), $\dot{\boldsymbol{r}}^x, \dot{\boldsymbol{r}}^z$ are root linear velocities on the floor, $\boldsymbol{r}^y$ is the root height. $\boldsymbol{j}^v \in \mathbb{R}^{L\times 3J}, \boldsymbol{j}^p \in \mathbb{R}^{L\times 3(J-1)}, \boldsymbol{j}^r \in \mathbb{R}^{L\times 6(J-1)}$ correspond to the velocity, position, and rotation (smpl pose) of local keypoints relative to the root, with $J = 22$ denoting the number of joints and $J - 1$ denoting all non-root joints. They are different representations of the same motion, Velocity aids foot stability, while Rotation helps maintain geometric consistency. Based on this, we perform implicit optimization during training and explicit guidance during inference to correct foot artifacts while preserving the original motion.

Self-supervised Training. To train the FRDM for dance, we aggregate several high-quality dance datasets captured using marker-based MoCap systems, including MotoricaDance [2], FineDance [20], DD100 [34], and InterDance [18]. All motions are retargeted to a standard SMPL-X body shape.

In classic diffusion, the model predicts the clean motion $\boldsymbol{x}_0$ from $\boldsymbol{x}_t$. As Fig. 3 (a) shows, the training process differs from standard diffusion pipelines. However, since we only aim to correct motion artifacts in the root, knee, and foot, we obtain $\acute{\boldsymbol{x}}_t$ by replacing the root, knee, and foot features in $\hat{\boldsymbol{x}}_0$ with those of $\boldsymbol{x}_t$. Then we denoise from $\acute{\boldsymbol{x}}_t$. This ensures the upper body remains unchanged during denoising.

To ensure that the corrected root, knee, and foot motions remain faithful to the original sequence, we introduce several MSE-based loss terms, such as $\mathcal{L}_{recon}(\hat{\boldsymbol{x}}_0, \boldsymbol{x}_0)$,$\mathcal{L}_{root}(\hat{\boldsymbol{r}}_0, \boldsymbol{r}_0)$. To encourage more stable foot–ground contact, we introduce a foot loss:

$$\hat{\boldsymbol{P}} = Rec(\hat{\boldsymbol{r}}_0, \hat{\boldsymbol{j}}_0^p), \tag{2}$$

$$\mathcal{L}_{\text{Foot}} = \sum_{k \in F} \left\| (\hat{\boldsymbol{P}}_k^{(i+1)} - \hat{\boldsymbol{P}}_k^{(i)}) \cdot \boldsymbol{b}_k^{(i)} \right\|_2^2, \tag{3}$$

where $Rec(\cdot)$ is the recovery function to get global joints position $\hat{\boldsymbol{P}}$, $k \in F$ means only select the foot joints, the $\boldsymbol{b}_k^{(i)}$ indicates the k-th foot joints whether contact with ground at frame i:

$$\boldsymbol{b}_k^{(i)} = \begin{cases} 1, & \text{if } \left\| \boldsymbol{P}_k^{(i+1)} - \boldsymbol{P}_k^{(i)} \right\|_2^2 < v_{\text{th}} \text{ and } h(\boldsymbol{P}_k^{(i)}) < H_{\text{th}}; \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where v_{th} and h_{th} are velocity and height thresholds, respectively. $h(\boldsymbol{P}_k^{(i)})$ is to get the height of root form $\boldsymbol{P}_k^{(i)}$. To constrain the generated $\hat{\boldsymbol{j}}_0^v$ and $\hat{\boldsymbol{j}}_0^p$ to be aligned, we introduce a loss $\mathcal{L}_{vp} = \left\| \text{cumsum}(\hat{\boldsymbol{j}}_0^v) - \hat{\boldsymbol{j}}_0^p \right\|_2^2$, where cumsum() integrates velocities over time to recover positions. Notably, we expect $\hat{\boldsymbol{j}}_0^r$ and $\hat{\boldsymbol{j}}_0^p$ to remain approximately aligned, which helps balance their distinct roles during optimization: $\hat{\boldsymbol{j}}_0^r$ aims to preserve geometric consistency between the repaired and original motions, while $\hat{\boldsymbol{j}}_0^p$ emphasizes accurate foot-ground contact for artifact correction. Therefore, we design an epsilon insensitive loss $\mathcal{L}_{\epsilon I-rp}$:

$$\mathcal{L}_{\epsilon I-rp} = \sum_{k \in KF} \max \left(\left\| FK(\hat{\boldsymbol{j}}_k^r) - \hat{\boldsymbol{j}}_k^p \right\|_2^2 - \epsilon,\ 0 \right)^2, \tag{5}$$

where $k \in KF$ means only select the knee and foot joints, $FK()$ is the forward kinematic function that calculates position from rotation, ϵ is a hyperparameter that defines the allowed error tolerance.

