DanceEditor: Towards Iterative Editable Music-driven Dance Generation with Open-Vocabulary Descriptions

Hengyuan Zhang[1,*], Zhe Li[1,*], Xingqun Qi[2,✉], Mengze Li[2], Muyi Sun[3], Siye Wang[3], Man Zhang[3,4], Sirui Han[2,✉]
[1] Peking University, [2] The Hong Kong University of Science and Technology
[3] Beijing University of Posts and Telecommunications, [4] Qinghai Institute of Technology

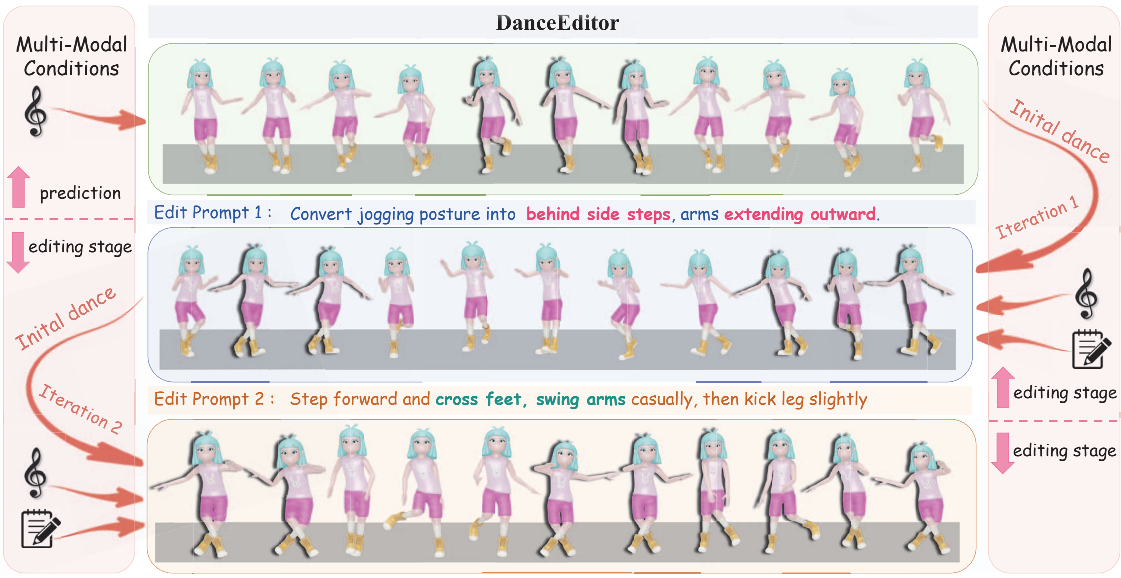


Figure 1. Our **DanceEditor** framework, pre-trained on a large-scale dataset, enables iterative and editable dance generation that is coherently aligned with the provided music signals. The highlighted texts and avatar shadow effects here specifically indicate edits related to body movements.

Abstract

Generating coherent and diverse human dances from music signals has gained tremendous progress in animating virtual avatars. While existing methods support direct dance synthesis, they fail to recognize that enabling users to edit dance movements is far more practical in real-world choreography scenarios. Moreover, the lack of high-quality dance datasets incorporating iterative editing also limits addressing this challenge. To achieve this goal, we first construct ***DanceRemix****, a large-scale multi-turn editable dance dataset comprising the prompt featuring over 25.3M dance frames and 84.5K pairs. In addition, we propose a novel framework for iterative and editable dance generation coherently aligned with given music signals, namely* ***DanceEditor****. Considering the dance motion should be both musical rhythmic and enable iterative editing by user descriptions, our framework is built upon a prediction-then-editing paradigm unifying multi-modal conditions. At the initial prediction stage, our framework improves the authority of generated results by directly modeling dance movements from tailored, aligned music. Moreover, at the subsequent iterative editing stages, we incorporate text descriptions as conditioning information to draw the editable results through a specifically designed* ***Cross-modality Editing Module (CEM)****. Specifically, CEM adaptively integrates the initial prediction with music and text prompts as temporal motion cues to guide the synthesized sequences. Thereby, the results display music harmonics while preserving fine-grained semantic alignment with text descriptions. Extensive experiments demonstrate that our method outperforms the state-of-the-art models on our newly collected DanceRemix dataset. Code is available at* https://lzvsdy.github.io/DanceEditor/.

* These authors contribute equally. ✉ Corresponding authors.

1. Introduction

The music-conditioned dance generation seeks to create coherent and diverse human dance movements synchronizing with musical rhythms. These non-verbal body languages transcend cultural boundaries to communicate emotions, ideas, and narratives in real life [7, 15]. Meanwhile, modeling music-driven dance movements has broad real-world applications, including choreography creation [5, 35, 50], embodied AI [29, 32], avatar animation [1, 36, 39], human-machine interaction [33, 34, 39], and virtual/augmented reality (AR/VR) [4, 13]. In recent years, many studies have been conducted to tackle this complex challenge.

While significant progress has been made in directly music-driven dance generation [5, 17, 18, 25, 46], dance editing remains rarely explored. Although some works synthesize the controllable dance sequences by incorporating text prompts and music jointly [10, 22, 24, 49], they mostly overlook producing iterative high-quality results that follow the user guidance in practice. Meanwhile, few researchers have devoted themselves to constructing datasets with multi-turn editable dance movements aligned with music conditions. For example, the choreographer may intend to iteratively edit the unsatisfactory generated initial dance by concise edit descriptions such as 'lift right knee higher" or 'kick left leg twice". In this work, we therefore introduce the new task of iterative editable dance generation conditioned on music and text descriptions, as shown in Figure 1.

There are two main challenges in this task: 1) High-quality dance datasets supporting multi-turn editing are scarce. Creating such a dataset composed of accurate text descriptions aligned with both iterative motions and music is quite difficult. The intrinsic subjective nature of dance makes it difficult to describe, combined with the costly labor required for annotations. 2) Modeling coherent and diverse dance motions aligned with musical rhythms and tempos is challenging, especially when utilizing iteratively open-vocabulary edit descriptions.

