PHYSIKDANCE: CONTACT- AND SLIDE-AWARE DUAL-SPACE MUSIC-TO-DANCE GENERATION VIA PROBABILISTIC INVERSE KINEMATICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-driven 3D dance generation is essential for virtual avatars and immersive digital experiences. Despite recent progress, methods that directly predict joint rotations often accumulate errors along the kinematic chain, leading to foot skating, ground penetration and inconsistent support. To tackle the issue, we propose PhysIKDance, a simple yet effective dual-space framework decouples physical reasoning from rotation reconstruction, integrating physically supervised geometric generation with music-conditioned rotation recovery for high-fidelity dance motion generation. In Joint Coordinate Space, a diffusion model jointly predicts root translation, joint coordinates, axial twist, and foot-interaction states. A state-aware objective couples root and local-foot motion to suppress drift during planted support while preserving intentional sliding, complemented by constraints on support height and ground penetration. In Joint Rotation Space, a music-conditioned, two-stage invertible inverse-kinematics module recovers joint swing and combines it with axial twist to produce SMPL-compatible rotations. Cross-space consistency supervision further connects geometric plausibility with accurate rotation reconstruction. Extensive experiments on AIST++ and FineDance demonstrate that PhysIKDance substantially improves foot–ground consistency and overall physical plausibility while achieving competitive motion quality, musical alignment, and motion diversity. Code will be released upon acceptance.

1 INTRODUCTION

Audio-driven 3D dance generation aims to synthesize high-fidelity human motion that is rhythmically, structurally, and stylistically aligned with music. This capability is fundamental to a wide range of applications, including character animation, virtual avatars, immersive entertainment, and computer-assisted choreography, for which manually authoring long and diverse dance sequences is both time-consuming and expensive. Recent advances in cross-modal sequence modelling (Tseng et al., 2023), learned choreographic representations (Siyao et al., 2022), and diffusion-based motion synthesis (Su et al., 2026) have substantially improved the realism, diversity, and musical coherence of generated dances.

Despite these advances, physically implausible foot–ground interactions remain prevalent, typically manifesting as foot skating, ground penetration, and unstable support. This limitation is closely related to the rotation-based motion representations adopted by most existing methods. Joint rotations are well suited to skeletal animation because they preserve bone lengths and can be directly integrated into standard rendering pipelines. However, they specify foot placement only indirectly through forward kinematics: rotations predicted at proximal joints are recursively composed along the skeletal hierarchy to determine the positions of distal joints. Consequently, as shown in Fig. 1 small rotational errors can propagate through the kinematic chain and produce substantial positional deviations at the feet. Existing approaches alleviate these artifacts through contact-aware objectives (Tseng et al., 2023), post-generation foot refinement (Li et al., 2024), or improved mappings between positional and rotational representations (Wang et al., 2026). Nevertheless, directly regulating foot–ground interactions through joint rotations remains difficult, particularly when stationary support must be distinguished from intentional sliding without suppressing natural root motion.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

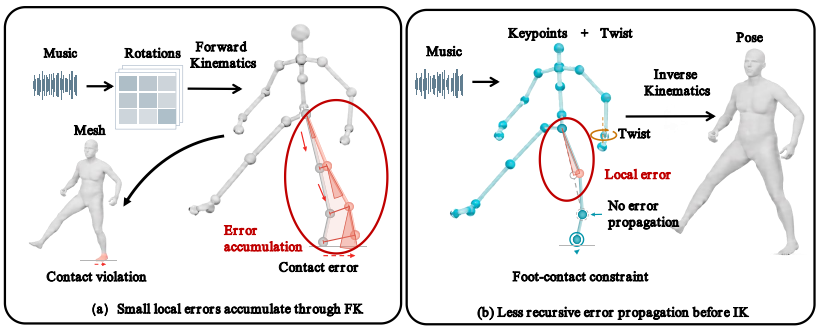


Figure 1: Motivation for dual-space dance generation. (a) Errors in predicted joint rotations accumulate along the kinematic chain through forward kinematics (FK), leading to inaccurate foot placement and implausible ground interactions. (b) PhysIKDance decouples motion generation in Joint Coordinate Space from rotation recovery in Joint Rotation Space, enabling direct geometric supervision of foot–ground interactions before inverse kinematics (IK) reconstructs animation-ready rotations from joint coordinates and axial twist.

This difficulty exposes a fundamental mismatch between the representations required for physical supervision and those required for character animation, as illustrated in Fig. 1. A straightforward solution is to generate joint coordinates directly, thereby allowing foot trajectories, support states, and ground constraints to be modelled explicitly in Euclidean space. Joint coordinates alone, however, do not inherently satisfy skeletal constraints; independently generated coordinates may therefore produce temporally varying bone lengths, invalid skeletal configurations, and insufficient rotational information for animation. An alternative is to predict the absolute orientation of each joint in a global or root-relative coordinate system, rather than recursively predicting parent-relative rotations. Although this formulation can reduce the direct propagation of local rotational errors, independently predicted absolute orientations may be incompatible with the underlying skeletal hierarchy and provide no explicit guarantee of plausible foot–ground interactions. Moreover, converting these orientations into local joint rotations can introduce additional reconstruction errors, while axial twist remains underdetermined by joint positions alone. Thus, neither representation is sufficient in isolation: joint coordinates provide a natural domain for physical reasoning, whereas joint rotations are indispensable for producing structurally valid and animation-ready motion.

Motivated by this complementarity, we propose **PhysIKDance**, a dual-space framework that decouples physically constrained motion generation from rotation recovery while maintaining consistency between the two processes. In **Joint Coordinate Space**, a diffusion model jointly generates root translation, root-local joint coordinates, axial twist, and foot-interaction states. These interaction states distinguish planted support from supported sliding, enabling a state-aware objective that couples root displacement with local foot motion. Specifically, it suppresses unintended drift during planted support while preserving the relative displacement required for intentional sliding. Complementary support-height and ground-penetration constraints further regulate the geometric relationship between the feet and the floor. The resulting geometry is subsequently transferred to **Joint Rotation Space** through a music-conditioned, two-stage invertible inverse-kinematics module. This module first recovers joint swing from the generated coordinates and then combines the recovered swing with axial twist to reconstruct SMPL-compatible rotations. Bidirectional reconstruction and forward-kinematics supervision establish cross-space consistency, while a teacher-to-predicted curriculum progressively adapts the rotation-recovery module from ground-truth coordinates to generated geometry. Together, these components preserve the physically plausible foot–ground relationships established in Joint Coordinate Space during animation-ready rotation reconstruction.

We evaluate PhysIKDance on AIST++ and FineDance in terms of motion quality, musical alignment, foot–ground consistency, motion activity, and rotation reconstruction fidelity. Experimental results show that PhysIKDance achieves superior performance over strong baselines. Our main contributions are threefold:

- We propose **PhysIKDance**, a dual-space framework that decouples physical motion generation in Joint Coordinate Space from animation-ready reconstruction in Joint Rotation

Space, mitigating foot–ground artifacts caused by errors propagated along the kinematic chain.

- We introduce a state-aware physical supervision scheme that distinguishes planted support from supported sliding, together with a music-conditioned, two-stage invertible inverse-kinematics module that recovers SMPL-compatible rotations under cross-space consistency supervision.
- Extensive experiments on AIST++ and FineDance demonstrate that PhysIKDance overperforms strong baselines, achieving improved foot–ground consistency and physical plausibility while preserving motion quality, musical alignment, and expressive movement.

2 Related Work

Music-Driven Dance Generation. Music-driven dance generation requires jointly modelling choreographic structure and music–motion correspondence. Early sequence-based methods establish this correspondence through temporal and cross-modal learning: Dance Revolution (Huang et al., 2020) adopts sequence-to-sequence generation with curriculum learning to mitigate long-term error accumulation, whereas FACT (Li et al., 2021b) introduces a full-attention cross-modal Transformer and the AIST++ benchmark. DanceFormer (Li et al., 2022) further decomposes generation into key-pose synthesis and parametric interpolation. Beyond continuous sequence modelling, Bailando (Siyao et al., 2022) constructs a discrete choreographic memory and employs an actor-critic GPT to compose beat-aligned motion units, while MDOT-Net (Wu et al., 2021) uses optimal transport to align music and motion distributions. More recently, diffusion-based approaches have improved generation quality, diversity, and controllability: EDGE (Tseng et al., 2023) supports editable generation through spatial and temporal constraints, POPDG (Luo et al., 2024) promotes motion diversity and musical alignment, and Lodge (Li et al., 2024) adopts coarse-to-fine modelling for long-form choreography. Meanwhile, FineDance (Li et al., 2023b) expands fine-grained full-body dance data, and SoulDance (Li et al., 2025b) extends generation to coordinated body, hand, and facial motion. Complementary to these advances in choreographic quality and musical coherence, our work focuses on explicitly modelling foot-ground interactions and preserving them when converting generated motion into skeletal rotations.

Physical Plausibility and Foot–Ground Interactions. Physical plausibility can be promoted either through simulation-based guidance or through kinematic constraints imposed directly on generated motion. PhysDiff (Yuan et al., 2023) follows the former paradigm, projecting intermediate diffusion predictions towards physically plausible trajectories through motion imitation in a physics simulator. Kinematic methods instead constrain observable quantities such as joint positions, foot displacement, and contact consistency, avoiding the computational overhead of simulation. For dance generation, however, contact supervision is complicated by the fact that ground contact may correspond either to planted support or to intentional sliding; indiscriminately penalizing foot velocity may therefore suppress valid choreography. EDGE (Tseng et al., 2023) addresses this distinction through contact-consistency supervision, Lodge (Li et al., 2024) introduces a dedicated foot-refinement module, and Lodge++ (Li et al., 2025a) additionally mitigates body self-penetration. PAMD (Wang et al., 2026) combines pose-manifold constraints with foot–ground refinement and further considers the discrepancy between joint-position optimization and rotation-based motion representations. Building on these observations, PhysIKDance applies state-dependent supervision during geometric generation, assigning distinct displacement objectives to planted support and supported sliding while coupling root and local-foot motion to regulate world-space foot displacement. We additionally evaluate motion activity to verify that improved contact consistency does not arise from trivially suppressing expressive movement.

