HiChor: Hierarchical Choreography Generation from Pop Music with Choreographic Primitives

Jungsu Kim[†1], Jungwoo Huh[†1], Jeongwook Choi[1], Wen-Huang Cheng[2], Jian-Yu Jiang-Lin[2], Weisi Lin[3], and Sanghoon Lee[*1]

[1] Yonsei University, Seoul, Republic of Korea {integer, gjwjddn9, yunakevin, slee}@yonsei.ac.kr
[2] National Taiwan University, Taipei, Taiwan wenhuang@csie.ntu.edu.tw jianyu@cmlab.csie.ntu.edu.tw
[3] Nanyang Technological University, Singapore wslin@ntu.edu.sg

Abstract. Choreography is a structured creative process grounded in musical form, rather than a mere sequence of improvised movements. Professional choreographers design choreography following a set of choreographic primitives: (1) organizing movements at the phrase level, (2) refining movements by aligning them with musical beats and energy, and (3) stylizing phrases through semantic text descriptions. However, existing dance generation methods overlook these primitives, leading to dances that feel structurally incomplete and resemble improvisation rather than professional choreography. Here, we present HiChor: Hierarchical Choreography generation from pop music, a framework inspired by the structured workflow of professional choreographers, which is grounded in choreographic primitives. HiChor hierarchically generates choreography from pop music by (1) generating phrase-level choreography aligned with beats, (2) enhancing choreography by aligning with music's energy, and (3) concatenating all phrase-level choreography and enhancing it through plausibility enhancement to ensure smooth transitions between phrases. Furthermore, we introduce a semantic text stylization module that converts semantic text descriptions into low-level choreography feature captions using an LLM, enabling semantic control. To assess how well the generated choreography adheres to choreographic principles, we introduce new metrics—Phrase Diversity, Feature Alignment, and Text Stylization Score. Experiments demonstrate that HiChor outperforms existing methods in both choreographic quality and semantic stylization.

Keywords: Choreography Generation · Diffusion Transformer · Dance Stylization

1 Introduction

Dance is one of the most fundamental forms of human motion, expressing human intentions and expressions. Over time, dance has evolved from spontaneous

[†] Equal contribution. [*] Corresponding author.

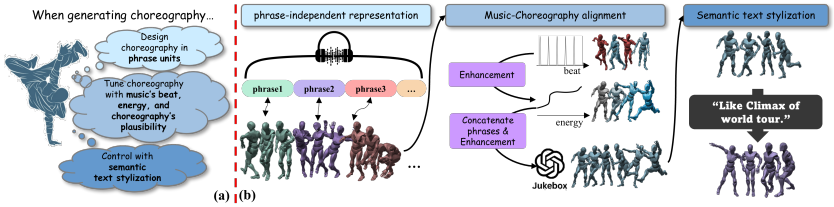

Fig. 1: Choreographic primitives and our framework. (a) Choreographic primitives used to design choreography based on theoretical principles. (b) Our framework generates choreography based on these primitives, achieving phrase-independent representation, music–choreography alignment, and semantic text stylization.

movements to a structured creative process. With the rise of human motion understanding in computer vision, dance generation has also become an active research topic, with significant advances. Recent advances in dance generation [17, 31, 33, 36, 38, 50, 51, 53, 55] have achieved remarkable progress in producing plausible dance with music. State-of-the-art works [17, 30, 31] extend beyond naturalness to tackle challenges such as generating longer sequences or ensuring strict beat alignment. However, prior approaches focus on tackling technical problems, such as synchronizing musical beats with human movements or generating longer dances by mirroring arbitrarily segmented, fixed-length dance sequences. As a result, the generated dances often resemble improvisation rather than professional choreography, where movements are intentionally created and arranged with musical structure. In this paper, we argue that moving beyond improvisational dance requires identifying the structural elements considered in choreographic theory and incorporating them into choreography generation.

Choreography is defined as "the art of designing and arranging the steps and movements in dances" [43]. This highlights that choreography is not merely improvisation but a creative process involving the structured design of movements aligned with music. Prominent choreographic theorists such as Blom, Humphrey, and Minton [2, 19, 41] described fundamental principles for structuring choreography, noting that professional choreography is built upon such compositional foundations, which we refer to in this work as *choreographic primitives*. First, choreography is structured into phrases, where each phrase forms an independent unit of motion. Phrases are aligned with the musical structure, typically spanning four beats, and then sequentially combined to form a complete choreography [2, 19, 41]. Second, the choreography should harmonize with the core elements of music, beat and energy [28], while maintaining the overall plausibility of movement. Third, choreographers incorporate subjective stylization into specific parts of the performance, reflecting their own style. To generate choreography that follows these primitives, we introduce *HiChor*, a hierarchical choreography generation framework that applies choreographic primitives at each stage to produce structured and musically aligned choreography.

Figure 1 illustrates the overall approach of HiChor. For phrase-independent representation, we slice the music into phrases and independently generate choreography for each phrase. This enables diverse representations across musical

phrases. For music-choreography alignment, we employ a hierarchical diffusion architecture that aligns musical features with choreography features while maintaining plausibility. By hierarchically generating and progressively enhancing movement based on musical features, HiChor injects the choreographic structure corresponding to each feature stage. Each musical feature is represented by its onset strength to capture beat information and RMS energy to represent musical intensity. This process is realized through the ChoreoEdit pipeline, which aligns multiple musical and choreographic features and also addresses the discontinuity between phrases. For semantic text stylization, we infer the choreographer's stylistic intent while preserving the underlying content of the original choreography. To achieve this, we leverage an LLM to interpret the semantic text describing the choreographer's intent and convert it into captions of low-level choreography features. We then use these feature-level captions as control signals in the ChoreoEdit pipeline to guide stylized choreography generation.

As this work targets professional choreography rather than simple dance movements, we introduce new metrics based on choreographic principles and an evaluation dataset captured in a dance studio: EvalPop. Unlike existing dance datasets [12, 32–34, 38] that provide short or genre-specific dance clips, EvalPop consists of full-length choreography performed by professional choreographers across entire pop songs. To effectively evaluate choreography generated under such realistic settings, we design new metrics that reflect the structural and creative principles of professional choreography. Beyond commonly used naturalness metrics [33] such as FID, we propose three measures of choreographic quality based on choreographic primitives. Specifically, our evaluation consists of three metrics: (1) Phrase Diversity, which measures the diversity of each phrase; (2) Music–Choreography Feature Alignment, which evaluates the alignment between music and choreography; and (3) Text Stylization Score, which quantifies how well the generated choreography reflects the intent of the semantic text.

