TeMuDance: Zero-Shot Textual Control for Music-Driven Dance Generation

Anonymous ACL submission

Abstract

Existing music-driven dance generation approaches demonstrate strong realism and effective alignment between audio and motion. However, they generally lack semantic controllability, making it difficult to guide specific movements through natural language descriptions. This limitation primarily stems from the absence of large-scale datasets that jointly align music, text, and motion, which prevents direct supervised learning of text-conditioned control. To address this challenge, we propose TeMuDance that enables zero-shot text-based control for music-conditioned dance generation. TeMuDance establishes a motion-centered bridging paradigm that aligns separate music-dance and text-motion datasets within a shared embedding space. Using motion as a pivot, we synthesize pseudo-triplets by retrieving and completing the missing modality for each corpus. Exploiting these synthesized priors, we train a text control branch that integrates semantic guidance into a frozen pretrained dance generation backbone, improving instruction compliance while preserving rhythmic consistency and motion realism. In addition, we introduce a motion-centered dual-stream fine-tuning strategy that jointly augments the two corpora and stabilises training in the presence of noisy pseudo annotations. Extensive experiments demonstrate that TeMuDance achieves competitive dance quality while substantially improving text-conditioned control over the existing methods. Code and quantitative results are available at: https://anonymous.4open.science/r/TeMuDance.

1 Introduction

In recent years, the media production industry has fueled a strong demand for automated, high-fidelity character animation (Mourot et al., 2022; Zhu et al., 2023). As a complex form of expressive motion, music-driven 3D dance generation has emerged as an important research domain, aiming to enable virtual characters to synthesize realistic movements

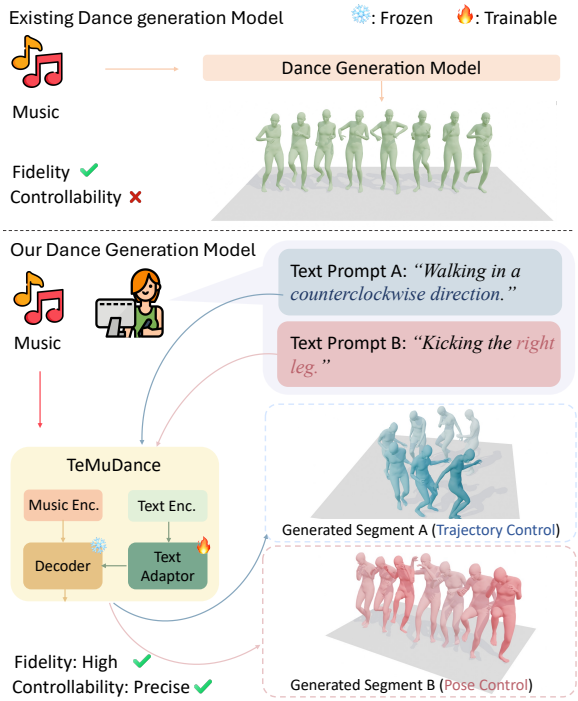

Figure 1: Our model generates dances conditioned on music and text jointly, producing sequences that are both rhythmically aligned and semantically controllable.

from music (Sun et al., 2020; Li et al., 2021). Despite the impressive realism achieved by current dance generation approaches (Kim et al., 2022; Siyao et al., 2022; Tseng et al., 2023; Li et al., 2024), a critical limitation remains: the lack of fine-grained semantic control. Most existing methods rely on coarse conditioning mechanisms, thus struggle to consistently follow explicit, intention-aligned instructions. This significantly reduces their practicality in real-world production settings.

To enable controllability, some methods rely on coarse cues such as global genre labels (Liu et al., 2025), which offer only high-level stylistic guidance and are unable to convey complex, continuous semantic intents. Efforts toward finer control, including text-guided editing (Zhang et al.,

2025a) and discrete codebook-based motion representations (Gong et al., 2023), introduce their own distinct trade-offs. Editing pipelines are often constrained by the supervision and edit distributions available in current datasets, which can limit zero-shot generalization such as spatial trajectories, whereas discrete quantization can lead to a noticeable disconnection between semantic actions and musical rhythm, preventing user-specified movements from being naturally integrated with music.

In essence, the challenge arises from the disjoint nature of existing datasets. Music–dance datasets provide rhythmic alignment but lack textual annotations, whereas text–motion datasets provide language supervision, but without accompanying music. As a result, the absence of music-text-motion triplets prevents existing models from jointly learning rhythmic coherence and semantic control.

To bridge this gap, we introduce TeMuDance, which enables zero-shot text-based control for music-conditioned 3D dance generation, as shown in Figure 1. To overcome the absence of paired triplets, we propose a Motion-Centered Bridging mechanism that uses motion as the shared anchor to align separate music–dance and text–motion datasets within a continuous unified embedding space. This facilitates retrieval-based completion of missing modalities, allowing the synthesis of high-quality pseudo triplets for end-to-end training while avoiding the artifacts associated with discrete quantization.

To maintain high-quality music-driven dance generation while incorporating textual control, we first pretrain a high-fidelity music-to-dance generation model and freeze it as the backbone, retaining its strong capabilities in rhythmic alignment and physical realism. Building on this backbone, we add a text-conditioned control branch that injects textual features into intermediate layers, guiding the generation towards the desired semantics, without modifying the backbone parameters. This design enables TeMuDance to achieve fine-grained textual control while preserving music-synchronised dance quality.

In addition, we employ a dual-stream training strategy that integrates mutual dataset augmentation with confidence-based noise filtering, thereby enhancing the precision of semantic control by effectively suppressing noise in pseudo conditions.

Overall, the main contributions of TeMuDance are summarised as follows.

(1) We propose a novel framework that enables zero-shot textual control in music-driven dance generation. By introducing a parameter-efficient adapter that injects semantic guidance into a frozen diffusion backbone, we achieve text-based control while preserving the high-fidelity rhythmic priors learned from music conditioning.

(2) We introduce motion-centered bridging that aligns separate music–dance and text–motion corpora into a shared latent space. This approach overcomes the lack of paired triplets and enables flexible, open-vocabulary semantic control.

(3) We design a dual-stream training strategy supported by a robust retrieval mechanism. Through mutually augmenting the separate datasets and performing confidence-based filtering to rigorously eliminate noisy samples, we ensure robust learning from synthesized supervision.

