Walk Before You Dance: High-fidelity and Editable Dance Synthesis via Generative Masked Motion Prior

Foram N Shah*, Parshwa N Shah*, Muhammad Usama Saleem, Ekkasit Pinyoanuntapong, Pu Wang, Hongfei Xue, Ahmed Helmy

University of North Carolina at Charlotte

{fshah3, pshah77, msaleem2, epinyoan, pwang13, hxue2, ahelmy1} @uncc.edu

Abstract

Recent advances in dance generation have enabled the automatic synthesis of 3D dance motions. However, existing methods still face significant challenges in simultaneously achieving high realism, precise dance-music synchronization, diverse motion expression, and physical plausibility. To address these limitations, we propose a novel approach that leverages a generative masked text-to-motion model as a distribution prior to learn a probabilistic mapping from diverse guidance signals, including music, genre, and pose, into high-quality dance motion sequences. Our framework also supports semantic motion editing, such as motion inpainting and body part modification. Specifically, we introduce a multi-tower masked motion model that integrates a text-conditioned masked motion backbone with two parallel, modality-specific branches: a music-guidance tower and a pose-guidance tower. The model is trained using synchronized and progressive masked training, which allows effective infusion of the pre-trained text-to-motion prior into the dance synthesis process while enabling each guidance branch to optimize independently through its own loss function, mitigating gradient interference. During inference, we introduce classifier-free logits guidance and pose-guided token optimization to strengthen the influence of music, genre, and pose signals. Extensive experiments demonstrate that our method sets a new state of the art in dance generation, significantly advancing the quality and editability over existing approaches.

Project Page — https://foram-s1.github.io/DanceMosaic/

1 Introduction

Music-conditioned dance motion generation has recently garnered significant attention due to the intuitive rhythmic structure and semantic richness that music provides in guiding human motion. This field has broad applications in choreography, animation, virtual and augmented reality, and robotics. However, existing approaches (Zhuang et al. 2022; Siyao et al. 2022; Li et al. 2023, 2024; Tseng, Castellon, and Liu 2023) still struggle to generate realistic, natural, and physically plausible 3D dance motions that not only reflect a specific dance genre, but also align precisely with musical cues such as rhythm, beat, and tempo.

State-of-the-art dance generation models fall primarily into two categories: autoregressive and motion-space diffusion models. Autoregressive approaches, such as Bailando (Siyao et al. 2022), employ GPT-style next-token prediction to foster rhythmic consistency while preserving motion diversity. Nevertheless, their sequential nature often results in a misalignment between musical beats and the generated movements. In contrast, diffusion-based models like FineNet, LODGE, and EDGE (Li et al. 2023, 2024; Tseng, Castellon, and Liu 2023) achieve better synchronization between music and motion but often fail to produce high-fidelity dance sequences that match the realism and naturalness of authentic human movement. Beyond fidelity and alignment, editability is a critical feature for motion generation models. The capacity for editing tasks, such as modifying the upper or lower body or performing motion interpolation, is vital for iterative refinement and prototyping. This enables choreographers to fine-tune their concepts and supports the creative process in crafting original dances.

To overcome these challenges, we introduce DanceMosaic, a novel high-fidelity dance generation framework supporting a wide range of motion editing tasks guided by multiple input signals, including music, pose, and textual prompts describing actions or dance genres. Inspired by the human process of acquiring motor skills, first mastering foundational movements before progressing to advanced, expressive actions such as dance, DanceMosaic's central innovation lies in a multi-tower generative motion model combined with specialized training and inference strategies. Specifically, our model integrates a text-conditioned masked-motion backbone with two parallel modality-specific towers: a music-guidance tower and a pose-guidance tower. The backbone model first acquires basic motion skills from textual instructions, establishing a robust motion prior. Building upon this learned foundation, the music and pose-guidance towers further refine their capability to generate complex dance movements conditioned by musical cues and predefined pose constraints. To effectively achieve this hierarchical skill transfer, we employ synchronized and progressive masked training, enabling the motion priors learned by the text-to-motion model to be seamlessly infused into dance synthesis. This approach allows each guidance tower to optimize independently through its modality-specific loss, mitigating gradient conflicts. Addi-

*These authors contributed equally.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Figure 1: DanceMosaic generates 3D dance motions based on multiple guidance signals. The top sequence showcases generated dance motions influenced by different text prompts, including genre-based or action-specific prompts. The color-coded figures represent different dance styles, synchronized with a music signal at the bottom. The pose signal allows further motion refinement, demonstrating the flexibility and precision of DanceMosaic.