Inference. We argue that solely relying on a foot loss during the training phase is insufficient to fully resolve foot-ground contact issues. If the weight of foot loss is too small, it fails to sufficiently suppress foot artifacts; conversely, if set too large, it over-constrains the motion, leading to less dynamic results, thereby degrading motion expressiveness. To address these issues, we propose Foot Contact Guidance to explicitly improve foot stability and propose Geometric Guidance to encourage geometric consistency during inference.

As illustrated in Fig. 3 (b), given a full-body motion x exhibiting foot jittering and skating artifacts, At each denoising step, the Foot Denoise Network predicts $\hat{\boldsymbol{x}}_0 = \boldsymbol{f}_{\boldsymbol{\theta}}(\hat{\boldsymbol{x}}_t, t)$. At the early denoising steps, $t > t_{th}$, t_{th} is a hyperparameter,

Geometric Guidance is then applied to enforce global consistency between the restored motion and the input, formulated as:

$$\begin{aligned}
&\hat{\boldsymbol{x}}_0 = [\hat{\boldsymbol{r}}_0, \hat{\boldsymbol{j}}_0^v, \hat{\boldsymbol{j}}_0^p, \hat{\boldsymbol{j}}_0^r], \\
&\tilde{\boldsymbol{j}}_0^r = (1 - w_t)\, \boldsymbol{j}^r + w_t\, \hat{\boldsymbol{j}}_0^r, w_t = t/T \\
&\tilde{\boldsymbol{j}}_0^p = \boldsymbol{b}(1 - w_t)\boldsymbol{j}_0^p + \boldsymbol{b}\, w_t\, \hat{\boldsymbol{j}}_0^p, \tilde{\boldsymbol{j}}_0^r = \hat{\boldsymbol{j}}_0^r.
\end{aligned} \tag{6}$$

At the final stage of denoising, $t < t_{th}$, we use Foot Contact Guidance to further ensure more stable foot-ground contact:

$$\begin{aligned}
&\tilde{\boldsymbol{j}}_0^v = (1 - w_t)\, \boldsymbol{j}^v + w_t\, \hat{\boldsymbol{j}}_0^v, \\
&\tilde{\boldsymbol{j}}_0^p = \mathrm{cumsum}(\boldsymbol{j}_0^v), \quad \tilde{\boldsymbol{j}}_0^r = \hat{\boldsymbol{j}}_0^r,
\end{aligned} \tag{7}$$

where $\boldsymbol{b}$ is a binary mask given by Equation 4. We then get $\tilde{\boldsymbol{x}}_0 = [\hat{\boldsymbol{r}}_0, \tilde{\boldsymbol{j}}_0^v, \tilde{\boldsymbol{j}}_0^p, \tilde{\boldsymbol{j}}_0^r]$.

4 Methodology

Our goal is to generate high-quality 3D dance motions that accurately match the tempo, style, and structural rhythm of in-the-wild music. To achieve this, we design ChoreoLLaMA, a scalable choreography framework composed of two key ideas: **(1) RAG-based Choreography:** Instead of relying solely on music embeds, we retrieve top-k reference dances that share similar musical attributes, providing strong choreographic priors for rare or unseen music. **(2) Cadence-MoE:** We decompose the reference motions into multiple frequency bands and process them via specialized Experts. This design mitigates data imbalance, enables structured motion fusion, and produces expressive, multi-frequency conditioning signals.

4.1 Tokenizer

To enable LLaMA to model the cross-modal correspondence between music and dance, we first design a task-specific tokenization strategy for both modalities. Specifically, for music tokenization, we adopt the pretrained MuQ model [49] to extract music features $\boldsymbol{m}_f \in \mathbb{R}^{N \times C_m}$, which are then projected through a linear layer into the final music embeddings $\boldsymbol{m}e \in \mathbb{R}^{N \times C_L}$. For dance tokenization, as shown in Fig. 5, we train a dance tokenizer following RVQ-VAE [9] to enhance motion quality and preserve fine-grained details. In the Dance Projection module, we first obtain three-layer discrete tokens $\boldsymbol{x}_{\text{idx}}$ and continuous quantized embeddings $\boldsymbol{x}_q \in \mathbb{R}^{N \times C_q}$. The quantized embeddings are then flattened and linearly projected to produce the final dance embeddings $\boldsymbol{x}_e \in \mathbb{R}^{N \times C_L}$.