(4) Extensive experiments demonstrate that TeMuDance produces high-fidelity, controllable dance motions, outperforming the existing approaches in semantic alignment while remaining competitive in rhythmic fidelity.

2 Related Work

2.1 3D Human Motion Synthesis

Traditionally, 3D human skeletal motion prediction relies on physics-based methods that explicitly model kinematics, dynamics, and physical constraints of the human body, which are often computationally complex and unstable (Loi et al., 2023). More recently, learning-based approaches leverage large-scale datasets to enable more efficient and accurate prediction of 3D motion trajectories. Specifically, early efforts primarily employ RNNs for this task (Martinez et al., 2017; Li et al., 2018; Liu et al., 2019). However, RNN-based models are susceptible to error accumulation, which can lead to discontinuities in predicted motion sequences (Gui et al., 2018). Ma et al. (Ma et al., 2022) propose a network composed of spatial dense GCNs and temporal dense GCNs, which alternates between spatial and temporal modules to extract spatiotemporal features over the global receptive field. Aksan et al. (Aksan et al., 2021) utilize a self-attention mechanism to learn high-dimensional joint embeddings and generate temporally coherent poses.

The Motion Diffusion Model (MDM) (Tevet et al., 2022) is the first to apply classifier-free diffusion to human motion generation, which inspires many subsequent diffusion-based approaches. MotionFix (Athanasiou et al., 2024) conditions diffu-

sion models on both source motion and edit text for seamless motion edits. Although prior work improves motion quality and diversity, dance generation remains challenging because it requires both precise beat synchrony and consistent genre-aligned style.

2.2 Music Driven Dance Generation

Early studies (Shiratori et al., 2006; Ofli et al., 2008; Fukayama and Goto, 2015) consider this task as a similarity-based retrieval problem. With the advent of deep learning, it is reframed as a supervised motion prediction problem, leveraging architectures such as CNN (Holden et al., 2016, 2015), RNN (Butepage et al., 2017; Chiu et al., 2019; Du et al., 2019), and Transformer (Fan et al., 2022; Huang et al., 2022; Li et al., 2022). However, these frame-by-frame prediction approaches often face challenges such as error accumulation and motion freezing (Zhuang et al., 2022).

Recent research shifts to a generative pipeline. While methods based on VQ-VAE (Gong et al., 2023; Siyao et al., 2022) have achieved outstanding performance, these systems are highly complex and involve multiple sub-networks. EDGE (Tseng et al., 2023) is the first method that employs a diffusion-based framework, featuring a single-model design optimized for a single objective. It also introduces a novel evaluation approach focusing on physical plausibility. Despite this progress in generation quality, a critical limitation persists: the lack of fine-grained semantic control.

2.3 Controllable Dance Generation

To enable controllability in dance generation, several approaches (Huang et al., 2022; Liu et al., 2025) utilize discrete genre embeddings to achieve coarse-grained style control. While effective for global stylization, these label-driven methods lack the granularity to specify concrete motion details.

To enable flexible semantic control, recent research increasingly explores text-driven generation and editing. For example, DanceEditor (Zhang et al., 2025a) proposes an iterative editing paradigm that leverages language guidance to progressively revise motions, enabling targeted modifications beyond coarse style switching. TM2D (Gong et al., 2023) takes a step towards finer control by introducing action-annotated data and explicitly modeling controllable action units. However, its VQ-VAE discretisation can hinder smooth transitions and seamless choreographic integration. In parallel, general-purpose multimodal motion generators, such as UniMomo (Kong et al., 2025), MotionAnything (Zhang et al., 2025b), and DanceChat (Wang et al., 2025), aim to unify motion synthesis under diverse conditioning signals, including text and music, within a single backbone. Despite richer conditioning, these generalist frameworks often treat text as a global cue, leading the model to follow instructions at the pose or clip level rather than to coherent choreography-level control.

3 The Proposed TeMuDance Method

We present the overall framework of TeMuDance in Figure 2, which comprises a high-fidelity music-driven dance generation backbone, a text-conditioned adapter that enables semantic control, and a motion-anchored bridging strategy for cross-modal alignment, thereby strengthening controllability while preserving motion quality.

3.1 Music-conditioned Dance Generator

Our approach builds upon a pretrained diffusion-based music conditioned dance generator that maps a music segment to a temporally coherent 3D dance sequence. Given a long music-dance pair, we partition it into 4-second clips and uniformly sample k segments per clip. Each segment is represented using the SMPL-X parameterization augmented (Loper et al., 2023). We denote a motion clip as $\boldsymbol{x} \in \mathbb{R}^{k \times F}$, where $F = 319$ denotes the dimensionality of the skeletal motion features. We provide details in Appendix A. The corresponding music clip is encoded into temporally aligned conditioning features $\boldsymbol{c}_M \in \mathbb{R}^{k \times C}$ using a pretrained music foundation model combined with low-level waveform descriptors following (Liu et al., 2025), where C denotes the music feature dimension.

We employ a diffusion-based dance generation backbone following the Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020). At each training step, we sample a timestep t and add noise to the clean motion clip to obtain $\boldsymbol{x}_t$:

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\, \boldsymbol{x}_0 + \sqrt{1 - \bar{\alpha}_t}\, \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.

As shown in Figure 3, the denoiser of the dance generation backbone comprises a Spatially Hierarchical Motion Encoder E and a Denoising Decoder D. To capture part-specific motion patterns while preserving whole-body coherence, the encoder E partitions the input channels into $M = 7$ body-part