In summary, our contributions are as follows:

1. We propose a choreography generation framework grounded in choreographic theory, taking a step beyond improvisational dance by modeling phrase-level structure, music–choreography alignment, and inter-phrase plausibility.
2. We develop a semantic text stylization method that leverages LLMs to convert high-level semantic descriptions into controllable low-level choreography features, enabling intuitive and effective text stylization.
3. We propose evaluation metrics that measure the creative quality of choreography in terms of diversity, alignment, and semantic text stylization.
4. We construct a high-quality evaluation dataset of professional choreographers performing pop music, providing a realistic benchmark for evaluation.

2 Related Work

2.1 Dance Generation

Research in dance generation has evolved into the creation of music-conditioned movements. Early dance generation models were primarily based on sequence ar-

chitectures such as RNN [1,16,52] and LSTM [6,27], which took musical features as input and produced dance sequences accordingly. With the introduction of VAE, GAN, and Transformer architectures, dance generation research [5, 21, 29] has rapidly expanded. FACT [33] contributed by releasing the AIST++ dance dataset and introducing a full-attention-based cross-modal transformer, providing a sequence-to-sequence generation method. Cheng et al. [5] proposed a transformer-based network with dance word embeddings. MNET [22] introduced a Transformer-GAN hybrid that includes a mapping network for dance generation. EDGE [50] demonstrated high-quality choreography by leveraging Jukebox AI [7] within a diffusion-based framework. Beyond approaches focused on generating plausible dance movements, LODGE [31] extended diffusion models to long-sequence choreography generation by combining global and local diffusion processes, ensuring coherence across long music tracks. BeatIt [17] proposed a beat-aware mask dilation method to encourage choreography that closely follows beat information. ChoreoNet [53] represents dances as combinations of Choreographic Action Units collected from multiple dance sequences and temporally arranged according to musical beats. It employs a coarse-to-fine network to synthesize intermediate motions and ensure smooth transitions between units, enabling coherent choreography generation.

2.2 Text Guided Motion Stylization

Research on text-guided motion stylization [4, 9, 11, 35, 56] has advanced significantly in recent years. Human motion generation was enabled by datasets such as HumanML3D [10] and KIT-ML [46], which allowed precise mapping from textual descriptions to motions. Earlier works on motion stylization focused mainly on stylization based on categorical labels such as "happy" or "sad." MotionCLIP [49] advanced this area by creating a shared embedding space for text and motion, and CLIP-Actor [54] extended MotionCLIP to enable text-based stylization beyond simple categorical labels. With diffusion models, research further expanded to complex style transfer. For instance, SMooDi [57] incorporated both text information and motion sequences as input to enhance stylized motion quality.

3 Method

3.1 Overview

Figure 2 illustrates the overall framework of HiChor. Following Humphrey and Blom [2,41], choreography is organized into phrases, often corresponding to a 4-beat unit. Using an All-in-One Analyzer [25], we divide a music track M into N phrases of 4 beats each, denoted as $M = \{\mathbf{m}^i\}_{i=1}^{N}$ with phrase length L. HiChor then proceeds through four hierarchical stages:

Phrase-level Beat-aware Generation. For each phrase $\mathbf{m}^i$, a beat-aware choreography sequence $\mathbf{x}^i_{\text{beat}} \in \mathbb{R}^{72\times L}$, represented as a sequence of SMPL [37] pose parameters, is generated independently from a beat-aware generation model

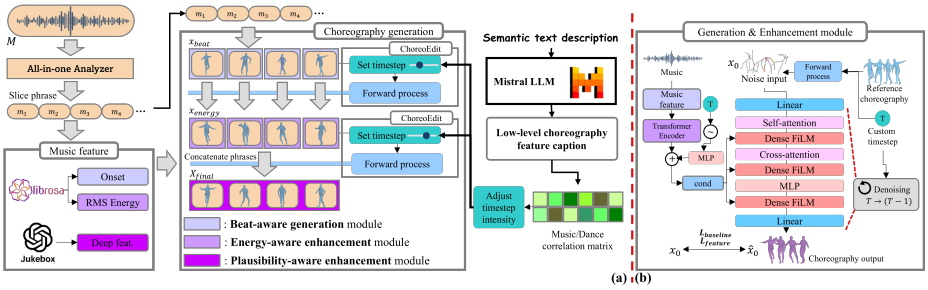

Fig. 2: Overview of the HiChor framework. (a) The input music M is segmented into phrase–level units, and each phrase conditions the beat–aware and energy–aware diffusion modules to generate phrase–level choreography. The generated phrases are then concatenated and refined by the plausibility–enhancement stage to resolve inter–phrase discontinuities. ChoreoEdit is used to enhance each stage and also enables text–based modulation for semantic stylization by adjusting diffusion timesteps. (b) Illustration of the generation and enhancement modules. When no reference choreography is provided, pure noise is used as the initial input.

$\mathcal{D}_{\text{beat}}$ using onset features $\mathbf{c}^i_{\text{beat}} \in \mathbb{R}^{1\times L}$ extracted from Librosa [39], ensuring phrase-independent representation and diversity.

Energy-aware Enhancement. The beat-aware choreography sequence $\mathbf{x}^i_{\text{beat}}$ is enhanced through an energy-aware enhancement model $\mathcal{D}_{\text{energy}}$, by incorporating RMS energy feature $\mathbf{c}^i_{\text{energy}} \in \mathbb{R}^{1\times L}$ extracted from Librosa, producing an energy-enhanced choreography sequence $\mathbf{x}^i_{\text{energy}}$.

Global Plausibility Enhancement. After concatenating the enhanced choreography sequences, we refine the full sequence to reduce inter-phrase discontinuities and improve naturalness using deep music features $\mathbf{c}_{\text{deep}} \in \mathbb{R}^{4800\times M}$ extracted using Jukebox AI [7], producing the final choreography $\mathbf{X}_{\text{final}} \in \mathbb{R}^{72\times M}$.

Choreography Stylization with Semantic Text. Given semantic text such as "Climax of world tour", we convert it into low-level choreography feature captions and modulate the strength of enhancements in each enhancement model to reflect the user's requirements.

Here, *enhancement* refers to generating a target choreography that injects a target condition while preserving the features of a reference choreography. HiChor achieves this by iteratively refining phrase-level choreography sequences using ChoreoEdit, a diffusion-based enhancement process motivated by SDEdit [40]. Further details are described in the following sections.