To address the issue of data scarcity, we construct a new large-scale multi-turn editable dance dataset comprising the prompt featuring over 25.3M dance frames and 84.5K pairs, dubbed **DanceRemix**, as shown in Figure 2. The key challenge is how to carefully build the iterative dance motions aligned with the given music signals. Inspired by the text-motion retrieval approach TMR [31], our insight is to leverage the pre-annotated text descriptions of dance motions for constructing pairs of similar dances. In particular, we first employ the TMR to conduct motion-to-motion retrieval, allowing us to obtain pairs of candidates for multiple rounds of editing. To ensure the retrieved dance motion rhythmically aligned with music beats, we use dynamic time warping to match motion and music beats as inspired by [49]. Once the motion pairs are acquired, we utilize the advanced MLLM [42] to derive dense descriptions of dance movements. Furthermore, based on these captions, we leverage ChatGPT [28] to produce the coherent transformation scripts from given sequences to similar ones.

Based on our DanceRemix dataset, we propose a novel framework, named **DanceEditor**, to model music-driven dance generation and enable iterative editing by open-vocabulary descriptions. To achieve this goal, our framework is built to unify multi-modal conditions by a prediction-then-editing paradigm. In the initial prediction stage, we introduce a transformer-based diffusion branch to directly produce the high-fidelity motion sequences according to tailored aligned music. The second stage aims to synthesize the multi-turn iterative edited motions, followed by the guidance of the text prompt and initial results.

To ensure the edited dance movements are fine-grained semantic alignment with edit prompts while preserving the rhythmic of the initial prediction with music signals, we present an editing paradigm called Cross-modality Editing Module (CEM), analogous to the famous ControlNet [52]. Specifically, CEM adaptively integrates the initial dance predictions with music and text prompts into a shared joint latent space. Then, we model the joint embedding of text and initial dance, as well as text and the current iterative motions, while thoroughly considering the temporal correlation between the text and dance. The learned joint embedding is leveraged as a soft weight to provide the temporal dependence of the generated current edited dance movements w.r.t. the initial ones. Here, we infuse it into the editing branch via an adaptive instance normalization (AdaIN) layer [48], thereby further enhancing high-fidelity dance generation that aligns well with the edited descriptions. Extensive experiments on our newly collected DanceEditor dataset verify the effectiveness of our method, demonstrating diverse and plausible dance generation and editing.

Overall, our contributions are summarized as follows:

- We introduce the new task of iterative, editable music-driven dance generation with one newly collected large-scale dataset named DanceRemix, significantly paving the way for diverse dance generation and editing.
- We propose DanceEditor, a novel framework that enables high-fidelity multi-turn iterative choreography following a prediction-then-editing paradigm.
- We present a custom-designed Cross-modality Editing Module (CEM) to encourage the temporal synchronization of dance movements w.r.t. multi-modal conditions, including music, and text prompts, thereby allowing open-vocabulary text-guided coherent dance motion generation.
- Extensive experiments show that our method outperforms state-of-the-art counterparts, displaying realistic and impressive iterative editable dance sequence generation.

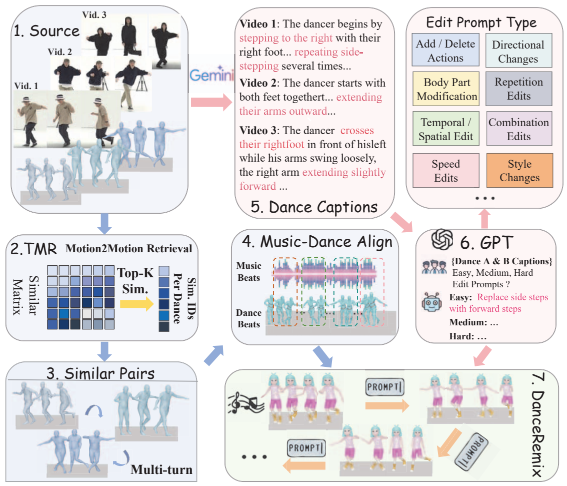


Figure 2. The workflow of DanceRemix dataset construction. Firstly, we perform motion-to-motion retrieval to obtain similar dance motion pairs. Then, we align the motion beats of the edited dance with the music beats. For aligned dance pairs, we use Gemini to generate dense dance captions for the dance videos. Next, based on the generated captions, we leverage ChatGPT to generate edit instructions. Through several motion pair retrievals, we obtain music, seed dance, a series of edit prompts, and corresponding edited dance motions.

2. Related Work

Music to Dance Generation: Early studies in music-to-dance synthesis focused on template-based approaches, including example-based alignment [9], statistical music-motion mapping [27], and music similarity-driven motion stitching [16]. While pioneering, these methods suffer from short motion duration, limited diversity, and unnatural transitions due to reliance on handcrafted rules. Subsequent deep learning approaches improve generalization: GrooveNet [2] and ChoreoMaster [5] leverage neural networks for real-time synthesis, while LSTM-autoencoders [41] and adversarial learning frameworks like DeepDance [40] enhance temporal coherence. Nevertheless, deterministic generation paradigms lead to rigid motions and restricted style diversity.

The advent of Transformers revolutionizes temporal modeling. Works like AI choreographer [17] and Music2Dance [56] utilize cross-modal attention to align music and motion features, achieving superior rhythmic synchronization. However, challenges like unnatural joint rotations and motion discontinuities persist in a complex choreography. Recent generative models further advance the field: GAN-based methods [14] improve continuity but face mode collapse, while diffusion models (*e.g.*, EDGE [46], Bailando [51]) achieve state-of-the-art quality through iterative denoising. Despite progress, critical limitations remain high computational costs, temporal inconsistency in

Table 1. To the best of our knowledge, our **DanceRemix** is the first large-scale dataset that enables **iterative** editable dance generation. Given each music segment, our dataset contains at least two different iterable editing sequence pairs (*i.e.*, $\times$ 2) with corresponding editing prompts.