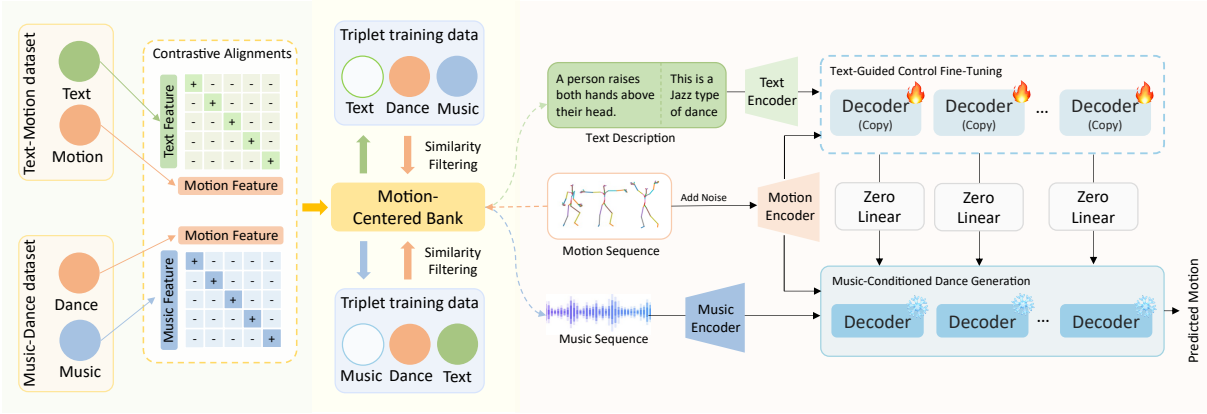

Figure 2: An overview of TeMuDance. We learn a motion-centered bank by contrastively aligning disjoint text–motion and music–dance datasets in a shared motion space, enabling similarity-filtered modality completion to form pseudo triplets. For generation, a pretrained diffusion Transformer is frozen as the backbone, while a text control branch steers denoising and produces rhythm-aligned, semantically controllable dances.

groups. Each group is processed by a hierarchical module to model local dynamics, followed by a fusion layer to capture inter-part dependencies, yielding the latent feature:

$$\boldsymbol{h}_t = E(\boldsymbol{x}_t) \in \mathbb{R}^{k \times H} \tag{2}$$

where H is the hidden feature dimension of the denoiser. Subsequently, these features are fed into the Denoising Decoder D to reconstruct the clean motion $\hat{\boldsymbol{x}}_0$. Each layer of D comprises a self-attention mechanism for temporal modeling, a cross-attention mechanism that integrates the music features $\boldsymbol{c}_M$, and a feed-forward network modulated by the timestep t through Feature-wise Linear Modulation (FiLM) layers (Perez et al., 2018). The network is trained to reconstruct the clean motion by the following objective:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\boldsymbol{x}_0, t}\left[\|\boldsymbol{x}_0 - D(\boldsymbol{h}_t, t, \boldsymbol{c}_M)\|_2^2\right] \tag{3}$$

In addition to $\mathcal{L}_{\text{diff}}$, following the settings of (Tseng et al., 2023; Tevet et al., 2022), we incorporate standard kinematic regularisers, including a joint position loss $\mathcal{L}_{\text{joint}}$, pose velocity and acceleration loss $\mathcal{L}_{\text{vel}}$, and foot contact loss $\mathcal{L}_{\text{contact}}$, to promote physically plausible and visually smooth motions:

$$\mathcal{L}_{\text{dance}} = \tau(\mathcal{L}_{\text{diff}}, \mathcal{L}_{\text{joint}}, \mathcal{L}_{\text{vel}}, \mathcal{L}_{\text{contact}}) \tag{4}$$

where $\tau(\cdot)$ aggregates multiple loss terms into a scalar objective. During pretraining, we adopt Aligned Multi-Task Learning (Aligned-MTL) (Senushkin et al., 2023) to stabilize the joint optimization of the loss terms. Specifically,

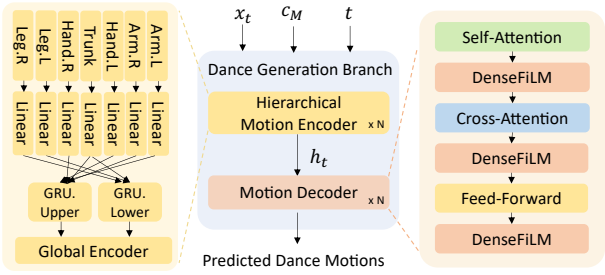

Figure 3: Architecture of the music-conditioned diffusion dance generator.

Aligned-MTL reduces gradient conflicts and prevents any single loss from dominating by aligning the gradients of different objectives, leading to more stable joint updates.

3.2 Text-Guided Control Fine-Tuning

The pretrained backbone is conditioned solely on music. Our objective is to enable free-form textual steering while preserving the motion quality and rhythmic fidelity of the dance generator. To this end, inspired by ControlNet (Zhang et al., 2023), we construct a trainable text-conditioned control branch by duplicating the denoiser of the pretrained backbone. The original music-conditioned denoiser is kept frozen, while the control branch is trained to predict layer-wise residual signals that are injected into the corresponding blocks of the frozen denoiser.

Given a text prompt, we use the BERT (Devlin et al., 2019) text encoder to extract contextual features, which are projected to yield the text condition embedding $\boldsymbol{c}_E \in \mathbb{R}^{N \times H}$, where N is the

token sequence length.

The control branch mirrors the structure of the first K blocks of the backbone. Analogous to the backbone described in Sec. 3.1, the control blocks employ cross-attention layers to inject the condition embeddings.

The frozen denoiser includes L stacked Transformer decoders $\{B^{(\ell)}\}_{\ell=0}^{L-1}$. We denote $\boldsymbol{h}_t^{(\ell)}$ as the hidden state serving as the input to the ℓ-th block, with $\boldsymbol{h}_t^{(0)} = E(\boldsymbol{x}_t)$. For the first K blocks (i.e., $\ell = 0, \ldots, K-1$), the control branch predicts a residual $\boldsymbol{\Delta}^{(\ell)}$ that is injected into the corresponding frozen block:

$$\boldsymbol{\Delta}^{(\ell)} = \mathcal{Z}^{(\ell)}\Big(B'^{(\ell)}(\boldsymbol{h}_t^{(\ell)}, \boldsymbol{c}_E, t)\Big) \tag{5}$$

$$\boldsymbol{h}_t^{(\ell+1)} = B^{(\ell)}(\boldsymbol{h}_t^{(\ell)}, \boldsymbol{c}_M, t) + \boldsymbol{\Delta}^{(\ell)} \tag{6}$$

where $B^{(\ell)}$ is the ℓ-th block of the frozen backbone, and $B'^{(\ell)}$ is its trainable counterpart in the control branch. $\mathcal{Z}^{(\ell)}$ represents a zero-initialized linear projection layer. The zero-initialization ensures that $\boldsymbol{\Delta}^{(\ell)}$ starts at zero, making the generator function-preserving at the beginning of fine-tuning. During fine-tuning, we update only the control branch and optimise it with the same dance objective as the backbone. We denote this training loss as $\mathcal{L}_{\text{text}}$.