Geometry-Based Rotation Recovery. Geometric and rotational representations offer complementary advantages for human motion modelling: joint coordinates expose spatial relationships for direct physical supervision, whereas joint rotations provide the articulated parameters required by skeletal animation. Recovering rotations from coordinates is nevertheless underdetermined because joint geometry does not uniquely specify axial twist. HybrIK (Li et al., 2021a) addresses this ambiguity by decomposing rotation into swing and twist, analytically deriving swing from joint geometry while predicting twist from visual features. KinePose (Gildea et al., 2022) instead em-

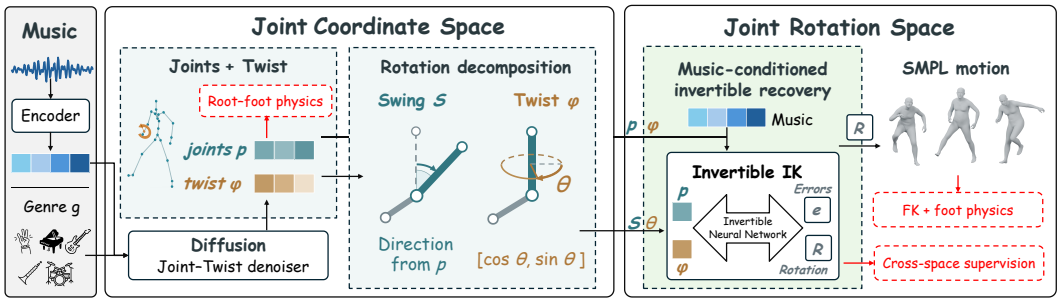


Figure 2: Overview of PhysIKDance. Music-conditioned diffusion generates joint coordinates and axial twist with foot–ground supervision; a two-stage invertible IK module then recovers SMPL rotations.

ploys optimization with temporal and biomechanical constraints to resolve rotational ambiguities, particularly for extended limbs. Since estimated or generated coordinates may themselves be imperfect, NIKI (Li et al., 2023a) further introduces invertible neural networks and bidirectional error modelling for robust rotation recovery under noisy geometric observations. Inspired by these developments, PhysIKDance employs a music-conditioned, two-stage invertible inverse-kinematics module to recover SMPL-compatible rotations from generated joint coordinates and axial twist. Bidirectional reconstruction, forward-kinematics supervision, and a teacher-to-predicted curriculum jointly maintain consistency between physically supervised geometry and the final animation-ready motion.

3 Method

3.1 Problem Formulation and Overview

Given frame-aligned music features $\mathbf{a} \in \mathbb{R}^{T\times D}$ and a dance genre label g, our goal is to generate an animatable motion sequence $\mathbf{x}^{\text{rot}} \in \mathbb{R}^{T\times 155}$ in *Joint Rotation Space*, where T denotes the number of frames and D is the music feature dimension. Each frame comprises four foot-contact values, four foot-slide values, a 3D root translation, and 24 joint rotations in the continuous 6D representation. Rather than directly modeling motion in Joint Rotation Space, PhysIKDance adopts a two-stage framework built around an intermediate *Joint Coordinate Space* representation:

$$\mathbf{y} = [\mathbf{c}, \mathbf{s}, \mathbf{r}, \mathbf{p}, \boldsymbol{\phi}] \in \mathbb{R}^{T\times 129}, \tag{1}$$

where $\mathbf{c}$ and $\mathbf{s}$ denote the foot-contact and foot-slide states, respectively, and $\mathbf{r}$ denotes the root translation. The root-local joint positions $\mathbf{p} \in \mathbb{R}^{T\times 24\times 3}$ describe the skeletal geometry in Joint Coordinate Space, while $\boldsymbol{\phi} \in \mathbb{R}^{T\times 23\times 2}$ encodes the axial twist of each non-root joint as $[\cos\theta, \sin\theta]$. The joint and twist dimensions are flattened when concatenated into $\mathbf{y}$. In the first stage, a diffusion model generates $\mathbf{y}$ in Joint Coordinate Space conditioned on the music features and genre label. In the second stage, two music-conditioned invertible networks map the generated representation into Joint Rotation Space by first recovering joint swing from the joint coordinates and then combining swing with axial twist. The recovered rotations are assembled with $\mathbf{c}$, $\mathbf{s}$, and $\mathbf{r}$ to form $\mathbf{x}^{\text{rot}}$. Fig. 2 summarizes the data flow and supervision. Next, we introduce our key designs in detail.

3.2 Joint Coordinate Generation with State-Aware Physics

Maintaining coherent foot–ground interactions requires distinguishing unwanted drift from supported sliding. A foot in contact with the ground is not necessarily stationary, and its world-space displacement depends on both root translation and local limb motion. We therefore generate joint coordinates together with foot-interaction states and supervise the relation between root and local-foot displacements according to the support state. Axial twist is generated alongside these quantities to provide the complementary orientation information required for subsequent recovery in Joint Rotation Space.

Joint coordinate and twist representation. To express the relation between body translation and foot motion, we decompose world-space joint positions into root translation and root-relative joint coordinates:

$$\mathbf{p}_{t,j} = \mathbf{f}^w_{t,j} - \mathbf{r}_t, \qquad \Delta\mathbf{f}^w_{t,j} = \Delta\mathbf{r}_t + \Delta\mathbf{p}_{t,j}, \tag{2}$$

where $\mathbf{f}^w_{t,j}$ denotes the world-space position of joint j and $\Delta\mathbf{v}_t = \mathbf{v}_{t+1} - \mathbf{v}_t$. Root-relative joint coordinates retain world-axis orientation rather than being rotated into a root-aligned coordinate frame. Consequently, a stationary supporting foot can be represented by compensating root and local-foot displacements, without requiring the root itself to remain stationary.

Joint coordinates describe skeletal geometry but do not uniquely specify rotation about a bone's axis. We retain this missing information through the swing–twist decomposition (Li et al., 2021a; 2023a):

$$\mathbf{R}_{t,j} = \mathbf{S}_{t,j}\mathcal{T}_j(\boldsymbol{\phi}_{t,j}), \qquad \boldsymbol{\phi}_{t,j} = [\cos\theta_{t,j}, \sin\theta_{t,j}], \tag{3}$$

where $\mathbf{S}_{t,j}$ denotes swing and $\mathcal{T}_j$ constructs a twist rotation around a skeleton-derived bone axis. This decomposition applies to non-root joints; root orientation is recovered directly in Joint Rotation Space without a separate twist pair. Together with contact and supported-slide states, these quantities form the intermediate Joint Coordinate Space representation $\mathbf{y}$ defined in Eq. equation 1.

Music-conditioned joint coordinate generation. We model the joint coordinates, axial twist, and interaction states jointly so that they are generated under the same musical context. Let $\mathbf{Y}_0$ denote the normalized sequence of states $\{\mathbf{y}_t\}_{t=1}^{T}$. Following DDPM (Ho et al., 2020), we perturb the sequence at diffusion step n:

$$\mathbf{Y}_n = \sqrt{\bar{\alpha}_n}\mathbf{Y}_0 + \sqrt{1-\bar{\alpha}_n}\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{4}$$

A bidirectional Mamba denoiser (Gu & Dao, 2023) predicts the clean state $\widehat{\mathbf{Y}}_0 = D_\theta(\mathbf{Y}_n, \mathbf{a}, g, n)$. Frame-aligned music features are incorporated through feature-wise affine modulation, together with genre and diffusion-step conditioning. We predict the clean state rather than noise so that geometric supervision can be applied directly to the denoised motion after reversing normalization. The generated twist pairs are normalized before being passed to the Joint Rotation Space recovery stage.

State-aware root–foot supervision. The desired foot displacement depends on the type of support. For planted support, root and local-foot displacements should cancel in the horizontal plane. For supported sliding, their sum should instead reproduce the reference sliding motion. We derive support labels from the height and vertical speed of four ankle/toe joints, and distinguish planted support from supported sliding using horizontal speed. These labels describe geometric support rather than choreographic intent; the thresholds and temporal filtering are specified in the appendix.

Let M_c and M_s select intervals whose endpoints are both planted or both sliding, respectively. Using z-up coordinates and $\Delta\widehat{\mathbf{f}}^w_{f,xy} = \Delta\widehat{\mathbf{r}}_{xy} + \Delta\widehat{\mathbf{p}}_{f,xy}$, we define

$$\mathcal{L}_{\mathrm{rf}} = \left\langle M_c, \left\|\Delta\widehat{\mathbf{f}}^w_{f,xy}\right\|_2^2 \right\rangle + \beta_c \left\langle \mathrm{sg}(\widehat{M}_c), \left\|\Delta\widehat{\mathbf{f}}^w_{f,xy}\right\|_2^2 \right\rangle + \beta_s \left\langle M_s, \left\|\Delta\widehat{\mathbf{f}}^w_{f,xy} - \Delta\mathbf{f}^w_{f,xy}\right\|_2^2 \right\rangle. \tag{5}$$

where f indexes the foot joints and $\langle M, z\rangle$ denotes a masked average. The first two terms penalize displacement during planted support, while the third matches reference displacement during sliding. The predicted planted confidence $\widehat{M}_c$ provides an additional weighting, with stop-gradient sg preventing direct reduction of the penalty through lowered confidence.

This objective constrains the combined world-space motion of the foot rather than either root or local-foot motion alone. We additionally supervise root translation and its temporal differences to discourage suppressing reference locomotion. Support-height and penetration terms regulate the vertical foot–ground relationship. These constraints are combined with joint-coordinate, velocity, and twist supervision in the overall training objective.

3.3 Music-Conditioned Recovery in Joint Rotation Space

The generated joint coordinates and axial twist provide complementary information for pose reconstruction but do not necessarily form a configuration that can be exactly realized by the prescribed skeleton. In particular, errors in Joint Coordinate Space can make the recovery of skeletal rotations inconsistent with the generated motion. We therefore learn a music-conditioned mapping from

Joint Coordinate Space to Joint Rotation Space. Building on NIKI (Li et al., 2023a), we introduce auxiliary error variables into two invertible mappings, allowing rotation recovery to accommodate imperfect joint coordinates while using the axial twist supplied by the first stage.

Conditioning on music and motion states. Recovery in Joint Rotation Space should account not only for joint coordinates but also for the motion context in which they occur. We construct a shared condition from music, foot-interaction states, and root motion:

$$\mathbf{h} = \mathrm{LN}\left(E_a(\mathbf{a}) + E_m\left([\widehat{\mathbf{c}}, \widehat{\mathbf{s}}, \widehat{\mathbf{r}}, \Delta\widehat{\mathbf{r}}]\right)\right), \tag{6}$$

where E_a encodes temporal music features and E_m embeds the motion states. This condition supplies musical and locomotion context to the rotation-recovery networks. It is held fixed when evaluating each forward–inverse pair and remains external to the invertible variables.