Dataset	Edit prompt	Motion	Music(hours)
Duolando[37]	✗	334	0.38
PopDanceSet[25]	✗	1,036	0.56
Finedance [18]	✗	211	14.6
AIST++[17]	✗	1,408	5.19
DanceRemix(Ours)	**84,523**	**84,523$\times$2**	**117.39**

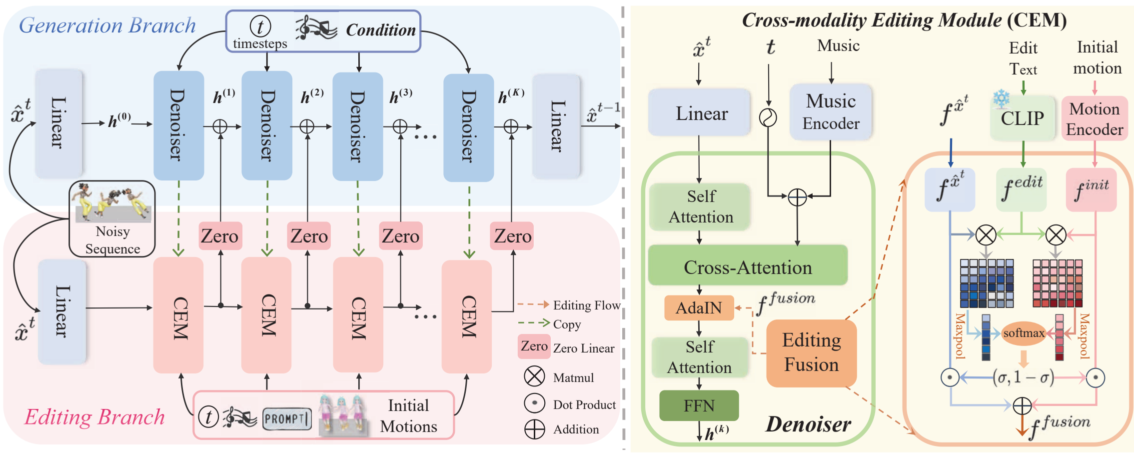
long sequences, and an inability to support iterative user editing, which is essential for practical choreography.

Text to Motion Generation: Text-to-motion research primarily follows two paradigms. Align-based methods [11, 30, 43] project text and motion into shared latent spaces but struggle with fine-grained correlations. Condition-based models like MotionDiffuse [53], MDM [44], and MLD [6] directly inject text features into generators via diffusion, yet neglect editability. Recent editing techniques attempt to bridge this gap: FineMoGen [54] edits spatiotemporal details but requires restrictive input formats. MotionFix [3] leverages paired data but fails in dance-specific edits. MotionLab [12] unifies generation/editing but lacks whole-body control. A key limitation across these works is the disruption of music-dance synchronization during editing which is also limited by data scarcity. Existing datasets (*e.g.*, AIST++ [47], FineDance [18]) focus on single-turn generation, lacking edit trajectories or textual instructions. This impedes the development of frameworks requiring iterative refinement. Our work addresses these gaps through the DanceRemix dataset and a multi-conditioned (music + text) editing framework, enabling coherent and user-controllable dance synthesis.

3. Proposed Method

3.1. Iterative Editable Dance Dataset Construction

Dataset Description: Due to the deficiency of single mapping of dance and music in existing datasets, we newly collect DanceRemix for iterative editable dance generation, upon prior works [17, 21, 25]. To be precise, This extensive dataset ensures diverse dance movements, allowing our model to learn intricate motion patterns and generate vivid dance motions. Additionally, to support our insight into iterative editable dance generation, our dataset has well-aligned dance motions and edit prompts. Specifically, several dance sequences corresponding to the same music piece are well-synchronized with musical rhythms and beats. The transitions among these dance segments are seamlessly aligned with the edit prompts. To the best of our



a) DanceEditor Framework b) Denoiser & CEM Module

Figure 3. In the initial prediction stage, a diffusion transformer-based **Generating Branch** takes music signals as input and synthesizes vivid dance motions. During the second stage, the **Editing Branch** that contains a **Cross-modality Editing Module (CEM)** adaptively incorporates the initial dance predictions with both the music and text prompts, guiding the generation of edited dance sequences.

knowledge, this is the first large-scale multi-turn editable dance dataset, as reported in Table 1.

Automatic Data Collection Workflow: Constructing datasets conducive to our task focuses on formulating one-to-many mapping (one music signal w.r.t. several iterative movement pairs), and synthesising fine-grained editing text prompts among pairs of movements. Thanks to advanced large multi-modality model approaches, we devise an automatic data collection workflow as illustrated in Figure 2. To obtain pairs of similar dance sequences, we employ a text-motion retrieval approach, TMR [31], to perform motion-to-motion retrieval based on re-organized motion sequences from prior datasets. Then we apply a top-k selection strategy to identify the similar pairs with an obvious but natural transition. To ensure the retrieved dance motions are rhythmically aligned with the same piece of music, we use dynamic time warping [49] to determine the optimal alignment between musical and motion beats. In this manner, movements that are not in harmony with the beat of the music are dropped out and manually checked.

Once several similar dance sequences well-aligned with the same music piece are acquired, we use Gemini [42] to generate fine-grained motion captions. Based on these captions, we leverage ChatGPT [28] to create coherent transformation scripts representing editing text prompts from one sequence to another similar one. Notably, the richness and detail of the captions enable diverse edit prompts, including additions, deletions, body part modifications, and temporal or spatial edits, as shown in Figure 2. By integrating these methods, we can automatically and cost-effectively generate a large-scale, multi-turn editable dance dataset with well-aligned motions and edit prompts.

3.2. Problem Formulation

Given a sequence of music signal $M = \{m_1, \ldots, m_N\}$, our DancerEditor framework aims to produce iterative editable human dance movements $X = \{x_1, \ldots, x_N\}$ with the aid of open-vocabulary descriptions. Here, N represents the total frame length of dance sequences. For dance posture x_i in each frame, we leverage the 24-joint SMPL [23] format to represent body structure, where each joint is represented in 6D rotation representation for better modeling. Besides, we introduce a 3-dimensional root position and a 4-dimensional binary foot contact to indicate the global body movements in a continuous sequence. Similar to the prior work [46], we utilize the advanced audio technique Jukebox [8] to map the raw music waves into rhythmic features.