3.2 Hierarchical Choreography Generation

Phrase Slicing and Beat-aware Generation. For phrase $\mathbf{m}^i$, we generate an initial choreography $\mathbf{x}^i_{\text{beat}}$ conditioned on beat features $\mathbf{c}^i_{\text{beat}}$ using a beat-aware conditional diffusion model $\mathcal{D}_{\text{beat}}$ with Gaussian noise initialization $\mathbf{x}_{\text{init}} = \mathbf{z} \sim \mathcal{N}(0, \mathbf{I}) \in \mathbb{R}^{72\times L}$:

$$\mathbf{x}^i_{\text{beat}} = \mathcal{D}_{\text{beat}}(\mathbf{x}_{\text{init}};\, \mathbf{c}^i_{\text{beat}},\, t_{\max}), \tag{1}$$

where $t_{\max}$ indicates the number of backward steps in the diffusion model; we set $t_{\max} = 1000$ following common practice. Since the duration of a 4-beat phrase varies with tempo, $\mathcal{D}_{\text{beat}}$ samples a fixed 5-second window. The target 4-beat interval is then selected using a Transformer-style mask derived from beat segmentation, enabling variable-length phrase generation. This stage ensures that each phrase forms an independent unit of choreography, promoting choreographic diversity with beat alignment.

Energy-aware Enhancement. We then enhance the beat-aware choreography $\mathbf{x}^i_{\text{beat}}$ with energy information using the ChoreoEdit pipeline. We take $\mathbf{x}^i_{\text{beat}}$ as the reference choreography and $\mathbf{c}^i_{\text{energy}}$ as the target condition. We first inject noise up to a custom timestep t_{energy} and denoise the noised choreography $\tilde{\mathbf{x}}^i_{\text{beat}}$ under $\mathbf{c}^i_{\text{energy}}$:

$$\begin{aligned} \tilde{\mathbf{x}}^i_{\text{beat}} &= \text{ChoreoEdit}(\mathbf{x}^i_{\text{beat}};\, t_{\text{energy}}), \\ \mathbf{x}^i_{\text{energy}} &= \mathcal{D}_{\text{energy}}(\tilde{\mathbf{x}}^i_{\text{beat}};\, \mathbf{c}^i_{\text{energy}},\, t_{\text{energy}}), \end{aligned} \tag{2}$$

where t_{energy} is a controllable timestep determining how much beat-aligned content is preserved versus how much energy-aligned content is added. This yields the energy-enhanced choreography $\mathbf{x}^i_{\text{energy}}$ for each phrase.

Global Plausibility Enhancement. We concatenate the outputs $\{\mathbf{x}^i_{\text{energy}}\}_{i=1}^N$ in sequential order to obtain the full choreography:

$$\mathbf{X}_{\text{concat}} = \text{Concat}(\{\mathbf{x}^i_{\text{energy}}\}_{i=1}^N) \in \mathbb{R}^{72\times M}. \tag{3}$$

However, simply concatenating each choreography sequence introduces discontinuities at phrase boundaries, and beat/energy alone do not capture high-level stylistic details. To resolve this, we perform a final enhancement conditioned on deep musical features $\mathbf{c}_{\text{deep}}$, which embed high-level musical context:

$$\begin{aligned} \tilde{\mathbf{X}}_{\text{concat}} &= \text{ChoreoEdit}(\mathbf{X}_{\text{concat}};\, t_{\text{deep}}), \\ \mathbf{X}_{\text{final}} &= \mathcal{D}_{\text{plausible}}(\tilde{\mathbf{X}}_{\text{concat}};\, \mathbf{c}_{\text{deep}},\, t_{\text{deep}}). \end{aligned} \tag{4}$$

Here $\mathbf{c}_{\text{deep}} \in \mathbb{R}^{4800\times M}$ and t_{deep} governs how much high-level content is added: smaller t_{deep} preserves more of the strict beat-energy structure, while larger t_{deep} emphasizes high-level stylistic information over structure. This final stage reduces inter-phrase discontinuity and enriches global detail.

3.3 ChoreoEdit Pipeline

Inspired by SDEdit [40], we propose the ChoreoEdit pipeline, which fuses a reference choreography with a target condition in diffusion models [14, 15] without any additional training. We extend Proposition 1 of SDEdit [40] to the choreography domain under standard Lipschitz and Gaussian assumptions. Figure 3(a) illustrates the concept of the ChoreoEdit pipeline, where the reference choreography is noised up to a custom timestep t so that only its features remain; then, at the same t, the target condition is injected and the model denoises for t steps to produce an output that fuses the reference features with the target

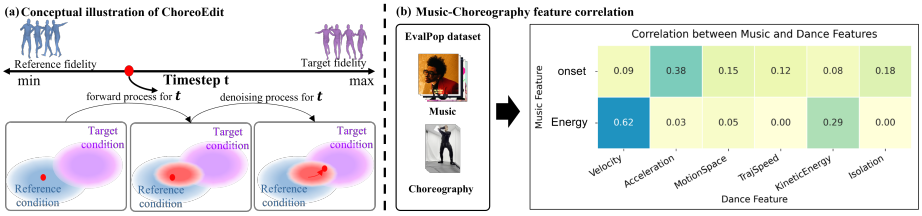

Fig. 3: Feature correlation and modulation. (a) The reference choreography is first noised up to a custom diffusion timestep t and then denoised for the same t steps under a target condition, enabling controlled choreography enhancement. (b) Correlation between music and choreography features, where beat exhibits correlation with acceleration, while energy correlates primarily with velocity and kinetic energy.

condition. Adjusting the diffusion timestep t implements a trade-off between reference choreography and target condition. The ChoreoEdit pipeline proceeds in two steps:

Step 1) Forward noising from 0 to t. Given a reference choreography $\mathbf{x}_{\text{ref}}$, we apply the forward process up to timestep t to inject a controlled amount of noise:

$$\mathbf{x}_{\text{ref}}(t) = \alpha(t)\,\mathbf{x}_{\text{ref}} + \sigma(t)\,\mathbf{z}, \qquad \mathbf{z} \sim \mathcal{N}(0, \mathbf{I}). \tag{5}$$

We summarize this as $\text{ChoreoEdit}(\mathbf{x}; t)$:

$$\tilde{\mathbf{x}}_{\text{ref}} = \mathbf{x}_{\text{ref}}(t) = \text{ChoreoEdit}(\mathbf{x}_{\text{ref}}; t). \tag{6}$$

Here, $\tilde{\mathbf{x}}_{\text{ref}}$ is the noisy sample at timestep t, and $\alpha(t)$, $\sigma(t)$ are the signal and noise scales. This converts the reference into a noisy sample that preserves its features.