Unlike prior approaches [12,22,45], which directly feed discrete token indices for both music and motion while ignoring their embedding representations, our design explicitly preserves continuous feature information. We argue that relying solely on token indices makes it difficult for the model to learn detailed and temporally aligned dependencies between music and motion.

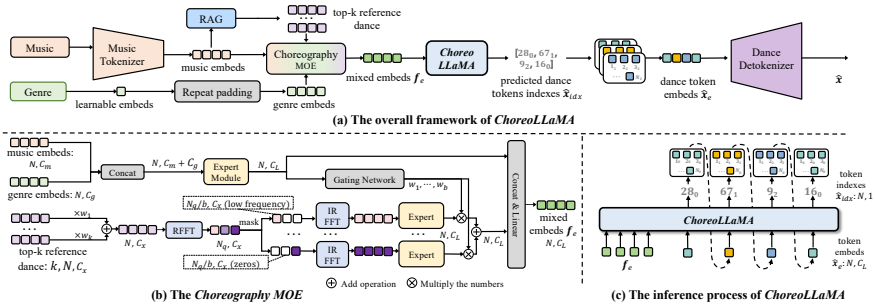

Fig. 4: (a) Given a music clip and a target genre, we first use RAG to retrieve the top-k most relevant reference dances. These retrieved dances, together with the given music and genre embeds, are then fed into the Cadence-MoE to produce fused embeds. ChoreoLLaMA then autoregressively predicts dance tokens that are decoded into dance sequences. (b) Each "Expert" is a neural module composed of linear layers, multi-head attention, and a Mamba block. The "RFFT" and "IRFFT" refer to the real-valued Fast Fourier Transform and its inverse, respectively. (c) At the inference phase, ChoreoLLaMA predicts dance token indices $\hat{\boldsymbol{x}}_{idx}$ one by one, we then lookup the quantized embeds $\hat{\boldsymbol{x}}_q$ in codebooks and project them into dance embeds $\hat{\boldsymbol{x}}_e$.

Given a music clip $\boldsymbol{m}$, previous methods first tokenize it into N discrete units and feed only the token indices $\boldsymbol{m}_{\text{idx}} \in \mathbb{R}^{N\times 1}$ to LLaMA. LLaMA must then learn the corresponding embeddings $\boldsymbol{m}_e \in \mathbb{R}^{N\times C_L}$ from scratch, where C_L denotes the embedding dimension. This inevitably discards important musical cues (especially low-level rhythmic structures), leading to misalignment between the generated fine-grained motion and the underlying music.

As shown in Fig. 4(c), ChoreoLLaMA instead operates on continuous embeddings rather than discrete token indices. This representation is temporally compact yet rich in expressive detail, allowing ChoreoLLaMA to more effectively model both global structures and local rhythmic dependencies between music and dance.

4.2 ChoreoLLaMA

To achieve scalable dance generation that is suitable for any given music, we propose ChoreoLLaMA, as illustrated in Fig. 4.

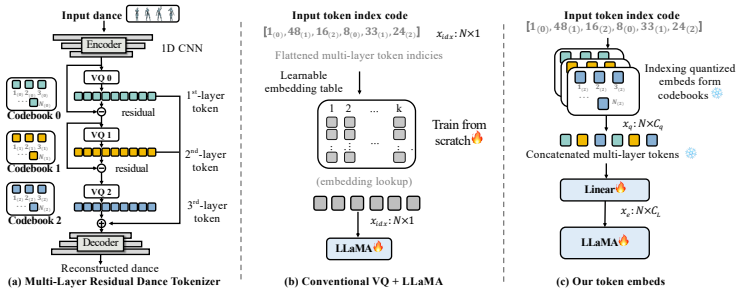

Fig. 5: (a) The residual Dance Tokenizer maintains multi-layer codebooks, each consisting of quantized embeds and corresponding token indices. **(b)** Previous methods directly feed the token indices into LLaMA, which loses the detailed features. The "$[48_{(1)}]$" indicates the token index 48 in the codebook 1. **(c)** We index the quantized embeds $\boldsymbol{x}_q$ and use a linear layer to project it into $\boldsymbol{x}_e$, then feed it into LLaMA. Overview of the Dance Tokenizer and our dance token embeds.

RAG-based Choreography. To improve generalization for diverse and even rare music, we propose a Retrieval Augmented Generation (RAG) based choreography method. We train a Music-Dance Cross-Modal Retrieval Network (MD-Retrieval), following the CLIP [31] architecture, where the Music Encoder and Dance Encoder utilize efficient attention [33], and the model is trained using the InfoNCE [30] loss on the training set of the InfiniteDance dataset.