Method	$FID_g \downarrow$	$Div_g \uparrow$	BAS $\uparrow$	Time(s)$\downarrow$	Editable
FACT	97.05	6.37	0.1831	9.46	×
MNET	90.31	6.14	0.1864	10.26	×
Bailando	28.17	6.25	0.2029	1.46	×
EDGE	50.38	6.45	0.2116	2.27	✓
LODGE	34.29	5.64	**0.2397**	8.16	×
Ours	**12.96**	**8.36**	**0.2190**	**0.8**	✓

Table 1: DanceMosaic outperforms SOTA methods in terms of Quality, Diversity, Run Time, and dance-music beat alignment, while allowing mutimodal editing with music, text, and pose guidance signals. (Red: best. Blue: runner-up)

tionally, during inference, we propose classifier-free logits guidance and pose-guided token optimization to further enhance the model's responsiveness to input signals, such as music rhythms, dance genres, and pose requirements.

- We introduce DanceMosaic, the first method leveraging multi-tower generative masked models to enable high-fidelity dance generation with flexible editability.
- We propose synchronized and progressive masked training along with inference-time multimodal guidance modules, facilitating the synthesis of realistic, natural, physically plausible, and editable dance movements.
- Experimental results show that DanceMosaic establishes a new state-of-the-art performance in dance generation, significantly advancing the quality and editability achieved by existing approaches shown in Table 1.

2 Related Work

2.1 Text-conditioned Motion Synthesis

Early methods relied on motion matching (Holden et al. 2020), while generative models have since advanced text-conditioned motion synthesis (Li et al. 2020a; Holden, Saito, and Komura 2016; Rempe et al. 2021; Zhou et al. 2019; Petrovich, Black, and Varol 2021; Guo et al. 2020; Ahuja and Morency 2019; Ghosh et al. 2021; Yamada, Matsunaga, and Ogata 2018; Petrovich, Black, and Varol 2022; Athanasiou et al. 2022; Guo et al. 2022a). Diffusion-based approaches refine motion through structured denoising (Kong et al. 2023; Kim, Kim, and Choi 2022; Zhang et al. 2022) but suffer from high training complexity and slow inference. Autoregressive models (Jiang et al. 2023; Pinyoanuntapong et al. 2024a) leverage causal transformers for improved realism and diversity but struggle with fine-grained control and temporal consistency. More recent frameworks like Momask (Guo et al. 2023) and MMM (Pinyoanuntapong et al. 2024b) adopt masked motion modeling for better motion quality and diversity. Nevertheless, the use of masked modeling in music-conditioned dance generation remains unexplored due to fundamental semantic differences between textual and musical modalities. Existing masked motion methods typically follow a single-tower architecture tailored for text-conditioned tasks. In contrast, our proposed DanceMosaic approach introduces a novel multi-tower masked model designed specifically to support high-quality, editable dance synthesis guided by diverse multimodal inputs, including music signals, pose constraints, and textual prompts.

2.2 Music-Conditioned Dance Synthesis

Music-driven motion generation requires precise temporal synchronization between movement and rhythm. Early methods used motion retrieval and rule-based approaches, yielding beat-aligned but repetitive dances. Autoregressive models like Bailando (Siyao et al. 2022) use GPT-like prediction to enhance rhythmic consistency while maintaining diversity. However, their sequential nature often causes misalignment between beats and movements. Recent diffusion-based models such as FineNet, LODGE, and EDGE (Li et al. 2023, 2024; Tseng, Castellon, and Liu 2023) improve dance-music alignment but struggle with generating diverse, high-fidelity dance. Their iterative denoising also leads to slow inference, limiting real-time usability. DanceMosaic overcomes these limitations by enabling high-quality and editable dance generation with real-time inference. Unlike diffusion models, it employs bidirectional BERT-like modeling (Chang et al. 2022a; Pinyoanuntapong et al. 2024b,a),

while supporting multimodal dance generation through a novel multi-tower conditioned masked transformer.

