A. Implementation Details

A.1. Architecture and Training Details

Our final implementation of OpenDanceNet comprises two main components: DDT and MCT.

Disentangled Dance Tokenizer (DDT). The DDT module adopts a hybrid architecture that combines convolutional layers with self-attention. A vector-quantized (VQ) bottleneck with a codebook size of 1024 is used, compressing the temporal length of the input sequence by a factor of 4. To enhance reconstruction stability, we impose strong penalties on motion velocity and foot-skating artifacts. For the three input types—joint rotations, global trajectories, and 2D keypoints—we use hidden dimensions of 512, 256, and 512, respectively, reflecting their differing modality complexity. Empirically, we observe that 2D keypoints are more difficult to model than 3D motion; accordingly, we introduce an additional regularization term on the keypoint reconstruction loss. The DDT is trained for 600 epochs with a batch size of 512 and a learning rate of 1×10^{-4}.

Multimodal-Condition Transformer (MCT). The MCT module performs early fusion over all style-conditioning signals by concatenating their embeddings, followed by a 10-layer Transformer with a hidden dimension of 512 and 8 self-attention heads. Since music serves as the main modality for guiding dance generation, we set the masking probability p_{mask} for the music stream to zero during training. However, the framework retains the flexibility to adjust this masking probability if desired. The MCT is trained for 600 epochs with a batch size of 128 and a learning rate of 1×10^{-4}.

A.2. Test-time Motion Optimization

At inference time (cf. Section 4.4), we first generate a sequence of discrete VQ tokens $\mathbf{z} = \{z_t\}_{t=1}^{T}$ using a Mask-Predict scheme. These tokens are then decoded by the DDT decoders into continuous motion features:

$$\mathbf{x} = f_{\text{dec}}(\mathbf{z}) = \{(\mathbf{r}_t, \mathbf{p}_t)\}_{t=1}^{T}, \tag{7}$$

where $\mathbf{r}_t \in \mathbb{R}^{24\times 6}$ denotes 6D joint rotations and $\mathbf{p}_t \in \mathbb{R}^3$ denotes the root joint position at time t. We convert $\mathbf{r}_t$ to axis–angle parameters $\boldsymbol{\theta}_t \in \mathbb{R}^{24\times 3}$ and apply a differentiable SMPL forward kinematics (FK) layer:

$$\mathbf{X}_t = \text{FK}(\boldsymbol{\theta}_t, \mathbf{p}_t) \in \mathbb{R}^{24\times 3}, \tag{8}$$

where $\mathbf{X}_t(j)$ is the 3D position of joint j at frame t.

Local joint refinement. During each sampling iteration (except the first), we introduce a small axis–angle increment $\Delta\boldsymbol{\theta}_t \in \mathbb{R}^{24\times 3}$ and update the joint rotations as

$$\mathbf{R}'_{t,j} = \Delta\mathbf{R}_{t,j}\,\mathbf{R}_{t,j}, \quad \Delta\mathbf{R}_{t,j} = \exp([\Delta\boldsymbol{\theta}_{t,j}]_\times), \tag{9}$$

where $\mathbf{R}_{t,j} \in SO(3)$ is the original rotation of joint j at frame t, and $[\cdot]_\times$ is the skew-symmetric matrix operator. We primarily apply these increments to lower-limb joints (e.g., knees and ankles), and recompute joint positions $\mathbf{X}'_t$ via FK.

Let F be the set of foot joints and u denote the vertical axis (either y or z). For each $j \in F$, we denote the height by $h_{t,j} = \mathbf{X}'_t(j)^{(u)}$, and the horizontal-plane coordinates by $\mathbf{X}'_t(j)^{(P)} \in \mathbb{R}^2$. The horizontal velocity between consecutive frames is

$$v_{t,j} = \left\|\mathbf{X}'_{t+1}(j)^{(P)} - \mathbf{X}'_t(j)^{(P)}\right\|_2 \cdot \text{fps}, \tag{10}$$

and we define a soft contact gate

$$w_{t,j} = \sigma\left(\frac{h_c - h_{t,j}}{\tau}\right), \tag{11}$$

where h_c is the contact height threshold, τ is a sharpness parameter, and $\sigma(\cdot)$ is the sigmoid.