During the training and inference of ChoreoLLaMA, we retrieve the top-k most relevant training-set reference dance $\{x^r\}_{r=1}^{k}$. Each x_r is processed through a linear projection operation to obtain dance embeds ${x^r}_e \in \mathbb{R}^{N \times C_L}$. The final reference embeds $\bar{\boldsymbol{x}}^r \in \mathbb{R}^{N \times C_L}$ is the weighted sum of $\boldsymbol{x}^r$.

Cadence MoE. To capture both high-frequency motion dynamics and low-frequency, graceful movements, and to effectively leverage choreographic priors from retrieved reference dances, we propose the Cadence-MoE Network. For genre g, we learn an embedding and repeat it to obtain $\boldsymbol{g}_e \in \mathbb{R}^{N \times C_g}$. As shown in Fig. 4(b), the reference dances are weighted by $[\omega_1, \ldots, \omega_k]$, where $k_i = i / \sum_{j=1}^{k} j$, and summed to $\bar{\boldsymbol{x}}_r$. We then apply the Real-valued Fast Fourier Transform (RFFT) to obtain frequency-domain features $\bar{\boldsymbol{x}}_f \in \mathbb{R}^{N_q \times C_x}$, where $N_q = N/2 + 1$ corresponds to the Nyquist frequency. A frequency mask divides the spectrum into b bands, each containing $(N_q/b, C_x)$ valid value and zeros elsewhere. Each band is processed by an Expert, and their outputs are combined using weights $[\gamma_1, \ldots, \gamma_b]$ predicted by a gating network consisting of a linear layer followed by a softmax. This design allows each expert to focus on different frequency characteristics, enabling the model to better adapt to various dance styles ranging from smooth, slow movements to fast, dynamic ones.

Inference Phase. At inference time, given an arbitrary music input, the genre can either be specified by the user or automatically selected as the genre of the retrieved top-k dance.

Table 2: Comparisons on the InfiniteDance dataset. "Our Wins" is the percentage of pairwise comparisons where participants preferred our results over others.

Method	Motion Quality			Motion Diversity		BAS↑	Our Wins↑
	$FID_k \downarrow$	$FID_g \downarrow$	FSR↓	$Div_k \uparrow$	$Div_g \uparrow$		
Ground Truth	2.55	0.60	5.09%	9.37	7.12	0.2332	$34.8 \pm 22.6\%$
Bailando [35]	117.38	82.37	15.56%	5.46	**5.28**	0.2137	$88.7 \pm 8.9\%$
EDGE [41]	96.07	63.53	14.15%	4.36	4.97	0.2321	$76.4 \pm 7.7\%$
FineDance [20]	94.39	62.29	13.53%	4.42	4.85	0.2318	$74.8 \pm 9.1\%$
Lodge [19]	89.52	60.38	6.72%	3.93	5.00	0.2329	$68.1 \pm 11.5\%$
ChoreoLLaMA(Ours)	**30.54**	**16.31**	**5.33%**	**6.23**	5.11	**0.2342**	-

5 Experiments of ChoreoLLaMA

5.1 Experimental Setup

Datasets. We train ChoreoLLaMA on a dataset collection combining our InfiniteDance dataset with several public datasets [20, 21]. InfiniteDance is split into training, validation, and test sets (85%, 5%, and 10%), with genre distributions kept consistent across splits. Public datasets follow their original splits.

Implementation details. Body motion was tokenized using an RVQ-VAE with 3 codebooks (512 entries, 1024-dim), trained on one A100 for 24 hours. For FRDM, we set ϵ=0.1, V_{th}=0.001, H_{th}^{toe}=0.05, H_{th}^{ankle}=0.08. For generation, ChoreoLLaMA is initialized from *LLaMA3.2-1B* and trained with batch size 8 and learning rate 3×10^{-4}; we retrieve top-10 references for RAG and use 2 frequency bands for Cadence-MoE, with C_m=1024, C_g=256, C_x=259, C_q=1024, C_L = 2048. At inference, we use temperature 0.85, top-k sampling (k=30), and top-p sampling (p=0.8).

5.2 Comparisons on the InfiniteDance dataset

As shown in Table 2, we evaluate our method against leading baselines. For a fair comparison, we retrained EDGE, LODGE, and Bailando on the InfiniteDance dataset using their official codes. However, we acknowledge that their default configurations may be sub-optimal for the unprecedented scale and complexity of our InfiniteDance dataset.