Error-augmented swing–twist recovery. Following the decomposition in Eq. equation 3, we factor the mapping to Joint Rotation Space into joint-to-swing and swing–twist-to-rotation mappings. Let $\mathbf{S}_6$ and $\mathbf{R}_6$ denote the flattened continuous 6D representations of swing and full rotations, respectively. We define the forward mappings as

$$F_J\left([\mathbf{S}_6, \mathbf{e}_S]; \mathbf{h}\right) = [\mathrm{vec}(\mathbf{p}), \mathbf{u}, \mathbf{e}_J], \tag{7}$$

$$F_\phi\left([\mathbf{R}_6, \mathbf{e}_R]; \mathbf{h}\right) = [\mathbf{S}_6, \mathrm{vec}(\boldsymbol{\phi})]. \tag{8}$$

Here, $\mathbf{u}$ is a learned per-joint uncertainty feature, and $\mathbf{e}_J$, $\mathbf{e}_S$, and $\mathbf{e}_R$ are auxiliary error variables. Their dimensions make the input and output of each mapping equal, providing additional capacity to represent discrepancies between Joint Coordinate Space and Joint Rotation Space. Both mappings use conditional affine coupling layers (Dinh et al., 2017). Invertibility is defined over these augmented feature spaces; it does not imply that joint coordinates uniquely determine physical rotations.

During inference, we predict $\widehat{\mathbf{u}}$ from the generated Joint Coordinate Space representation and music. Starting with zero input error, we first recover swing:

$$[\widehat{\mathbf{S}}_6, \widehat{\mathbf{e}}_S] = F_J^{-1}\left([\mathrm{vec}(\widehat{\mathbf{p}}), \widehat{\mathbf{u}}, \mathbf{0}]; \mathbf{h}\right). \tag{9}$$

We then combine the recovered swing with the normalized twist pairs generated by the first stage:

$$[\widehat{\mathbf{R}}_6, \widehat{\mathbf{e}}_R] = F_\phi^{-1}\left([\widehat{\mathbf{S}}_6, \mathrm{vec}(\widehat{\boldsymbol{\phi}})]; \mathbf{h}\right). \tag{10}$$

The root block of $\mathbf{S}_6$ retains the full root orientation and has no corresponding twist pair. The recovered rotations are assembled with $\widehat{\mathbf{c}}$, $\widehat{\mathbf{s}}$, and $\widehat{\mathbf{r}}$ to obtain $\widehat{\mathbf{x}}^{\mathrm{rot}}$ in Joint Rotation Space. Thus, the second stage reconstructs skeletal articulation through two inverse mappings rather than an additional diffusion process.

4 Experiments

4.1 Experimental Setup

We evaluate PhysIKDance on AIST++ (Li et al., 2021b) and FineDance (Li et al., 2023b), following the EDGE split for AIST++ and the official cross-genre split for FineDance. Both datasets are processed at 30 FPS using a fixed 24-joint SMPL skeleton and segmented into 5-second training clips. We use 35-dimensional frame-aligned audio descriptors and compute normalization statistics exclusively from the corresponding training sets. PhysIKDance and all internal ablations are trained for 1,000 epochs using Adan with a learning rate of 10^{-4} and a batch size of 64, followed by 50-step DDIM sampling at inference. The physical-loss warm-up and inverse-kinematics curriculum each span 200 epochs. We compare PhysIKDance with FACT (Li et al., 2021b), Bailando (Siyao et al., 2022), EDGE (Tseng et al., 2023), FineDance (Li et al., 2023b), Lodge (Li et al., 2024), BADM (Zhang et al., 2024), and FlowerDance (Yang et al., 2025); InfiniteDance (Li et al., 2026) is reported separately because it uses additional training data. For a transparent evaluation, we distinguish results obtained by controlled retraining under our unified data-processing and audio-feature settings from those computed using officially released outputs. We assess motion quality and diversity using FID $* g/k$ and Div $* g/k$, musical alignment using BeatAlign, and physical plausibility using PFC and contact-specific measures of skating and penetration. Motion activity statistics complement these measures to identify improvements associated with reduced movement.

Table 1: Motion quality evaluation results. Bold and underlined values indicate the best and second-best results, respectively, within each dataset, excluding GT. ↓/↑ denote lower/higher is better; ↔ denotes proximity to GT. AIST++ geometric scores use raw features and are therefore not directly comparable to those on FineDance.

Method	FID_k ↓	FID_g ↓	Div_k ↔	Div_g ↔	Active joint ratio ↔	Root disp. (m) ↔	Motion beats (s^{-1}) ↔
FineDance							
GT	0.000	0.000	9.709	7.179	0.850	0.363	0.798
FineNet	16.041	<u>15.119</u>	7.277	<u>6.546</u>	0.774	0.232	0.758
InfiniteDance	22.287	31.071	<u>7.897</u>	25.227	0.880	0.268	0.873
FlowerDance	26.740	15.701	5.322	6.065	<u>0.828</u>	0.266	0.879
LODGE	45.090	27.545	6.011	6.517	0.883	0.038	0.953
MEGADance	<u>13.182</u>	34.403	7.414	6.160	0.877	<u>0.444</u>	<u>0.818</u>
PhysIKDance (Ours)	**13.162**	**5.239**	**8.829**	**6.894**	**0.833**	**0.324**	**0.795**
AIST++ (raw geometric scores)							
GT	0.000	0.000	9.749	1.440	0.863	0.269	1.031
EDGE	<u>35.957</u>	<u>0.821</u>	7.143	0.796	<u>0.910</u>	0.326	<u>0.892</u>
Bailando	52.654	1.506	6.805	**1.235**	0.703	0.172	0.841
FlowerDance	37.570	1.010	<u>7.382</u>	<u>1.222</u>	**0.890**	<u>0.295</u>	**1.024**
PhysIKDance (Ours)	**31.522**	**0.801**	**7.641**	1.149	0.946	**0.287**	0.883

Table 2: Physical quality comparison results on the test set. GT is excluded from ranking; ↓ indicates lower is better, and ↔ indicates proximity to GT. Bold and underlined values denote the best and second-best results, respectively.

Method	PFC ↓	Grounded foot speed (m/s) ↓	Foot skating ratio ↓	Support ratio ↔
GT	0.158	0.134	0.346	0.5768
FineNet	<u>0.262</u>	<u>0.197</u>	<u>0.801</u>	0.4897
InfiniteDance	0.361	0.302	0.938	**0.5872**
FlowerDance	0.353	0.249	0.929	0.6276
LODGE	0.288	0.249	0.888	0.4595
MEGADance	0.558	0.308	0.941	0.4969
PhysIKDance (Ours)	**0.239**	**0.151**	**0.587**	<u>0.5659</u>

4.2 Main Results

Motion quality and diversity comparison. We first evaluate motion quality, diversity, and activity of our method and the baselines on both FineDance and AIST++, with the results summarized in Table 1. On FineDance, PhysIKDance achieves the best performance across all seven reported metrics. In particular, it reduces FID_g from 15.119 for the strongest baseline to 5.239 and marginally improves FID_k over MEGADance (13.162 vs. 13.182). PhysIKDance also yields the highest kinetic diversity and a geometric diversity score closest to that of the ground-truth motions. Moreover, its active-joint ratio, root displacement, and motion-beat frequency most closely match the corresponding reference statistics, indicating that the improvements in distributional quality do not compromise motion activity or temporal dynamics. On AIST++, PhysIKDance remains competitive and obtains the lowest kinetic and geometric FID among all evaluated methods. The improvements are less pronounced than those on FineDance, which may be attributed to noise in the video-reconstructed motions of AIST++; inaccuracies in foot trajectories and ground contacts can weaken contact supervision and thereby limit the effectiveness of our physical objectives.

Physical plausibility. We assess the physical plausibility of our method and the baselines on FineDance by measuring foot–ground consistency, using its high-fidelity motion references to ensure reliable contact-sensitive evaluation. As reported in Table 2, PhysIKDance achieves the lowest

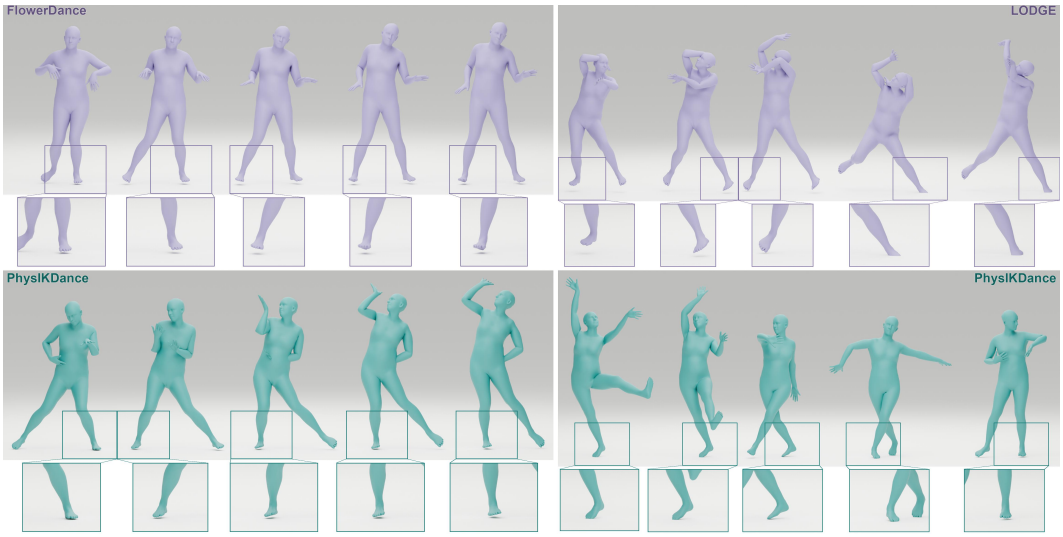


Figure 3: Qualitative comparisons on FineDance at matched time steps. Foot close-ups highlight floating and ground penetration in the baseline examples and more consistent contact in PhysIKDance.

Table 3: Ablation study on FineDance.

Method	$\text{FID}_k \downarrow$	$\text{FID}_g \downarrow$	Div_k	Div_g	PFC↓
w/o twist	28.991	26.944	5.255	5.133	0.7008
w/o IK	14.111	904.857	8.813	16.296	**0.2024**
Global rotation prediction	14.158	67.998	9.307	9.319	0.5243
w/o direct IK music	13.425	29.448	9.020	9.145	0.5939
w/o rotation-to-joints training	15.690	20.577	8.988	8.418	0.3676
PhysIKDance	**13.162**	**5.239**	8.829	6.894	0.2390

physical foot-contact score (PFC), grounded foot speed, and foot-skating ratio among all evaluated methods. Compared with FineNet, the strongest baseline under these metrics, PhysIKDance reduces PFC from 0.262 to 0.239, grounded foot speed from 0.197 to 0.151 m/s, and the foot-skating ratio from 0.801 to 0.587. Meanwhile, its support ratio of 0.5659 closely approximates the ground-truth value of 0.5768, suggesting that the reduction in foot motion artifacts does not arise from underestimating ground contact. When considered alongside the motion-activity statistics in Table 1, these results demonstrate that PhysIKDance substantially improves foot stability and ground-contact consistency without trivially suppressing overall movement.