3.3. Prediction-then-Editing Paradigm

Considering the iterative editable dance movement synthesis should preserve coherent alignment with both the given music signals and editing text prompts, we present a novel prediction-then-editing paradigm to unify the incorporation of these multimodal conditions. As shown in Figure 3, the prediction-then-editing paradigm is collaboratively composed of an initial prediction stage and an iterative editing stage. In the initial prediction stage, we utilize a music-conditioned dance diffusion model to synthesize dance movements from tailored, aligned music. During the iterative editing stage, we intend to exploit the text prompts as guidance for editable dance motion production.

3.3.1. Music-conditioned Initial Prediction

As illustrated in Figure 3, the initial prediction stage is composed of a diffusion transformer-based generation branch.

This branch is capable of effectively learning both the beat and tempo of the dance, thus ensuring synchronization and alignment with the accompanying music signals. During training, our denoiser is trained to generate continuous dance motions from noisy dance sequences x^t, conditioned on the diffusion timesteps t and the music condition c_m. We first feed the noisy dance sequences into a linear layer to obtain the hidden state of encoded dance movement, indicated as h^k, where $k \in \{0, ..., K\}$ is the stacked denoiser blocks for K times. The denoising process is guided by the simple objective:

$$\mathcal{L}_{simple} = \mathbb{E}_{\mathrm{x},t,c_m,\epsilon\sim\mathcal{N}(0,1)} \left[\left\| \mathrm{x} - \mathcal{D}_c(\mathrm{x}^t, t, c_m) \right\|_2^2 \right], \quad (1)$$

where $\mathcal{D}_c$ represents our music-conditional denoiser, $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the added random Gaussian noise, and $\mathrm{x}^t = \mathrm{x} + \sigma_t \epsilon$ is the gradually noise adding process at step t. Here, $\sigma_t \in (0, 1)$ is a constant hyperparameter. Moreover, following the settings of [11, 45], we adopt the velocity loss $\mathcal{L}_{vel}$ and foot contact loss $\mathcal{L}_{foot}$ to improve the smoothness and physical plausibility of the generated dance motions. Accordingly, the overall objective is defined as:

$$\mathcal{L}_{total} = \lambda_{simple}\mathcal{L}_{simple} + \mathcal{L}_{vel} + \mathcal{L}_{foot}, \quad (2)$$

where λ_{simple} is trade-off weight coefficients.

3.3.2. Iterative Editable Dance Generation

In the following editing stage, we incorporate the editing transformation prompts to achieve iterative editable dancer generation. Inspired by ControlNet [52], we introduce an editing branch to inject the condition signals into the dance motion generation, as depicted in Figure 3.

Cross-modality Editing Module: To ensure the edited dance movements have fine-grained semantic alignment with edit prompts while preserving the rhythm of the initial prediction with music signals, we design a Cross-modality Editing Module (CEM). Considering that the editing prompts usually display as the temporal-wise effects upon the final results, our CEM effectively captures the temporal relevance between the extracted text prompt features and the motion embeddings of iteratively edited dance. In particular, we first include the addition of a denoising timestep and music condition as the query feature to match the key feature and value feature belonging to the current noisy motion via a cross-attention mechanism, thereby obtaining the music-coherent current sequence embedding.

Then, to achieve the fine-grained semantic control of the updated current sequence embedding, we introduce an editing fusion block. Here, we calculate the temporal correlation matrix between the updated current dancer embedding and extracted text embedding as $M^{edit} \in \mathbb{R}^{N\times N}$. This matrix represents the temporal variants of the editing impact upon the current dance sequence. Subsequently, we conduct a similar operation to acquire the matrix that denotes the temporal deformation of the editing impact upon the initial dance sequence, as $M^{init} \in \mathbb{R}^{N\times N}$. Furthermore, we obtain the two learnable parameters represented by dance fusion weights of the initial dances and the current iterative dance motions, as follows

$$(\sigma, 1-\sigma) = \mathrm{Softmax}(AdPool(M^{edit}), AdPool(M^{init})), \quad (3)$$

where $AdPool$ indicates the AdaptiveMaxPooling operation along the temporal dimension.

Once we acquire these weight parameters, we leverage them to obtain the fused motion embedding by adaptively integrating the current motion sequence features and the initial ones, formulated as:

$$f^{fusion} = \sigma \cdot f^{init} + (1-\sigma) \cdot f^{\hat{x}_t}, \quad (4)$$

where f^{init} denotes the initial motion features and $f^{\hat{x}_t}$ indicates the current iterative movement features. Through this fashion, the obtained motion fusion embedding contains the collaborative guidance representation with the help of edit cues. The collaborative guidance representation is then exploited to boost the current iterative motion features by an adaptive instance normalization (AdaIN) layer [48].

$$f_h'^{\hat{x}_t} = \mathrm{AdaIN}(f^{\hat{x}_t}, f^{fusion}), \quad (5)$$

where $f_h'^{\hat{x}_t}$ represents edited motion embedding. In this fashion, CEM allows the editing branch to effectively integrate temporal motion cues from the initial dance movements and edit prompts, enabling the iterative generation of editable dances through prompt adjustments. For the iterative editing stage of text prompt conditioned dance generation, our training objective is consistent with the previous initial prediction stage.

4. Experiments

4.1. Datasets and Experimental Setting

DanceRemix Dataset: Since the existing dance generation datasets fail to provide accurate edit descriptions aligned with both dance motions and music, we contribute a new dataset named DanceRemix to evaluate our approach, upon prior work [17, 21, 25]. The dance motions of our DanceRemix are collected from online sources and existing dance datasets. We then employ TMR to retrieve similar dance pairs, ensuring their alignment with music beats through dynamic time warping. To generate coherent and natural transformation scripts for motions in these similar

*Here, the editing transformation prompts usually represent the movement changes from the initial sequence to the target one, such as “lifting your right leg higher”, or “swing your arms wider”.