3 Proposed Method: DanceMosaic

Given an input music signal A, a text prompt T, and pose constraints on specific joints or body parts P, our goal is to synthesize a physically plausible 3D dance sequence that aligns with the music and text prompts, while satisfying the pose constraints. To achieve this, we propose a multi-tower masked motion model, which incorporates music and pose guidance modules into the text-conditioned masked motion model. The overview of DanceMosaic, shown in Fig. 2, consists of a dance motion tokenizer and three integrated masked transformers, including a Text-to-Motion (T2M) model, Music Tower (MT), and Pose Tower (PT). Motion tokenizer transforms the raw motion sequence into categorical discrete motion tokens within a learned codebook (Sec 3.1). The T2M model learns to predict masked motion tokens, conditioned on textual prompts, like dance genres and action descriptions. The music and pose towers in Sec. 3.2 are designed to introduce the influence of music and pose signals to the T2M model via synchronized and progressive multimodal training (Sec. 3.3). During inference, the multimodal guidance modules, including classifier-free logits guidance and inference-time token optimization (ITTO), steer the token sampling process so that generated dance sequence is aligned with diverse prompt signals (Sec. 3.4).

3.1 Text-to-Motion Model as a Generative Prior

Motion Tokenizer. The goal of this module is to encode continuous dance motions into discrete categorical pose tokens using a learned codebook based on VQ-VAE (Esser, Rombach, and Ommer 2020). This discretization is a critical step, as it enables subsequent generative masked models to learn the per-token categorical distribution conditioned on diverse input modalities. To unify the representation space, we train the motion tokenizer on a mixture of text-to-motion and music-to-motion datasets, enforcing a shared latent space across modalities. This joint training not only enhances the expressiveness of learned pose tokens but also allows us to leverage text-to-motion model as a strong motion prior for music-conditioned motion generation.

Given a motion sequence $M = [m_1, m_2, \ldots, m_N]$, where each frame $m_i \in \mathbb{R}^D$ represents a 3D skeletal pose, where D is the dimension of human pose representations. The encoder compresses it into a latent representation $Z \in \mathbb{R}^{n \times d}$ with a temporal downsampling factor of N/n. The latent features $Z = [z_1, z_2, ..., z_n]$ are then quantized into discrete tokens $\bar{Z} = [\bar{z}_1, \bar{z}_2, ..., \bar{z}_n]$ from a learned codebook $\mathcal{C} = \{c_l\}_{l=1}^K$, consisting of K unique code entries. The best-matching code is determined by minimizing the Euclidean distance $\bar{z}_k = \text{argmin}_l \|z_k - c_l\|_2^2$. The tokenizer is trained using the following loss function:

$$\mathcal{L}_{\text{RVQ}} = \|M - \hat{M}\|_1 + \|\text{sg}(Z) - \bar{Z}\|_2^2 + \|Z - \text{sg}(\bar{Z})\|_2^2 \quad (1)$$

where $\hat{M}$ is the reconstructed motion and $\text{sg}(\cdot)$ represents the stop-gradient operation. To minimize the quantization errors, we adopt Residual Vector Quantization (RVQ) (Guo et al. 2023), which encodes a raw motion sequence into multiple token sequences within the latent space, with each sequence produced by a distinct quantizer. Each quantizer focuses on capturing the residual error from its predecessor. More details can be found in the supplementary material.