3.3 Motion-bridging Cross-Modal Alignment

A key challenge in our setting is the absence of paired music–text–motion triplets. Since direct supervision is unavailable, we propose a *motion-centered bridging framework* that operates in two stages. First, we use motion as a pivot to embed disjoint datasets into a shared latent space, establishing a unified foundation for cross-modal retrieval. Second, we introduce a dual-stream training strategy, balancing text controllability with the rhythmic fidelity of generated motion.

3.3.1 Motion-Centered Latent Alignment

To enable cross-modal semantic transfer without paired music–text–motion triplets, we adopt a motion-centered contrastive formulation. Specifically, we use FineDance (Li et al., 2023) for music–dance supervision and HumanML3D (Guo et al., 2022) for text–motion supervision. We then learn a unified embedding space with two contrastive streams and a motion-level regulariser to to remain domain-consistent across datasets.

For the music–dance stream, we optimise a queue-based InfoNCE loss (He et al., 2020). Given a paired sample $(\boldsymbol{c}_M, \boldsymbol{x}_0^{\text{Da}})$, we define the ℓ_2-normalised query and key as:

$$\begin{aligned} \boldsymbol{q} &= \text{norm}(P_{\text{mus}}(\rho_M(\boldsymbol{c}_M))) \\ \boldsymbol{k} &= \text{norm}\big(P_{\text{mot}}(\rho_X(\bar{E}(\boldsymbol{x}_0^{\text{Da}})))\big) \end{aligned} \tag{7}$$

where $\bar{E}$ is an EMA teacher used to compute stable motion keys, and $\rho_M(\cdot)$ and $\rho_X(\cdot)$ are temporal pooling operators that map token sequences to global vectors. Let $\mathcal{Q}_{\text{Da}} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{K_q}] \in \mathbb{R}^{H \times K_q}$ be a momentum-updated queue of K_q negative motion keys. We minimise:

$$\begin{aligned} \ell_{\text{mus}} = -\, s_{\text{mus}} \boldsymbol{q}^\top \boldsymbol{k} + \log\Big(\exp\big(s_{\text{mus}} \boldsymbol{q}^\top \boldsymbol{k}\big) \\ + \sum_{j=1}^{K_q} \exp\big(s_{\text{mus}} \boldsymbol{q}^\top \boldsymbol{u}_j\big)\Big) \end{aligned} \tag{8}$$

where $s_{\text{mus}} = \exp(\alpha_{\text{mus}})$, α_{mus} is a learnable scalar logit-scale parameter and $\mathcal{L}_{\text{mus}} = \mathbb{E}[\ell_{\text{mus}}]$.

In parallel, for the text–motion stream, we use the same contrastive form $\mathcal{L}_{\text{text}}$ to align text descriptions with their corresponding motion embeddings.

Although both streams share the motion encoder, the motion distributions of FineDance and HumanML3D are inherently different, training them independently can separate the two motion domains in the embedding space, breaking the semantic bridge between music and text. To reduce domain drift, we regularise motion embeddings by aligning their batch-wise mean and covariance across the two domains:

$$\mathcal{L}_{\text{bridge}} = \|\boldsymbol{\mu}_{\text{Da}} - \boldsymbol{\mu}_{\text{Mo}}\|_2^2 + \|\boldsymbol{\Sigma}_{\text{Da}} - \boldsymbol{\Sigma}_{\text{Mo}}\|_F^2 \tag{9}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ denote the batch-wise mean vector and covariance matrix of motion embeddings, the subscripts $_{\text{Da}}$ and $_{\text{Mo}}$ denote the dance and motion domains of the FineDance and HumanML3D datasets, respectively, and $\|\cdot\|_F$ is the Frobenius norm. The overall alignment objective is $\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{mus}} + \mathcal{L}_{\text{text}} + \lambda\, \mathcal{L}_{\text{bridge}}$. This regulariser preserves a coherent semantic bridge in the shared embedding space for cross-modal alignment.

3.3.2 Motion-Centered Dual-Stream Training

Although the contrastive alignment brings the unpaired datasets into a shared latent space, a key challenge in dual-stream fine-tuning remains: we aim to learn a jointly music–text conditioned generator, but the available supervision consists only of disjoint text-motion and music-dance pairs. To

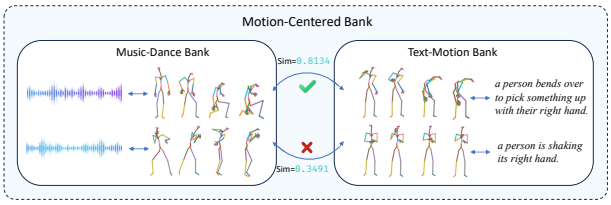

Figure 4: Illustration of the Motion-Centered Bank.

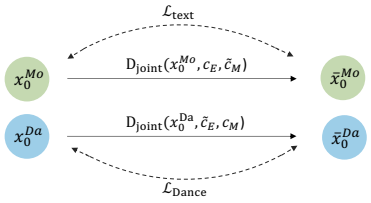

Figure 5: Visual pipeline of Dual-Stream Training.

address this missing-modality issue, we construct Motion-Centred Banks that enable motion-bridging cross-domain retrieval. Specifically, we freeze the encoders and index the datasets into two Motion-Centred Banks, denoted as $\mathcal{B}_{\text{MD}}$ (music-dance) and $\mathcal{B}_{\text{TM}}$ (text-motion).