Step 2) Conditional denoising from t to 0. We then denoise the noised reference $\tilde{\mathbf{x}}_{\text{ref}}$ under a target condition $\mathbf{c}_{\text{tar}}$, using a learned score function s_θ:

$$\mathbf{x}_{\text{tar}} = \mathcal{D}(\tilde{\mathbf{x}}_{\text{ref}};\, \mathbf{c}_{\text{tar}},\, s_\theta,\, t)\,, \tag{7}$$

where $\mathcal{D}$ denotes the conditional diffusion sampling process, with classifier-free guidance [15]. The resulting $\mathbf{x}_{\text{tar}}$ fuses the features of $\tilde{\mathbf{x}}_{\text{ref}}$ and the condition $\mathbf{c}_{\text{tar}}$.

3.4 Feature-aware Diffusion Model

Our diffusion models $\mathcal{D}_{\text{beat}}$, $\mathcal{D}_{\text{energy}}$, and $\mathcal{D}_{\text{plausible}}$ are transformer-based diffusion models [45, 50] for choreography generation and enhancement. Figure 2(b) shows the details of each model, which takes musical features as conditionals and uses classifier-free guidance. Additionally, we train feature-aware losses for $\mathcal{D}_{\text{beat}}$ and $\mathcal{D}_{\text{energy}}$ so that the generated choreography follows musically meaningful feature relationships.

Motivation from Music–Choreography Correlation. To design losses and connect modalities, we first analyze correlations between musical and choreography features using EvalPop, which consists of 10 K-pop/Pop songs with

choreography performed by professional choreographers. Music features consist of the beat B, represented as a raised-cosine activation around detected onsets, and the RMS energy E computed with Librosa's frame-wise RMS, stacked as $\mathbb{R}^{F\times L}$ with $F = 2$. Choreography features are extracted by converting SMPL [37] pose parameters to 3D joint coordinates via forward kinematics. Following Laban Movement Analysis [26] and prioritizing temporally expressive characteristic measures, we consider six features as candidates: acceleration A, velocity V, root trajectory speed, kinetic energy K, motion space (3D spatial volume), and isolation (salient distal acceleration). Each is computed per frame, forming a tensor in $\mathbb{R}^{F\times L}$ with $F = 6$.

Let the set of music and choreography feature sequences be represented as $\mathcal{M} = \{M^i\}_{i=1}^T$ and $\mathcal{C} = \{C^i\}_{i=1}^T$. We then compute phrase-wise Pearson correlations between music and choreography feature sequences:

$$\mathrm{Corr_P}(\mathcal{M}, \mathcal{C}) = \frac{\sum_{i=1}^{T}(M^i - \bar{M})(C^i - \bar{C})}{\sqrt{\sum_{i=1}^{T}(M^i - \bar{M})^2}\,\sqrt{\sum_{i=1}^{T}(C^i - \bar{C})^2}} \tag{8}$$

where M^i and C^i are music and choreography feature values of the i-th frame in a phrase, $\bar{M}$ and $\bar{C}$ are their means, and T is the phrase length. We then softmax-normalize correlations across the six choreography features to obtain relative association strengths:

$$p_{n,m} = \frac{\exp\big(\tau \cdot \mathrm{Corr_P}(\mathcal{M}_n, \mathcal{C}_m)\big)}{\sum_{l=1}^{6}\exp\big(\tau \cdot \mathrm{Corr_P}(\mathcal{M}_n, \mathcal{C}_l)\big)}, \quad n \in \{1, 2\},\ m \in \{1, \ldots, 6\}. \tag{9}$$

where τ is the softmax scaling strength. As shown in Figure 3(b), *beat correlates most strongly with acceleration*, and *RMS energy correlates strongly with velocity and kinetic energy*, supporting the intuition that beats trigger impulsive actions while energy modulates speed and effort. We use these associations both as evidence and as guidance for loss design and text stylization.

Feature-aware Loss Function. Based on the observation above, we enforce generated choreography to follow these correlations via a feature-aware loss function $\mathcal{L}_{\text{feature}}$, which minimizes negative Pearson correlations between specific choreography features and the corresponding music features. For the correlation between the beat and acceleration, we design the following loss function:

$$\mathcal{L}_{\text{beat}} \;=\; -\,\mathrm{Corr_P}\big(\mathcal{B}, \mathcal{A}\big). \tag{10}$$

Here, the beat $\mathcal{B}$ uses a 7-frame raised-cosine window centered at each beat frame, which encourages increased acceleration on nearby beats. This loss term enforces a strong alignment between beats and acceleration, which initially results in an exaggerated choreography. However, through the subsequent enhancement stages, only the intended beat–acceleration relation is preserved, allowing the choreography to become natural. For correlations between RMS energy and velocity-related choreography features, we design the following loss functions:

$$\begin{gathered}\mathcal{L}_{\text{energy-KE}} = -\mathrm{Corr_P}(\mathcal{E}, \mathcal{K}), \quad \mathcal{L}_{\text{energy-vel}} = -\mathrm{Corr_P}(\mathcal{E}, \mathcal{V}), \\ \mathcal{L}_{\text{energy}} = \alpha\,\mathcal{L}_{\text{energy-KE}} + \beta\,\mathcal{L}_{\text{energy-vel}}.\end{gathered} \tag{11}$$

Here, $\mathcal{E}, \mathcal{K}, \mathcal{V}$ denote the RMS energy, kinetic energy, and velocity sequences; each normalized to $[0, 1]$ after FK-based conversion to the position domain. This encourages kinetic energy and velocity to follow their correlation with the energy. Then, we define $\mathcal{L}_{\text{feature}}$ as follows to be selectively used in beat- and energy-aware diffusion models:

$$\mathcal{L}_{\text{feature}} = \begin{cases} \mathcal{L}_{\text{beat}}, & \text{(beat-aware)}, \\ \mathcal{L}_{\text{energy}}, & \text{(energy-aware)}. \end{cases} \tag{12}$$

In addition, following EDGE [50], we use a baseline loss $\mathcal{L}_{\text{baseline}}$ that includes reconstruction and physical plausibility terms commonly adopted in prior work:

$$\begin{aligned} &\mathcal{L}_{\text{recon}} = \mathbb{E}_{x_0,\epsilon,t}[\|\epsilon - \epsilon_\theta(x_t, t, c)\|_2^2], \qquad \mathcal{L}_{\text{vel}} = \frac{1}{N-1}\sum_{i=1}^{N-1} \|\Delta x^i - \Delta \hat{x}^i\|_2^2, \\ &\mathcal{L}_{\text{joint}} = \frac{1}{N}\sum_{i=1}^{N} \|\text{FK}(x^i) - \text{FK}(\hat{x}^i)\|_2^2, \quad \mathcal{L}_{\text{contact}} = \frac{1}{N-1}\sum_{i=1}^{N-1} \|\Delta \text{FK}(\hat{x}^i) \odot b^i\|_2^2. \end{aligned} \tag{13}$$