The contact-weighted foot-velocity loss and ground-penetration loss are

$$\mathcal{L}_{\text{vel}} = \frac{1}{(T-1)\,|F|}\sum_{t=1}^{T-1}\sum_{j\in F} w_{t,j}\, v_{t,j}^2, \tag{12}$$

$$\mathcal{L}_{\text{ground}} = \frac{1}{T\,|F|}\sum_{t=1}^{T}\sum_{j\in F} \max(0, -h_{t,j}). \tag{13}$$

To keep the incremental rotations small and temporally smooth, we further add

$$\mathcal{L}_{\text{smooth}} = \frac{1}{T-1}\sum_{t=1}^{T-1}\left\|\Delta\boldsymbol{\theta}_{t+1} - \Delta\boldsymbol{\theta}_t\right\|_2^2, \tag{14}$$

$$\mathcal{L}_{\text{mag}} = \frac{1}{T}\sum_{t=1}^{T}\left\|\Delta\boldsymbol{\theta}_t\right\|_2^2. \tag{15}$$

The local optimization objective in each iteration is

$$\mathcal{L}_{\text{local}} = \lambda_v\mathcal{L}_{\text{vel}} + \lambda_g\mathcal{L}_{\text{ground}} + \lambda_s\mathcal{L}_{\text{smooth}} + \lambda_m\mathcal{L}_{\text{mag}}, \tag{16}$$

which we minimize with a small number of Adam steps, while keeping the discrete tokens and network weights fixed.

Latent-level refinement. After the full token sequence is generated, we perform a second-stage refinement directly in the DDT latent space: We treat the quantized latents $(\boldsymbol{\ell}^{\text{rot}}, \boldsymbol{\ell}^{\text{kpt}}, \boldsymbol{\ell}^{\text{pos}})$ as optimization variables, decode them at each step, and recompute the same FK-based losses. We use an objective of the form

$$\mathcal{L}_{\text{latent}} = \lambda_v\mathcal{L}^{\text{lat}}_{\text{vel}} + \lambda_g\mathcal{L}^{\text{lat}}_{\text{ground}} + \lambda_s\mathcal{L}^{\text{lat}}_{\text{smooth}} + \lambda_m\mathcal{L}^{\text{lat}}_{\text{mag}}, \tag{17}$$

where the terms mirror those above but are computed from the decoded latents, with additional temporal smoothness

and magnitude regularizers on ℓ^{rot} and ℓ^{pos}. In practice, a few gradient steps with a small learning rate are sufficient to reduce foot sliding and ground penetration, while preserving the high-level semantics and diversity encoded by the discrete tokens.

B. Metric Details

Frechet Inception Distance (FID) Score. The FID score evaluates the proximity between synthesized and real dance distributions. Following [37], we use a style classifier to derive motion features. FID_k computes kinematic attributes (kinetic energy, acceleration) over the full sequence, while FID_g uses per-frame geometric pose descriptors (distances, angles, spatial relationships).

Diversity. Diversity assesses variance among generated sequences by computing mean feature distance using the same classifier as FID [37]. We denote kinetic and geometric diversity as Div_k and Div_g.

Beat Alignment Score (BAS). BAS quantifies music-motion congruence by computing the average temporal distance between each musical beat and the nearest dance beat.

Physical Foot Contact (PFC) Score. PFC [41] evaluates physical plausibility by examining consistency between character acceleration and foot-ground contacts.

In-depth Analysis of FID Results FID_k and FID_g assess different motion feature levels. As shown in Table 3, after test-time refinement, FID_k degrades (23.19 $\rightarrow$ 37.40) while FID_g improves (11.89 $\rightarrow$ 7.72). This illustrates the refinement trade-off: corrections that alleviate foot sliding introduce localized kinematic changes (degrading FID_k) while increasing geometric similarity to ground truth (improving FID_g).

C. More Experiments

We conduct all ablation experiments in a smaller subset (about 10 %) of the proposed OpenDanceSet to facilitate fast evaluation.

C.1. Ablation of DDT

To validate the architectural design of DDT (Section 4.1), we conduct an ablation study comparing our disentangled VQ-VAE tokenizer against a baseline (where all modalities are fused early in a single VQ-VAE) across various codebook size, as shown in Table 8.

