Text Dictates, Music Decorates: Energy-based Attention for Editable Dance Motion Generation

Seong Jong Yoo, Siyuan Peng, Felix Gu, Stratis Aloimonos, and Cornelia Fermüller

University of Maryland, College Park, MD, 20742, USA
Perception and Robotics Group

Abstract. Choreographic motion generation poses unique challenges for AI, demanding precise semantic control over complex, temporally structured, and expressive full-body dynamics. While existing models can synthesize motion from music, they remain largely black boxes. Conversely, attempting to condition generation on both text and music frequently leads to modality collapse, where dense acoustic rhythms overwhelm sparse semantic text prompts, destroying user controllability. To resolve this spatial-temporal conflict, we propose STREAM (Structural-Temporal Rhythmic Energy-based Attention for Motion), a modality-decoupled diffusion transformer. STREAM strictly separates conditioning pathways: global text semantics dictate the kinematic structure via Adaptive Layer Normalization (AdaLN), while a novel Bimodal Energy-Based Attention Module (BEAM) routes these features to the musical beat without overwriting the semantics. We further introduce Motorica++, a newly curated dataset enriched with domain-specific dance vocabulary and frame-level semantic annotations from existing Motorica dataset. Additionally, to rigorously quantify zero-shot editability, we propose the Exchange Evaluation Protocol and Editable Dance Score (EDS). Through extensive experiments, STREAM achieves state-of-the-art alignment between motion and music while preserving choreographic semantics, positioning AI not merely as a reactive synthesizer, but as a controllable, collaborative partner for artistic direction. The source code and datasets are available at https://github.com/SeongJong-Yoo/STREAM.

1 Introduction

Dance is a universal phenomenon across cultures, ethnicities, and eras [4, 5, 54]. Humans possess a strong innate desire to move their bodies with music and rhythm [28]. When these spontaneous movements are shaped into structured, intentional forms, they become artistic expressions crafted through choreographic design [2]. Professional choreographers translate musical ideas into movement, balancing artistic vision with the physical realities of the human body.

Although AI-driven dance motion generation has achieved significant advances in realism and synchronization, current systems still lack one crucial capability: semantically meaningful control [1, 33, 36, 46]. Most generative models operate

as black boxes dominated by given music conditions, offering limited access to the nuanced, interpretable directions that choreographers rely on. Even existing controllable (editable) approaches typically fall into one of three categories: micro-level edits [63] (*e.g.*, joint-wise adjustments, which are tedious and counter-intuitive), temporal inpainting [26,63] (which cannot specify semantic intent), and high-level conditioning [1] (*e.g.*, genre labels that offer minimal fine-grained control). These constraints make current systems misaligned with creative workflows in which choreographers require precise, expressive, and concept-driven manipulation of movement.

In contrast to dance-specific generative models, controllability in text-to-motion generation has advanced rapidly [3,14,25,60,68]. These models achieve fine-grained control through textual descriptions, enabled by large-scale text-motion datasets such as HumanML3D [17], KIT-ML [49], and Motion-X [37]. Recent efforts further attempt to unify multiple conditioning modalities [32] or combine general text-motion datasets with dance datasets to increase controllability [15, 65,67]. However, the dynamics of everyday human motions differ fundamentally from the structure, expressiveness, and rhythmic constraints of dance, as shown in Fig. 1. When standard cross-attention architectures attempt to fuse these modalities, they often suffer from **Modality Collapse**, in which the dense, high-frequency rhythmic signals of the music overwhelm the sparse, high-level semantic signals of the text, causing the network to ignore the user's prompt and revert to music-conditioned dance generation.

To bridge this gap, we propose a system built on three components: a professionally curated dataset, a novel energy-based architecture for controllable dance generation that is directed by text and modulated by music, and a new metric to quantify controllability in dance generation task. The system is designed to speak the language of choreographers by providing a direct, semantically meaningful interface while preserving high fidelity and strong alignment with musical structure. By making dance generation controllable, learnable, and artist-friendly, we aim to empower creators rather than replace them.

Specifically, we first **annotate the Motorica dance dataset [1]** with frame-level dance-technique labels curated by a professional dancer and detailed human-motion text descriptions. While datasets such as AIST++ [36], Motorica [1], FineDance [34], and DanceRemix [67] contain high-quality music-motion pairs, they lack fine-grained text-motion annotations. Our Motorica++ dataset addresses this deficiency by linking dance motion with dance-specific textual descriptions. Second, we propose **STREAM**, a Structural-Temporal Rhythmic Energy-based Attention Module for the dance Motion generation pipeline. To prevent modality collapse, STREAM strictly disentangles the conditioning pathways, *text dictates*, while *music decorates*. Specifically, we inject global text semantics via Adaptive Layer Normalization (AdaLN) to define the spatial kinematic manifold, while our novel **Bimodal Energy-based Attention Module (BEAM)** injects raw acoustic energy directly into the attention logits. By construction, the music pathway conveys only the temporal self-similarity structure of the audio: the alignment drift term depends on the music solely through S_m, which encodes

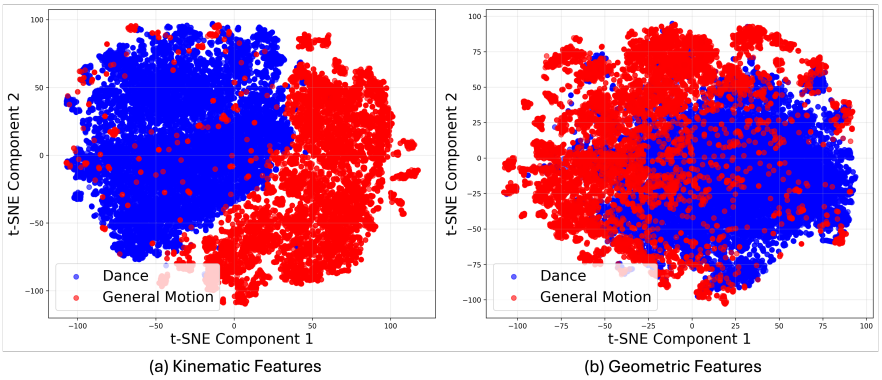

Fig. 1: t-SNE visualization of dance (Motorica) and general motion (HumanML3D) of kinematic features [43] and geometric features [41].

when the motion should repeat or vary, but carries no semantic content that could override the text conditioning. Finally, we propose the **Exchange Evaluation Protocol** and a unified metric, the **Editable Dance Score (EDS)**. Existing editable dance generation pipelines rarely evaluate zero-shot editability under conflicting conditions (*e.g.*, forcing a slow semantic dance onto a fast acoustic beat). By evaluating models on mismatched text-music pairs, EDS computes the harmonic mean of semantic preservation and rhythmic adaptation, rigorously penalizing models that suffer from modality collapse. In summary, our contributions are as follows:

1. We curate a domain-specific dance dataset, labeled with frame-level dance-technique annotations and detailed motion text descriptions collected by a professional dancer, addressing the lack of fine-grained text-motion annotations in existing dance datasets.
2. We propose STREAM, a text-controlled, music-decorated pipeline for generating dance motions, designed with a novel Bimodal Energy-based Attention Module. STREAM enables semantic control of dance motion through text while decorating it to follow musical structure and beats.
3. We introduce a new metric, called Editable Dance Score (EDS). This experiment and metric are designed to measure a new task for semantically controlled yet musically aligned motion generation.