Text-to-Motion Model (T2M). Our T2M model employs a standard multilayer transformer, whose inputs are the concatenation of motion tokens $x_{1:t}$ from the tokenizer with t as the sequence length, and the embedding x_0 from the pre-trained CLIP model (Radford et al. 2021) that takes both the dance prompt "A person dances on [genre] style" from dance dataset and the general text prompt from HumanML3D dataset. Due to the nature of self-attention in transformers, all motion tokens are learned in relation to the text embedding. Given the discrete dance token sequence $\bar{Z} = [\bar{z}_1, \bar{z}_2, ..., \bar{z}_n]$, a subset of tokens is masked, forming a corrupted sequence $Z_{\mathbf{M}}$. This sequence, along with a text signal T, is processed by a text-conditioned masked motion transformer to recover the original dance motion. The model is trained to maximize the likelihood of correctly predicting masked tokens (Pinyoanuntapong et al. 2024b; Guo et al. 2023), i.e., minimizing the cross-entropy (CE) loss:

$$\mathcal{L}_{\text{CE}}^{T2M} = -\mathbb{E}_{\bar{Z}} \sum_{k \in \Omega} \log p_\theta(\bar{z}_k \mid Z_{\mathbf{M}}, T), \quad (2)$$

where Ω denotes the set of masked indices and $p_\theta(\bar{z}_k \mid Z_{\mathbf{M}}, T)$ is the parameterized probability of each motion token conditioned on $Z_{\mathbf{M}}$ and T.

3.2 Multi-tower Masked Motion Model

Parallel Guidance Towers. Building on the strong motion prior of the pre-trained T2M model, our model is designed to generate dance sequences conditioned on a music control signal A, a text prompt T, and pose guidance P. In particular, we integrate the text-conditioned masked motion model with two parallel modality-specific masked models, including a music-guidance tower and a pose-guidance tower. Each tower is implemented as a trainable counterpart to the original text-conditioned transformer. Each self-attention layer in the modality-specific tower is paired with a corresponding layer in the T2M model, connected via a zero-initialized linear layer (ϕ). This design, inspired by diffusion ControlNet (Zhang, Rao, and Agrawala 2023), ensures that the learned text-conditioned motion distribution does not interfere with music-to-dance mapping during early training.

Guidance Signals and Per-frame Alignment. For the music tower, we adopt a frozen Jukebox model (Dhariwal et al. 2020) to provide a per-frame music embedding sequence, enabling precise synchronization between dance movements and musical rhythm. For pose tower, we introduce a spatial editing signal $P \in \mathbb{R}^{N \times J \times 3}$ into the masked token reconstruction, allowing targeted modifications to upper-body, lower-body or joint-specific movements. The pose signal P encodes the (x, y, z) positions of each of the J skeleton joints over N frames. A subset of joint locations is provided as spatial constraints and remains fixed during generation, while the positions of the remaining joints, which are set to zero, are editable. To effectively integrate these guidance signals, we leverage learnable music and pose encoders to