We observe an improvement in reconstruction quality (Recon Loss) as the codebook size increases from 512 to 2048, indicating that a larger vocabulary capacity is essential for capturing diverse dance motions. Then the disentangled design demonstrates significant superiority over the "Single Tokenizer" baseline, particularly at larger codebook size. While the single model struggles to balance the reconstruction of global trajectories and local pose details within a shared latent space, our disentangled approach ensures that each modality (joint rotations, keypoints, and trajectories) is encoded by specialized codebooks.

Notably, with codebook size $N = 2048$, our disentangled model achieves the best performance, reducing the Reconstruction Loss to 0.0288 and PA-MPJPE to 51.59, respectively. This confirms that explicit modality separation provides a more robust and precise foundation for motion representation compared to early fusion strategies (single tokenizer).

However, as shown in Tab. 9, increasing the codebook size does not lead to a consistent or significant improvement in generation quality. A likely reason is that, as the codebook size grows, the MCT must discriminate among many more entries, making it harder to retrieve the most appropriate code indices and potentially increasing quantization error. In practice, we therefore set the codebook size to $N = 1024$, which offers a favorable trade-off between performance and stability.

Table 8. Ablation on DDT reconstruction. We compare the single tokenizer (non-disentangled) baseline and our disentangled design across different codebook sizes (N).

Disentangled	Codebook Size	Recon Loss ↓	MPJPE ↓	PA-MPJPE ↓
Single Tokenizer	512	0.0505	84.17	66.92
Single Tokenizer	1024	0.0466	79.50	62.88
Single Tokenizer	2048	0.0426	75.48	59.47
DDT (Ours)	512	0.0355	87.08	69.03
DDT (Ours)	1024	0.0315	74.88	57.11
DDT (Ours)	2048	**0.0288**	**66.42**	**51.59**

Table 9. Comparison of VQ tokenizers on AIST++ and OpenDanceSet for different codebook sizes (N).

Dataset	N	PFC ↓	FID_k ↓	FID_g ↓	DIV_k ↑	DIV_g ↑	BAS ↑
AIST++	GT	1.332	17.10	10.60	9.44	7.31	0.2403
	512	**0.999**	32.92	13.88	5.38	**6.31**	0.2387
	1024	1.140	24.82	**12.54**	5.24	5.29	0.2513
	2048	1.643	**14.47**	15.41	**6.35**	4.86	**0.2538**
OpenDanceSet	GT	0.0494	29.63	1.73	8.10	7.49	0.2397
	1024	**0.3462**	**23.19**	**11.89**	**7.82**	**6.41**	0.2472
	2048	0.3720	27.92	13.28	7.27	6.34	0.2509
	3072	0.9819	32.14	17.28	7.23	5.93	**0.2567**

Table 10. Ablation on text control effectiveness.

Text	CLIP Score ↑	FID_k ↓	FID_g ↓	Div_k →	Div_g →	BAS ↑
w/o text	0.7538	49.31	21.09	5.77	7.86	0.2182
w. text	**0.9273**	48.46	21.79	5.95	8.00	0.2288

Notably, with codebook size $N = 2048$, our disentangled model achieves the best performance, reducing the Reconstruction Loss to 0.0288 and PA-MPJPE to 51.59, respectively. This confirms that explicit modality separation provides a more robust and precise foundation for motion representation compared to early fusion strategies (single tokenizer).

However, as shown in Tab. 9, increasing the codebook size does not lead to a consistent or significant improvement in generation quality. A likely reason is that, as the codebook size grows, the MCT must discriminate among many more entries, making it harder to retrieve the most appropriate code indices and potentially increasing quantization error. In practice, we therefore set the codebook size to $N = 1024$, which offers a favorable trade-off between performance and stability.

C.2. Ablation of Sampling Steps

We compare different sampling step counts with and without test-time refinement in Table 12. Without refinement,

Table 11. User study with 95% confidence intervals.

Metrics	Bailando [37]	TM2D [12]	EDGE [41]	Ours
Natural Perform.	1.9% ± 1.56%	13.3% ± 5.40%	22.4% ± 3.84%	**62.2%** ± **4.26%**
Text Follow.	2.2% ± 1.56%	12.4% ± 4.20%	26.4% ± 2.56%	**58.8%** ± **5.11%**

increasing steps from 32 to 100 yields marginal improvements. With refinement, 50 steps achieve the best FID_g (5.53) while 32 steps offer the best speed-quality trade-off.