Qualitative comparison. We qualitatively compare PhysIKDance with FlowerDance and LODGE on FineDance in Fig. 3, presenting five temporally matched frames for each comparison. For a consistent visualization of foot–ground interactions, we apply the same ground-estimation procedure to all methods, defining a fixed reference plane using the median of the minimum sole heights over the entire sequence. All poses are oriented toward the camera and displayed chronologically from left to right. The magnified foot regions show that FlowerDance produces visible gaps between the support foot and the estimated ground, whereas LODGE exhibits both foot floating and ground penetration. In contrast, PhysIKDance maintains more consistent support-foot contact throughout the sequences while preserving diverse and dynamic movements, including lateral steps, arm gestures, and leg lifts.

4.3 Ablation Study

Table 3 examines the main components of PhysIKDance on FineDance. Removing twist increases FID_k from 13.162 to 28.991, FID_g from 5.239 to 26.944, and PFC from 0.2390 to 0.7008. Re-

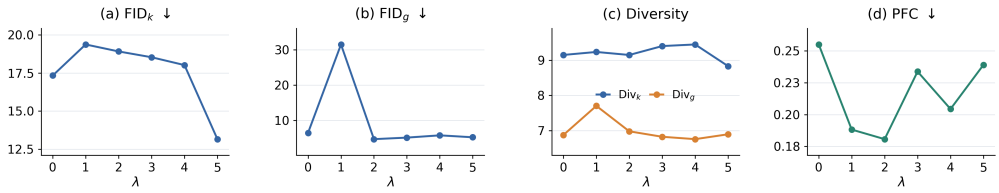


Figure 4: Effect of the root–foot coupling weight λ on motion fidelity, diversity, and physical consistency on FineDance.

moving IK yields a much larger FID$_g$ of 904.857, although its PFC is lower at 0.2024; physical consistency alone therefore does not capture motion quality. Predicting global rather than local rotations increases FID$_g$ to 67.998 and PFC to 0.5243. Removing direct music conditioning from IK yields FID$_g$ $=$ 29.448 and PFC$=$ 0.5939, while removing the rotation-to-joints training constraint yields FID$_g$ $=$ 20.577 and PFC$=$ 0.3676. These variants retain different parts of the pipeline, but each has substantially worse geometric fidelity than the full model.

Figure 4 shows how the root–foot coupling weight affects the results. Every evaluated nonzero weight reduces PFC relative to $\lambda = 0$, with the lowest PFC of 0.1807 at $\lambda = 2$. The same weight achieves the lowest FID$_g$ of 4.712, whereas $\lambda = 5$ achieves the lowest FID$_k$ of 13.162. The trends are not monotonic: for example, FID$_g$ rises to 31.533 at $\lambda = 1$. We therefore use $\lambda = 5$ for its kinetic fidelity while acknowledging that no single weight optimizes all metrics.

4.4 User Study

We conduct a user study to compare our method with four representative baselines in terms of expressiveness and physical plausibility. As shown in Fig. 5, our method achieves an overall win rate above 50% against every baseline, demonstrating a consistent preference for our results. In particular, our method is preferred for expressiveness in all comparisons, with win rates ranging from 62.2% to 75.7%. It also outperforms all competing methods in physical plausibility, achieving win rates between 54.0% and 67.9%. These results indicate that our method generates dance motions that are both more expressive and more physically plausible than those produced by existing approaches. The full study protocol, questionnaire design, and aggregation details are provided in the appendix.

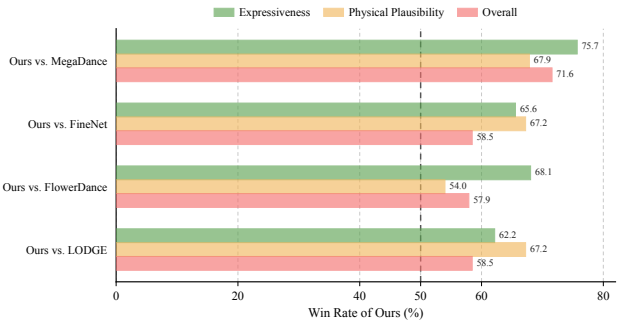


Figure 5: User study results. Win rates of our method against four baselines.

5 Conclusion

We present PhysIKDance, a dual-space framework that decouples physical motion generation from rotation reconstruction for music-driven dance synthesis. The framework distinguishes planted support from intentional sliding to supervise foot–ground interactions in joint coordinate space, and recovers SMPL-compatible rotations through a music-conditioned, two-stage invertible inverse-kinematics module. Experiments on AIST++ and FineDance support its effectiveness in improving motion quality, with additional improvements in foot–ground consistency on FineDance. The user study further supports the expressiveness and perceived physical plausibility of the generated dances. Future work will improve music–motion synchronization while preserving physically plausible foot interactions.

6 GENERATIVE AI USE DISCLOSURE

In this study, we used generative AI tools to assist in designing and discussing ablation experiments; developing and revising training, evaluation, and plotting code; analyzing and presenting experimental results; and editing parts of the manuscript. These tools also assisted in creating figures and tables from experimental results.

The authors reviewed AI-assisted text, code, figures, tables, and numerical claims against the relevant implementations and recorded experimental results, making revisions where necessary. The authors are responsible for the experimental findings and scientific conclusions. We take full responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

REFERENCES

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/pdf?id=HkpbnH9lx.

Kevin Gildea, Clara Mercadal-Baudart, Richard Blythman, Aljosa Smolic, and Ciaran Simms. Kinepose: A temporally optimized inverse kinematics technique for 6dof human pose estimation with biomechanical constraints. *arXiv preprint arXiv:2207.12841*, 2022.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Daxin Jiang. Dance revolution: Long-term dance generation with music via curriculum learning. In *International conference on learning representations*, 2020.

Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 1272–1279, 2022.

Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. 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 (CVPR)*, pp. 3383–3393, 2021a.

Jiefeng Li, Siyuan Bian, Qi Liu, Jiasheng Tang, Fan Wang, and Cewu Lu. NIKI: Neural inverse kinematics with invertible neural networks for 3D human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12933–12942, 2023a.

Ronghui Li, Junfan Zhao, Yachao Zhang, Mingyang Su, Zeping Ren, Han Zhang, Yansong Tang, and Xiu Li. FineDance: A fine-grained choreography dataset for 3D full body dance generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10234–10243, 2023b. URL https://openaccess.thecvf.com/content/ICCV2023/html/Li_FineDance_A_Fine-grained_Choreography_Dataset_for_3D_Full_Body_Dance_ICCV_2023_paper.html.

Ronghui Li, YuXiang Zhang, Yachao Zhang, Hongwen Zhang, Jie Guo, Yan Zhang, Yebin Liu, and Xiu Li. Lodge: A coarse to fine diffusion network for long dance generation guided by the characteristic dance primitives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1524–1534, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Li_Lodge_A_Coarse_to_Fine_Diffusion_Network_for_Long_Dance_CVPR_2024_paper.html.

Ronghui Li, Hongwen Zhang, Yachao Zhang, Yuxiang Zhang, Youliang Zhang, Jie Guo, Yan Zhang, Xiu Li, and Yebin Liu. Lodge++: High-quality and long dance generation with robust choreography patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Ronghui Li, Zhongyuan Hu, Li Siyao, Youliang Zhang, Haozhe Xie, Mingyuan Zhang, Jie Guo, Xiu Li, and Ziwei Liu. InfiniteDance: Scalable 3D dance generation towards in-the-wild generalization. arXiv preprint arXiv:2603.13375, 2026.

Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. AI Choreographer: Music conditioned 3D dance generation with AIST++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13401–13412, 2021b. URL https://openaccess.thecvf.com/content/ICCV2021/html/Li_AI_Choreographer_Music_Conditioned_3D_Dance_Generation_With_AIST_ICCV_2021_paper.html.

Xiaojie Li, Ronghui Li, Shukai Fang, Shuzhao Xie, Xiaoyang Guo, Jiaqing Zhou, Junkun Peng, and Zhi Wang. Music-aligned holistic 3d dance generation via hierarchical motion modeling. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14420–14430. IEEE, 2025b.

Zhenye Luo, Min Ren, Xuecai Hu, Yongzhen Huang, and Li Yao. Popdg: Popular 3d dance generation with popdanceset. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26974–26983. IEEE, 2024.

Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3D dance generation by actor-critic GPT with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11050–11059, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Siyao_Bailando_3D_Dance_Generation_by_Actor-Critic_GPT_With_Choreographic_Memory_CVPR_2022_paper.html.

Ming-Cong Su, Wei-Lun Huang, and Tse-Yu Pan. Danceldm: latent-based diffusion model for dance generation and editing conditioned on music and text prompt. *Multimedia Syst.*, 32(4), May 2026. ISSN 0942-4962. doi: 10.1007/s00530-026-02311-z. URL https://doi.org/10.1007/s00530-026-02311-z.

Jonathan Tseng, Rodrigo Castellon, and Karen Liu. EDGE: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 448–458, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Tseng_EDGE_Editable_Dance_Generation_From_Music_CVPR_2023_paper.html.

Hongsong Wang, Ying Zhu, Qiuxia Lai, Yang Zhang, Guo-Sen Xie, and Xin Geng. Pamd: plausibility-aware motion diffusion model for long dance generation. *Visual Intelligence*, 4(1): 20, 2026.

Shuang Wu, Shijian Lu, and Li Cheng. Music-to-dance generation with optimal transport. *arXiv preprint arXiv:2112.01806*, 2021.

Kaixing Yang, Xulong Tang, Ziqiao Peng, Xiangyue Zhang, Puwei Wang, Jun He, and Hongyan Liu. Flowerdance: Meanflow for efficient and refined 3d dance generation. *arXiv preprint arXiv:2511.21029*, 2025.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15964–15975. IEEE, 2023.

Canyu Zhang, Youbao Tang, Ning Zhang, Ruei-Sung Lin, Mei Han, Jing Xiao, and Song Wang. Bidirectional autoregessive diffusion model for dance generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 687–696, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhang_Bidirectional_Autoregessive_Diffusion_Model_for_Dance_Generation_CVPR_2024_paper.html.

A Appendix

A.1 Representation and Coordinate Conventions

This appendix supplements the main paper with the representation conventions, foot-state construction, network architecture, complete training objectives, inference procedure, evaluation definitions, user-study design and vote aggregation, and additional qualitative comparisons of PhysIKDance. The description concerns the 24-joint body model and the joint–twist diffusion configuration.