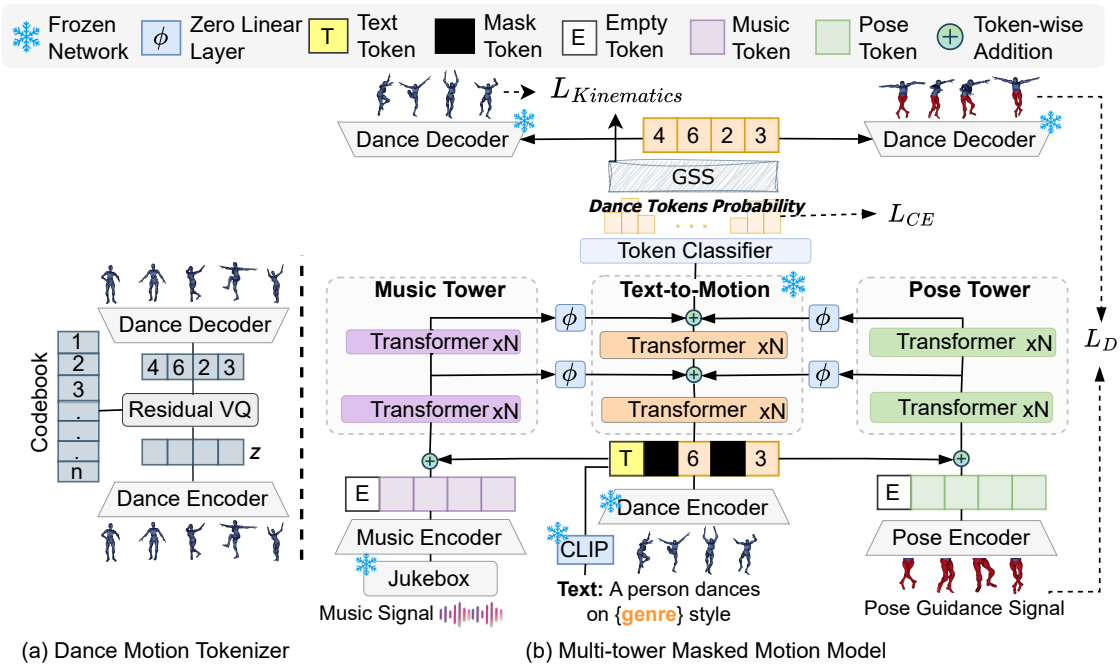

Figure 2: Overview of DanceMosaic's training phase. (a) The process involves encoding dance motions into discrete token sequences using a dance motion tokenizer. (b) These tokens are then processed through a multi-tower masked motion model, where each tower (music, text, and pose) is used to learn the probabilistic mappings from modality-specific guidance signals to motion tokens. The model is trained using a progressive training strategy to integrate music, text, and pose signals.

transform the respective inputs into latent tokens matching the dimensionality of motion tokens. These music and pose latent tokens are then precisely aligned with motion tokens on a per-frame basis via token-wise addition, ensuring coherent synchronization across dance, music, and pose inputs.

3.3 Synchronized Progressive Masked Training

Synchronized Masking. The music and pose-guidance towers are trained using synchronized masked training, where both the T2M model and guidance towers receive identically masked motion token. This synchronized masking ensures that the guidance towers can effectively leverage the motion distributions learned by the T2M model. The music tower is trained to recover the masked motion tokens, conditioned on the text prompt T, the music signal A, and the corrupted motion token sequence $Z_{\mathbf{M}}$ by minimizing the CE loss:

$$\mathcal{L}_{\text{CE}}^{MT} = -\mathbb{E}_{\bar{Z}} \sum_{k \in \Omega} \log p_\theta(\bar{z}_i \mid Z_{\mathbf{M}}, T, A), \tag{3}$$

where Ω represents the masked indices.

Besides CE loss $\mathcal{L}_{\text{CE}}^{MT}$, we adopt additional kinematics losses, including joints position loss $\mathcal{L}_{\text{pos}}$, velocity loss $\mathcal{L}_{\text{vel}}$, acceleration loss in $\mathcal{L}_{\text{acc}}$, and foot loss $\mathcal{L}_{\text{foot}}$. These losses measure the difference in motion dynamics between the generated dance and ground-truth motion. Incorporating these losses can improve the physical plausibility and naturalness of diffusion-based dance generation (Tseng, Castellon, and Liu 2023). However, integrating these kinematics losses into generative masked model training is difficult because it requires converting discrete pose tokens from the model's latent space into continuous Euclidean space. This conversion requires sampling the categorical distribution of motion tokens during training, which is non-differentiable. To address this challenge, we employ a straight-through Gumbel-Softmax strategy (Li et al. 2020b) to make token sampling process differentiable by approximating discrete categorical distribution with continuous Gumbel-Softmax distribution. The total loss function $\mathcal{L}_{\text{MT}}$ is the weighted sum of the CE loss $\mathcal{L}_{\text{CE}}^{MT}$ and kinematics losses ($\mathcal{L}_{Kinematics}$), i.e., $\mathcal{L}_{MT} = \mathcal{L}_{\text{CE}}^{MT} + \lambda_{pos}\mathcal{L}_{\text{pos}} + \lambda_{vel}\mathcal{L}_{\text{vel}} + \lambda_{acc}\mathcal{L}_{\text{acc}} + \lambda_{foot}\mathcal{L}_{\text{foot}}$. The details of $\mathcal{L}_{\text{MT}}$ are listed in supplementary materials.