C.3. Ablation of Text Control

To quantify text control effectiveness, we remove the text condition during training and employ a re-annotation evaluation protocol: generated sequences are re-captioned by an LLM [5], and we compute the CLIP [34] score between original prompts and re-generated descriptions.

As shown in Table 10, the CLIP Score rises from 0.7538 (w/o text) to **0.9273** (w. text), demonstrating effective semantic integration. Text guidance also improves BAS from 0.2182 to 0.2288, suggesting that explicit genre constraints help select motion patterns better aligned with musical rhythm.

C.4. User Study of OpenDanceNet

Participants compare side-by-side 2D skeleton projections and 3D motion visualizations (Figure 6), answering: (1) Which dance has more natural alignment with the music beat? (2) Which better matches the textual description? We invite 28 professional dancers, each evaluating 40 pairs.

Table 11 reports preferences with 95% confidence intervals. Our model is chosen 62.2%±4.26% for natural performance—nearly three times higher than EDGE (22.4%±3.84%). For text-following, we achieve 58.8%±5.11%, far exceeding EDGE (26.4%±2.56%).

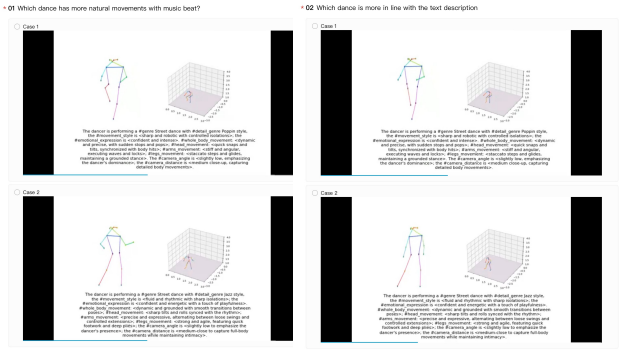

Figure 6. A screenshot of the user study interface. Left: 2D keypoints for loco-movements; Right: 3D motion for physically plausible performance.

D. Dataset Details

D.1. Text Annotation Details

OpenDanceSet provides text annotations covering genre, detailed genre, movement style, and camera metadata, curated by professional artists, human annotators, and VLM. Below we show the Vision–Language prompt used to generate captions:

> **System**: You are a professional artist who helps to generate the caption for the dance sequence.
> **User**: The #genre is {genre} and the #detail_genre is {detail_genre}. The COCO keypoints in seconds are {kpts2d}. Please generate a caption in English following the format:
>
> > The dancer is performing a #genre {genre} dance with #detail_genre {detail_genre} style,
> > the #movement_style is {}, #emotional expression is {}.
> > #whole_body movement:{}, #head movement:{}, #arms movement:{}, #legs movement:{};
> > the #camera_angle is {}, the #camera_distance is {}.
>
> Fill in all {} and keep the format.
> **Assistant**: The dancer is performing a ...

D.2. User Study of OpenDanceSet Quality

To evaluate perceptual quality of OpenDanceSet compared to AIST++ [21], we randomly sampled 20 motion-music pairs from both test sets. Participants selected the better motion based on "Motion Realism" and "Diversity". As shown in Table 13, OpenDanceSet outperforms AIST++ with win rates of 62.4% for Motion Realism and 58.8% for Diversity, confirming that our post-optimization pipeline yields high-quality 3D motion from in-the-wild videos.

E. Broader Impacts

Our work on controllable, multimodal dance generation has several positive impacts:

- By lowering the barrier to create professional-quality dance motions, our model can empower artists, educators, and game/VR developers without access to expensive motion-capture studios.
- The ability to specify high-level textual, kinematic, or positional conditions enables novel forms of artistic expression, interactive storytelling, and educational tools for dance training.
- A large, genre-diverse dataset encourages exploration of global dance styles, fostering cultural appreciation and promoting diversity in digital media.

Table 12. Comparison with baselines and ablation study on sampling steps and test-time refinement. We report the generation quality and inference time (in seconds) across different step settings.