2 Related Works

2.1 Dance Motion Generation

The key challenges include ensuring physical plausibility, producing diverse and expressive motions, achieving fine-grained controllability, and modeling complex human-body interactions. To address them, prior work has explored conditioning

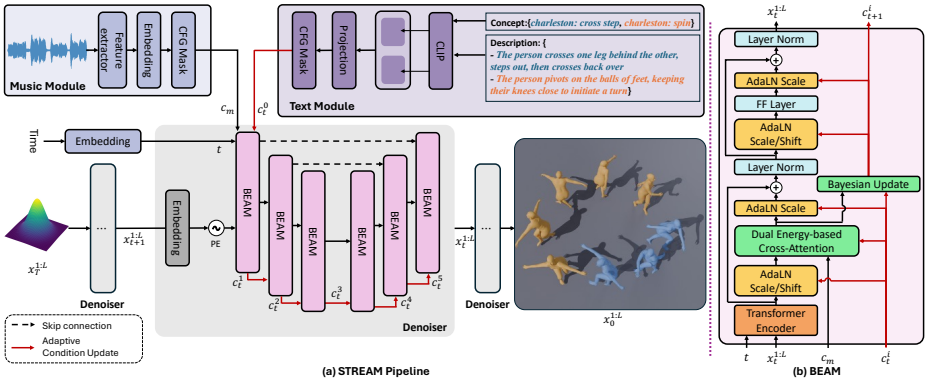

Fig. 2: Overview of STREAM. (a) **Left**: STREAM is first controlled by text (high-level concept and a low-level detailed description). Second, the music condition modulates the motion via the music alignment energy function, transforming general motion into *dance-like* motion aligned to musical beats. (b) **Right**: The Bimodal Energy Attention Module (BEAM) adaptively updates the text conditions (red line) via MAP estimation, and global text information is applied through AdaLN modulation.

modalities such as music [36, 59], style (genre) [1, 33], and text [15, 65]. **Music-driven dance generation** focuses on synthesizing motion coherent with musical structure. For example, EDGE [63] uses a transformer-based diffusion model conditioned on music features extracted from Jukebox [7], and [53] developed DanceMosaic, which incorporates music in its masked motion modeling. **Style-conditioned dance generation** aims to generate motion aligned with specific dance styles (*e.g.*, ballet, hip-hop). [64] proposed DGSDP, a diffusion model guided by textual style prompts. LDA [1] conditioned on both music and style, applied classifier-free guidance [23] to enable smooth interpolation between different styles. **Language-Based Dance Motion Generation** leverages detailed motion descriptions from general text-motion datasets, such as HumanML3D [17]. For instance, TM2D [16] presented a novel approach to generating 3D dance movements using a VQ-VAE conditioned on text and music modalities. DanceEditor created dance pairs with difference descriptions, allowing text prompted dance motion editing [67].

Despite these advances, existing methods still struggle with limitations such as insufficient fine-grained control [46, 53], and a persistent semantic gap between textual descriptions and the nuanced expressiveness of dance [15, 47, 65]. Our method bridges this gap by introducing a controllable, text-guided framework that captures fine-level dance semantics while maintaining strong alignment with music. Also, STREAM enables interpretable motion control, preserves dance structure, and produces expressive movements that better reflect choreographic intent.

3 Preliminaries

Energy-Based Models (EBMs) define a probability distribution over data via a scalar energy function, expressing the density in the exponential family form:

$$p_\theta(x) = \frac{1}{Z(\theta)} \exp(-E_\theta(x)), \tag{1}$$

where θ denotes model parameters, $E_\theta(x) : \mathbb{R}^d \to \mathbb{R}$ is the scalar energy function, and $Z(\theta) = \int \exp(-E_\theta(x))dx$ is the partition function. Training aims to shape $E_\theta(x)$ so that data samples obtain low energy while unobserved configurations map to high energy. Learning an explicit energy landscape also enables compositional reasoning, as distinct concepts can be combined through Boolean-like operations [9, 21, 68]. Furthermore, this formulation allows for inference-time optimization to drive iterative refinement [10, 13]. Finally, because EBMs impose no structural assumptions on the form of E_θ, they offer strong modeling flexibility [57].

A major challenge is the intractable partition function, which complicates sampling and training [8]. To mitigate this, several approximations are used: (1) Contrastive Divergence (CD) employs MCMC sampling to estimate likelihood gradients [21, 42]; (2) Score Matching (SM) learns the score function $s_\theta(x) = \nabla_x \log p_\theta(x) = -\nabla_x E_\theta(x)$, avoiding the partition term [27, 58]; and (3) Noise Contrastive Estimation (NCE) reformulates learning as classifying data versus noise samples [10, 18].

Relation to Hopfield Networks and Self-Attention: [51] showed that self-attention can be viewed as minimizing an energy function of a modern Hopfield network:

$$E(X, \xi) = -\text{lse}(\beta X^\top \xi) + \tfrac{1}{2}\xi^\top \xi + C, \tag{2}$$

where $\text{lse}(X) = \log \sum_i \exp(x_i)$ and $\xi \in \mathbb{R}^{M \times d}$ is the state. Minimizing E using the Concave–Convex Procedure (CCCP) [66] yields the update

$$\xi_{\text{new}} = X\text{Softmax}(\beta X^\top \xi), \tag{3}$$

which corresponds exactly to the self-attention mechanism, with X as key/value, ξ as query, and $\beta = 1/\sqrt{d}$. This establishes a direct theoretical link between EBMs and attention-based architectures. Adapting this perspective, we propose a new energy functions with both text and music conditions. By minimizing the new energy function, STREAM is able to have semantic control while preserving musical alignment.

4 Method

In this section, we present the pipeline for the diffusion-based Structural-Temporal Energy-based Attention for dance Motion (STREAM) generation pipeline (Fig. 2). STREAM strictly enforces the spatial and semantic structure of the human motion

(the "What") as specified by the text descriptions. The model explicitly aligns the generated motion with the acoustic cues (the "When"), ensuring that the music modulates the temporal dynamics of the semantic motion. By decoupling the structural semantics from the rhythmic timing via the Bimodal Energy-based Attention Module (BEAM), STREAM prevents modality collapse and enables semantically controllable, yet musically aligned, motion generation.

4.1 Problem Statement

Our goal is to generate a human dance motion $x^{1:L}$ of length L, conditioned on music c_m and text $c_t = \{c_h, c_l\}$, where c_h represents a high-level concept such as style, dance technique, or action label, and c_l is a low-level detailed description of human motion. The two modalities play distinct roles: the text conditions the motion behavior, while the music modulates temporal rhythm and beat alignment. We model the conditional distribution $p(x^{1:L}|c_t, c_m)$ using a denoising diffusion probabilistic model [22] formulated from an energy-based perspective [44, 51], which allows the system to capture structured dependencies between motion, text, and music cues.

Motion Representation: We represent a motion sequence of length L as $x^{1:L} = \{x^1, \cdots, x^L\} \in \mathbb{R}^{L \times D}$, sampled at 30 FPS over 5 second clips. Each frame x^i follows the SMPL parameterization [38], defined as $x^i = (t_g^i, \alpha^i, \beta^i)$, where $\alpha^i \in \mathbb{R}^{24\times 6}$ denotes joint rotations in 6D representation [70], $t_g^i \in \mathbb{R}^3$ is the global root translation, and $\beta^i \in \mathbb{R}^{10}$ encodes the body shape. Following prior works [3, 29, 30, 61], all motions are canonicalized with respect to the first frame x^1, oriented to face the forward direction, and aligned such that the floor lies on the positive xy-plane.