The pose tower is trained to learn the motion distribution $p_\theta(\bar{z}_i \mid Z_{\mathbf{M}}, T, P)$ conditioned on corrupted motion tokens $Z_{\mathbf{M}}$, text prompt T, and pose signal P.

$$\mathcal{L}_{\text{CE}}^{PT} = -\mathbb{E} \sum_{k \in \Omega} \log p_\theta(\bar{z}_i \mid Z_{\mathbf{M}}, T, P), \tag{4}$$

To further amplify the influence of pose signals, we extract the pose control signals from the generated dance sequence via Gumbel-Softmax sampling and minimize the discrepancy $D(P, \hat{P})$ between input pose signals P and those extracted from the output $\hat{P}$, i.e.,

$$D(P, \hat{P}) = \frac{\sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{J}} I(i,j) \|\hat{P}_{i,j} - P_{i,j}\|_2^2}{\sum_{i \in \mathcal{N}} \sum_{j \in \mathcal{J}} I(i,j)} \tag{5}$$

where $I(i,j)$ is a binary value indicating whether the pose signal P contains a valid value at frame i for joint j. The total

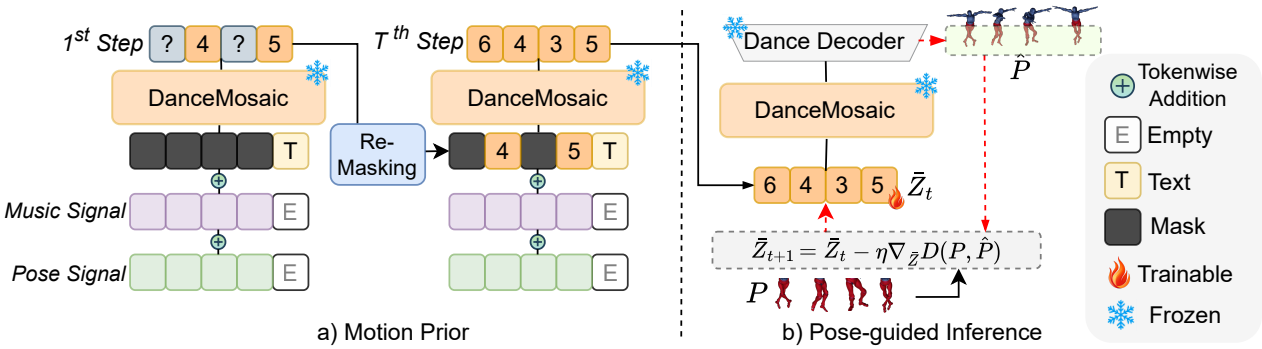

Figure 3: Overview of DanceMosaic's inference phase. (a) Motion Prior: It parallelly encodes music, pose, and text conditions, passing the conditions through respective towers, which generates guidance for each modality. (b) Pose-guided Inference: At the final stage, we utilize inference time computing to refine the generated pose to align closely, if pose conditions are provided.

loss function, $\mathcal{L}_{\text{PT}} = \mathcal{L}_{\text{CE}}^{PT} + \lambda_D \mathcal{L}_D$, is the weighted sum of CE loss and discrepancy loss $\mathcal{L}_D = \mathbb{E}[D(P, \hat{P})]$,

Progressive Training. Integrating music, text, and pose control signals into a dance generation model is challenging due to conflicting loss functions that create competing gradients during training. In particular, the kinematic terms in the music tower loss ($\mathcal{L}_{MT}$) aim to recover the full body skeleton joints, while the discrepancy term in the pose tower loss ($\mathcal{L}_{PT}$) focuses on restoring only the non-editable joints. To address this, we adopt multimodal progressive training, which incrementally incorporates modality-specific signals. This approach allows each modality of the model to be trained according to its respective loss function, mitigating gradient conflicts. The training process begins with training T2M model on text-to-motion dataset using textual prompts that describe a broad range of human actions. Next, we introduce the music tower, trained on dance-only datasets while keeping the T2M model frozen. This stage incorporates both music and genre prompts. At last, pose tower is trained with the frozen T2M using pose and genre prompts. Although the pose tower is not explicitly trained with music signal A, integrating the music and pose towers with the T2M allows the pose P and music A signals to simultaneously manipulate the prior motion distribution learned by the T2M.