Here, $\mathbf{x}^i$ and $\hat{\mathbf{x}}^i$ are ground-truth and generated pose parameters at frame i, N is the number of frames per training clip, $\text{FK}(\cdot)$ maps poses to global 3D joint positions, $\mathbf{b}^i \in \{0, 1\}^J$ is a contact mask with ones on the feet joints, and $\odot$ is element-wise product. The total loss for each model is computed as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{baseline}} + \lambda_{\text{feature}} \mathcal{L}_{\text{feature}}. \tag{14}$$

3.5 Choreography Stylization with Semantic Text

For semantic text stylization, HiChor consists of two steps: semantic text decomposition and timestep modulation. First, semantic text decomposition translates high-level semantic text into structured low-level choreography feature targets. Then, based on these feature targets, timestep modulation controls the diffusion timesteps in the ChoreoEdit pipeline. By modulating the timesteps, HiChor controls how strongly the desired features are reflected in the stylized choreography.

Semantic Text Decomposition. Given the user text and music analysis, we prompt Mistral-7B [20] to output structured directives: (i) which phrases to modify and (ii) the desired intensity of each choreography feature as `High` / `Medium` / `Low` (Table 1). The underlying motivation of discrete feature labeling is provided in Figure 4(a): stylized choreographies performed by professional dancers exhibit consistent *directional* changes in feature trends compared to their original versions, suggesting that high-level textual descriptions can be expressed as variations of low-level features and provide sufficient cues to guide text-driven stylization.

Timestep Modulation. Figure 4(b) illustrates that the diffusion timestep controls a complementary trade-off: smaller timesteps preserve beat-sensitive structure, whereas larger timesteps emphasize energy-related characteristics such as

Table 1: Semantic-to-feature decomposition via an LLM. A high-level text prompt ("Climax of world tour") is mapped to controllable low-level features (e.g., *Acceleration: High*) and target phrases.

Role	Content
System	"Extract human choreography features (acceleration, velocity, etc.) from the user's request. Specify the intensity as [High], [Medium], or [Low]. Output the target phrases and a list of features. For example: Phrases: [1–4], Acceleration: [High], Velocity: [Medium]."
Assistant	(Music analysis JSON from All-in-One analyzer is provided here.)
User	"At the first highlight part, I want a choreography like climax of world tour."
Output	**{Phrases: [8–11], Acceleration: [High], Velocity: [High], KineticEnergy: [High]** ... }

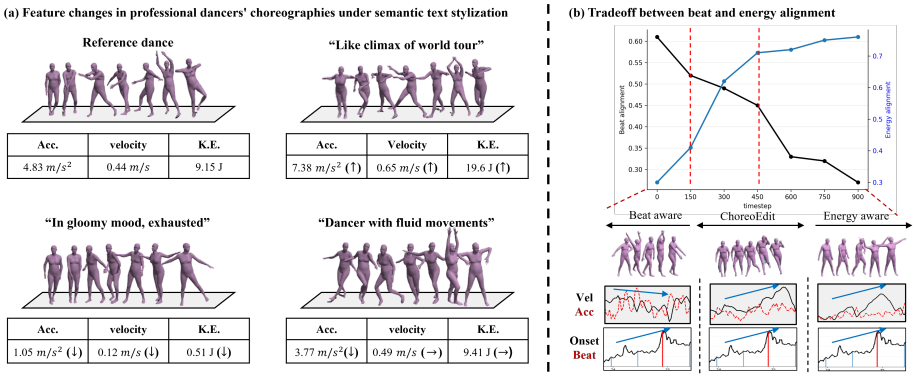

Fig. 4: Feature differences induced by text stylization. (a) The dancer performs choreography according to the given text guidance, and the resulting feature changes exhibit the directional trends. (b) Effect of the diffusion timestep on beat–energy alignment, illustrating how different timestep values modulate the trade-off between beat-aware and energy-aware choreography.

velocity and kinetic energy. Based on this observation, semantic text is first decomposed by an LLM into choreography feature targets. The LLM outputs six choreography features with discrete intensity levels $\{\text{Low}, \text{Medium}, \text{High}\}$, which we encode as numerical weights $\mathbf{s} \in \{0, 0.5, 1\}^6$. To map these choreography features into the music conditions used by our diffusion models ($\mathcal{D}_{\text{beat}}$ and $\mathcal{D}_{\text{energy}}$), we use a pre-defined music–choreography correlation map $\mathbf{C} \in \mathbb{R}^{2\times 6}$ (Figure 3(b)), where C_{mj} denotes the correlation strength between music feature $m \in \{\text{beat}, \text{energy}\}$ and choreography feature j. We compute the contribution of each music feature as:

$$p_m = \sum_{j=1}^{6} C_{mj} \cdot s_j, \qquad m \in \{\text{beat}, \text{energy}\}. \tag{15}$$

We then modulate the timesteps based on these contributions:

$$t_{\text{energy}} = t_{\min} + (t_{\max} - t_{\min}) \cdot (p_{\text{energy}} - p_{\text{beat}}),$$
$$t_{\text{deep}} = t_{\min} + (t_{\max} - t_{\min}) \cdot (1 - p_{\text{beat}} - p_{\text{energy}}). \tag{16}$$

We set $t_{\min} = 100$ and $t_{\max} = 500$, and clip both $(1-p_{\text{beat}}-p_{\text{energy}})$ and $(p_{\text{energy}}-p_{\text{beat}})$ to $[0, 1]$ to avoid negative values. Intuitively, emphasizing beat-related targets increases p_{beat}, yielding smaller timesteps that preserve beat-driven structure. Conversely, emphasizing energy-related targets increases p_{energy}, resulting in larger timesteps and stronger energy enhancement. When both cues are weak, t_{deep} becomes larger, encouraging global plausibility smoothing.

3.6 Choreography Evaluation Dataset

Existing datasets such as AIST++ [33], FineDance [32], and PopDG [38] are mainly limited to genre-specific settings or short dance sequences. AIST++ and FineDance provide short dance sequences paired with music from a specific genre, and although PopDG is based on Pop songs, it is also limited to short highlight-oriented motion clips of approximately 15–30 seconds. Consequently, these datasets are not well suited for evaluating full-length choreography generation for in-the-wild music. To address this limitation, we construct EvalPop, a custom test dataset collected in a professional dance capture studio. EvalPop contains full-length performances of 10 K-pop/Pop songs, recorded by four professional dancers using an OptiTrack [42] motion capture system [18, 23, 24, 44]. Although the number of songs is relatively small, each performance exceeds 200 seconds, resulting in 7,332 seconds of motion capture data in total. This enables long-horizon evaluation of full-length choreography generation, including phrase-level continuity, music–choreography alignment, and overall motion plausibility.