Method / Steps	Refine	PFC ↓	FID_k ↓	FID_g ↓	Div_k →	Div_g →	BAS ↑	Time (s)
GT	-	0.0494	29.63	1.73	8.10	7.49	0.2397	-
TM2D	-	2.3345	69.86	36.64	2.59	1.15	0.2078	6.41
MoMask	-	0.0945	78.80	23.00	2.79	1.59	0.2337	**0.14**
EDGE	-	**0.0902**	57.07	17.57	5.73	**9.32**	0.2322	9.47
FineDance	-	0.1329	51.00	9.05	**6.80**	8.10	0.2365	9.36
OpenDanceNet 32	×	0.1829	41.71	10.32	5.97	7.32	0.2408	0.42
OpenDanceNet 50	×	0.1858	41.43	10.31	6.00	7.33	**0.2412**	0.63
OpenDanceNet 100	×	0.1865	**41.36**	10.31	6.04	7.37	0.2371	1.27
OpenDanceNet 32	✓	0.1371	49.29	5.66	5.45	7.43	0.2324	20.24
OpenDanceNet 50	✓	0.1309	51.19	**5.53**	5.33	7.49	0.2373	33.44
OpenDanceNet 100	✓	0.1306	51.24	**5.53**	5.33	7.51	0.2352	58.34

Table 13. **User Study on Dataset Quality.** Win rates (%) where participants preferred OpenDanceSet over AIST++ [21] on Ground Truth data.

Dataset	Motion Realism	Diversity
AIST++	37.6%	41.2%
OpenDanceSet (Ours)	**62.4%**	**58.8%**

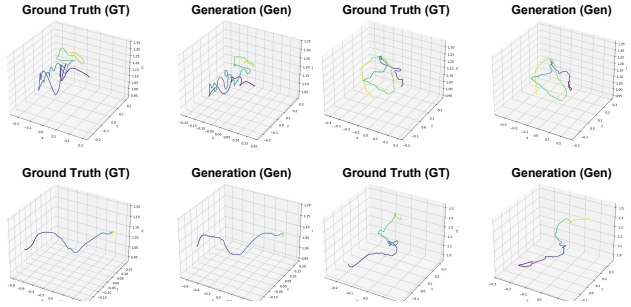

Figure 7. Ground-truth vs. generated global trajectories. Lighter color indicates earlier time steps.

F. Additional Visualizations

We render samples from OpenDanceSet to demonstrate data quality (Figure 11) and generated dance results under different control signals (Figure 8, Figure 9). We also visualize cross-dataset generalization by training on OpenDanceSet and evaluating on AIST++ [21] (Figure 10).

Trajectory Visualization. Figure 7 compares ground-truth and generated global trajectories. The generated trajectories closely follow the ground-truth paths, with high-frequency displacements corresponding to beat-aligned dance movements rather than noise.

Failure Case. Despite strong controllability, limitations remain (Figure 12). OpenDanceNet struggles with unreasonable or highly complex spatial trajectories, such as too-low or wavy paths, where the generated motion may deviate from the target.