Table 2. Comparison of our DanceEditor framework and the state-of-the-art methods on our DanceRemix dataset and POPDG dataset. ↑ denotes the higher the better, and ↓ indicates the lower the better.

Models	DanceRemix Dataset				POPDG Dataset			
	FID ↓	BAS ↑	Diversity ↑	PFC ↓	FID ↓	BAS ↑	Diversity ↑	PFC ↓
EDGE [46] $_{CVPR'23}$	3.91	0.2519	2.29	1.635	4.26	0.2491	2.14	6.156
TM2D [10] $_{ICCV'23}$	3.84	0.2470	2.16	1.327	4.15	0.2466	2.09	5.340
Lodge [19] $_{CVPR'24}$	3.57	0.2545	2.92	1.559	4.03	0.2473	2.57	5.788
POPDG [25] $_{CVPR'24}$	4.02	0.2513	2.64	1.122	4.32	0.2502	2.33	4.875
DanceEditor	**2.83**	**0.2560**	**3.12**	**0.784**	**2.87**	**0.2535**	**3.05**	**3.514**

dance pairs, we leverage the state-of-the-art MLLM[42] to extract dense captions from dance videos. Subsequently, we utilize ChatGPT[28] to generate diverse editing instructions. This data collection workflow takes more than three months, acquiring over 84.5 K motion pairs with corresponding music/edit descriptions. To further advance multi-turn editable dance generation, we extend DanceRemix to DanceRemix-X. By transforming dance pairs, we also construct three-level edit descriptions representing different granularities. For more details, please refer to the supplementary materials.

Implementation Details: We set the motion and music sequence length to 5 seconds, namely $N = 150$ frames. Additionally, we apply first-frame canonicalization to ensure that the motions face the same direction and have identical initial global positions, similar to the approaches in [3, 30]. Temporally, our DanceEditor synthesizes the 5-second dance motions including 24 joints in practice. Each joint is converted to the 6D rotation representation [55] for better modeling in the experiments.

In the initial prediction stage, we set $\lambda_{simple} = 10$, empirically. The total diffusion time step is $1,000$ with the cosine noisy schedule [26]. The initial learning rate is set as 1×10^{-4} with AdamW optimizer. Our model is trained on 8 NVIDIA H800 GPUs with a batch size of 128. The total training process takes a total of 250 epochs, accounting for 1 day. During the following editing stage, we train the CEM module with a batch size of 96 for 200 epochs. The initial learning rate is set as 1×10^{-5}. We take the DDIM [38] sampling strategy with 50 denoising timesteps for inference.

Evaluation Metrics:

- **FID:** We compute the Fréchet Inception Distance (FID) using features extracted from an independently trained motion autoencoder, following previous works [5, 18, 22]. A lower FID indicates that the generated dance motion better matches the ground truth distribution.
- **BAS:** The Beat Alignment Score (BAS) measures the temporal alignment between music and dance by calculating the correspondence of music beats with sequence movements. The higher score means the better music-dance rhythmic alignment.
- **Diversity:** The Diversity score measures the distance between generated motion features conditioned on different music inputs. The higher score means the better diversity of generated results.
- **PFC:** The Physical Foot Contact (PFC) score evaluates the plausibility of dance movements based on hip acceleration and foot velocity, where a lower score signifies more realistic foot-ground contact.
- **Motion-Editing Text Align Score (MEAS):** Inspired by [20], we measure the distance between dance motion pairs and editing texts using a custom-trained CLIP-based model. A lower score indicates better alignment between the edited dance and the editing descriptions.

4.2. Quantitative Evaluation

Comparisons with SOTA Methods: To the best of our knowledge, we are the first to explore iterative, editable music-driven dance generation with open-vocabulary descriptions. To thoroughly validate the superiority of our method, we retrain various state-of-the-art (SOTA) Music-to-Dance **(M2D)** approaches on our DanceRemix dataset setting. For diffusion-based methods—EDGE [46], Lodge [19], and POPDG [25]—as well as VQ-VAE models like TM2D [10], we leverage the official implementations and adhere to the published hyperparameter configurations.

As shown in Table 2, we adopt the FID, BAS, Diversity, and PFC for a well-rounded view of comparison. Our DanceEditor outperforms all the competitors by a large margin on the DanceRemix dataset. Remarkably, our method even achieves more than 21% $(i.e., (3.57 - 2.83)/3.57 \approx 21\%)$ improvement over the sub-optimal counterpart in FID. We observe that both Lodge [19] and ours synthesize the authority dance motions with much higher diversity than others. This is because both benefit from the two-stage paradigms, which separately focus on music-aligned global dance patterns and highly effective local dance quality. However, Lodge shows lower performance on FID due to its lack of learning fine-grained dance transformations from given sequences to similar ones. In terms of BAS, our method achieves much better results than other counterparts. This strongly aligns with our insight into the prediction-then-editing paradigm and CEM, which encour-

Table 3. Ablation on Prediction-then-Editing Paradigm. ↑ denotes the higher the better, and ↓ indicates the lower the better.

Models	DanceRemix Dataset			
	FID ↓	BAS ↑	Diversity ↑	MEAS ↓
Ours (Initial)	**2.83**	**0.2560**	3.12	✗
Ours (Iteration #1)	2.85	0.2553	3.16	**0.784**
Ours (Iteration #2)	2.91	0.2541	3.23	0.786
Ours (Iteration #3)	3.04	0.2524	**3.35**	0.793

Table 4. Ablation on Cross-modaliy Editing Module. ↑ denotes the higher the better, and ↓ indicates the lower the better.

Models	DanceRemix Dataset			
	FID ↓	BAS ↑	Diversity ↑	MEAS ↓
w/o Editing Brach	3.95	0.2514	2.32	1.351
Editing Branch w/o CEM	3.68	0.2537	2.69	1.024
DanceEditor (full version)	**2.85**	**0.2553**	**3.16**	**0.784**

ages our model to further learn the local variations of temporal motions and their interactions with music beats and rhythms.