Motion Quality. To evaluate the generated dance quality, we adopt the FID metric introduced in [19,35], which compares motion features between generated and ground-truth sequences using Frechet Inception Distance (**FID**) [11]. We further evaluate foot-ground contact quality using the Foot Skating Ratio (**FSR**) [19], which measures foot sliding during contact. Table 2 shows that our method yields significantly lower FID and FSR than prior works.

Motion Diversity. Following [35], we evaluate diversity in feature space, where Div_k and Div_g reflect kinematic and geometric variation. As shown in Table 2,

ChoreoLLaMA achieves the highest Div_k, while its Div_g is slightly below Bailando's, likely because our emphasis on plausibility limits spatial variation that Bailando's high foot skating instead inflates.

Beat Alignment Score (BAS). We evaluate dance-music alignment using BAS [21]. Our method achieves the best BAS of 0.2342, exceeding the Ground Truth slightly. This occurs because in slow-paced dances (e.g., classical), humans prioritize emotional expression over strict beat matching, resulting in low GT BAS. Conversely, our model sometimes generates slightly faster, beat-aligned motions for slow music, which artificially inflates the overall score.

User study. 50 participants viewed 40 random pairs, each pairing our result with a baseline or ground truth, with win rates reported in Table 2.

5.3 Generalization to In-the-Wild Music

To evaluate in-the-wild generalization, we test models trained on InfiniteDance under two cross-dataset settings, AIST++ [21] and FineDance [20], and one OOD setting, an unseen-music set with BPMs outside the training range and rare instruments/styles (*e.g.*, theremin, ambient, body percussion). As shown in Table 3, ChoreoLLaMA consistently outperforms Lodge [19] across all settings. More visual results for OOD testing are available on our project page.

Table 3: Generalization experiments. Both models are trained on InfiniteDance and evaluated on cross-dataset and OOD settings.

Method	Test Setting	$\text{FID}_k \downarrow$	$\text{Div}_k \uparrow$	BAS ↑
Lodge [19]	AIST++ [21] (cross-dataset)	48.73	4.26	0.2364
ChoreoLLaMA (Ours)	AIST++ [21] (cross-dataset)	**35.45**	**5.79**	**0.2378**
Lodge [19]	FineDance [20] (cross-dataset)	106.85	4.14	0.2317
ChoreoLLaMA (Ours)	FineDance [20] (cross-dataset)	**59.38**	**5.81**	**0.2382**
Lodge [19]	Unseen Music (OOD)	119.66	5.13	**0.2332**
ChoreoLLaMA (Ours)	Unseen Music (OOD)	**56.22**	**5.52**	0.2315

Table 4: Ablation study of different components. We progressively add each component to ChoreoLLaMA and evaluate on the InfiniteDance test set.

Components				Metrics		
Token Indices	Token Embeds	RAG	MoE	$\text{FID}_k \downarrow$	$\text{Div}_k \uparrow$	BAS ↑
✓	✗	✗	✗	79.84	**13.09**	0.2073
✗	✓	✗	✗	62.87	5.49	0.2269
✗	✓	✓	✗	38.74	6.16	0.2325
✗	✓	✓	✓	**30.54**	**6.23**	**0.2342**

5.4 Ablation Studies

Token Embeds Inputs. We compare two LLaMA inputs (both without reference dance or Cadence-MoE): (i) raw token indices fed directly to LLaMA, and (ii) embeds extracted by MuQ and the Dance Tokenizer via their projection layers. As shown in Table 4, token indices give higher diversity but significantly degrade motion quality, especially BAS, as low-level details are lost; the embeds-based input yields better fidelity and beat alignment.
Cadence-MoE. Adding Cadence-MoE (rows 3–4 of Table 4) further improves the metrics: by assigning experts to different frequency bands and dance patterns, it alleviates generation bias and strengthens style consistency across musical tempos, yielding more rhythmically coherent dances.
FRDM Post-processing. Applying FRDM on the full model (last row of Table 4) reduces FSR from 8.89% to **5.33%** while keeping FID and BAS competitive, removing foot-skating artifacts without harming choreography quality.
RAG based Choreography. As shown in Table 4, incorporating reference dances sharply improves all metrics, even without Cadence-MoE. This confirms that reference priors significantly enhance natural, diverse, and rhythmically aligned generation. Notably, our retrieval set is restricted to the *training* split. To rule out data leakage from high-similarity retrieval, we vary the reference similarity rank from Top-10 to Top-1000 in Table 5. Even the least-similar Top-1000 references consistently outperform the No-RAG baseline (*e.g.*, FID_k 58.97 vs. 62.87). This proves our gains stem from generalized stylistic priors rather than copying ground-truth duplicates.