A.1.1 Motion channels and spatial coordinates

We use a right-handed, z-up coordinate system, with positions measured in metres and motion sampled at 30 frames per second. Root-relative coordinates follow Eq. (2) in Section 3.2 of the main paper. Table 4 summarizes the motion channels.

Table 4: Per-frame representation of the 24-joint model. The contact and slide channels refer to the same four ankle/toe joints.

Quantity	Dimension	Symbol	Interpretation
Planted contact	4	$\mathbf{c}$	Geometrically supported, approximately stationary feet
Supported sliding	4	$\mathbf{s}$	Supported feet outside the planted state
Root translation	3	$\mathbf{r}$	World-space pelvis trajectory
Joint coordinates	72	$\mathbf{p}$	24×3 root-relative coordinates
Axial twist	46	$\boldsymbol{\phi}$	23 non-root pairs $[\cos\theta, \sin\theta]$
Geometric state	129	$\mathbf{y}$	$[\mathbf{c}, \mathbf{s}, \mathbf{r}, \mathrm{vec}(\mathbf{p}), \mathrm{vec}(\boldsymbol{\phi})]$
Rotation state	155	$\mathbf{x}^{\mathrm{rot}}$	$[\mathbf{c}, \mathbf{s}, \mathbf{r}, \mathbf{R}_6]$, with $\mathbf{R}_6 \in \mathbb{R}^{144}$

Here, $\mathbf{c}$ denotes *planted contact*, rather than all supported contact. The union of planted contact and supported sliding defines support. The target states are mutually exclusive, whereas their continuous predictions need not satisfy exact exclusivity before thresholding. This convention distinguishes supported sliding from a loss that treats every grounded foot as stationary.

A.1.2 Normalization and rotation conversion

The first 83 channels of $\mathbf{y}$ are normalized componentwise using a min–max transform fitted on training motion. The transform maps its fitted range to $[-1, 1]$; inverse transformation clips normalized values to that interval. The remaining 46 twist channels retain their cosine–sine representation rather than receiving an additional fitted scaler. Training and evaluation reuse the training normalizers. Physical objectives are computed after inverse transformation, in metric coordinates.

The invertible module operates on normalized features. Its joint-coordinate input is additionally centered by subtracting the root-joint feature from each joint feature. We distinguish this internal feature centering from the physical coordinate definition in Eq. (2) of the main paper: the internal flow variables should not be interpreted directly as distances in metres. Root displacement in the context encoder is likewise computed from its normalized root input; root velocities used for physical evaluation are computed after inverse normalization.

For a 6D rotation represented by two three-vectors $(\mathbf{a}_1, \mathbf{a}_2)$, the conversion to a rotation matrix orthonormalizes the two vectors:

$$\mathbf{b}_1 = \frac{\mathbf{a}_1}{\|\mathbf{a}_1\|_2}, \quad \mathbf{b}_2 = \frac{\mathbf{a}_2 - (\mathbf{b}_1^\top \mathbf{a}_2)\mathbf{b}_1}{\|\mathbf{a}_2 - (\mathbf{b}_1^\top \mathbf{a}_2)\mathbf{b}_1\|_2}, \quad \mathbf{b}_3 = \mathbf{b}_1 \times \mathbf{b}_2. \tag{11}$$

The vectors are stacked according to the rotation convention used by the conversion routine, consistently for encoding and decoding. Numerical normalization uses a small denominator floor. Forward kinematics (FK) then places the recovered local joint rotations on the fixed skeleton. This enforces the prescribed bone lengths; it does not alone enforce correct foot placement or contact.

A.2 Construction of Planted and Sliding States

A.2.1 Floor estimation and velocity tests

The foot set is $\mathcal{F} = \{7, 8, 10, 11\}$ in zero-based SMPL indexing, corresponding to the left/right ankles and toes. State construction uses world-space joints before normalization. For each sequence, the floor estimate is

$$h_0 = Q_{0.02}\big(\{f^w_{t,f,z} : t = 1, \dots, T,\ f \in \mathcal{F}\}\big), \tag{12}$$

where $Q_{0.02}$ is the second percentile. This estimate reduces the influence of an isolated low outlier compared with taking the minimum height.

Let $\nu = 30$ denote the frame rate. For $t \geq 2$, state-label velocities are $\mathbf{v}_{t,f} = \nu(\mathbf{f}^w_{t,f} - \mathbf{f}^w_{t-1,f})$; the first frame repeats the first available difference. A one-frame sequence receives zero velocity. Preliminary support is

$$q^0_{t,f} = \mathbb{1}[f^w_{t,f,z} \leq h_0 + \tau_h]\,\mathbb{1}[|v_{t,f,z}| \leq \tau_z]. \tag{13}$$

The foot-state thresholds are $\tau_h = 0.07$ m, $\tau_z = 0.20$ m/s, and a planted-speed threshold $\tau_p = 0.12$ m/s. Expressing the speed thresholds per second avoids changing their physical meaning with the frame rate.

A.2.2 Temporal filtering and interval masks

Let $\mathcal{M}_3$ denote a three-frame majority filter, with a binary output of one when at least two entries are active. The temporal filter uses zero padding at sequence boundaries and skips filtering for sequences shorter than three frames. The final masks are

$$q = \mathcal{M}_3(q^0), \tag{14}$$

$$c = q \wedge \mathcal{M}_3(q \wedge \mathbb{1}[\|\mathbf{v}_{xy}\|_2 \leq \tau_p])\,, \tag{15}$$

$$s = q \wedge \neg c. \tag{16}$$

The order matters: support is filtered first, planted contact is filtered second and intersected with support, and sliding is assigned last. Consequently, $c \wedge s = 0$ and $c \vee s = q$. Because of temporal filtering, the final label at a frame is not determined by that frame's horizontal speed alone.

A displacement loss uses intervals rather than isolated frames. With $\Delta\mathbf{z}_t = \mathbf{z}_{t+1} - \mathbf{z}_t$, the target masks are

$$M^c_{t,f} = c_{t,f}c_{t+1,f}, \qquad M^s_{t,f} = s_{t,f}s_{t+1,f}. \tag{17}$$

The sliding-acceleration loss uses a three-frame mask $M^{sss}_{t,f} = s_{t,f}s_{t+1,f}s_{t+2,f}$. These masks avoid imposing a persistent-support constraint across a state transition. The labels describe geometric support and motion, not annotated choreographic intent; low aerial motion or inaccurate reconstructed geometry may therefore be ambiguous.

A.3 Architecture and Rotation Recovery

A.3.1 Joint–twist diffusion denoiser

The denoiser projects the 129D motion state and 35D music features to 512 channels. Four bidirectional Mamba blocks encode music, and eight bidirectional Mamba blocks process motion. Each bidirectional block runs separate forward and backward Mamba branches and sums their temporally aligned outputs. The Mamba state size is 16, the local convolution width is 4, the expansion factor is 2, and dropout is 0.1.

A learned genre embedding modulates the music representation through a gated residual update. If $\mathbf{a}'$ is a projected music feature and $\mathbf{g}$ a projected genre embedding, the gate has the form

$$\mathbf{a}'' = \mathbf{a}' + \mathrm{sigmoid}(A([\mathbf{a}', \mathbf{g}])) \odot D([\mathbf{a}', \mathbf{g}]), \tag{18}$$

where A and D are two-layer MLPs. The genre embedding table has 16 entries. The gate's final bias is initialized to -4, giving a small initial genre-dependent residual.

The diffusion timestep is embedded sinusoidally and transformed by an MLP. Each motion block applies timestep-dependent feature-wise affine modulation to its Mamba output and to the encoded

music before residual addition. For feature $\mathbf{z}$, the modulation is $(1 + \boldsymbol{\gamma}_n) \odot \mathbf{z} + \boldsymbol{\beta}_n$. The motion and music paths have separate modulation parameters. When their lengths differ, projected music features are linearly interpolated to the motion length. The final linear layer predicts the clean 129D state. The sampling procedure described here uses conditional predictions directly, without an additional classifier-free-guidance combination.

A.3.2 SWING–TWIST TARGETS

For each non-root joint, a reference bone axis $\mathbf{a}_j$ is derived from the skeleton offsets. A joint with children uses the offset of its longest outgoing bone; a leaf uses its incoming offset. The axis is normalized. Write a unit rotation quaternion as $q_j = (w_j, \mathbf{v}_j)$. The twist quaternion is the normalized projection

$$q_j^{\mathrm{tw}} = \mathcal{N}_4\left(w_j, (\mathbf{v}_j^\top \mathbf{a}_j)\mathbf{a}_j\right), \qquad q_j^{\mathrm{sw}} = q_j \otimes (q_j^{\mathrm{tw}})^{-1}. \tag{19}$$

If the projected quaternion norm is at most 10^{-8}, we substitute the identity twist. For $q_j^{\mathrm{tw}} = (w_j^{\mathrm{tw}}, s_j\mathbf{a}_j)$, its full-angle twist pair is

$$\boldsymbol{\phi}_j = \left[(w_j^{\mathrm{tw}})^2 - s_j^2,\ 2w_j^{\mathrm{tw}} s_j\right]. \tag{20}$$

This representation avoids an explicit angular discontinuity at $\pm\pi$. The root keeps its full orientation in the swing block and has no separate twist pair. Predicted pairs are normalized as $\mathcal{N}_2(\mathbf{z}) = \mathbf{z}/\max(\|\mathbf{z}\|_2, 10^{-8})$ before recovery. A zero pair remains degenerate under this numerical convention; the unit-norm training term discourages such predictions.

A.3.3 CONTEXT, UNCERTAINTY, AND AUGMENTED DIMENSIONS

A two-layer temporal Transformer encodes music with width 512, eight attention heads, feed-forward width 2048, GELU activation, dropout 0.1, and pre-layer normalization. An MLP embeds the 14D concatenation $[\mathbf{c}, \mathbf{s}, \mathbf{r}, \Delta\mathbf{r}]$. Their sum is layer-normalized to obtain the shared context $\mathbf{h}$. The first root difference is zero in this encoder.

A separate joint-and-music Transformer has the same width, depth, and attention-head count. It predicts a 24D uncertainty feature

$$\mathbf{u} = 0.01\,\mathrm{sigmoid}\big(W_u\,\mathrm{LN}(\mathbf{z}_{\mathrm{joint,music}}) + \mathbf{b}_u\big). \tag{21}$$

This is a learned conditioning feature, not a calibrated confidence interval. The rotation-recovery module receives the normalized twist generated by diffusion.