4.2 STREAM Pipeline

Diffusion Framework: Following the great success of diffusion frameworks for human motion generation tasks [1, 6, 60, 63, 68], STREAM uses the DDPM [22] framework to model $p(x|c_t, c_m)$. The forward diffusion process progressively adds noise to a data sample, $x_0 \sim p(x)$, according to a predefined noise scheduler, α_t, yielding $x_T \sim \mathcal{N}(0, \mathbf{I})$ at time T:

$$p(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I}) \tag{4}$$

Then the reverse process, parameterized by θ, denoises samples step-by-step using a Gaussian transition [22, 56]:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\mathbf{I}) \tag{5}$$

Our denoiser adopts a UNet-like architecture [6, 52] augmented with multiple Bimodal Energy-based Attention Module (BEAM) layers. The BEAM consists of three components: (1) a Text Adaptive Layer Normalization (Text-AdaLN),

(2) a Dual Energy-based Cross-Attention (D-EBCA), and (3) a Bayesian Update Module. Text-AdaLN modulates global text-condition information into motions, and then the D-EBCA minimizes the energy function defined by the text, music, and motions. Finally, the abstract text condition is optimized via MAP estimation [44, 68] as shown in Fig. 2. (b).

We further employ classifier-free guidance (CFG) [23] for multi-conditioned motion generation. During training, we randomly mask the text condition and music condition with probabilities p_{cfg_t}, p_{cfg_m}, respectively. To ensure the music condition *modulates* the rhythm without overpowering the semantic structure during inference, we employ Hierarchical Classifier-Free Guidance, by guiding text semantic with rhythmic delta. In this way, we guarantee that the semantic manifold is established before rhythmic modulation is applied. Specifically, our inference time CFG equation is:

$$\hat{x}_{t-1} = \hat{x}_\theta(x_t) + \lambda_t(\hat{x}_\theta(x_t, c_t) - \hat{x}_\theta(x_t)) + \lambda_m(\hat{x}_\theta(x_t, c_t, c_m) - \hat{x}_\theta(x_t, c_t)) \quad (6)$$

where λ_t and λ_m are guidance hyperparameters of text and music, respectively with condition of $\lambda_t > \lambda_m$.

Condition Modules: As outlined in the previous section, the music and text conditions serve different purposes: the text controls semantic generation, and the music refines the motion's temporal and stylistic details, transforming it into a dance. The music module extracts features from each music clip using a pretrained Jukebox model [7] sampled at 30 Hz [39, 63]. The extracted features are projected to form the music condition embedding $c_m \in \mathbb{R}^{L\times D}$. On the other hand, the text condition comprises both high-level information and low-level descriptions, denoted by $c_t = \{c_h, c_l\}$. We encode c_h and c_l using a pretrained CLIP text encoder [50], concatenate to project them onto an initial embedding $c_t^0 \in \mathbb{R}^{L\times D}$, as shown in Fig. 2. This embedding is iteratively refined through the BEAM layers, enabling hierarchical alignment between linguistic cues and motion features. Consequently, c_t effectively captures both global motion semantics (*e.g.*, "Charleston cross step") and detailed movement description (*e.g.*, "The person crosses one leg behind the other" shown with blue letters in Fig. 2).

4.3 Bimodal Energy-based Attention Module (BEAM)

Text contains abstract and high-level information. In choreography, body movements are highly dense and complex, unlike everyday human motion, *e.g.*, HumanML3D [17]. Dancers do not move randomly; their movements follow well-defined structural conventions, commonly referred to as dance technique. While music provides dynamic, fine-grained local cues, the dancers respond to instantaneously, it often carries stronger correlational signals than text in dance motion generation tasks. As a result, models tend to ignore the textual condition and rely predominantly on musical cues. To address this imbalance, we propose the Bimodal Energy-Based Attention Module, which structurally enforces the disentanglement of text and music conditions through an inductive bias.

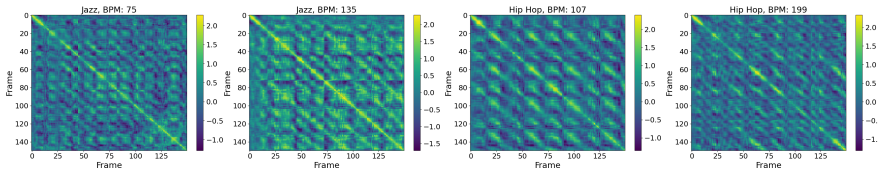

Fig. 3: Visualization example of self-similarity matrices of different music and BPM at first layer of D-EBCA.

Text-AdaLN: Our text condition comprises both high-level information and low-level descriptions of body movements. Therefore, we apply text conditions globally to motion query through AdaLN [45]. Specifically, the motion query x is modulated as $x' = \gamma(c_t) \odot \text{LayerNorm}(x) + \beta(c_t)$, where the scale γ and shift β are linearly projected from the text condition. In this way, the global text information physically transforms the query latent vector onto the correct semantic manifold. We apply text-AdaLn twice: once after the Transformer encoder with previous text conditions c_t^i, and the other after Dual Energy-based Cross-Attention with newly updated text conditions c_t^{i+1}.

Dual Energy-Based Cross Attention: The Dual Energy-Based Cross Attention (D-EBCA) forms the core of STREAM, coupling motion dynamics with textual semantics and aligning them musically through an energy-minimization formulation. Inspired by [44], we reinterpret cross-attention as an *energy-based model (EBM)* [51], where attention weights are optimized to minimize an energy function that encodes both *semantic alignment* and *structural regularity*. Specifically, query, key, and value are defined as

$$Q = xW_Q, \quad K_t = c_t W_{K^t}, \quad K_m = c_m W_{K^m}, \quad V_t = c_t W_V,$$

where x denotes motion features, c_t is the previous text embedding, and c_m is music embedding. We define the energy functions as:

$$\begin{aligned} E(Q; K_t, K_m) = -\frac{1}{\beta} \sum_{l=1}^{L} \text{lse}(A_l, \beta) + \mathcal{R}(Q, K_m) \\ + \frac{\alpha_t}{2}\|K_t\|^2 + \frac{\alpha_m}{2}\|K_m\|^2 + \frac{1}{2}\|Q\|^2 \end{aligned} \tag{7}$$

$$\mathcal{R}(Q, K_m) = -\sum_{i,j} (S_m)_{i,j} \cdot (q_i^\intercal q_j) \tag{8}$$

where $A = QK_t^\intercal + \gamma_m QK_m^\intercal$, $\text{lse}(\cdot)$ denotes the log-sum-exp operator, and S_m is the self-similarity matrix [12] of K_m. $\mathcal{R}(Q, K_m)$ is the music alignment energy

function, measuring how well musical information is aligned with the given motion query. Minimizing Eq. (7) yields the forward path:

$$Q_{new} = Q - \eta \nabla_Q E(Q, K_t, K_m) \approx \underbrace{\text{Softmax}(d^{-1/2}A)V_t}_{\text{Dual Attractor}} - \underbrace{\gamma_d \nabla_Q \mathcal{R}(Q, K_m)}_{\text{Music Alignment Drift}} \quad (9)$$

with $\eta = 1$ and $\nabla_{q_i} \mathcal{R}(Q, K_m) = -2 \sum_j (S_m)_{i,j} \cdot q_j$ where q_j is the j-th row vector of Q and γ_d is the hyperparameter. While the exact optimization yields the key matrix K_t, following standard Transformer architectures [51], we relax this by projecting the states into a separate value space V_t, enhancing the network's expressive capacity.