To verify that the model does not simply copy retrieved references, we evaluate their feature-space distances (Table 6). The results show that generated-to-retrieved distances (Gen.–Retr.) substantially exceed intra-reference distances (Retr.–Retr.), confirming that the model leverages choreographic priors without direct replication.

Table 5: RAG retrieval sensitivity. Even low-similarity references outperform No-RAG, indicating no test leakage.

Ref. Dances	$\text{FID}_k \downarrow$	$\text{FID}_m \downarrow$	$\text{Div}_k \uparrow$	BAS ↑
Top-10	**38.74**	**13.48**	**6.92**	**0.2340**
Top-100–110	47.16	17.34	6.56	0.2339
Top-500–510	51.76	18.73	6.58	0.2335
Top-1000–1010	58.97	18.92	6.39	0.2327
No RAG	62.87	147.41	4.68	0.2269

Table 6: Copying-behavior analysis. Generated (Gen.) motions stay far from retrieved (Retr.) references, showing influence without copying. K/G/J: kinematic/geometric/joint.

Distance Type	K	G	J
Retr.–Retr.	11.67	7.97	27.39
Gen.–Retr.	23.52	10.11	43.13
Gen.–GT	297.94	20.87	57.11

5.5 Experiments of Motion Collection Pipeline

To evaluate the effectiveness of our motion collection pipeline, we compare our generated motions against FineDance [20], a high-quality marker-based MoCap

dataset. As shown in Table 7, our final pipeline achieves a lower Foot Skating Ratio (FSR) and Penetration Rate than FineDance, demonstrating highly competitive physical fidelity. We further ablate each step of our pipeline to validate its necessity. Raw 3D motions estimated by GVHMR exhibit severe foot skating (FSR 28.63%) and mesh penetration (0.7864%). While applying physics-based imitation (PHC) sharply reduces penetration, it inadvertently introduces severe leg jitter, increasing the jitter metric from 31.89 to 78.60. A naive linear smoothing approach can mitigate this jitter, but it comes at the expense of increased foot skating (FSR rises to 14.29%). In contrast, our proposed FRDM effectively overcomes this trade-off. It achieves the lowest FSR (5.09%) and significantly suppresses jitter (14.33) while maintaining a minimal penetration rate (0.0559%), successfully elevating video-sourced motions to a quality comparable to professional marker-based MoCap.

Table 7: Motion quality of InfiniteDance vs. marker-based MoCap, and ablation of our collection pipeline.

Method	FSR↓	Jitter↓	Penetration↓
FineDance [20] (MoCap)	6.22%	12.69	0.6954%
GVHMR	28.63%	31.89	0.7864%
GVHMR+PHC	8.87%	78.60	0.0536%
GVHMR+PHC+Smooth	14.29%	15.39	0.0561%
GVHMR+PHC+FRDM (Ours)	**5.09%**	**14.33**	**0.0559%**

6 Conclusion and Limitation

We present a scalable framework for 3D dance generation advancing both data and model design: a pipeline that yields the InfiniteDance dataset, and ChoreoLLaMA, which improves quality and generalization through RAG-based Choreography and Cadence-MoE. Limitations remain: The BAS metric is not suitable for slow-paced dances, and under long silent passages or abrupt genre switches the dance follows the music less reliably, and long-sequence structure weakens as retrieval captures mainly intra-clip similarity.

Acknowledgements

This study is supported in part by the Shenzhen Key Laboratory of next generation interactive media innovative technology (No. ZDSYS20210623092001004), in part by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOE-T2EP20223-0002), and in part by cash and in-kind funding from NTU S-Lab and industry partner(s).