As illustrated in Figure 4, these banks serve as the foundation for cross-domain retrieval, enabling us to synthesize pseudo-triplets by imputing missing modalities in the subsequent fine-tuning stage. For a mini-batch sampled from the text-motion dataset, we have paired text and motion $(\boldsymbol{c}_E, \boldsymbol{x}_0^{\text{Mo}})$ but no music. We retrieve a rhythmically compatible music condition by querying the music-dance bank with the motion embedding. Concretely, we compute motion embeddings $E(\boldsymbol{x}_0^{\text{Mo}})$, perform nearest-neighbour search in $\mathcal{B}_{\text{MD}}$ with cosine similarity, and obtain the corresponding music features:

$$\tilde{\boldsymbol{c}}_M = \mathcal{B}_{\text{MD}}(E(\boldsymbol{x}_0^{\text{Mo}})) \tag{10}$$

where matches falling below a similarity threshold are replaced by a null condition. This yields pseudo-triplets $(\boldsymbol{x}_0^{\text{Mo}}, \tilde{\boldsymbol{c}}_M, \boldsymbol{c}_E)$ to train the text-control branch to steer denoising under music-compatible priors. For a mini-batch sampled from the music–dance dataset, we analogously impute the missing text condition by querying the text–motion bank with the motion embedding to obtain $(\boldsymbol{x}_0^{\text{Da}}, \boldsymbol{c}_M, \tilde{\boldsymbol{c}}_E)$. To bridge the gap between specific motion semantics and global musical style, we construct a composite instruction by concatenating the retrieved description with coarse music genre tags, encouraging the control branch to follow both fine-grained actions and global style cues.

As illustrated in Figure 5, we fine-tune the joint denoiser D_{joint} by alternating between two streams utilizing the objectives defined in Sec. 3.1 and Sec. 3.2. Specifically, the text–motion stream optimizes $\mathcal{L}_{\text{text}}$ for semantic control, while the music–dance stream optimizes $\mathcal{L}_{\text{dance}}$ to preserve the rhythmic prior. Accordingly, the final fine-tuning objective is formulated as a weighted combination:

$$\mathcal{L}_{\text{ft}} = (1 - \lambda_p)\,\mathcal{L}_{\text{text}} + \lambda_p\,\mathcal{L}_{\text{dance}} \tag{11}$$

where λ_p is a trade-off hyperparameter.

In inference, we apply the classifier-free guidance (Ho and Salimans, 2022) to the music-conditioned backbone to continuously regulate the influence of music on the generated motion. Adjusting the music guidance scale yields a smooth continuum of behaviours, ranging from text-only generation under a null music condition, to music-only generation without text residual injection, and to joint text–music generation when both conditioning pathways are active. This inference mechanism provides a controllable trade-off between rhythmic fidelity and semantic steering, as illustrated in Figure 6.

4 Experiments and Results

4.1 Experimental Setup

Datasets. Given that triplet-level supervision was not available for this task, we formulated the setting under disjoint supervision and leveraged a music–motion dataset together with a text–motion dataset for training and evaluation. For music–dance supervision, we used FineDance (Li et al., 2023), which provides paired music and 52-joint 3D SMPL-X motions at 30 FPS over 16 genres. For text–motion supervision, we used HumanML3D (Guo et al., 2022), which contains natural-language descriptions paired with SMPL-based 3D motions. Implementation details are provided in Appendix B.

Evaluation metrics. We evaluated our method in terms of motion quality using the Fréchet Inception Distance (FID) between feature distributions of generated and real motions (Li et al., 2021, 2020; Heusel et al., 2017), diversity using the diversity score adopted in Bailando (Siyao et al., 2022), music–motion synchronization using the Beat Alignment Score (BAS) (Siyao et al., 2022), and physical plausibility using Physical Foot Contact (PFC) and Physical Body Contact (PBC) (Tseng et al., 2023; Luo et al., 2024).

Table 1: A quantitative comparison on FineDance. The best results are in **bold** and the second-best are underlined. ↓ indicates lower is better ↑ indicates higher is better, and → indicates closer to the ground truth is better. ∗ marks abnormally high diversity values caused by discontinuous motions (Li et al., 2021).

	Motion Quality		Motion Diversity		PFC↓	PBC→	BAS↑
	FID_hand↓	FID_body↓	Div_hand↑	Div_body↑			
GT	/	/	11.82 ± 0.1314	10.18 ± 0.1327	/	5.23 ± 0.16	0.2318 ± 0.0070
DanceRevolution (Huang et al., 2020)	219.52 ± 18.32	99.83 ± 7.79	1.85 ± 0.60	4.49 ± 0.25	6.81 ± 0.81	23.39 ± 2.03	0.2104 ± 0.0057
MNET (Kim et al., 2022)	195.56 ± 5.04	154.79 ± 2.80	6.79 ± 0.20	8.25 ± 0.39*	2.98 ± 0.11	12.21 ± 0.15	0.1792 ± 0.0014
Bailando (Siyao et al., 2022)	55.60 ± 8.15	57.77 ± 6.01	6.40 ± 0.68	4.27 ± 0.43	0.34 ± 0.01	3.09 ± 0.06	0.2152 ± 0.0028
EDGE (Tseng et al., 2023)	25.37 ± 3.24	51.56 ± 3.62	8.29 ± 0.30	5.88 ± 0.32	0.21 ± 0.03	7.78 ± 0.07	0.2171 ± 0.0056
FineNet (Li et al., 2023)	26.88 ± 3.09	23.59 ± 3.56	8.30 ± 0.45	6.64 ± 0.28	**0.12 ± 0.01**	3.35 ± 0.11	0.2066 ± 0.0046
DGFM (Liu et al., 2024)	20.699 ± 3.52	24.63 ± 3.14	8.77 ± 0.41	6.77 ± 0.75	0.20 ± 0.01	4.23 ± 0.06	0.2153 ± 0.0054
LODGE (Li et al., 2024)	18.36 ± 2.10	47.56 ± 1.37	8.57 ± 0.36	5.41 ± 0.27	0.13 ± 0.01	3.46 ± 0.06	0.2327 ± 0.0050
TeMuDance	**15.90 ± 3.28**	**23.41 ± 1.78**	**9.15 ± 0.37**	**6.89 ± 0.36**	0.19 ± 0.01	**4.95 ± 0.10**	**0.2342 ± 0.0057**

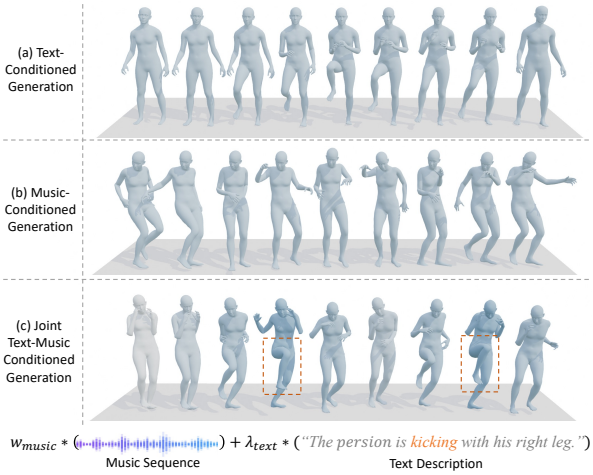

Figure 6: Text–music controllability at inference.