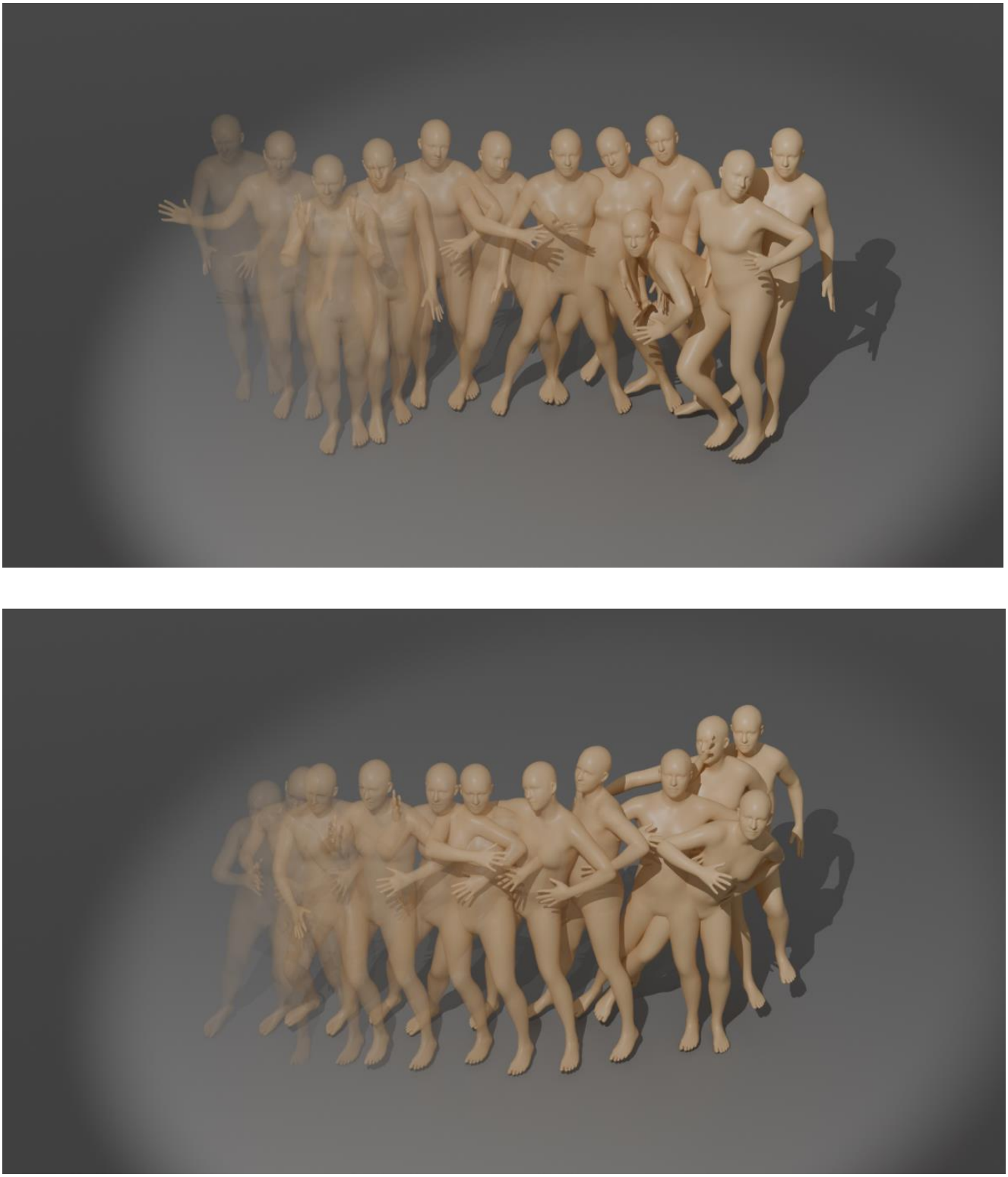

Figure 8. Visualization of generated dance motions from **full modality control**, including music, text, keypoints, and trajectory.