The two conditional mappings have matched input and output dimensions:

$$F_J: \quad [\mathbf{S}_6, \mathbf{e}_S] \in \mathbb{R}^{144+32} \longleftrightarrow [\boldsymbol{\psi}, \mathbf{u}, \mathbf{e}_J] \in \mathbb{R}^{72+24+80}, \tag{22}$$

$$F_\phi: \quad [\mathbf{R}_6, \mathbf{e}_R] \in \mathbb{R}^{144+46} \longleftrightarrow [\mathbf{S}_6, \mathrm{vec}(\boldsymbol{\phi})] \in \mathbb{R}^{144+46}, \tag{23}$$

where $\boldsymbol{\psi}$ denotes the centered joint features described in Section A.1. Thus, F_J is 176-dimensional and F_ϕ is 190-dimensional. Invertibility holds in these augmented Euclidean feature spaces, conditional on fixed $\mathbf{h}$. It does not imply that joint positions uniquely determine a physical rotation or that projection from 6D features to $SO(3)$ is invertible.

A.3.4 CONDITIONAL AFFINE COUPLING

Each mapping contains 16 affine coupling layers with alternating even/odd channel masks. For binary mask $\mathbf{m}$, define $\mathbf{z}_A = \mathbf{m} \odot \mathbf{z}$. A three-linear-layer MLP with two 512D hidden layers and GELU activations predicts scale and translation from $[\mathbf{z}_A, \mathbf{h}]$. The scale is bounded by $\mathbf{s} = 0.1\tanh(\mathbf{s}_{\mathrm{raw}})$. On the unmasked channels,

$$\mathbf{y}_A = \mathbf{z}_A, \qquad \mathbf{y}_B = \mathbf{z}_B \odot \exp(\mathbf{s}) + \mathbf{t}, \tag{24}$$

$$\mathbf{z}_A = \mathbf{y}_A, \qquad \mathbf{z}_B = (\mathbf{y}_B - \mathbf{t}) \odot \exp(-\mathbf{s}). \tag{25}$$

The inverse traverses the layers in reverse order. The same external context must be used for both directions. The training objective combines reconstruction and error regularization; it does not include an exact maximum-likelihood term.

A.4 Complete Training Objectives

Let MSE denote the mean of squared differences over all included scalar entries. For a mask M and values Z, define $\langle Z \rangle_M$ as the weighted sum divided by the sum of the broadcast weights, with the denominator clamped below by one. An empty binary mask therefore contributes zero. Temporal differences in training losses are *per-frame displacements*; they are not multiplied by the frame rate. This differs from the velocities in state labeling and evaluation.

A.4.1 Geometric-space reconstruction and root motion

The clean-state diffusion loss compares the prediction and target in normalized 129D space. Auxiliary position and velocity terms operate in physical space:

$$\mathcal{L}_{\text{joint}} = \text{MSE}(\widehat{\mathbf{p}}, \mathbf{p}), \tag{26}$$

$$\mathcal{L}_{\text{vel}} = \text{MSE}(\Delta\widehat{\mathbf{f}}^{w}, \Delta\mathbf{f}^{w}), \tag{27}$$

$$\mathcal{L}_{\text{root}} = \text{MSE}(\widehat{\mathbf{r}}, \mathbf{r}) + 2\,\text{MSE}(\Delta\widehat{\mathbf{r}}, \Delta\mathbf{r}) + 0.5\,\text{MSE}(\Delta^2\widehat{\mathbf{r}}, \Delta^2\mathbf{r}). \tag{28}$$

The root terms discourage reducing foot motion simply by suppressing locomotion.

A.4.2 Foot and root–foot objectives

State reconstruction uses

$$\mathcal{L}_{\text{state}} = \text{MSE}(\widehat{\mathbf{c}}, \mathbf{c}) + \text{MSE}(\widehat{\mathbf{s}}, \mathbf{s}). \tag{29}$$

On supported frames, height supervision matches the reference foot height:

$$\mathcal{L}_{\text{height}} = \langle |\widehat{f}^{w}_{f,z} - f^{w}_{f,z}| \rangle_q. \tag{30}$$

It does not force every supported ankle and toe to the same zero height. The remaining terms are

$$\mathcal{L}_{\text{slide-acc}} = \langle (\Delta^2\widehat{\mathbf{f}}^{w}_{f,xy})^2 \rangle_{M^{sss}}, \tag{31}$$

$$\mathcal{L}_{\text{pen}} = \text{mean}\left[\max(h_0 - \widehat{f}^{w}_{f,z}, 0)^2\right], \tag{32}$$

$$\mathcal{L}_{\text{foot}} = \mathcal{L}_{\text{state}} + 0.5\mathcal{L}_{\text{height}} + 0.25\mathcal{L}_{\text{slide-acc}} + 0.5\mathcal{L}_{\text{pen}}. \tag{33}$$

Here h_0 comes from the target feet. Sliding acceleration is regularized only on persistent sliding intervals; the sliding velocity itself is not forced to zero.

With $\widehat{\mathbf{d}} = \Delta\widehat{\mathbf{r}}_{xy} + \Delta\widehat{\mathbf{p}}_{f,xy}$ and $\mathbf{d} = \Delta\mathbf{r}_{xy} + \Delta\mathbf{p}_{f,xy}$, the implemented root–foot objective is

$$\mathcal{L}_{\text{rf}} = \langle \widehat{\mathbf{d}}^{\,2} \rangle_{M^c} + 0.25\langle \widehat{\mathbf{d}}^{\,2} \rangle_{\text{sg}(\widehat{M}^c)} + 0.5\langle (\widehat{\mathbf{d}} - \mathbf{d})^2 \rangle_{M^s}. \tag{34}$$

Squares denote elementwise squared values before masked averaging. For predictions clipped to $[0, 1]$, the soft planted-interval confidence is

$$\widehat{M}^c_{t,f} = \min(\widehat{c}_{t,f}, \widehat{c}_{t+1,f})\,[1 - \max(\widehat{s}_{t,f}, \widehat{s}_{t+1,f})]\,. \tag{35}$$

The entire confidence is detached. This blocks the direct gradient route that would reduce the consistency term by lowering the predicted contact confidence.

A.4.3 Twist supervision and Stage-1 total

Let $\widehat{\boldsymbol{\phi}}^{\text{raw}}$ be the raw twist prediction and $\widehat{\boldsymbol{\phi}} = \mathcal{N}_2(\widehat{\boldsymbol{\phi}}^{\text{raw}})$. The twist objective is

$$\begin{aligned}\mathcal{L}_{\text{tw}} = {} & \text{MSE}(\widehat{\boldsymbol{\phi}}, \boldsymbol{\phi}) + 0.25\,\text{MSE}(\Delta\widehat{\boldsymbol{\phi}}, \Delta\boldsymbol{\phi}) \\ & + 0.1\,\text{mean}\left[(\|\widehat{\boldsymbol{\phi}}^{\text{raw}}_{t,j}\|_2 - 1)^2\right].\end{aligned} \tag{36}$$

For the documented coefficient setting, the complete geometric loss is

$$\mathcal{L}_{\text{geo}} = 0.636\mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{joint}} + 2\mathcal{L}_{\text{vel}} + \mathcal{L}_{\text{root}} + 5w_e\mathcal{L}_{\text{foot}} + 5w_e\mathcal{L}_{\text{rf}} + \mathcal{L}_{\text{tw}}, \tag{37}$$

where w_e is the physical warm-up factor. The coefficients in Eq. equation 37 specify the numerical objective described in this appendix.

A.4.4 BIDIRECTIONAL RECOVERY AND ERROR REGULARIZATION

The predicted-input inverse path recovers $(\widehat{\mathbf{S}}_6, \widehat{\mathbf{e}}_S)$ and $(\widehat{\mathbf{R}}_6, \widehat{\mathbf{e}}_R)$ from the current geometric input with zero joint-error input. A second, clean inverse path uses target joints and target twist. Denote its rotation and swing losses by $\mathcal{L}_R^{\text{clean}}$ and $\mathcal{L}_S^{\text{clean}}$.

In the forward path, $F_\phi([\mathbf{R}_6, \mathbf{0}]; \mathbf{h}^{\text{clean}})$ predicts swing and twist, and F_J maps that predicted swing with zero swing-error input to joint features, uncertainty, and joint error. Forward supervision includes joint-feature reconstruction $\mathcal{L}_J^{\rightarrow}$ and swing/twist reconstruction $\mathcal{L}_S^{\rightarrow} + \mathcal{L}_\phi^{\rightarrow}$. These terms supervise reconstruction of swing and twist. All these reconstruction terms use MSE on the corresponding features.

Clean-path error regularization is

$$\mathcal{L}_0 = \text{mean}\big[(\mathbf{e}_S^{\text{clean}})^2\big] + \text{mean}\big[(\mathbf{e}_R^{\text{clean}})^2\big] + \text{mean}\big[(\mathbf{e}_J^{\rightarrow})^2\big]. \tag{38}$$

Predicted-path errors are compared with standard Gaussian samples using a multiscale MMD kernel

$$k(\mathbf{x}, \mathbf{y}) = \sum_{\sigma \in \{0.05, 0.2, 0.9\}} \frac{\sigma^2}{\sigma^2 + \|\mathbf{x} - \mathbf{y}\|_2^2}. \tag{39}$$

The empirical estimate averages $k(X, X) + k(Y, Y) - 2k(X, Y)$, including diagonal entries. At most 256 flattened frame samples are selected at evenly spaced indices. The error-distribution term $\mathcal{L}_{\text{MMD}}$ covers $\mathbf{e}_S$ and $\mathbf{e}_R$. An additional term $\mathcal{L}_{\text{ind}}$ compares concatenated target-pose/independent-Gaussian samples against predicted-pose/predicted-error samples, with pose features detached. This encourages errors to model discrepancy rather than encode pose, without proving exact statistical independence.

For the same documented configuration, the recovery loss is

$$\begin{aligned}\mathcal{L}_{\text{rec}} = \mathcal{L}_R + 0.5\mathcal{L}_S + \mathcal{L}_\phi + 0.5\mathcal{L}_R^{\text{clean}} + 0.25\mathcal{L}_S^{\text{clean}} \\ + \mathcal{L}_J^{\rightarrow} + 0.5(\mathcal{L}_S^{\rightarrow} + \mathcal{L}_\phi^{\rightarrow}) + 0.1\mathcal{L}_0 \\ + 0.1(1 - \rho_e)(\mathcal{L}_{\text{MMD}} + \mathcal{L}_{\text{ind}}).\end{aligned} \tag{40}$$

Here ρ_e is the teacher-input ratio defined below. The direct twist term compares the twist consumed by recovery against target twist; when diffusion supplies that twist, this is not a separate twist estimate produced by the invertible mappings.