Baselines. We considered two evaluation settings. (i) Music-driven dance generation. We benchmarked our method against representative and recent state-of-the-art music-conditioned dance generators on FineDance, including DanceRevolution (Huang et al., 2020), MNET (Kim et al., 2022), Bailando (Siyao et al., 2022), EDGE (Tseng et al., 2023), FineNet (Li et al., 2023), DGFM (Liu et al., 2024), and LODGE (Li et al., 2024), following their standard evaluation protocols whenever available. (ii) Text–music controlled generation. We qualitatively compared with TM2D (Gong et al., 2023), as it similarly combines music–dance and text–motion datasets to enable text-and-music conditioned dance generation.

4.2 Results and Analysis

4.2.1 Evaluation on Music-text Conditioned Dance Generation

We validated music-driven dance generation on the FineDance test set, with results summarised in Table 1. TeMuDance achieves the best overall motion quality, attaining the lowest FID for both hands and the body, indicating the closest match to the real-motion distribution. It also provides the strongest diversity on both hand and body motions. Beyond motion quality metrics, TeMuDance achieves the best physical body-contact score and the highest beat-alignment score, demonstrating that gains in realism and diversity are accompanied by improved physical plausibility and music–motion synchronisation. Although TeMuDance is not the top-performing method on PFC, it remains competitive and exhibits low foot-contact violations. Overall, TeMuDance demonstrates an excellent trade-off across realism, diversity, physical plausibility, and music–motion consistency.

4.2.2 Controllable Generation using Text Description

Figure 7 presents a qualitative comparison between TM2D (Gong et al., 2023) and our method for joint music–text controlled dance generation. It shows that TM2D often separates the generated dance motion from the text-controlled action, making the instruction appear as an isolated segment rather than being fused into the choreography. For example, in the "clockwise direction" case, TM2D first generates several dance-like poses before briefly switching to a walking-and-turning pattern mid-sequence. This behaviour is consistent with the VQ-VAE-based discrete codebook representation in TM2D. In particular, quantised motion tokens from two datasets with different distributions are jointly used for training, which can encourage piecewise composition rather than continuous cross-modal fusion. In contrast, our method maintains the dance characteristics while enforcing the textual instruction throughout, resulting in more coherent joint control under combined music and text conditioning.

4.3 Ablation Study

To assess the necessity of each proposed component, we qualitatively compared the full model with

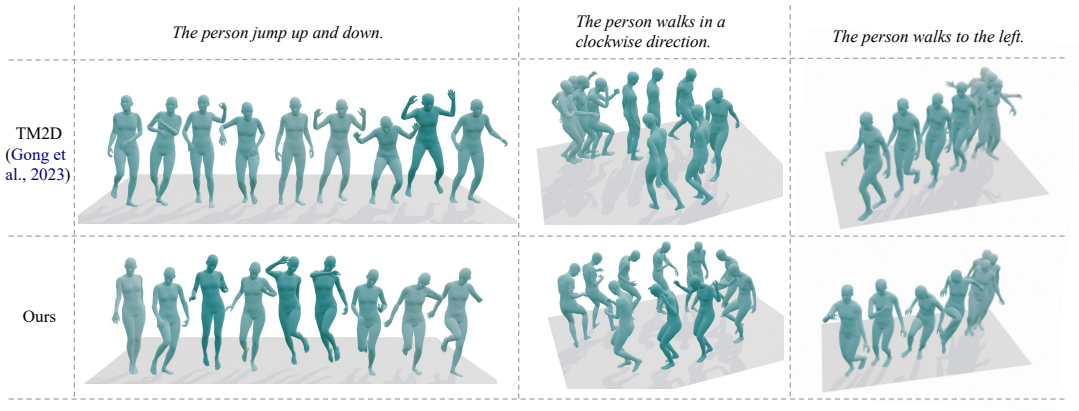

Figure 7: Visual comparison of the generated dance between the proposed method and TM2D (Gong et al., 2023).

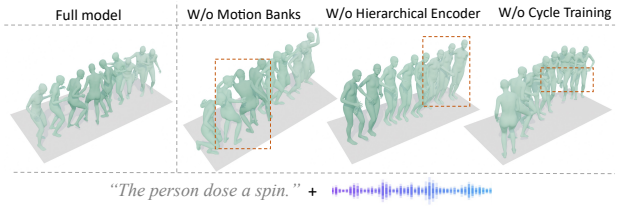

Figure 8: Visual comparisons of the ablation designs and our model.

Table 2: Perceptual evaluation of generated samples

Comparison	Ours Win (%)
Music-to-Dance Generation Quality	
vs. Bailando (Siyao et al., 2022)	85.4
vs. EDGE (Tseng et al., 2023)	77.9
vs. FineDance (Li et al., 2023)	62.5
vs. LODGE (Li et al., 2024)	69.2
Text-Driven Controllability (vs. TM2D (Gong et al., 2023))	
Choreographic Coherence	75.0
Semantic Controllability	61.9

three ablated variants. As shown in Figure 8, removing the retrieval mechanism significantly degrades generation quality. The resulting motion is lethargic and lacks rhythmic dynamism; instead of executing the requested "spin," the model produces a slow, partial rotation. This indicates that motion banks provide essential priors for both semantic controllability and beat-aligned dynamics. Similarly, without the Hierarchical Encoder, the model struggles with precise semantic control. While the character attempts a turning motion, the execution is stiff and mechanically flawed, as highlighted by the orange box, lacking the fluidity and definition of a well-controlled action. This confirms that hierarchical modeling is necessary to enable fine-grained control over complex motion units. Finally, the model trained without the dual-stream strategy tends to over-prioritize textual instructions at the expense of dance fidelity. This suggests that the dual-stream strategy is vital for balancing strong semantic guidance with the inherent physical and coherence of music-driven dance.