Figure 5 illustrates the data acquisition process. During capture, each dancer performed the full choreography while wearing an OptiTrack suit, and the resulting motions were exported in BVH format. Since our generation and evaluation pipeline represents human motion using SMPL parameters, we convert the captured BVH motions into the SMPL representation through a BVH-to-SMPL fitting process. The converted SMPL sequences are then used as the reference motions for evaluating generated full-length choreography.

3.7 Implementation Details

We use AIST++ [33] for training and reserve EvalPop only for evaluation. AIST++ contains 1,408 dance sequences captured at 30 FPS from more than 10 professional dancers. We follow the official AIST++ split, without using a separate validation split. Our framework employs three diffusion models in total: the beat-aware model $\mathcal{D}_{\text{beat}}$, the energy-aware model $\mathcal{D}_{\text{energy}}$, and the plausibility-aware model $\mathcal{D}_{\text{plausible}}$. For reproducibility, we use the publicly available pre-trained EDGE [50] model as $\mathcal{D}_{\text{plausible}}$ without further fine-tuning, and apply it in a frozen evaluation mode to improve cross-phrase continuity and full-length motion plausibility. For training, we use motion sequences of 5 seconds, corresponding to 150 frames at 30 FPS, as training samples. The beat-aware and

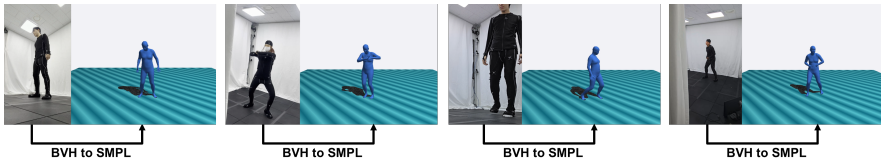

Fig. 5: Dataset acquisition pipeline of EvalPop. Four professional dancers perform full-length K-pop/Pop choreographies in an OptiTrack motion capture studio.

energy-aware diffusion models, $\mathcal{D}_{\text{beat}}$ and $\mathcal{D}_{\text{energy}}$, are trained on NVIDIA A6000 GPUs for 24 hours with a batch size of 64. For timestep modulation, we empirically set $t_{\text{energy}} = 300$ and $t_{\text{deep}} = 150$, which provide a stable balance between feature-level stylization and preservation of the overall choreographic structure. To provide a clearer view of the computational cost, we further measure the inference complexity, latency, and GPU memory usage of each component. The beat-aware and energy-aware modules together have 47M parameters, require 3.2 seconds per phrase on average, and consume 3,413MB of GPU memory. Given an average of 110 phrases per song, this results in approximately 6 minutes of inference time for a full song. The plausibility-aware module has 49M parameters, requires 4.3 seconds per phrase on average, and consumes 19,411MB of GPU memory, resulting in approximately 8 minutes of inference time per full song.

4 Experiments

4.1 Choreography Metrics

Phrase Diversity. Higher phrase-level diversity indicates that the generated choreography avoids repetition and more closely reflects how real choreography is structured. For comparison, we adopt the FACT [33] metrics, which define kinetic and geometric feature space distances as Div_k and Div_g. We adapt these metrics for phrase-level evaluation by computing L2 distances between phrases within the same song. We denote the resulting scores as Phrase-Div_k and Phrase-Div_g and combine them into a single metric, Phrase Diversity.

Feature Alignment. Following Burger et al. [3], we compute Pearson correlations between music features (beat onset and RMS energy) and choreography features extracted from motion to measure music–choreography alignment. Guided by our correlation analysis (Figure 3(b)), we evaluate beat–acceleration and energy–velocity alignments, as these feature pairs exhibit the strongest correlations. Both phrase-level and global-level energy–velocity alignment are measured to assess whether the generated choreography aligns with the overall musical structure.

Text Stylization. We evaluate whether semantic text stylization induces the intended feature changes in generated choreography. For the same choreography, we provided five semantic text prompts and asked professional choreographers to produce corresponding stylized versions. We extracted choreography

features from these human-stylized choreographies and treated the resulting feature changes as ground-truth stylization targets. We then applied text stylization methods and measured whether the generated feature changes followed the same directions as the professional references. The final stylization score ranges from 0 to 1, where a score of 1 indicates that all feature changes are consistent with the ground-truth directions.

FID. To assess the naturalness of generated motion, we computed Fréchet Inception Distance [13] on both geometric and kinematic dance features, as commonly used in prior work. For each generated choreography and each ground-truth choreography in the EvalPop dataset, we extract kinematic features $z_k \in \mathbb{R}^{72}$ and geometric features $z_g \in \mathbb{R}^{33}$. The FID is then computed separately in the two feature spaces.

4.2 Quantitative Results

For evaluation, we compare state-of-the-art models [8, 22, 31, 33, 34, 47, 48, 50, 57], our model, and ground-truth choreography from EvalPop. Using the metrics defined in Section 4.1, we evaluate whether HiChor outperforms other methods in choreography generation, as shown in Table 2. First, **Phrase Diversity** demonstrates that HiChor effectively generates phrase-independent representations, achieving diversity levels higher than other generative models. Second, **Feature Alignment** shows that HiChor captures both beat–acceleration and energy–velocity relationships effectively, achieving balanced alignment patterns that are closer to those of real choreography. The phrase-level and global-level alignment demonstrates that HiChor captures both local phrase dynamics and the overall intensity flow of the music. Third, **Text Stylization** score confirms that semantic text prompts translate into feature-level changes as intended, outperforming recent text-guided motion stylization approaches. Finally, despite introducing additional structure-aware mechanisms, HiChor maintains FID scores comparable to those of state-of-the-art models, indicating that improved choreographic structure does not compromise the naturalness of the generated motion. Together, these results show that HiChor successfully adheres to choreographic principles while preserving the motion naturalness comparable to state-of-the-art methods.

4.3 Qualitative Results

Figure 6 presents a qualitative comparison between existing methods and HiChor. In the visualization, red and blue vertical lines indicate phrase and beat timings, while black curves represent RMS energy (music) and mean joint velocity (choreography). EDGE exhibits moderate beat alignment but tends to repeat similar movement patterns across phrases, indicating limited phrase-level diversity and weaker responsiveness to changes in musical energy. In contrast, HiChor generates more phrase-independent choreography, as distinct movement patterns emerge across different phrases rather than repeating a uniform motion style throughout the sequence. At the same time, HiChor shows stronger beat

Table 2: Comparison with state of the art. Evaluation on both dance quality and choreography structure quality on the EvalPop dataset.