A.4.5 FK CONSISTENCY AND THE JOINT OBJECTIVE

Let $(\widetilde{\mathbf{r}}, \widetilde{\mathbf{p}}, \widetilde{\mathbf{c}}, \widetilde{\mathbf{s}})$ be the physical-space version of the current recovery input. The reference geometry is $\widetilde{\mathbf{f}}^w = \widetilde{\mathbf{r}} + \widetilde{\mathbf{p}}$, and the realized geometry is $\widehat{\mathbf{f}}^{w,\text{FK}} = \text{FK}(\widehat{\mathbf{R}}, \widetilde{\mathbf{r}})$. Their MSE gives $\mathcal{L}_{\text{FK}}$; first-difference MSE gives $\mathcal{L}_{\text{FKvel}}$.

The IK root–foot term uses planted and sliding masks thresholded from $\widetilde{\mathbf{c}}, \widetilde{\mathbf{s}}$ at 0.5. It contains planted cancellation plus 0.5 times sliding-displacement reconstruction against $\widetilde{\mathbf{f}}^w$, without the extra predicted-confidence term of Eq. equation 34. The IK penetration floor is estimated from these input feet. Thus,

$$\mathcal{L}_{\text{cs}} = \mathcal{L}_{\text{FK}} + 0.5\mathcal{L}_{\text{FKvel}} + 5w_e\mathcal{L}_{\text{rf}}^{\text{IK}} + 2\mathcal{L}_{\text{pen}}^{\text{IK}}, \qquad \mathcal{L} = \mathcal{L}_{\text{geo}} + \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{cs}}. \tag{41}$$

The outer weights of the recovery and cross-space consistency objectives are both one in this specification.

A.5 OPTIMIZATION AND TRAINING PROCEDURE

A.5.1 SCHEDULES AND CONFIGURATION

For epoch index e starting at one, the physical warm-up and teacher ratio are

$$w_e = \min(1, e/E_{\text{phys}}), \qquad \rho_e = \max\big(0, 1 - (e-1)/E_{\text{teach}}\big). \tag{42}$$

A zero physical-warm-up duration gives $w_e = 1$. With $E_{\text{teach}} = 200$, recovery uses clean inputs at epoch 1 and fully predicted inputs from epoch 201. For the normalized physical prefix $\mathbf{v}$,

$$\widetilde{\mathbf{v}} = \rho_e \mathbf{v} + (1 - \rho_e)\widehat{\mathbf{v}}, \qquad \widetilde{\boldsymbol{\phi}} = \mathcal{N}_2\left(\rho_e \boldsymbol{\phi} + (1 - \rho_e)\widehat{\boldsymbol{\phi}}\right). \tag{43}$$

Context and uncertainty are recomputed from the mixed input. The predicted component comes from a one-step clean-state estimate at a randomly sampled diffusion timestep, not from unrolling the entire sampling chain during training.

Table 5: Training and inference settings. The first block lists the experimental settings reported in the main paper; the second lists the documented model configuration.

Setting	Value
Frame rate / training sequence length	30 FPS / 150 frames
Frame-aligned audio dimension	35
Optimizer / learning rate	Adan / 10^{-4}
Main training epochs / batch size	1,000 / 64
Physical warm-up / teacher-input schedule	200 / 200 epochs
DDIM inference steps	50
Weight decay	0.02
Forward diffusion steps / schedule	1,000 / cosine
Denoiser EMA decay	0.9999
DDIM stochasticity η	0
Motion/music hidden width	512
Motion/music Mamba blocks	8 / 4
Coupling layers per invertible mapping	16
Conditional Transformer layers / heads	2 / 8
Overlap for window assembly	75 frames

During joint training, recovery gradients propagate through the predicted geometric input to the geometric generator.

A.5.2 Training algorithm

For each minibatch, the joint-training procedure is:

1. Convert target rotations and root translations to world joints. Construct root-relative joints and foot states, and extract non-root twist targets.
2. Normalize the physical channels, append twist, and sample a diffusion timestep and Gaussian noise. Predict the clean state from noisy motion, music, genre, and timestep.
3. Invert normalization for geometric losses. Compute Eq. equation 37, including normalized-pair twist reconstruction and raw-pair norm regularization.
4. Mix target and predicted inputs according to ρ_e. Normalize mixed twist pairs and compute the shared context and uncertainty features.
5. Apply F_J^{-1} with zero joint-error input, then F_ϕ^{-1} with the supplied twist. Evaluate the clean inverse and supervised forward paths for Eq. equation 40.
6. Run FK on predicted rotations with the mixed root trajectory and compute Eq. equation 41. Backpropagate the total loss, update the optimizer, and update the denoiser EMA.

The MMD weight increases as the teacher ratio decreases. The distributional regularizer therefore gains weight as recovery is exposed to generated geometry.

A.6 Sampling and Long-Sequence Assembly

A.6.1 Short-sequence inference

Sampling starts from Gaussian noise and uses the EMA denoiser. For diffusion indices $n > m$, the deterministic DDIM update is

$$\widehat{\boldsymbol{\epsilon}}_n = \frac{\mathbf{Y}_n - \sqrt{\bar{\alpha}_n}\widehat{\mathbf{Y}}_0}{\sqrt{1-\bar{\alpha}_n}}, \qquad \mathbf{Y}_m = \sqrt{\bar{\alpha}_m}\widehat{\mathbf{Y}}_0 + \sqrt{1-\bar{\alpha}_m}\widehat{\boldsymbol{\epsilon}}_n. \tag{44}$$

The predicted clean state is clipped to $[-1, 1]$ in normalized space. The final update returns the clean prediction. Although $\eta = 0$ makes the updates deterministic conditional on their inputs, different initial noise samples can still produce different dances.

After sampling, twist pairs are normalized. The recovery encoder predicts $\widehat{\mathbf{u}}$ and constructs $\mathbf{h}$, and rotation recovery follows Eqs. (9)–(10) in Section 3.3 of the main paper with zero joint-error input $\mathbf{e}_J = 0$. Root translation and state predictions are assembled with the recovered rotations. Zero input error does not imply zero predicted output errors.

A.6.2 Overlapping-window generation

The overlapping-window procedure divides each music sequence into windows of $W = 150$ frames with overlap $O = 75$. The stride is $W - O$, and final padding repeats the last music-feature frame. Windows from one source are processed together; overlap constraints never connect different songs.

At each reverse update, the overlapping contact/slide channels, root height, root-relative joints, and twist features are synchronized. Horizontal root coordinates require different treatment because each training clip may have its own origin. Their *increments* are matched while preserving the next window's initial horizontal root value. Copying absolute root coordinates would repeatedly pull a later window toward an earlier origin.

After sampling and inverse normalization, let $\mathbf{r}^{\mathrm{old}}$ be the assembled trajectory and $\mathbf{r}^{\mathrm{new}}$ the next window. Its horizontal translation is shifted by

$$\mathbf{d}_{xy} = \mathbf{r}^{\mathrm{old}}_{-O,xy} - \mathbf{r}^{\mathrm{new}}_{0,xy}. \tag{45}$$

The shared interval is linearly blended with weights increasing from zero to one. Twist blending operates on cosine–sine pairs followed by pair normalization, rather than directly averaging wrapped angles. The unshared suffix is appended, and padding is removed. The stitched physical state is normalized again for recovery. Because normalization clips to its fitted range, exceptionally long trajectories outside the training root range remain a limitation of this path.

A.7 Evaluation Definitions and Comparison Protocol

A.7.1 Kinetic and geometric distribution measures

The distribution metrics compare kinetic and geometric feature distributions of generated and ground-truth motion. For feature means $\boldsymbol{\mu}_g, \boldsymbol{\mu}_r$ and covariances Σ_g, Σ_r,

$$\mathrm{FID} = \|\boldsymbol{\mu}_g - \boldsymbol{\mu}_r\|_2^2 + \mathrm{tr}\left(\Sigma_g + \Sigma_r - 2(\Sigma_g\Sigma_r)^{1/2}\right). \tag{46}$$

Diversity is the mean Euclidean distance over distinct pairs of features:

$$\mathrm{Div} = \frac{2}{N(N-1)}\sum_{i<j}\|\mathbf{z}_i - \mathbf{z}_j\|_2. \tag{47}$$

A larger diversity value is not sufficient evidence of realistic variation: extreme or invalid motions can also increase distances. It should be interpreted with FID and the ground-truth diversity.

Feature scaling is part of the metric definition. The standardized FlowerDance-style evaluator applies ground-truth feature mean and standard deviation, with 10^{-4} added to the standard deviation. These are evaluation-feature statistics, distinct from the training motion normalizers. The main paper labels AIST++ geometric scores as *raw*; those scores must remain distinct from standardized geometric scores. In particular, raw AIST++ geometric FID cannot be compared numerically

against FineDance scores under another scaling. Nearly constant geometric-feature dimensions can make standardized scores sensitive to the variance convention and denominator floor, so both the raw/standardized flag and evaluator version are necessary for reproducing a table.

A.7.2 BEAT ALIGNMENT

Music beats are read from the last channel (zero-based index 34) of the aligned 35D audio features. Motion beats are local minima of mean joint displacement magnitude after Gaussian smoothing with standard deviation five frames. With music-beat indices B_m and motion-beat indices B_d,

$$\mathrm{BA} = \frac{1}{|B_m|} \sum_{b \in B_m} \exp\left[-\frac{\min_{d \in B_d} (b-d)^2}{2 \times 3^2} \right]. \tag{48}$$

The score is computed per clip and then averaged. The score is defined as zero if either beat set is empty. Motion-beat indices follow the first-difference array convention.

A.7.3 PHYSICAL FOOT-CONTACT SCORE

For root acceleration, the vertical component is clamped below at zero before taking the norm. This norm is divided by its maximum within the clip, giving $\bar{a}_t$; a clip with negligible maximum acceleration receives zero normalized acceleration. Let

$$d_t^L = \min_{f \in \{7,10\}} \|\Delta \mathbf{f}_{t,f,xy}^w\|_2, \qquad d_t^R = \min_{f \in \{8,11\}} \|\Delta \mathbf{f}_{t,f,xy}^w\|_2. \tag{49}$$

The score is

$$\mathrm{PFC} = 10^4 \, \mathrm{mean}_t[\bar{a}_t d_t^L d_t^R]. \tag{50}$$

Root acceleration uses seconds, but d_t^L, d_t^R retain the evaluator's per-frame displacement convention. Replacing them by speeds changes the numerical scale. PFC does not use a contact mask and can become small for nearly static motion, motivating the activity measures reported alongside it.

A.7.4 SUPPORT AND SKATING STATISTICS

Evaluation recomputes support from each motion's geometry using a per-clip second-percentile floor, a 0.07 m height band, and a 0.20 m/s vertical-speed threshold. Unlike the training labels in Section A.2, the activity evaluator does not apply the three-frame majority filter. Grounded foot speed is the mean horizontal speed over these supported foot frames. Foot-skating ratio is the fraction exceeding 0.05 m/s, and support ratio is the fraction of all foot frames classified as supported.