4.4 User Study

We conducted a user preference study with 20 participants. For the general music-to-dance assessment, we randomly sampled 12 music clips from the test set. For the text-driven controllability task, we selected 8 test cases focusing on specific action instructions. For dance generation quality, we obtain preference rates of 62.5%–85.4%, indicating that participants generally favour the motions synthesized by our dance generation backbone in terms of fidelity and physical plausibility. Compared with TM2D, our model achieves a 75.0% preference on Choreographic Coherence, indicating more temporally continuous choreography with fewer clip-level composition discontinuities. Meanwhile, we maintain strong semantic controllability, suggesting that improved coherence dose not come at the expense of instruction following.

5 Conclusion

We presented *TeMuDance*, which enables fine-grained motion control for music-driven 3D dance generation. This was achieved by learning a shared continuous motion space from disjoint music–dance and text–motion datasets via contrastive alignment, enabling zero-shot textual control without paired music–text–motion triplets. Built on a pretrained music-to-dance backbone with a text-conditioning adapter, our framework injects semantic guidance while preserving physical realism and rhythmic alignment. Experiments showed that TeMuDance markedly improved textual controllability over existing baselines while maintaining competitive dance quality and music alignment.

6 Limitations

While TeMuDance has demonstrated promising results in enabling zero-shot text control for music-conditioned dance generation, several limitations remain.

First, a principled automatic evaluation protocol for joint music–text conditioning is still underdeveloped. Existing metrics typically assess either text–motion semantic consistency or music–motion rhythm/style alignment in isolation, and there is no unified measure that captures their interaction, including how well the model resolves potential conflicts between the two conditions. Consequently, we still rely on human evaluation to validate joint controllability, which is costly and may limit reproducibility across studies. Second, our control branch mainly provides choreography-level semantic steering; it does not explicitly enforce strict geometric or physical constraints, such as precise spatial trajectories or temporally localised edits. Such fine-grained constraints may require additional structured controllers or constraint-aware objectives. Finally, our language supervision largely follows existing text–motion HumanML3D dataset, where descriptions are predominantly English, short, and action-centric. Therefore, the generalization to multilingual prompts or more complex instructions, such as those involving multi-step, compositional, and long-form tasks, has not been fully validated.

We view these limitations as opportunities for future work, including developing joint evaluation metrics, adding constraint-aware control mechanisms, expanding multilingual instruction coverage, and improving retrieval robustness under domain shift.

References

Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. 2021. A spatio-temporal transformer for 3d human motion prediction. In *2021 International Conference on 3D Vision (3DV)*, pages 565–574. IEEE.

Nikos Athanasiou, Alpár Cseke, Markos Diomataris, Michael J Black, and Gül Varol. 2024. Motionfix: Text-driven 3d human motion editing. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11.

Judith Butepage, Michael J Black, Danica Kragic, and Hedvig Kjellstrom. 2017. Deep representation learning for human motion prediction and classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6158–6166.

Hsukuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles. 2019. Action-agnostic human pose forecasting. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1423–1432. IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 1 (Long and Short Papers)*, pages 4171–4186.

Xiaoxiao Du, Ram Vasudevan, and Matthew Johnson-Roberson. 2019. Bio-lstm: A biomechanically inspired recurrent neural network for 3-d pedestrian pose and gait prediction. *IEEE Robotics and Automation Letters*, 4(2):1501–1508.

Di Fan, Lili Wan, Wanru Xu, and Shenghui Wang. 2022. A bi-directional attention guided cross-modal network for music based dance generation. *Computers and Electrical Engineering*, 103:108310.

Satoru Fukayama and Masataka Goto. 2015. Music content driven automated choreography with beat-wise motion connectivity constraints. *Proceedings of SMC*, pages 177–183.

Kehong Gong, Dongze Lian, Heng Chang, Chuan Guo, Zihang Jiang, Xinxin Zuo, Michael Bi Mi, and Xinchao Wang. 2023. Tm2d: Bimodality driven 3d dance generation via music-text integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9942–9952.

Liang-Yan Gui, Yu-Xiong Wang, Xiaodan Liang, and José MF Moura. 2018. Adversarial geometry-aware human motion prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 786–803.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. 2022. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5152–5161.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30.

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

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Daniel Holden, Jun Saito, and Taku Komura. 2016. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)*, 35(4):1–11.

Daniel Holden, Jun Saito, Taku Komura, and Thomas Joyce. 2015. Learning motion manifolds with convolutional autoencoders. In *SIGGRAPH Asia 2015 technical briefs*, pages 1–4. Association for Computing Machinery, New York, NY, USA.

Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Daxin Jiang. 2020. Dance revolution: Long-term dance generation with music via curriculum learning. *arXiv preprint arXiv:2006.06119*.

Yuhang Huang, Junjie Zhang, Shuyan Liu, Qian Bao, Dan Zeng, Zhineng Chen, and Wu Liu. 2022. Genre-conditioned long-term 3d dance generation driven by music. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4858–4862. IEEE.

Jinwoo Kim, Heeseok Oh, Seongjean Kim, Hoseok Tong, and Sanghoon Lee. 2022. A brand new dance partner: Music-conditioned pluralistic dancing controlled by multiple dance genres. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3490–3500.

Xiangzhe Kong, Zishen Zhang, Ziting Zhang, Rui Jiao, Jianzhu Ma, Wenbing Huang, Kai Liu, and Yang Liu. 2025. Unimomo: Unified generative modeling of 3d molecules for de novo binder design. *arXiv preprint arXiv:2503.19300*.

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

Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. 2018. Convolutional sequence to sequence model for human dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5226–5234.

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

Ronghui Li, YuXiang Zhang, Yachao Zhang, Hongwen Zhang, Jie Guo, Yan Zhang, Yebin Liu, and Xiu Li. 2024. 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)*, pages 1524–1534.

Ronghui Li, Junfan Zhao, Yachao Zhang, Mingyang Su, Zeping Ren, Han Zhang, Yansong Tang, and Xiu Li. 2023. Finedance: A fine-grained choreography dataset for 3d full body dance generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10234–10243.

Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13401–13412.

Xinran Liu, Xu Dong, Diptesh Kanojia, Wenwu Wang, and Zhenhua Feng. 2025. Gcdance: Genre-controlled 3d full body dance generation driven by music. *arXiv preprint arXiv:2502.18309*.