Additionally, our method achieves a significantly lower PFC compared to others, indicating more physically realistic motions.

We further evaluate our method on the POPDG dataset, with the results presented in Table 2. These findings further validate the superiority of our proposed method.

Ablation Study: Here, to evaluate the effectiveness of our proposed prediction-then-editing paradigm and CEM Module, we conduct extensive ablation studies separately.

Ablation on Prediction-then-Editing Paradigm: To verify the effectiveness of our proposed prediction-then-editing paradigm, we formulate the ablation on multi-turn iterative editing, and the results are reported in Table 3. Given a period of music signals, our framework enables the generation of high-quality sequences while preserving harmonic rhythm with the beat of the music. The initial prediction is produced directly by the generation branch from music signals. Then, we integrate the initial prediction with the corresponding music & text editing prompts into the editing branch to synthesize the editable dance movements (*i.e.*, denoted as Iteration #1). The subsequent iterations are generated in a similar way. The results demonstrate that our prediction-then-editing paradigm effectively promotes the framework to model the dance movements in a carefully learned joint distribution space.

In particular, with continuous iterative editing, the performance of our model always tends to a stable state. The metrics Diversity clearly attain much better results than the initial ones. This indicates that the editing text prompts in our dataset significantly facilitate the framework to learn prolific dynamic features. The MEAS metric slightly decreases due to minor motion shifts from the previous editing iteration, but the overall visual quality remains excellent. Although the FID displays slightly worse scores than the initial ones due to the sampled open-vocabulary text descriptions, our method still achieves the optimal result. Meanwhile, we observe that the BAS scores still perform better than the state-of-the-art counterparts. This applies the effectiveness of our editing paradigm that can adaptively model the temporal dependency between the music beats and iterative dance movements. Our prediction-then-editing paradigm demonstrates a superior ability to generate iterative editable dance motions with a unified framework.

Ablation on the CEM module: To further validate the effectiveness of our Cross-Modality Editing Module (CEM), we conduct ablation studies to assess the necessity of the editing branch and the effectiveness of CEM. Specifically, we evaluate experimental settings where both music and editing text are used as inputs to generate dance motions. As reported in 4, we design three models for comparison: (1) only the generation branch, (2) two branches without the CEM module, and (3) DanceEditor with both prediction and editing. The first two models use simple concatenation to fuse music and textual features. To be specific, the generation-only branch struggles to directly learn the complex interactions between dance motions, music, and text. As a result, implementing the model without an editing branch leads to worse performance across all metrics. Meanwhile, our CEM brings significant improvements across all metrics compared to the editing branch without CEM. This demonstrates that CEM effectively enhances the temporal dependency between music beats and iterative dance movements, leading to vivid and coherent motions.

4.3. Qualitative Evaluation

Visualization: To showcase the superior performance of our method, we present visualized keyframes from our DanceEditor framework alongside other approaches on our DanceRemix dataset. Unlike others, which generate dance solely from music, our method showcases both the initial prediction and iteratively edited dance motions As shown in Figure 4, our method generates vivid and diverse dance movements compared to others. Specifically, EDGE and TM2D tend to produce stiff and unnatural results. While Lodge generates relatively varied motions, it suffers from foot drifting and abrupt changes. POPDG, though more natural, lacks expressiveness. In contrast, our DanceRemix not only generates natural and expressive dance movements but also seamlessly edits motions while maintaining coherence with the edit instructions, aligning well with our prediction-then-editing paradigm. For more visualization demos, please refer to the videos available on our project website.

User Study: To assess the quality of synthesized dance

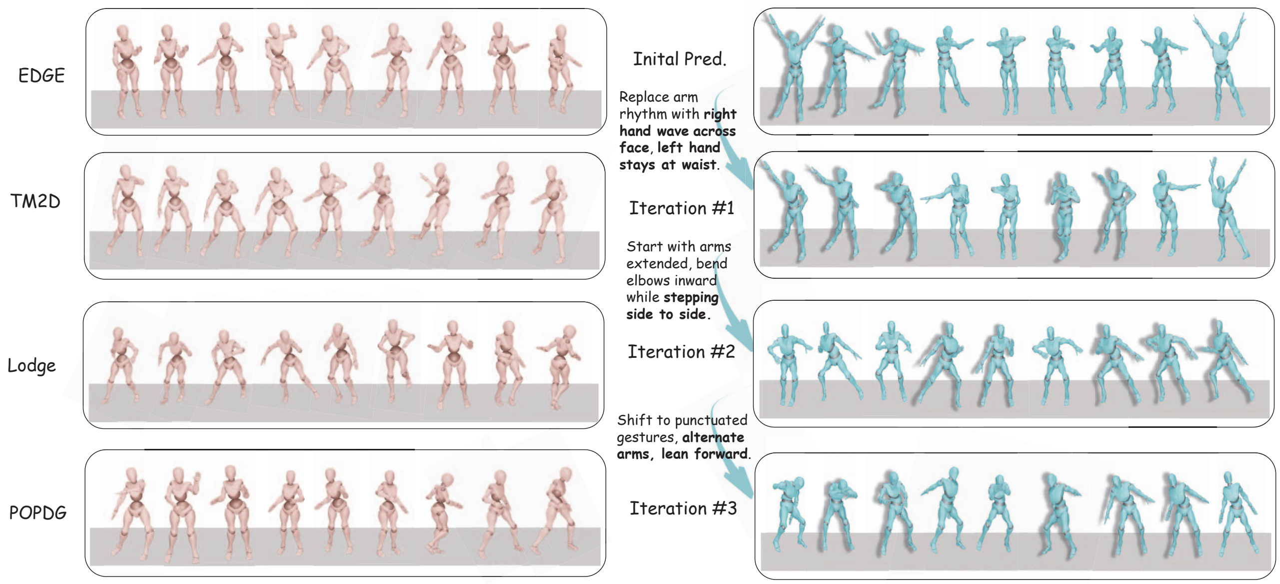


Figure 4. Given the same music segments, the results generated by our **DanceEditor** framework and other SOTA comparison methods.