References

1. Aaron Grattafiori, Abhimanyu Dubey, e.a.: The llama 3 herd of models (2024), https://arxiv.org/abs/2407.21783 3
2. Alexanderson, S., Nagy, R., Beskow, J., Henter, G.E.: Listen, denoise, action! audio-driven motion synthesis with diffusion models. ACM Trans. Graph. **42**(4), 44:1–44:20 (2023). https://doi.org/10.1145/3592458 4, 6, 7
3. Berman, A., James, V.: Kinetic imaginations: exploring the possibilities of combining ai and dance. In: Twenty-Fourth International Joint Conference on Artificial Intelligence. p. 2431–2437 (2015) 4
4. Chen, K., Tan, Z., Lei, J., Zhang, S.H., Guo, Y.C., Zhang, W., Hu, S.M.: Choreomaster: Choreography-oriented music-driven dance synthesis. ACM Transactions on Graphics (TOG) **40**(4), 1–13 (2021) 4, 6
5. Ciolfi Felice, M., Alaoui, S.F., Mackay, W.E.: How do choreographers craft dance? designing for a choreographer-technology partnership. In: Proceedings of the 3rd International Symposium on Movement and Computing. pp. 1–8 (2016) 4
6. Cohan, S., Tevet, G., Reda, D., Peng, X.B., van de Panne, M.: Flexible motion in-betweening with diffusion models (2024), https://arxiv.org/abs/2405.11126 4
7. Erez, T., Tassa, Y., Todorov, E.: Simulation tools for model-based robotics: Comparison of bullet, havok, mujoco, ode and physx. In: 2015 IEEE international conference on robotics and automation (ICRA). pp. 4397–4404. IEEE (2015) 3
8. Goel, P., Wang, K.C., Liu, C.K., Fatahalian, K.: Iterative motion editing with natural language. In: Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24. p. 1–9. SIGGRAPH '24, ACM (Jul 2024). https://doi.org/10.1145/3641519.3657447, http://dx.doi.org/10.1145/3641519.3657447 4
9. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1900–1910 (2024) 9
10. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5152–5161 (2022) 6
11. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium (2018), https://arxiv.org/abs/1706.08500 12
12. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language (2023), https://arxiv.org/abs/2306.14795 9
13. Kim, J., Oh, H., Kim, S., Tong, H., Lee, S.: 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. pp. 3490–3500 (2022) 4
14. Le, N., Pham, T., Do, T., Tjiputra, E., Tran, Q.D., Nguyen, A.: Music-driven group choreography (2023), https://arxiv.org/abs/2303.12337 6
15. Li, B., Zhao, Y., Zhelun, S., Sheng, L.: Danceformer: Music conditioned 3d dance generation with parametric motion transformer. Proceedings of the AAAI Conference on Artificial Intelligence **36**, 1272–1279 (06 2022). https://doi.org/10.1609/aaai.v36i2.20014 4, 6
16. Li, J., Bian, S., Xu, C., Chen, Z., Yang, L., Lu, C.: Hybrik-x: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025) 4

17. Li, J., Xu, C., Chen, Z., Bian, S., Yang, L., Lu, C.: Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3383–3393 (2021) 4
18. Li, R., Zhang, Y., Zhang, Y., Zhang, Y., Su, M., Guo, J., Liu, Z., Liu, Y., Li, X.: Interdance:reactive 3d dance generation with realistic duet interactions (2024), https://arxiv.org/abs/2412.16982 6, 7
19. Li, R., Zhang, Y., Zhang, Y., Zhang, H., Guo, J., Zhang, Y., Liu, Y., Li, X.: Lodge: A coarse to fine diffusion network for long dance generation guided by the characteristic dance primitives (2024), https://arxiv.org/abs/2403.10518 2, 4, 12, 13
20. Li, R., Zhao, J., Zhang, Y., Su, M., Ren, Z., Zhang, H., Tang, Y., Li, X.: Finedance: A fine-grained choreography dataset for 3d full body dance generation (2023), https://arxiv.org/abs/2212.03741 4, 6, 7, 12, 13, 14, 15
21. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13401–13412 (2021) 4, 6, 12, 13
22. Ling, Z., Han, B., Li, S., Shen, H., Cheng, J., Zou, C.: Motionllama: A unified framework for motion synthesis and comprehension. arXiv preprint arXiv:2411.17335 (2024) 9
23. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 851–866 (2023) 4
24. Luo, M., Hou, R., Li, Z., Chang, H., Liu, Z., Wang, Y., Shan, S.: M3 gpt: An advanced multimodal, multitask framework for motion comprehension and generation. arXiv preprint arXiv:2405.16273 (2024) 4
25. Luo, Z., Cao, J., Winkler, A., Kitani, K., Xu, W.: Perpetual humanoid control for real-time simulated avatars (2023), https://arxiv.org/abs/2305.06456 3
26. Luo, Z., Cao, J., Winkler, A.W., Kitani, K., Xu, W.: Perpetual humanoid control for real-time simulated avatars. In: International Conference on Computer Vision (ICCV) (2023) 5
27. Luo, Z., Ren, M., Hu, X., Huang, Y., Yao, L.: Popdg: Popular 3d dance generation with popdanceset (2024), https://arxiv.org/abs/2405.03178 4, 6
28. Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., et al.: Isaac gym: High performance gpu-based physics simulation for robot learning. arXiv preprint arXiv:2108.10470 (2021) 3
29. Ofli, F., Erzin, E., Yemez, Y., Tekalp, A.M.: Learn2dance: Learning statistical music-to-dance mappings for choreography synthesis. IEEE Transactions on Multimedia **14**(3), 747–759 (2011) 4
30. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018) 11
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021) 11
32. Shen, Z., Pi, H., Xia, Y., Cen, Z., Peng, S., Hu, Z., Bao, H., Hu, R., Zhou, X.: World-grounded human motion recovery via gravity-view coordinates. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024) 5