Xinran Liu, Zhenhua Feng, Diptesh Kanojia, and Wenwu Wang. 2024. Dgfm: Full body dance generation driven by music foundation models. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*.

Zhenguang Liu, Shuang Wu, Shuyuan Jin, Qi Liu, Shijian Lu, Roger Zimmermann, and Li Cheng. 2019. Towards natural and accurate future motion prediction of humans and animals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10004–10012.

Iliana Loi, Evangelia I Zacharaki, and Konstantinos Moustakas. 2023. Machine learning approaches for 3d motion synthesis and musculoskeletal dynamics estimation: a survey. *IEEE transactions on Visualization and Computer Graphics*, 30(8):5810–5829.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2023. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866.

Zhenye Luo, Min Ren, Xuecai Hu, Yongzhen Huang, and Li Yao. 2024. Popdg: Popular 3d dance generation with popdanceset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26984–26993.

Tiezheng Ma, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. 2022. Progressively generating better initial guesses towards next stages for high-quality human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6437–6446.

Julieta Martinez, Michael J Black, and Javier Romero. 2017. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2891–2900.

Lucas Mourot, Ludovic Hoyet, François Le Clerc, François Schnitzler, and Pierre Hellier. 2022. A survey on deep learning for skeleton-based human animation. In *Computer Graphics Forum*, volume 41, pages 122–157. Wiley Online Library.

Ferda Ofli, Yasemin Demir, Yücel Yemez, Engin Erzin, A Murat Tekalp, Koray Balcı, İdil Kızoğlu, Lale Akarun, Cristian Canton-Ferrer, Joëlle Tilmanne, and 1 others. 2008. An audio-driven dancing avatar. *Journal on Multimodal User Interfaces*, 2:93–103.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Dmitry Senushkin, Nikolay Patakin, Arseny Kuznetsov, and Anton Konushin. 2023. Independent component alignment for multi-task learning. In *CVPR*, pages 20083–20093.

Takaaki Shiratori, Atsushi Nakazawa, and Katsushi Ikeuchi. 2006. Dancing-to-music character animation. In *Computer Graphics Forum*, volume 25, pages 449–458. Wiley Online Library.

Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. 2022. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11050–11059.

Guofei Sun, Yongkang Wong, Zhiyong Cheng, Mohan S Kankanhalli, Weidong Geng, and Xiangdong Li. 2020. Deepdance: music-to-dance motion choreography with adversarial learning. *IEEE Transactions on Multimedia*, 23:497–509.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. 2022. Human motion diffusion model. *Preprint*, arXiv:2209.14916.

Jonathan Tseng, Rodrigo Castellon, and Karen Liu. 2023. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 448–458.

Qing Wang, Xiaohang Yang, Yilan Dong, Naveen Raj Govindaraj, Gregory Slabaugh, and Shanxin Yuan. 2025. Dancechat: Large language model-guided music-to-dance generation. *arXiv preprint arXiv:2506.10574*.

Xingyu Xie, Pan Zhou, Huan Li, Zhouchen Lin, and Shuicheng Yan. 2024. Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Hengyuan Zhang, Zhe Li, Xingqun Qi, Mengze Li, Muyi Sun, Siye Wang, Man Zhang, and Sirui Han. 2025a. Danceeditor: Towards iterative editable music-driven dance generation with open-vocabulary descriptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12158–12168.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847.

Zeyu Zhang, Yiran Wang, Wei Mao, Danning Li, Rui Zhao, Biao Wu, Zirui Song, Bohan Zhuang, Ian Reid, and Richard Hartley. 2025b. Motion anything: Any to motion generation. *arXiv preprint arXiv:2503.06955*.

Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. 2023. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2430–2449.

Wenlin Zhuang, Congyi Wang, Jinxiang Chai, Yangang Wang, Ming Shao, and Siyu Xia. 2022. Music2dance: Dancenet for music-driven dance generation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2):1–21.

A Detailed Motion Representation

In this section, we detail the composition of the motion representation $\boldsymbol{x} \in \mathbb{R}^{k \times F}$ ($F = 319$) derived from the SMPL-X parameterization (Loper et al., 2023). The feature vector comprises three parts: (1) Joint Rotations: The poses of 52 skeletal joints are transformed into a continuous 6-dimensional rotation representation, yielding a 312-dimensional vector; (2) Root Translation: A 3-dimensional vector representing the global trajectory in world space; and (3) Foot Contact: Following (Tseng et al., 2023), we append a 4-dimensional binary signal encoding the heel and toe contact states. Together, these components constitute the final feature dimension of $312 + 3 + 4 = 319$.

B Implementation Details

We set the motion and music sequence length to 4 seconds, corresponding to $N = 120$ frames. In practice, TeMuDance generates 4-second dance clips with 52 joints. We train the model on two NVIDIA GeForce RTX 3090 GPUs. In the initial training stage, we use Adan (Xie et al., 2024) with a learning rate of 2×10^{-4} and an L_2 reconstruction objective. We train for 1000 epochs with a batch size of 128, taking 3 days.

For the text-guided control fine-tuning, the stage is trained for 200 epochs with a batch size of 96. During inference, we employ the standard DDPM sampler with $T = 1,000$ steps. We apply classifier-free guidance with a scale of 3, a value empirically determined to offer the optimal trade-off between text consistency and the naturalness of the generated dance motions.

C User Study Instruction

We recruited 20 participants for a user preference study. Participation was voluntary and anonymous. We collected no personally identifying information, and we report results only in aggregate. Most participants were PhD students or senior-level researchers. In terms of background, 70% reported research experience in multimodal learning, 25% in computer vision, and 5% in art or media-related fields. The gender distribution was 60% male and 40% female. Although none were professional dancers, all participants were familiar with generative models and motion evaluation. In each trial, participants were shown two candidate videos, denoted as Video A and Video B, generated from the same conditioning inputs, with the left–right order randomized. Participants selected the better result under one of three criteria: For music-to-dance generation, Dance Quality measures motion realism, physical plausibility, temporal smoothness, and overall naturalness. For joint music–text generation, Choreographic Coherence assesses whether the motion forms a coherent choreography that remains rhythmically consistent with the music, with clear beat-synchronous accents and stable tempo following, while Semantic Controllability measures whether the motion follows the text instruction, including whether the required action is present and consistently maintained.