While the theoretical music alignment drift $\nabla_Q \mathcal{R}$ provides a principled mechanism for music alignment, naively applying it within a diffusion framework introduces instabilities, such as magnitude explosion and high-frequency noise injection. To mitigate instability, we normalize music alignment drift term before applying it. Furthermore, we mask diagonal part of self-similarity matrix as it has strong bias signals from how they are computed (see at Fig. 3).

Bayesian Update Module: Because the text condition contains abstract, high-level information, it is underspecified early in the diffusion process [44]. During the generation process, as motion takes shape (*e.g.*, a specific type of "happy jump dance"), the text embedding should be refined to focus on that specific mode of the distribution. Following previous work [44, 68], we update the text conditions via MAP estimation. The gradient of the log-posterior is approximated as

$$\nabla_{K_t} \log p(K_t | Q, K_m) = -(\nabla_{K_t} E(Q; K_t, K_m) + \nabla_{K_t} E(K_t)) \quad (10)$$

where $E(K_t) := \text{lse}(\frac{1}{2}\text{diag}(K_t K_t^\intercal), 1)$. We update the text embedding with $\alpha_t = 0$:

$$c_t^i = c_t^{i-1} + \Big(\delta_a \text{Softmax}(\frac{1}{\sqrt{d}}A)Q - \delta_r \mathcal{D}\big(\text{Softmax}(K_t')\big)K_t\Big)W_K^\intercal \quad (11)$$

where $\mathcal{D}(\cdot)$ denotes diagonalization operators and $K_t' = \text{diag}(K_t K_t^\intercal)$ and δ_a, δ_r are the hyperparameters to control update rate.

4.4 Loss Functions

Our primary training objective follows the simplified DDPM formulation [22], in which the model estimates the clean motion sample [60] directly, rather than predicting noise [1] or velocity [40]. The main loss is defined as

$$\mathcal{L}_m = \mathbb{E}_{x_0,t}[\|x_0 - \hat{x}_\theta(x_t, t, c_t, c_m)\|_2^2] \quad (12)$$

where $x_0 \sim p(x_0|c_t, c_m), t \sim [1, T]$. Following prior works [1,60,63], we incorporate auxiliary objectives to improve physical plausibility and temporal smoothness, including joint position loss, foot-skating loss, and velocity loss. Further implementation details are provided in the supplementary material.

Table 1: Comparison of selected single dance motion datasets. *Text* refers to whether the dataset has corresponding text pairs. On the other hand, *Dance Technique* is a domain-specific frame-level label, such as 'Charleston cross step'.

Dataset Name	Total hours	# Genres	MoCap	Text	Dance Technique
DanceRevolution [24]	12 h	3	x	x	x
AIST++ [36]	5.19 h	10	x	x	x
PopDanceSet [39]	3.56 h	19	x	x	x
DanceNet [71]	0.96 h	2	✓	x	x
ChoreoSpectrum3D [19]	70.32 h	4	✓	x	x
Motorica Dance [1]	6.22 h	8	✓	x	x
FineDance [34]	14.6 h	22	✓	x	x
DanceRemix [67]	117.39 h	10	x	✓	x
Motorica++ (Ours)	**4.61 h**	**8**	✓	✓	✓

5 Experiments

5.1 Experimental Settings

Datasets: We evaluate STREAM on AIST++ [35] and Motorica++ (an extension of Motorica [1]) for dance motion generation. Because the two datasets use different human-body representations, we align Motorica to the SMPL [38] representation using LBFGS optimization. As AIST++ doesn't have corresponding text information, we train the model only with music condition.

Motorica++: To improve text–motion alignment, we augment Motorica [1] with fine-grained dance annotations. Since dance motions differ significantly from general human movement (see Fig. 1), a professional dancer labeled the Motorica frames by genre, dance technique, and detailed descriptions of body movements. Due to missing musical references, 34 samples were excluded, leaving 97 fully annotated sequences, equivalent to 4.62 hours (see Tab. 1). To the best of our knowledge, this is the first dataset specifically annotated with frame-level domain-specific dance techniques and text-based motion descriptions. Please check the supplementary material for more details.

Evaluation Metrics: We follow traditional dance motion generation metrics, including both kinetic and geometric features, as proposed by [35]. We report FID_k, FID_g, Dist_k, and Dist_g [20,41,43,55], as well as the Beat Alignment Score (BAS) [35]. However, these metrics capture only the quality of the generated motion with respect to the music. Therefore, we propose a new experimental setup and metric to express dance editability.

Editable Dance Score (EDS): To properly assess whether the system can generate motion conditioned on text while faithfully following the given musical cues, we need to design a specialized experimental setup, which we call the

Table 2: Comparison with SoTAs on the AIST++ dataset and Motorica++, where BAS refers to Beat-Alignment Score. → means closer to ground truth is better. 'A' and 'T' represent audio and text modality, respectively. **Bold** and underline indicate the best and 2nd results. TM2D* is trained with both audio and text.

	Method	Modality	Motion Quality		Motion Diversity		BAS ↑	EDS		
			FID_k ↓	FID_g ↓	$Dist_k$ →	$Dist_g$ →		S_{text} ↑	S_{music} ↑	EDS ↑
AIST++	Ground Truth	A	-	-	10.03	7.38	0.2633	-	-	-
	Li et al. [31]	A	86.43	43.46	6.85	3.32	0.1607	-	-	-
	DanceNet [71]	A	69.18	25.49	2.86	2.85	0.1430	-	-	-
	DanceRevolution [24]	A	73.42	25.92	3.52	4.87	0.1950	-	-	-
	FACT [36]	A	35.35	22.11	5.94	6.18	0.2209	-	-	-
	Bailando [55]	A	**28.16**	**9.62**	7.83	6.34	0.2332	-	-	-
	EDGE [63]	A	42.16	22.12	3.96	4.61	0.2334	-	-	-
	Lodge [33]	A	37.09	18.79	5.58	4.85	**0.2513**	-	-	-
	STREAM (ours)	A	29.58	11.54	**8.74**	**7.63**	0.2312	-	-	-
Motorica++ (ours)	Ground Truth	A + T	-	-	10.54	7.33	0.2413	-	-	-
	EDGE [63]	A	67.52	18.34	4.70	7.20	0.2052	0.8318	0.3904	0.5272
	POPDG [39]	A	27.02	8.56	6.73	5.80	0.2345	0.8294	**0.5365**	0.6501
	DanceFusion [62]	A	31.34	10.53	7.43	7.33	0.2076	0.8164	0.3383	0.4780
	Danceba [11]	A	41.24	11.36	7.36	5.99	0.1947	0.8363	0.3983	0.5352
	MDM [60]	T	37.29	13.66	12.46	9.68	-	0.9684	0.4879	0.6467
	MLD-5 [6]	T	87.51	56.54	16.56	11.82	-	**1.0**	0.3727	0.5423
	ReMoDiffuse [69]	T	338.73	468.33	18.42	19.37	-	0.8053	0.4860	0.6048
	TM2D* [15]	T	313.08	308.58	18.38	13.02	-	0.8936	0.4818	0.6252
	TM2D* [15]	A + T	126.46	570.84	**10.82**	15.23	**0.2744**	0.8391	0.4625	0.5952
	UniMuMo [65]	A + T	28.10	169.14	8.26	11.06	0.2360	0.7514	0.3945	0.5162
	STREAM (ours)	A	14.89	10.69	8.67	7.01	0.2249	0.8464	0.4526	0.5853
	STREAM (ours)	T	7.80	**6.15**	9.99	**7.08**	-	**1.0**	0.4533	0.6207
	STREAM (ours)	A + T	**7.32**	6.87	10.30	6.55	0.2528	**1.0**	0.4870	**0.6539**