Method	Dance Quality		Choreography Structure Quality				
	$FID_k \downarrow$	$FID_g \downarrow$	Phrase Div.↑	Beat align.↑	Energy align.(Phrase)↑	Energy align.(Global)↑	Text stylization↑
FACT [33]	121.59	80.21	5.10	0.205	0.372	0.301	–
Bailando [47]	75.27	22.75	5.59	0.226	0.458	0.420	–
MNET [22]	115.62	66.52	4.70	0.231	0.455	0.377	–
EDGE [50]	96.70	**18.08**	5.36	0.261	0.541	0.512	–
LODGE [31]	65.12	18.22	4.62	0.287	0.528	0.495	–
SoulDance [34]	58.11	22.68	5.91	0.265	0.551	0.529	–
MEGADance [48]	60.26	19.07	6.29	0.277	0.603	0.581	–
TM2D [8]	76.79	21.59	5.51	0.211	0.450	0.421	0.60
SmooDi [57]	–	–	–	–	–	–	0.55
GroundTruth	–	–	8.63	0.542	0.692	0.711	–
Ours	**45.32**	18.20	**6.44**	**0.366**	**0.642**	**0.558**	**0.85**

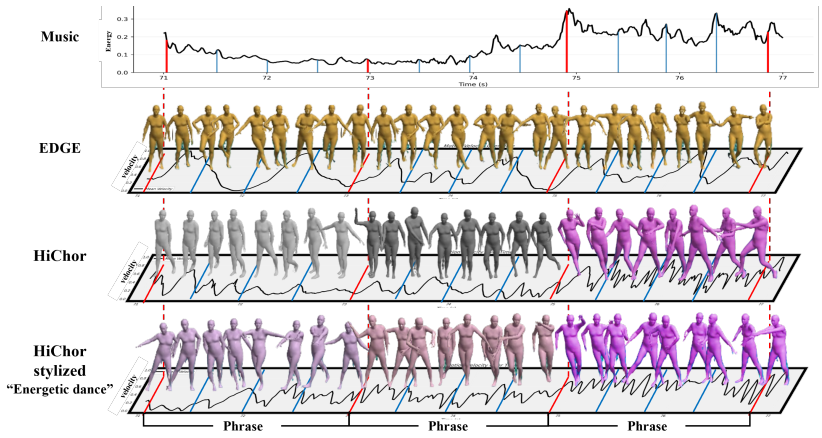

Fig. 6: Qualitative Results. Comparison of EDGE, HiChor, and text-stylized HiChor results generated from the same music. HiChor produces phrase-independent choreography with clearer beat and energy alignment, while text stylization further modulates intensity.

alignment while adapting motion intensity according to RMS energy. With semantic stylization "Energetic dance," HiChor further increases velocity compared to the original choreography, reflecting the intended energetic emphasis.

4.4 Ablation Study

We conducted an ablation study on the three hierarchical components of HiChor. By removing each module, we evaluated changes in beat alignment, energy alignment, and plausibility. As summarized in Table 3, removing each component introduces trade-offs across individual metrics, but overall performance consistently degrades compared to the full model. Figure 7 further illustrates that absence of the final plausibility stage causes visible discontinuities at phrase boundaries, resulting in fragmented motion transitions across consecutive phrases. In contrast, the plausibility stage smooths these transitions and restores a more

Table 3: Ablation on HiChor.

Ablation	Alignment		Quality
	Beat↑	Energy↑	FID_k ↓
w/o $\mathcal{D}_{\text{beat}}$	0.261	0.541	96.70
w/o $\mathcal{D}_{\text{energy}}$	0.722	0.490	175.63
w/o $\mathcal{D}_{\text{plausible}}$	0.448	0.710	152.37
Full (Ours)	**0.366**	**0.642**	**45.32**

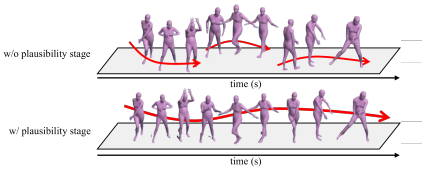

Fig. 7: Discontinuity Visualization. Without the plausibility stage, discontinuities are observed between phrases.

continuous trajectory over the full sequence, highlighting its importance in producing seamless full-length choreography.

5 Conclusion and Limitations

In this work, we presented HiChor, a hierarchical choreography generation framework grounded in structured choreographic primitives and designed to follow how human choreographers construct movement. Our evaluation shows that HiChor produces choreography closer to professionally created work than existing baselines. However, the primary contribution of this framework lies in moving beyond improvisational dance generation toward structure-aware choreography guided by choreographic principles, rather than replicating the highly polished performances observed in real idol dances. As a result, HiChor does not yet achieve the full performance quality of professional choreography. In future work, we aim to further improve choreography quality by scaling up training data through additional dance capture sessions using the same studio setup as EvalPop. By expanding EvalPop to a large-scale training dataset, we expect to better capture real-world choreographic patterns and advance the model toward more professional-level choreography generation.

Acknowledgements

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2024 (RS-2024-00398413, Contribution Rate: 50%) and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2025-02216328, Contribution Rate: 50%).

References

1. Alemi, O., Françoise, J., Pasquier, P.: Groovenet: Real-time music-driven dance movement generation using artificial neural networks. networks **8**(17), 26 (2017)
2. Blom, L.A., Chaplin, L.T.: The intimate act of choreography. University of Pittsburgh Pre (1982)