33. Shen, Z., Zhang, M., Zhao, H., Yi, S., Li, H.: Efficient attention: Attention with linear complexities. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 3531–3539 (2021) 11
34. Siyao, L., Gu, T., Yang, Z., Lin, Z., Liu, Z., Ding, H., Yang, L., Loy, C.C.: Duolando: Follower gpt with off-policy reinforcement learning for dance accompaniment (2024), https://arxiv.org/abs/2403.18811 6, 7
35. Siyao, L., Yu, W., Gu, T., Lin, C., Wang, Q., Qian, C., Loy, C.C., Liu, Z.: Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11050–11059 (2022) 4, 12
36. Sun, G., Wong, Y., Cheng, Z., Kankanhalli, M.S., Geng, W., Li, X.: Deepdance: Music-to-dance motion choreography with adversarial learning. IEEE Transactions on Multimedia **23**, 497–509 (2021). https://doi.org/10.1109/TMM.2020.2981989 6
37. Sun, H., Zheng, R., Huang, H., Ma, C., Huang, H., Hu, R.: Lgtm: Local-to-global text-driven human motion diffusion model. In: Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24. p. 1–9. SIGGRAPH '24, ACM (Jul 2024). https://doi.org/10.1145/3641519.3657422, http://dx.doi.org/10.1145/3641519.3657422 4
38. Tang, T., Jia, J., Hanyang, M.: Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis. In: ACM International Conference on Multimedia. pp. 1598–1606 (2018). https://doi.org/10.1145/3240508.3240526 6
39. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models (2023), https://arxiv.org/abs/2302.13971 3
40. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models (2023), https://arxiv.org/abs/2307.09288 3
41. Tseng, J., Castellon, R., Liu, K.: Edge: Editable dance generation from music. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 448–458 (2023) 4, 12
42. Varghese, R., Sambath, M.: Yolov8: A novel object detection algorithm with enhanced performance and robustness. In: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS). pp. 1–6. IEEE (2024) 5
43. Wang, Z., Jia, J., Sun, S., Wu, H., Han, R., Li, Z., Tang, D., Zhou, J., Luo, J.: Dancecamera3d: 3d camera movement synthesis with music and dance (2024), https://arxiv.org/abs/2403.13667 4, 6
44. Yin, W., Cai, Z., Wang, R., Zeng, A., Wei, C., Sun, Q., Mei, H., Wang, Y., Pang, H.E., Zhang, M., et al.: Smplest-x: Ultimate scaling for expressive human pose and shape estimation. arXiv preprint arXiv:2501.09782 (2025) 5

45. Zhang, J., Zhang, Y., Cun, X., Huang, S., Zhang, Y., Zhao, H., Lu, H., Shen, X.: T2m-gpt: Generating human motion from textual descriptions with discrete representations (2023), https://arxiv.org/abs/2301.06052 9
46. Zhang, Y., Li, R., Zhang, Y., Pan, L., Wang, J., Liu, Y., Li, X.: A plug-and-play physical motion restoration approach for in-the-wild high-difficulty motions. arXiv preprint arXiv:2412.17377 (2024) 3
47. Zhang, Z., Wang, Y., Mao, W., Li, D., Zhao, R., Wu, B., Song, Z., Zhuang, B., Reid, I., Hartley, R.: Motion anything: Any to motion generation. arXiv preprint arXiv:2503.06955 (2025) 4
48. Zhang, Z., Liu, R., Aberman, K., Hanocka, R.: Tedi: Temporally-entangled diffusion for long-term motion synthesis (2023), https://arxiv.org/abs/2307.15042 4
49. Zhu, H., Zhou, Y., Chen, H., Yu, J., Ma, Z., Gu, R., Luo, Y., Tan, W., Chen, X.: Muq: Self-supervised music representation learning with mel residual vector quantization (2025), https://arxiv.org/abs/2501.01108 9
50. Zhuang, H., Lei, S., Xiao, L., Li, W., Chen, L., Yang, S., Wu, Z., Kang, S., Meng, H.: Gtn-bailando: Genre consistent long-term 3d dance generation based on pre-trained genre token network. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023) 4
51. Zhuang, W., Wang, C., Chai, J., Wang, Y., Shao, M., Xia, S.: Music2dance: Dancenet for music-driven dance generation. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **18**(2), 1–21 (2022) 6