Exchange Evaluation Protocol. First, we select from the test dataset the samples longer than 3 seconds text-music pairs. Then we change the paired music to another based on BPM (Beat Per Minute) difference (we categorize three different tempo ranges- low, medium, and high). Lastly, we generate new text-music pairs, where the dance motion (from text) totally mismatches the music, *e.g.*, fast hip hop dance with slow jazz music.

To properly evaluate the exchange protocol, we propose a new metric,EDS, the harmonic mean of semantic preservation (S_{text}) and rhythmic adaptation (S_{music}). Specifically, we use the finetuned normalized TMR [48] CLIP score as S_{text} and normalized BAS as S_{music}, then we define EDS as $\text{EDS} := \frac{2 \cdot S_{text} \cdot S_{music}}{S_{text} + S_{music}}$. EDS can capture whether the generated motions are semantically and rhythmically well-aligned. For more details, please check supplementary material.

Implementation Details: All datasets are segmented into 5-second clips with a 1-second hop size and sampled at 30 FPS for both motion and audio, yielding sequences of length $L = 150$. For long-term dance generation, we stitch 2.5-second segments in an overlapping manner using the dynamics CFG and latent blending, similar to EDGE [63] (please see the supplementary material for more details). The model employs 9 BEAM layers with an embedding dimension of $D = 512$.

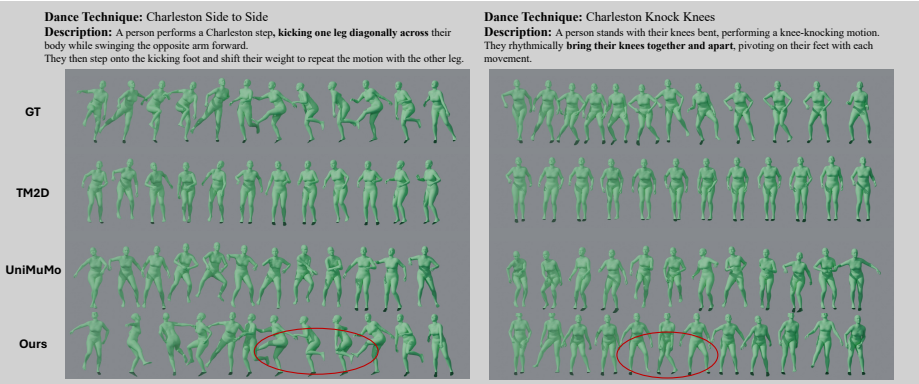

Fig. 4: Qualitative dance motion generation results with text and music conditioned, compared with other SoTA models.

Baselines: Since we propose a 'text + music to motion generation pipeline' with a new dataset, we retrain state-of-the-art dance models [39, 62, 63], text to motion models [6, 60, 69], and text+music-conditioned models [15, 65], following their original training recipes.

5.2 Evaluation of Dance Motion Generation

We evaluate STREAM on two datasets: 1. AIST++ [36], and 2. Motorica++. Since, AIST++ doesn't have text annotations, we train our model using only music conditioning and evaluate it with standard metrics such as FID, Dist, and BAS. In contrast, Motorica++ contains high-quality text annotations. Therefore, we train both music-to-motion and text-to-motion models, including models conditioned on both music and text. Because, most text-motion models are trained and evaluated using text-motion pairs (with motion length varying depending on the text prompt), we evaluate the entire Motorica++ based on text-motion pairs, where motion durations range from 0.5 seconds to 10 seconds. Quantitative results are reported in Tab. 2. On Motorica++, our method achieves state-of-the-art performance across FID, Dist, and BAS. On AIST++, STREAM attains the best motion diversity (Dist) while remaining competitive with specialized music-to-dance models on FID and BAS. Fig. 4 presents qualitative comparisons with SoTA methods [15, 65] under text conditioning. The results show that our model can generate sophisticated dance motions that closely follow the provided text prompts.

5.3 Evaluation of Editable Dance Generation

We evaluate the editability of dance motion generation using the proposed *Exchange Evaluation Protocol* with EDS metric in Tab. 2, which measures the joint alignment performance w.r.t. both text and music. Most single-modality methods

Fig. 5: Visualization of dance motion editing example. The original motion contains two dance techniques: Charleston Opposites (green) and Charleston Messaround (blue). We first edit Charleston Opposites to Charleston Side to Side (red) while preserving the other. Similarly, we can edit one more time to Charleston Knock Knees (yellow).

exhibit a trade-off between S_{music} and S_{text}. For example, POPDG achieves (0.8294 vs 0.5365) and MLD (1.0 vs 0.3727), indicating strong performance in one modality but weaker alignment in the other. Furthermore, existing audio-text multimodal pipelines often fail to capture both modalities effectively due to the way the two modalities are combined. For instance, TM2D [15] primarily uses audio features and incorporates text features through window-based late fusion. Because of this design, the features are conflicted in some samples, showing TM2D text-only shows better result. In contrast, STREAM disentangles text and music modalities through the proposed energy function and Hierarchical CFG. As result, the model achieves strong text controllability while maintaining accurate musical alignment, enabling effective editing of dance motions based on text prompts. Fig. 5 shows qualitative results of editing quality. The initial dance text conditions are *Opposites* and *Messaround*. We edit only Opposites to *Side to Side* or *Knock Knees*. This demonstrates fine-grained, text-conditioned edits with high motion fidelity.

5.4 Ablation Study

Proposed Modules: We conduct ablation studies to evaluate the effectiveness of the proposed modules by varying γ_m and γ_d in the energy function in Eq. (9), as well as the normalization applied to the music alignment drift term ($\nabla_Q \mathcal{R}$). Compared with the Abl-4 model in Tab. 3, removing the music alignment energy function mainly reduces the beat alignment score (BAS). Although other quality metrics remain largely unchanged, the BAS decreases, indicating that the model relies more heavily on textual information. Similarly, removing AdaLN leads to a significant drop in BAS. This suggests that the model attempts to capture global structure from the music features, placing excessive reliance on them and limiting

Table 3: Ablation study on different γ and normalization at Eq. (9), AdaLN and Bayesian music context c_m update. Note that $\gamma_m = \gamma_d = 0$ means the model only uses text information.