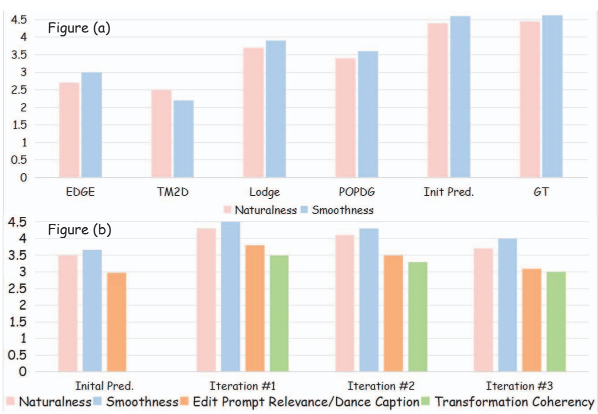


Figure 5. User study with dance naturalness, motion smoothness, Motion-Editing Text Alignment, and motion transformation coherency on our dataset.

movements, we conducted a user study with 15 anonymously recruited participants from diverse academic backgrounds. For each comparison method and our prediction stage, we randomly selected two generated videos. Additionally, for our iterative editing results, we presented two editing processes, each with an initial prediction and three iterations. Each participant should watch 18 videos, each lasting 5 seconds. They rated the dance results for the M2D task based on naturalness and motion smoothness. For our iterative editable dance generation, we introduced two additional evaluation criteria: edit prompt relevance and transformation coherence. The visualized videos are randomly selected, ensuring at least two samples per method. The results, shown in Figure 5, are rated on a scale of 0 to 5 (higher is better). Our framework outperforms all competitors, particularly excelling in smoothness, highlighting the effectiveness of our prediction-then-editing paradigm. Even after multiple editing iterations, edit prompt relevance and transformation coherence decline only slightly, with CEM ensuring motions remain well-aligned with the edit texts.

5. Conclusion

In this paper, we introduce DanceEditor, a novel framework for iterative and editable dance generation coherently aligned with given music signals. To fulfill this goal, we first newly collect a large-scale multi-turn editable dance dataset, DanceRemix, featuring over 25.3M dance frames and 84.5K pairs. Along with this dataset, our approach generates vivid dance motions and enables coherent editing guided by open-vocabulary descriptions. To further capture the temporal synchronization of dance movements w.r.t. multimodal conditions, we propose a Cross-modality Editing Module that adaptively integrates the initial dance predictions with both music and texts to guide edited dance genreation. Extensive experiments conducted on DanceRemix show the superiority of our model.

Limitation Currently, we support only the generation and editing of body movements in dance sequences. In the future, we will explore semantic descriptions and editing of facial expressions and gestures to create more expressive and realistic dance performances coherent with music.

Acknowledgements. This work is funded in part by the HKUST Start-up Fund (R9911), Theme-based Research Scheme grant (T45-205/21-N), the InnoHK funding for Hong Kong Generative AI Research and Development Center, Hong Kong SAR, and the research funding under HKUST-DXM AI for Finance Joint Laboratory (DXM25EG01).

References

[1] Saliha Akbas, Asim Evren Yantac, Terry Eskenazi, Kemal Kuscu, Sinem Semsioglu, Onur Topal Sumer, and Asli Ozturk. Virtual dance mirror: A functional approach to avatar representation through movement in immersive vr. In *Proceedings of the 8th International Conference on Movement and Computing*, pages 1–4, 2022. 2

[2] Omid Alemi, Jules Françoise, and Philippe Pasquier. Groovenet: Real-time music-driven dance movement generation using artificial neural networks. *networks*, 8(17):26, 2017. 3

[3] Nikos Athanasiou, Alpár Cseke, Markos Diomataris, Michael J Black, and Gül Varol. Motionfix: Text-driven 3d human motion editing. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 3, 6

[4] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5933–5942, 2019. 2

[5] Kang Chen, Zhipeng Tan, Jin Lei, Song-Hai Zhang, Yuan-Chen Guo, Weidong Zhang, and Shi-Min Hu. Choreomaster: choreography-oriented music-driven dance synthesis. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 2, 3, 6

[6] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 3

[7] Jane Desmond. *Meaning in motion: New cultural studies of dance*. Duke University Press, 1997. 2

[8] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020. 4

[9] Rukun Fan, Songhua Xu, and Weidong Geng. Example-based automatic music-driven conventional dance motion synthesis. *IEEE transactions on visualization and computer graphics*, 18(3):501–515, 2011. 3

[10] Kehong Gong, Dongze Lian, Heng Chang, Chuan Guo, Zihang Jiang, Xinxin Zuo, Michael Bi Mi, and Xinchao Wang. Tm2d: Bimodality driven 3d dance generation via music-text integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9942–9952, 2023. 2, 6

[11] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5152–5161, 2022. 3, 5

[12] Ziyan Guo, Zeyu Hu, Na Zhao, and De Wen Soh. Motionlab: Unified human motion generation and editing via the motion-condition-motion paradigm. *arXiv preprint arXiv:2502.02358*, 2025. 3

[13] Javid Iqbal and Manjit Singh Sidhu. Acceptance of dance training system based on augmented reality and technology acceptance model (tam). *Virtual Reality*, 26(1):33–54, 2022. 2

[14] Jinwoo Kim, Heeseok Oh, Seongjean Kim, Hoseok Tong, and Sanghoon Lee. 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*, pages 3490–3500, 2022. 3

[15] Kimerer LaMothe. The dancing species: how moving together in time helps make us human. *Aeon, June*, 1(1):1–2, 2019. 2

[16] Minho Lee, Kyogu Lee, and Jaeheung Park. Music similarity-based approach to generating dance motion sequence. *Multimedia tools and applications*, 62:895–912, 2013. 3

[17] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021. 2, 3, 5

[18] Ronghui Li, Junfan Zhao, Yachao Zhang, Mingyang Su, Zeping Ren, Han Zhang, Yansong Tang, and Xiu Li. Finedance: A fine-grained choreography dataset for 3d full body dance generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10234–10243, 2023. 2, 3, 6