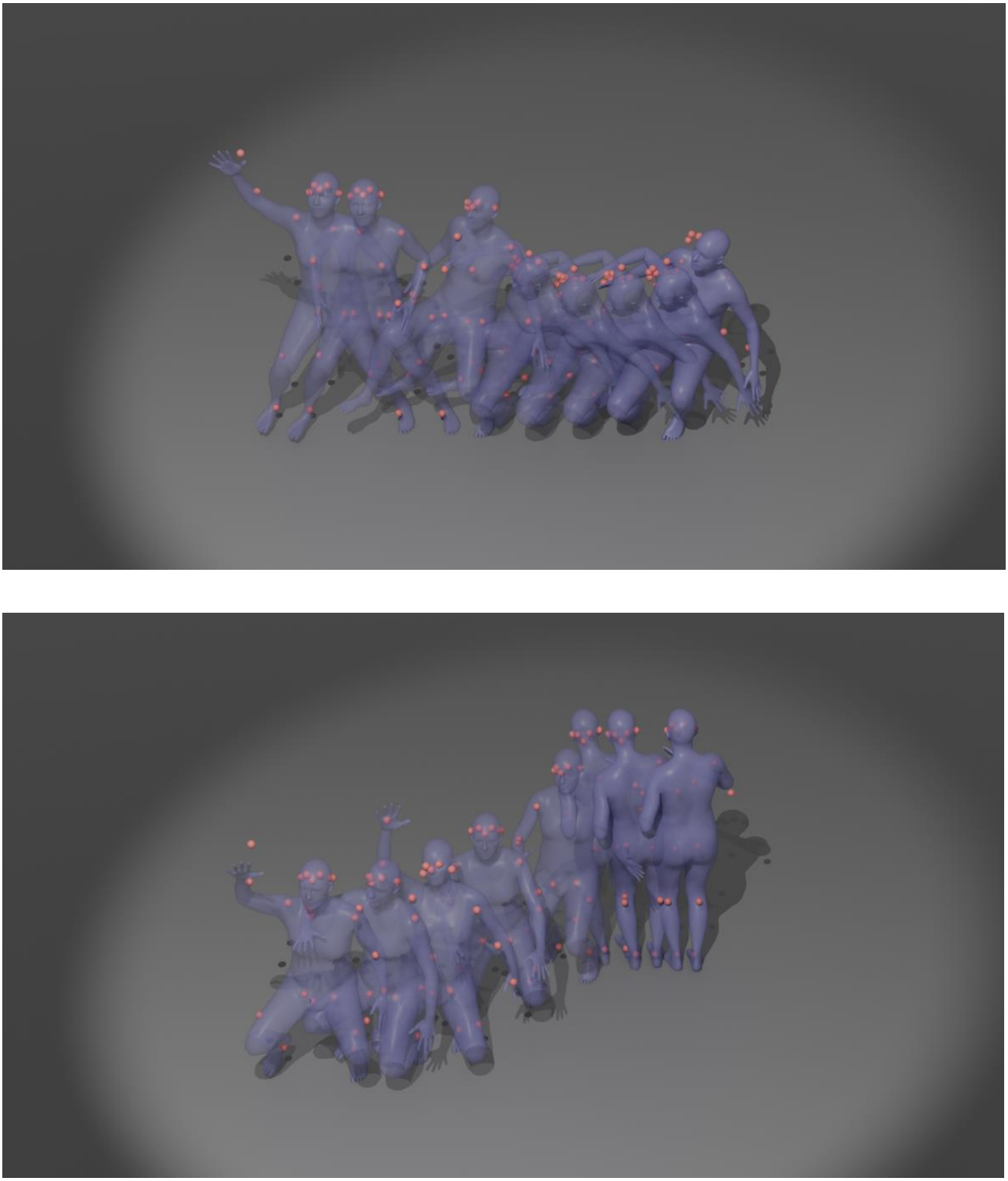

Figure 9. Visualization of generated dance motions under **Music+Keypoint** Control. The red sphere indicates the keypoint positions projected to 3D space.

Figure 10. Visualization of cross-dataset generated dance motions. OpenDanceNet is trained on OpenDanceSet and evaluated using music from AIST++.

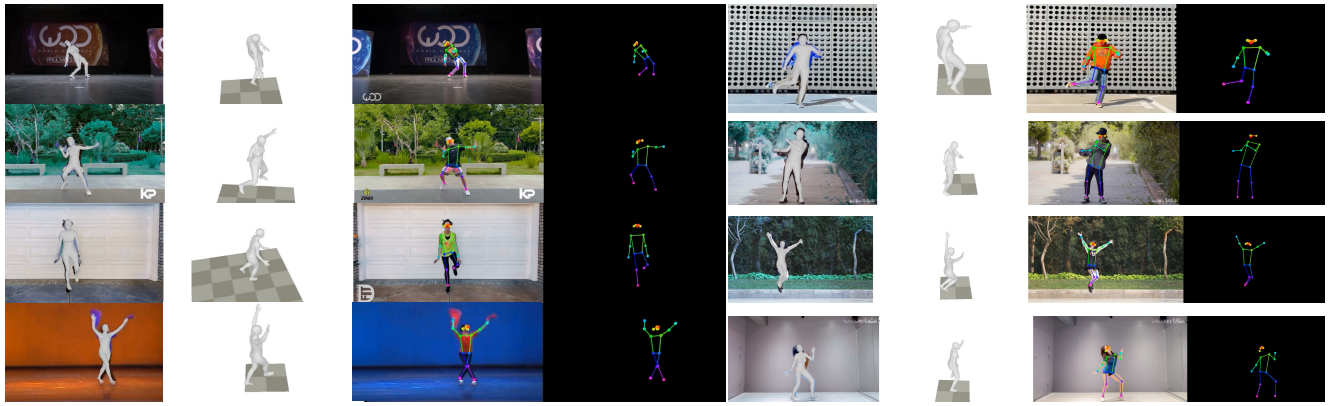

Figure 11. OpenDanceSet dance data demonstration, with accurate 3D global motion estimation and 2D keypoint estimation results.

Figure 12. Failure case. The character lies on the ground because of a low-height trajectory condition.