Name	$\mathbf{FID}_k \downarrow$	$\mathbf{FID}_g \downarrow$	$\mathbf{Dist}_k \rightarrow$	$\mathbf{Dist}_g \rightarrow$	BAS $\uparrow$
STREAM($\gamma_m = 1.0, \gamma_d = 1.0$)	7.32	6.87	10.30	6.55	0.2528
Abl-1 ($\gamma_m = 1.0, \gamma_d = 0.5$)	9.33	5.31	10.79	6.42	0.2405
Abl-2 ($\gamma_m = 1.0, \gamma_d = 0.0$)	9.10	6.19	10.21	6.38	0.2369
Abl-3 ($\gamma_m = 0.0, \gamma_d = 0.5$)	8.32	6.58	10.60	6.43	0.2468
Abl-4 ($\gamma_m = 0.0, \gamma_d = 0.0$)	10.38	5.75	9.04	6.38	0.2301
w/o Norm ($\gamma_m = 1.0, \gamma_d = 0.5$)	10.37	10.17	10.34	7.30	0.2251
w/o AdaLN ($\gamma_m = 1.0, \gamma_d = 1.0$)	7.79	4.28	10.54	7.33	0.2285
c_m update	7.16	7.32	10.51	6.58	0.2372

its ability to capture fine-grained local beat information. Finally, removing the normalization at $\nabla_Q \mathcal{R}$ degrades both the overall generation quality and the BAS.

Music Context Update: Similar to text context Bayesian update at Eq. (3), one could theoretically update the music condition via MAP estimation:$c_m^i = c_m^{i-1} - \nabla_{K_m}\big(E(Q; K_t, K_m) + E(K_m)\big)$. However, we empirically find this symmetric update degrades performance. While the text condition provides abstract, global semantics that benefit from iterative refinement as the motion materializes, the music condition provides strict, fine-grained temporal anchors. If the music context is updated via MAP estimation, the model tends to minimize the energy landscape by shifting the music condition to align with the currently generated motion, rather than forcing the motion to adapt to the true musical beats. Consequently, the context slowly deviates from the ground-truth acoustic properties, deteriorating both overall motion generation quality and Beat Alignment Scores (BAS), as demonstrated in Tab. 3. Therefore, STREAM employs an asymmetric update strategy: iteratively refining the abstract text semantics while keeping the rhythmic music condition strictly frozen.

6 Conclusion and Limitation

In this work, we present STREAM, a Structural-Temporal Rhythmic Energy-based Attention for dance Motion generation pipeline that jointly leverages music and text conditions via a novel energy function. By disentangling these modalities, text conditions govern the global motion structure and semantics of the motion, while music conditions enrich local temporal details, enabling controlled and expressive dance synthesis. In addition, we introduce Motorica++, a dance-specialized text-motion dataset that enables the generation of fine-grained dance techniques through language descriptions. A current limitation of our system is its focus on single-person dance generation. Many choreographic scenarios involve group performance, where dancers exhibit complex spatial formations,

coordinated timing, and inter-person interactions. Expanding STREAM to multi-dancer settings thus represents an important direction for future work.

7 Acknowledgements

We gratefully acknowledge support from the National Science Foundation under Grant BCS-2318255 and from the University of Maryland through the AIM Research Seed Award Program.

References

1. Alexanderson, S., Nagy, R., Beskow, J., Henter, G.E.: Listen, Denoise, Action! Audio-Driven Motion Synthesis with Diffusion Models. ACM Trans. Graph. **42**(4), 44:1–44:20 (2023). https://doi.org/10.1145/3592458, https://dl.acm.org/doi/10.1145/3592458
2. Angelov, V.: You, the Choreographer: Creating and Crafting Dance. Routledge, London (Aug 2023). https://doi.org/10.4324/9781003009764
3. Athanasiou, N., Ceske, A., Diomataris, M., Black, M.J., Varol, G.: MotionFix: Text-Driven 3D Human Motion Editing (Sep 2024). https://doi.org/10.48550/arXiv.2408.00712, http://arxiv.org/abs/2408.00712
4. Basso, J.C., Satyal, M.K., Rugh, R.: Dance on the Brain: Enhancing Intra- and Inter-Brain Synchrony. Frontiers in Human Neuroscience **14**, 584312 (Jan 2021). https://doi.org/10.3389/fnhum.2020.584312, https://pmc.ncbi.nlm.nih.gov/articles/PMC7832346/
5. Cameron, D.J., Bentley, J., Grahn, J.A.: Cross-cultural influences on rhythm processing: Reproduction, discrimination, and beat tapping. Frontiers in Psychology **6** (Apr 2015). https://doi.org/10.3389/fpsyg.2015.00366, https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2015.00366/full
6. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing Your Commands via Motion Diffusion in Latent Space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18000–18010 (2023), https://openaccess.thecvf.com/content/CVPR2023/html/Chen_Executing_Your_Commands_via_Motion_Diffusion_in_Latent_Space_CVPR_2023_paper.html
7. Dhariwal, P., Jun, H., Payne, C., Kim, J.W., Radford, A., Sutskever, I.: Jukebox: A generative model for music. arXiv preprint arXiv:2005.00341 (2020), https://assets.pubpub.org/2gnzbcnd/11608661311181.pdf
8. Du, Y., Durkan, C., Strudel, R., Tenenbaum, J.B., Dieleman, S., Fergus, R., Sohl-Dickstein, J., Doucet, A., Grathwohl, W.: Reduce, Reuse, Recycle: Compositional Generation with Energy-Based Diffusion Models and MCMC (Sep 2024). https://doi.org/10.48550/arXiv.2302.11552, http://arxiv.org/abs/2302.11552
9. Du, Y., Li, S., Mordatch, I.: Compositional Visual Generation with Energy Based Models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6637–6647. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper/2020/hash/49856ed476ad01fcff881d57e161d73f-Abstract.html
10. Du, Y., Mao, J., Tenenbaum, J.B.: Learning Iterative Reasoning through Energy Diffusion (Jun 2024). https://doi.org/10.48550/arXiv.2406.11179, http://arxiv.org/abs/2406.11179

11. Fan, C., Guan, J., Zhao, X., Xu, D., Lin, Y., Ye, T., Feng, P., Pan, H.: Align your rhythm: Generating highly aligned dance poses with gating-enhanced rhythm-aware feature representation. arXiv preprint arXiv:2503.17340 (2025)
12. Foote, J.: Visualizing music and audio using self-similarity. In: Proceedings of the Seventh ACM International Conference on Multimedia (Part 1). pp. 77–80. MULTIMEDIA '99, Association for Computing Machinery, New York, NY, USA (Oct 1999). https://doi.org/10.1145/319463.319472, https://dl.acm.org/doi/10.1145/319463.319472
13. Gladstone, A., Nanduru, G., Islam, M.M., Han, P., Ha, H., Chadha, A., Du, Y., Ji, H., Li, J., Iqbal, T.: Energy-Based Transformers are Scalable Learners and Thinkers (Jul 2025). https://doi.org/10.48550/arXiv.2507.02092, http://arxiv.org/abs/2507.02092
14. Goel, P., Wang, K.C., Liu, C.K., Fatahalian, K.: Iterative Motion Editing with Natural Language. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–9. SIGGRAPH '24, Association for Computing Machinery, New York, NY, USA (Jul 2024). https://doi.org/10.1145/3641519.3657447, https://dl.acm.org/doi/10.1145/3641519.3657447
15. Gong, K., Lian, D., Chang, H., Guo, C., Jiang, Z., Zuo, X., Mi, M.B., Wang, X.: TM2D: Bimodality Driven 3D Dance Generation via Music-Text Integration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9942–9952 (2023), https://openaccess.thecvf.com/content/ICCV2023/html/Gong_TM2D_Bimodality_Driven_3D_Dance_Generation_via_Music-Text_Integration_ICCV_2023_paper.html
16. Gong, K., Lian, D., Chang, H., Guo, C., Jiang, Z., Zuo, X., Mi, M.B., Wang, X.: Tm2d: Bimodality driven 3d dance generation via music-text integration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9942–9952 (2023)
17. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating Diverse and Natural 3D Human Motions From Text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5152–5161 (2022), https://openaccess.thecvf.com/content/CVPR2022/html/Guo_Generating_Diverse_and_Natural_3D_Human_Motions_From_Text_CVPR_2022_paper.html
18. Gutmann, M., Hyvärinen, A.: Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. pp. 297–304. JMLR Workshop and Conference Proceedings (Mar 2010), https://proceedings.mlr.press/v9/gutmann10a.html
19. Han, B., Ren, Y., Peng, H., Zhang, T., Ling, Z., Yin, X., Han, F.: Enchantdance: Unveiling the potential of music-driven dance movement. arXiv preprint arXiv:2312.15946 (2023)
20. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html
21. Hinton, G.E.: Training Products of Experts by Minimizing Contrastive Divergence. Neural Computation **14**(8), 1771–1800 (Aug 2002). https://doi.org/10.1162/089976602760128018, https://ieeexplore.ieee.org/abstract/document/6789337

22. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html
23. Ho, J., Salimans, T.: Classifier-Free Diffusion Guidance (Jul 2022). https://doi.org/10.48550/arXiv.2207.12598, http://arxiv.org/abs/2207.12598
24. Huang, R., Hu, H., Wu, W., Sawada, K., Zhang, M., Jiang, D.: Dance Revolution: Long-Term Dance Generation with Music via Curriculum Learning (Sep 2023). https://doi.org/10.48550/arXiv.2006.06119, http://arxiv.org/abs/2006.06119
25. Huang, Y., Wan, W., Yang, Y., Callison-Burch, C., Yatskar, M., Liu, L.: Como: Controllable motion generation through language guided pose code editing. In: European Conference on Computer Vision (ECCV) (2024)
26. Huang, Z., Xu, X., Xu, C., Zhang, H., Zheng, C., Qin, J., He, S.: Beat-it: Beat-synchronized multi-condition 3d dance generation. In: European Conference on Computer Vision. pp. 273–290. Springer (2024)
27. Hyvärinen, A., Dayan, P.: Estimation of non-normalized statistical models by score matching. Journal of Machine Learning Research **6**(4) (2005), https://www.jmlr.org/papers/volume6/hyvarinen05a/hyvarinen05a.pdf
28. Kaeppler, A.L.: Dance Ethnology and the Anthropology of Dance. Dance Research Journal **32**(1), 116–125 (2000). https://doi.org/10.2307/1478285, https://www.jstor.org/stable/1478285
29. Kulkarni, N., Rempe, D., Genova, K., Kundu, A., Johnson, J., Fouhey, D., Guibas, L.: NIFTY: Neural Object Interaction Fields for Guided Human Motion Synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 947–957 (2024), https://openaccess.thecvf.com/content/CVPR2024/html/Kulkarni_NIFTY_Neural_Object_Interaction_Fields_for_Guided_Human_Motion_Synthesis_CVPR_2024_paper.html
30. Lee, T., Moon, G., Lee, K.M.: MultiAct: Long-Term 3D Human Motion Generation from Multiple Action Labels (Feb 2023). https://doi.org/10.48550/arXiv.2212.05897, http://arxiv.org/abs/2212.05897
31. Li, J., Yin, Y., Chu, H., Zhou, Y., Wang, T., Fidler, S., Li, H.: Learning to Generate Diverse Dance Motions with Transformer (Aug 2020). https://doi.org/10.48550/arXiv.2008.08171, http://arxiv.org/abs/2008.08171
32. Li, J., Cao, J., Zhang, H., Rempe, D., Kautz, J., Iqbal, U., Yuan, Y.: GENMO: A GENeralist Model for Human MOtion (May 2025). https://doi.org/10.48550/arXiv.2505.01425, http://arxiv.org/abs/2505.01425
33. Li, R., Zhang, Y., Zhang, Y., Zhang, H., Guo, J., Zhang, Y., Liu, Y., Li, X.: Lodge: A Coarse to Fine Diffusion Network for Long Dance Generation Guided by the Characteristic Dance Primitives (Apr 2024). https://doi.org/10.48550/arXiv.2403.10518, http://arxiv.org/abs/2403.10518
34. Li, R., Zhao, J., Zhang, Y., Su, M., Ren, Z., Zhang, H., Tang, Y., Li, X.: FineDance: A Fine-grained Choreography Dataset for 3D Full Body Dance Generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10234–10243 (2023), https://openaccess.thecvf.com/content/ICCV2023/html/Li_FineDance_A_Fine-grained_Choreography_Dataset_for_3D_Full_Body_Dance_ICCV_2023_paper.html
35. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++ (2021)
36. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Learn to dance with aist++: Music conditioned 3d dance generation (2021)

37. Lin, J., Zeng, A., Lu, S., Cai, Y., Zhang, R., Wang, H., Zhang, L.: Motion-X: A Large-scale 3D Expressive Whole-body Human Motion Dataset. Advances in Neural Information Processing Systems **36**, 25268–25280 (Dec 2023), https://proceedings.neurips.cc/paper_files/paper/2023/hash/4f8e27f6036c1d8b4a66b5b3a947dd7b-Abstract-Datasets_and_Benchmarks.html
38. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM Transactions on Graphics **34**(6), 248:1–248:16 (2015). https://doi.org/10.1145/2816795.2818013, https://dl.acm.org/doi/10.1145/2816795.2818013
39. Luo, Z., Ren, M., Hu, X., Huang, Y., Yao, L.: POPDG: Popular 3D Dance Generation with PopDanceSet. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26984–26993 (2024)
40. Meng, Z., Han, Z., Peng, X., Xie, Y., Jiang, H.: Absolute Coordinates Make Motion Generation Easy (May 2025). https://doi.org/10.48550/arXiv.2505.19377, http://arxiv.org/abs/2505.19377
41. Müller, M., Röder, T., Clausen, M.: Efficient content-based retrieval of motion capture data. In: ACM SIGGRAPH 2005 Papers. pp. 677–685. SIGGRAPH '05, Association for Computing Machinery, New York, NY, USA (Jul 2005). https://doi.org/10.1145/1186822.1073247, https://dl.acm.org/doi/10.1145/1186822.1073247
42. Neal, R.M.: MCMC Using Hamiltonian Dynamics. In: Handbook of Markov Chain Monte Carlo. Chapman and Hall/CRC (2011)
43. Onuma, K., Faloutsos, C., Hodgins, J.K.: FMDistance: A Fast and Effective Distance Function for Motion Capture Data. Eurographics (Short Papers) **7**(10) (2008), http://reports-archive.adm.cs.cmu.edu/anon/2007/CMU-CS-07-164.pdf
44. Park, G.Y., Kim, J., Kim, B., Lee, S.W., Ye, J.C.: Energy-Based Cross Attention for Bayesian Context Update in Text-to-Image Diffusion Models. Advances in Neural Information Processing Systems **36**, 76382–76408 (Dec 2023), https://proceedings.neurips.cc/paper_files/paper/2023/hash/f0878b7efa656b3bbd407c9248d13751-Abstract-Conference.html
45. Peebles, W., Xie, S.: Scalable Diffusion Models with Transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023), https://openaccess.thecvf.com/content/ICCV2023/html/Peebles_Scalable_Diffusion_Models_with_Transformers_ICCV_2023_paper.html
46. Peng, S., Ladenheim, K., Shrestha, S., Fermüller, C.: Choreographing the digital canvas: A machine learning approach to artistic performance. arXiv preprint arXiv:2404.00054 (2024)
47. Peng, S., Ladenheim, K., Shrestha, S., Fermüller, C.: Generation of novel fall animation with configurable attributes. In: Proceedings of the 9th International Conference on Movement and Computing. MOCO '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3658852.3659087, https://doi.org/10.1145/3658852.3659087
48. Petrovich, M., Black, M.J., Varol, G.: TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In: International Conference on Computer Vision (ICCV) (2023)
49. Plappert, M., Mandery, C., Asfour, T.: The KIT Motion-Language Dataset. Big Data **4**(4), 236–252 (Dec 2016). https://doi.org/10.1089/big.2016.0028, https://www.liebertpub.com/doi/abs/10.1089/big.2016.0028
50. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable

Visual Models From Natural Language Supervision (Feb 2021). https://doi.org/10.48550/arXiv.2103.00020, http://arxiv.org/abs/2103.00020
51. Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Adler, T., Gruber, L., Holzleitner, M., Pavlović, M., Sandve, G.K., Greiff, V., Kreil, D., Kopp, M., Klambauer, G., Brandstetter, J., Hochreiter, S.: Hopfield Networks is All You Need (Apr 2021). https://doi.org/10.48550/arXiv.2008.02217, http://arxiv.org/abs/2008.02217
52. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer International Publishing, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
53. Shah, F.N., Shah, P., Saleem, M.U., Pinyoanuntapong, E., Wang, P., Xue, H., Helmy, A.: Dancemosaic: High-fidelity dance generation with multimodal editability. arXiv preprint arXiv:2504.04634 (2025)
54. Sievers, B., Polansky, L., Casey, M., Wheatley, T.: Music and movement share a dynamic structure that supports universal expressions of emotion. Proceedings of the National Academy of Sciences of the United States of America **110**(1), 70–75 (Jan 2013). https://doi.org/10.1073/pnas.1209023110, https://pmc.ncbi.nlm.nih.gov/articles/PMC3538264/
55. Siyao, L., Yu, W., Gu, T., Lin, C., Wang, Q., Qian, C., Loy, C.C., Liu, Z.: Bailando: 3D Dance Generation by Actor-Critic GPT With Choreographic Memory. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11050–11059 (2022), https://openaccess.thecvf.com/content/CVPR2022/html/Siyao_Bailando_3D_Dance_Generation_by_Actor-Critic_GPT_With_Choreographic_Memory_CVPR_2022_paper.html
56. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In: Proceedings of the 32nd International Conference on Machine Learning. pp. 2256–2265. PMLR (Jun 2015), https://proceedings.mlr.press/v37/sohl-dickstein15.html
57. Song, Y., Kingma, D.P.: How to Train Your Energy-Based Models (Feb 2021). https://doi.org/10.48550/arXiv.2101.03288, http://arxiv.org/abs/2101.03288
58. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-Based Generative Modeling through Stochastic Differential Equations (Feb 2021). https://doi.org/10.48550/arXiv.2011.13456, http://arxiv.org/abs/2011.13456
59. Sun, J., Wang, C., Hu, H., Lai, H., Jin, Z., Hu, J.F.: You Never Stop Dancing: Non-freezing Dance Generation via Bank-constrained Manifold Projection. Advances in Neural Information Processing Systems **35**, 9995–10007 (Dec 2022), https://proceedings.neurips.cc/paper_files/paper/2022/hash/40bfe6177e8aed33c982264cf9e6e62c-Abstract-Conference.html
60. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human Motion Diffusion Model (Oct 2022). https://doi.org/10.48550/arXiv.2209.14916, http://arxiv.org/abs/2209.14916
61. Tripathi, S., Taheri, O., Lassner, C., Black, M.J., Holden, D., Stoll, C.: HUMOS: Human Motion Model Conditioned on Body Shape (Sep 2024). https://doi.org/10.48550/arXiv.2409.03944, http://arxiv.org/abs/2409.03944
62. Truong-Thuy, T.V., Bui-Le, G.C., Nguyen, H.D., Le, T.N.: Rethinking Sampling for Music-Driven Long-Term Dance Generation. In: Proceedings of the Asian Conference on Computer Vision. pp. 2667–2683 (2024), https://openaccess.

thecvf.com/content/ACCV2024/html/Truong-Thuy_Rethinking_Sampling_for_Music-Driven_Long-Term_Dance_Generation_ACCV_2024_paper.html
63. Tseng, J., Castellon, R., Liu, K.: Edge: Editable dance generation from music. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 448–458 (2023)
64. Wang, H., Zhu, Y., Geng, X.: Flexible music-conditioned dance generation with style description prompts. arXiv preprint arXiv:2406.07871 (2024)
65. Yang, H., Su, K., Zhang, Y., Chen, J., Qian, K., Liu, G., Gan, C.: UniMuMo: Unified Text, Music, and Motion Generation. Proceedings of the AAAI Conference on Artificial Intelligence **39**(24), 25615–25623 (Apr 2025). https://doi.org/10.1609/aaai.v39i24.34752, https://ojs.aaai.org/index.php/AAAI/article/view/34752
66. Yuille, A.L., Rangarajan, A.: The Concave-Convex Procedure (CCCP). In: Advances in Neural Information Processing Systems. vol. 14. MIT Press (2001), https://proceedings.neurips.cc/paper/2001/hash/a012869311d64a44b5a0d567cd20de04-Abstract.html
67. Zhang, H., Li, Z., Qi, X., Li, M., Sun, M., Zhang, M., Han, S.: DanceEditor: Towards Iterative Editable Music-driven Dance Generation with Open-Vocabulary Descriptions (Aug 2025). https://doi.org/10.48550/arXiv.2508.17342, http://arxiv.org/abs/2508.17342
68. Zhang, J., Fan, H., Yang, Y.: EnergyMoGen: Compositional Human Motion Generation with Energy-Based Diffusion Model in Latent Space (Dec 2024). https://doi.org/10.48550/arXiv.2412.14706, http://arxiv.org/abs/2412.14706
69. Zhang, M., Guo, X., Pan, L., Cai, Z., Hong, F., Li, H., Yang, L., Liu, Z.: ReMoDiffuse: Retrieval-Augmented Motion Diffusion Model (Apr 2023). https://doi.org/10.48550/arXiv.2304.01116, http://arxiv.org/abs/2304.01116
70. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the Continuity of Rotation Representations in Neural Networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5745–5753 (2019), https://openaccess.thecvf.com/content_CVPR_2019/html/Zhou_On_the_Continuity_of_Rotation_Representations_in_Neural_Networks_CVPR_2019_paper.html
71. Zhuang, W., Wang, C., Chai, J., Wang, Y., Shao, M., Xia, S.: Music2dance: Dancenet for music-driven dance generation. ACM Trans. Multimedia Comput. Commun. Appl. **18**(2) (Feb 2022). https://doi.org/10.1145/3485664, https://doi.org/10.1145/3485664