[19] Ronghui Li, YuXiang Zhang, Yachao Zhang, Hongwen Zhang, Jie Guo, Yan Zhang, Yebin Liu, and Xiu Li. Lodge: A coarse to fine diffusion network for long dance generation guided by the characteristic dance primitives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1524–1534, 2024. 6

[20] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, 132(9):3463–3483, 2024. 6

[21] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36: 25268–25280, 2023. 3, 5

[22] Zeyu Ling, Bo Han, Shiyang Li, Hongdeng Shen, Jikang Cheng, and Changqing Zou. Motionllama: A unified framework for motion synthesis and comprehension. *arXiv preprint arXiv:2411.17335*, 2024. 2, 6

[23] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 4

[24] Mingshuang Luo, Ruibing Hou, Zhuo Li, Hong Chang, Zimo Liu, Yaowei Wang, and Shiguang Shan. M^3 gpt: An advanced multimodal, multitask framework for motion comprehension and generation. *arXiv preprint arXiv:2405.16273*, 2024. 2

[25] Zhenye Luo, Min Ren, Xuecai Hu, Yongzhen Huang, and Li Yao. Popdg: Popular 3d dance generation with popdanceset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26984–26993, 2024. 2, 3, 5, 6

[26] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 6

[27] Ferda Ofli, Engin Erzin, Yücel Yemez, and A Murat Tekalp. Learn2dance: Learning statistical music-to-dance mappings for choreography synthesis. *IEEE Transactions on Multimedia*, 14(3):747–759, 2011. 3

[28] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 2, 4, 6

[29] Hua Peng, Changle Zhou, Huosheng Hu, Fei Chao, and Jing Li. Robotic dance in social robotics—a taxonomy. *IEEE Transactions on Human-Machine Systems*, 45(3):281–293, 2015. 2

[30] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. 3, 6

[31] Mathis Petrovich, Michael J Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9488–9497, 2023. 2, 4

[32] Xingqun Qi, Chen Liu, Muyi Sun, Lincheng Li, Changjie Fan, and Xin Yu. Diverse 3d hand gesture prediction from body dynamics by bilateral hand disentanglement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4616–4626, 2023. 2

[33] Xingqun Qi, Chen Liu, Lincheng Li, Jie Hou, Haoran Xin, and Xin Yu. Emotiongesture: Audio-driven diverse emotional co-speech 3d gesture generation. *IEEE Transactions on Multimedia*, 26:10420–10430, 2024. 2

[34] Xingqun Qi, Jiahao Pan, Peng Li, Ruibin Yuan, Xiaowei Chi, Mengfei Li, Wenhan Luo, Wei Xue, Shanghang Zhang, Qifeng Liu, et al. Weakly-supervised emotion transition learning for diverse 3d co-speech gesture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10424–10434, 2024. 2

[35] Xingqun Qi, Hengyuan Zhang, Yatian Wang, Jiahao Pan, Chen Liu, Peng Li, Xiaowei Chi, Mengfei Li, Wei Xue, Shanghang Zhang, et al. Cocogesture: Toward coherent co-speech 3d gesture generation in the wild. *arXiv preprint arXiv:2405.16874*, 2024. 2

[36] Xingqun Qi, Yatian Wang, Hengyuan Zhang, Jiahao Pan, Wei Xue, Shanghang Zhang, Wenhan Luo, Qifeng Liu, and Yike Guo. Co^{3}gesture: Towards coherent concurrent co-speech 3d gesture generation with interactive diffusion. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[37] Li Siyao, Tianpei Gu, Zhitao Yang, Zhengyu Lin, Ziwei Liu, Henghui Ding, Lei Yang, and Chen Change Loy. Duolando: Follower gpt with off-policy reinforcement learning for dance accompaniment. *arXiv preprint arXiv:2403.18811*, 2024. 3

[38] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 6

[39] Marina Stergiou and Spyros Vosinakis. Exploring costume-avatar interaction in digital dance experiences. In *Proceedings of the 8th International Conference on Movement and Computing*, pages 1–6, 2022. 2

[40] Guofei Sun, Yongkang Wong, Zhiyong Cheng, Mohan S Kankanhalli, Weidong Geng, and Xiangdong Li. Deepdance: music-to-dance motion choreography with adversarial learning. *IEEE Transactions on Multimedia*, 23:497–509, 2020. 3

[41] Taoran Tang, Jia Jia, and Hanyang Mao. Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1598–1606, 2018. 3

[42] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2, 4, 6

[43] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022. 3

[44] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. *ArXiv*, abs/2209.14916, 2022. 3

[45] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2022. 5

[46] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 448–458, 2023. 2, 3, 4, 6

[47] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, pages 501–510, Delft, Netherlands, 2019. 3

[48] Canqian Yang, Meiguang Jin, Xu Jia, Yi Xu, and Ying Chen. Adaint: Learning adaptive intervals for 3d lookup tables on real-time image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17522–17531, 2022. 2, 5

[49] Han Yang, Kun Su, Yutong Zhang, Jiaben Chen, Kaizhi Qian, Gaowen Liu, and Chuang Gan. Unimumo: Unified text, music and motion generation. *arXiv preprint arXiv:2410.04534*, 2024. 2, 4

[50] Hye-jeong Yun, Kwang-il Kim, Jeong-hun Lee, and Hae-Yeoun Lee. Development of experience dance game using kinect motion capture. *KIPS transactions on software and data engineering*, 3(1):49–56, 2014. 2

[51] Canyu Zhang, Youbao Tang, Ning Zhang, Ruei-Sung Lin, Mei Han, Jing Xiao, and Song Wang. Bidirectional au-

toregessive diffusion model for dance generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 687–696, 2024. 3

[52] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2, 5

[53] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 3

[54] Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. Finemogen: Fine-grained spatio-temporal motion generation and editing. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[55] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753. 6

[56] Wenlin Zhuang, Congyi Wang, Jinxiang Chai, Yangang Wang, Ming Shao, and Siyu Xia. Music2dance: Dancenet for music-driven dance generation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2):1–21, 2022. 3