3. Burger, B., Thompson, M.R., Luck, G., Saarikallio, S., Toiviainen, P.: Influences of rhythm-and timbre-related musical features on characteristics of music-induced movement. Frontiers in psychology **4**, 183 (2013)
4. Cha, J., Kim, J., Yoon, J.S., Baek, S.: Text2hoi: Text-guided 3d motion generation for hand-object interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1577–1585 (2024)
5. Cheng, Y., Jiang, Y., Wang, Y.: Music-stylized hierarchical dance synthesis with user control. Virtual Reality & Intelligent Hardware **6**(5), 339–357 (2024)
6. Crnkovic-Friis, L., Crnkovic-Friis, L.: Generative choreography using deep learning. arXiv preprint arXiv:1605.06921 (2016)
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)
8. 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)
9. Guo, C., Mu, Y., Zuo, X., Dai, P., Yan, Y., Lu, J., Cheng, L.: Generative human motion stylization in latent space. arXiv preprint arXiv:2401.13505 (2024)
10. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5152–5161 (2022)
11. Guo, Z., Lee, Y.Y., Liu, J., Ben-Shabat, Y., Zordan, V., Kapadia, M.: Stylemotif: Multi-modal motion stylization using style-content cross fusion. arXiv preprint arXiv:2503.21775 (2025)
12. Gupta, P., Fotso-Puepi, J.A., Li, Z., Mehta, J., Bera, A.: Mdd: A dataset for text-and-music conditioned duet dance generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13932–13941 (2025)
13. 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. Advances in neural information processing systems **30** (2017)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
15. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
16. Huang, R., Hu, H., Wu, W., Sawada, K., Zhang, M., Jiang, D.: Dance revolution: Long-term dance generation with music via curriculum learning. In: International conference on learning representations (2020)
17. 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)
18. Huh, J., Park, Y., Kim, S., Kim, J., Cheng, W.H., Lee, S.: Biomas: A wear-free biomechanical motion assessor for physical self-training. In: 2025 IEEE International Conference on Consumer Electronics (ICCE). pp. 1–4. IEEE (2025)
19. Humphrey, D.: The art of making dances (1959)
20. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b (2023), https://arxiv.org/abs/2310.06825
21. Kim, J., Kwon, B., Kim, J., Lee, S.: Mnet++: Music-driven pluralistic dancing toward multiple dance genre synthesis. IEEE transactions on pattern analysis and machine intelligence **45**(12), 15036–15050 (2023)

22. Kim, J., Oh, H., Kim, S., Tong, H., Lee, S.: A brand new dance partner: Music-conditioned pluralistic dancing controlled by multiple dance genres. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3490–3500 (2022)
23. Kim, J., Huh, J., Park, Y., Kim, S., Choi, J., Lee, S.: Permission to dance: An end-to-end dance enhancement system from dance capture to analysis. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 13531–13533 (2025)
24. Kim, S., Huh, J., Park, Y., Kim, J., Lee, S.: Dancemimic: Awaken your dancing instinct through a real-time dance imitation capture system. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 11267–11269 (2024)
25. Kim, T., Nam, J.: All-in-one metrical and functional structure analysis with neighborhood attentions on demixed audio (2023), https://arxiv.org/abs/2307.16425
26. Laban, R.v., Ullmann, L.: The language of movement: A guidebook to choreutics. (No Title) (1966)
27. Lee, I., Kim, D., Lee, S.: 3-d human behavior understanding using generalized ts-lstm networks. IEEE Transactions on Multimedia **23**, 415–428 (2020)
28. Lerdahl, F., Jackendoff, R.S.: A Generative Theory of Tonal Music, reissue, with a new preface. MIT press (1996)
29. Li, B., Zhao, Y., Zhelun, S., Sheng, L.: Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 1272–1279 (2022)
30. Li, R., Zhang, H., Zhang, Y., Zhang, Y., Zhang, Y., Guo, J., Zhang, Y., Li, X., Liu, Y.: Lodge++: High-quality and long dance generation with robust choreography patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
31. 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. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1524–1534 (2024)
32. Li, R., Zhao, J., Zhang, Y., Su, M., Ren, Z., Zhang, H., Tang, Y., Li, X.: Finedance: A fine-grained choreography dataset for 3d full body dance generation (2023), https://arxiv.org/abs/2212.03741
33. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13401–13412 (2021)
34. Li, X., Li, R., Fang, S., Xie, S., Guo, X., Zhou, J., Peng, J., Wang, Z.: Music-aligned holistic 3d dance generation via hierarchical motion modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14420–14430 (2025)
35. Li, Z., Cheng, K., Ghosh, A., Bhattacharya, U., Gui, L., Bera, A.: Simmotionedit: Text-based human motion editing with motion similarity prediction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27827–27837 (2025)
36. Liu, X., Dong, X., Kanojia, D., Wang, W., Feng, Z.: Gcdance: Genre-controlled 3d full body dance generation driven by music. arXiv preprint arXiv:2502.18309 (2025)
37. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 851–866 (2023)

38. 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)
39. McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E., Nieto, O.: librosa: Audio and music signal analysis in python. SciPy **2015**, 18–24 (2015)
40. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
41. Minton, S.C.: Choreography: a basic approach using improvisation. Human Kinetics (2017)
42. NaturalPoint, Inc.: Optitrack: Precision motion capture systems. `https://optitrack.com/` (2025), accessed: 2025-10-04
43. Oxford University Press: Choreography, n., sense 2. Oxford English Dictionary, Oxford University Press (September 2025). `https://doi.org/10.1093/OED/1203098853`, retrieved October 4, 2025, from urlhttps://doi.org/10.1093/OED/1203098853
44. Park, Y., Yoon, H., Huh, J., Kim, J., Choi, J., Lee, S.: Sdas: Semantic data acquisition system for minimizing redundancy and maximizing diversity. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 29679–29681 (2025)
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)
46. Plappert, M., Mandery, C., Asfour, T.: The kit motion-language dataset. Big data **4**(4), 236–252 (2016)
47. 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)
48. Tang, X., Peng, Z., Hu, Y., He, J., Liu, H., et al.: Megadance: Mixture-of-experts architecture for genre-aware 3d dance generation. Advances in Neural Information Processing Systems **38**, 10947–10969 (2026)
49. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. In: European Conference on Computer Vision. pp. 358–374. Springer (2022)
50. 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)
51. Wang, Q., Yang, X., Dong, Y., Govindaraj, N.R., Slabaugh, G., Yuan, S.: Dancechat: Large language model-guided music-to-dance generation. arXiv preprint arXiv:2506.10574 (2025)
52. Yalta, N., Watanabe, S., Nakadai, K., Ogata, T.: Weakly-supervised deep recurrent neural networks for basic dance step generation. In: 2019 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2019)
53. Ye, Z., Wu, H., Jia, J., Bu, Y., Chen, W., Meng, F., Wang, Y.: Choreonet: Towards music to dance synthesis with choreographic action unit. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 744–752 (2020)
54. Youwang, K., Ji-Yeon, K., Oh, T.H.: Clip-actor: Text-driven recommendation and stylization for animating human meshes. In: European Conference on Computer Vision. pp. 173–191. Springer (2022)
55. Zhang, C., Tang, Y., Zhang, N., Lin, R.S., Han, M., Xiao, J., Wang, S.: Bidirectional autoregessive diffusion model for dance generation. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 687–696 (2024)
56. Zhang, J., Chen, X., Yu, G., Tu, Z.: Generative motion stylization of cross-structure characters within canonical motion space. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 7018–7026 (2024)
57. Zhong, L., Xie, Y., Jampani, V., Sun, D., Jiang, H.: Smoodi: Stylized motion diffusion model. In: European Conference on Computer Vision. pp. 405–421. Springer (2024)