The 0.05 m/s skating threshold is deliberately separate from the 0.12 m/s planted-label threshold. The skating statistic also counts supported slides; consequently, ground-truth dances can have a nonzero skating ratio. It is an operational foot-motion measure, not a semantic detector of unintended skating. Support ratio and motion activity are needed to interpret it. A clip with no supported frames has undefined grounded speed and skating ratio; aggregation excludes undefined values.

The qualitative visualization uses a separate reference-floor convention, detailed in Section A.9.

A.7.5 ACTIVITY AND RECONSTRUCTION DIAGNOSTICS

Root displacement is the Euclidean distance between the initial and final horizontal root positions, whereas root path length sums all horizontal root increments. The former can be small for an energetic dance that returns to its starting location. Mean root-relative joint speed removes global translation, and active-joint ratio measures the fraction of joint–time entries above 0.05 m/s. Motion-beat frequency divides the number of detected motion beats by the clip duration.

Additional activity diagnostics available in the evaluator include root-speed percentiles, mean root acceleration, mean squared root-relative joint speed, and a freeze ratio. The latter uses the tenth percentile of ground-truth frame-level mean root-relative speed as a shared threshold. These quantities help distinguish improved foot stability from suppressed motion.

For paired recovery evaluation, MPJPE is the mean joint distance between corresponding reconstructed and target poses. Rotation error can be computed as

$$\theta(\widehat{R}, R) = \arccos\left(\mathrm{clip}\left[\frac{\mathrm{tr}(\widehat{R}^\top R) - 1}{2}, -1, 1 \right] \right). \tag{51}$$

Such paired measures assess recovery of a particular target. They should not be interpreted as complete measures of music-to-dance quality, since one musical segment can admit multiple valid dances.

A.8 User Study Design and Vote Aggregation

A.8.1 Comparison design and response collection

We compared PhysIKDance with MEGADance, FineNet, FlowerDance, and LODGE using 12 video pairs per baseline. For each baseline, four pairs assessed beat alignment, four assessed expressiveness, and four assessed physical plausibility, giving 48 video-pair questions in total. Beat alignment concerns the correspondence between movement timing and music; expressiveness concerns the perceived richness of the dance; physical plausibility concerns the perceived naturalness of body motion and foot–ground interactions.

Each respondent answered four randomly sampled questions. Recruitment through social networks yielded 183 questionnaires and $183 \times 4 = 732$ judgments. All recorded completion times exceeded one minute. Because only a subset of questions was answered in each questionnaire, the number of judgments differs across criteria. The preference table in the main paper reports the corresponding vote counts.

A.8.2 Vote counts and pooled preference rates

Let $w_{b,c}$ denote the number of judgments preferring PhysIKDance over baseline b for criterion c, and let $n_{b,c}$ denote the corresponding total number of judgments. The criterion-specific preference rate and the overall rate are

$$P_{b,c} = 100\frac{w_{b,c}}{n_{b,c}}, \qquad P_b^{\text{overall}} = 100\frac{\sum_c w_{b,c}}{\sum_c n_{b,c}}. \tag{52}$$

The overall rate pools votes across criteria rather than taking an unweighted mean of the three percentages. Percentages are rounded to two decimal places after aggregation.

For example, the MEGADance comparison pools $40+53+38 = 131$ preferences for PhysIKDance over $57 + 70 + 56 = 183$ judgments, yielding 71.58%.

The preference rates describe perceived quality on the sampled video comparisons. The 732 judgments are repeated responses from 183 questionnaires, rather than 732 independent participants; the reported percentages are descriptive and do not establish statistical significance.

A.9 Additional Qualitative Comparisons

A.9.1 Rendering and example selection

Figures 6 and 7 provide four additional paired examples comparing PhysIKDance with MEGADance and FineNet on FineDance, complementing the FlowerDance and LODGE comparisons in the main paper. Each panel shows five temporally ordered poses and enlarged foot regions. Baselines are shown in purple and PhysIKDance in teal. Within each pair, the methods use matching source segments and frame indices, a common body model, and the same camera and lighting settings. The views align the displayed facing direction and spread the poses horizontally to expose their temporal evolution; horizontal display positions therefore do not represent the original root trajectory.

The reference floor is fixed within each complete 150-frame clip using the median of its per-frame minimum sole heights. The same estimation rule is applied separately to each method. No per-frame vertical correction or pose editing is applied. This visualization reference is not an independently calibrated ground plane and differs from the joint-based floor estimate used by the quantitative evaluator. Foot-region insets are crops of the corresponding full-body renderings.

The examples were selected to make local foot–ground differences visible, rather than sampled to estimate their frequency. For MEGADance and FineNet, candidate screening covered 348 and 366 paired clips, respectively, followed by mesh inspection and selection of candidate examples. Two

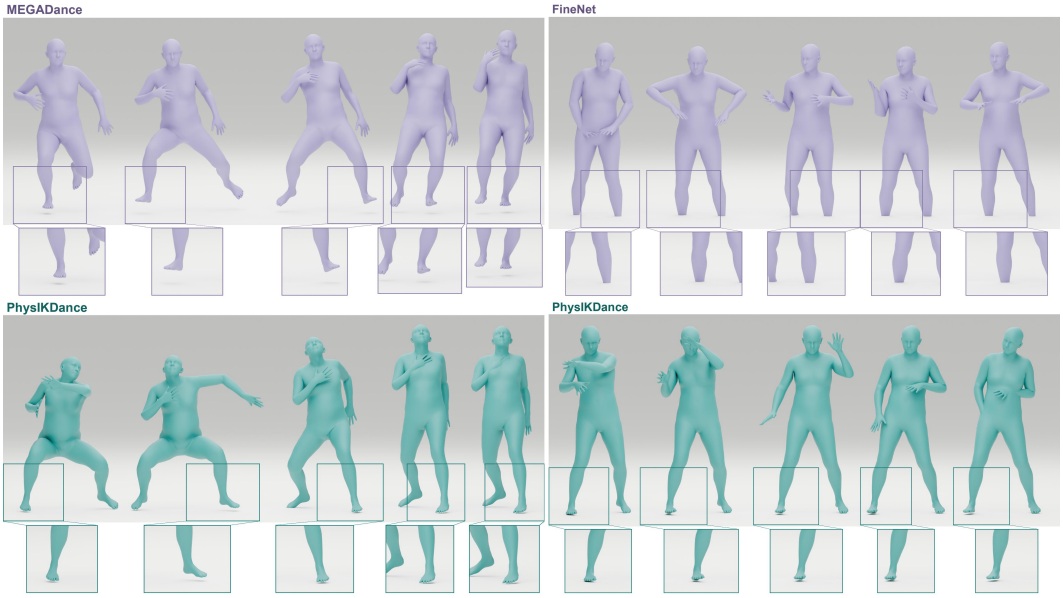


Figure 6: Additional comparisons, set 1. Left: MEGADance and PhysIKDance. Right: FineNet and PhysIKDance. Foot crops show above-plane sole positions in the selected MEGADance excerpt and below-plane positions in the FineNet excerpt.

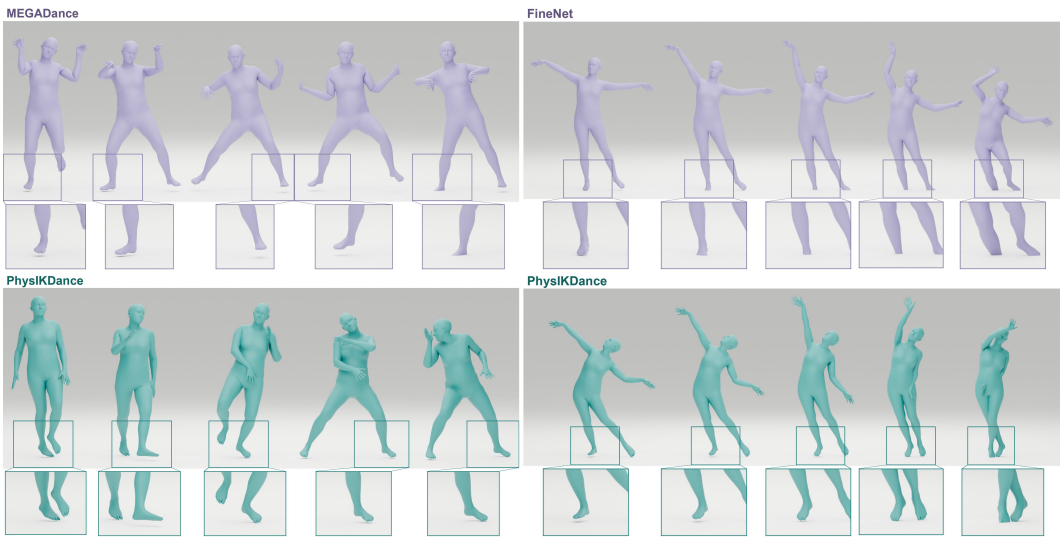


Figure 7: Additional comparisons, set 2. Left: MEGADance and PhysIKDance. Right: FineNet and PhysIKDance. These further examples show changing foot height relative to the fixed reference floor. All panels follow the conventions in Section A.9.

examples per baseline are included here. These qualitative examples complement the full-test-set metrics; they do not establish average performance or replace temporal evaluation of foot skating. Raised feet can be part of intentional dance motion, so isolated snapshots should be interpreted in their sequence context.

A.9.2 ADDITIONAL EXAMPLES WITH MEGADANCE AND FINENET

The following two figures show different paired motion excerpts under the same rendering convention. The insets expose both above-plane and below-plane foot positions in the selected examples; PhysIKDance also retains visible deviations in some frames.

A.10 LIMITATIONS

The method's physical supervision is kinematic. It constrains positions, displacements, support states, and penetration, but does not solve for contact forces, torques, friction cones, dynamic balance, or biomechanical feasibility. Low kinematic error should therefore not be read as proof of physically executable motion. Geometry-derived support labels can be imperfect, particularly for floor-adjacent aerial motion, noisy reconstruction, or nonplanar terrain.

The body representation contains 24 joints and does not generate articulated fingers or facial motion. The geometric generator can violate bone lengths before recovery, and fixed-skeleton recovery can shift foot positions while satisfying articulation constraints. Auxiliary errors provide capacity to accommodate inconsistent inputs; they do not guarantee that recovery will preserve every generated contact. Finally, overlapping windows encourage local continuity but do not explicitly model song-level choreography, and normalized linear twist blending is numerically delicate for nearly antipodal pairs. These limitations motivate evaluation of motion activity and temporal consistency in addition to contact-sensitive scores.