3.4 Multimodal Guidance at Inference

Multimodal Classifier-free Logits Guidance. To enhance the influence of music, genre, and pose signals, we adopt Multimodal Classifier-free Logits Guidance, inspired by Classifier-Free Diffusion Guidance (Ho and Salimans 2022), within our masked dance model. As shown in Fig. 2, the final logits from token classifier are computed as a linear combination of unconditional logits and conditional logits informed by text, music, and pose inputs. These combined logits are transformed through a softmax function to yield the token distribution $p_\theta(\bar{z}_k \mid Z_{\text{M}}, T, A, C)$. During inference, motion tokens are iteratively generated using this learned distribution as a decoding confidence measure. Starting with a fully masked sequence of length L, the model performs T decoding iterations. At each iteration t, masked tokens are sampled from $p_\theta(\bar{z}_k \mid Z_{\text{M}}, T, A, C)$. Tokens with low confidence are remasked and resampled, with the number of remasked tokens n_M following a decaying schedule $n_M = L \cos\left(\frac{\pi}{2}\frac{t}{T}\right)$ (Chang et al. 2022b; Pinyoanuntapong et al. 2024b). This schedule applies a higher masking ratio in early iterations when confidence is low and gradually reduces it as more context is available, shown in Fig. 3(a).

Pose-guided Token Optimization. The sparse pose control signals have a subtler influence on motion distribution compared to textual and musical cues. Thus, the pose tower alone is not sufficient to accurately incorporate the editing constraints. To enhance motion quality and enforce dance motion alignment with sparse pose guidance, we further refine motion token embeddings $\bar{Z}$ via gradient descent to minimize the discrepancy function in Eq. (5):

$$\bar{Z}^+ = \arg\min_{\bar{Z}} D(P, \hat{P}) \tag{6}$$

where motion tokens are iteratively updated as,

$$\bar{Z}_{t+1} = \bar{Z}_t - \eta \nabla_{\bar{Z}} D(P, \hat{P}) \tag{7}$$

During inference-time optimization illustrated in Fig. 3(b), the network remains frozen, with only input motion tokens $\bar{Z}$ being trainable, thus accelerating the optimization process. Optimized input tokens $\bar{Z}^+$ pass through classifier to alter the logits and manipulate the sampling probabilities.

3.5 Applications

Spatial Body Part Editing. As shown in Fig. 4(b), DanceMosaic enables precise joint-specific control by progressively refining motion sequences through pose tower and inference-time optimization. It first generates a base motion sequence, then iteratively adjusts designated editable joints while preserving global coherence. By integrating music-aware and pose control signals, DanceMosaic achieves fine-grained upper- and lower-body refinements, adapting seamlessly to diverse music styles and text prompts for expressive, rhythmically aligned dance synthesis.

Temporal Body Part Editing. As shown in Fig. 4(c) and Fig. 4(a), DanceMosaic leverages its generative masking to interpolate missing motion segments while ensuring temporal consistency and rhythmic alignment with text-conditioned motion generation. By strategically placing

Methods	Motion Quality ↓		Foot Skating Ratio ↓	Motion Diversity ↑		BAS ↑	Run Time (s) ↓
	FID_k	FID_g	FSR	Div_k	Div_g		
FACT (Li et al. 2021)	113.38	97.05	28.44%	3.36	6.37	0.1831	9.46
MNET (Kim et al. 2022)	104.71	90.31	39.36%	3.12	6.14	0.1864	10.26
Bailando (Siyao et al. 2022)	82.81	28.17	18.76%	7.74	6.25	0.2029	1.46
EDGE (Tseng, Castellon, and Liu 2023)	94.34	50.38	20.04%	8.13	6.45	0.2116	2.27
LODGE (Li et al. 2024)	45.56	34.29	**5.01 %**	6.75	5.64	**0.2397**	8.16
DanceMosaic (ours)	**22.33**	**12.96**	**7.58%**	**8.54**	**8.36**	**0.2190**	**0.8**

Table 2: Performance comparison on FineDance dataset. Red bold indicates the best and blue underline the second-best results.

Figure 4: Various Applications Using DanceMosaic

[MASK] tokens, it reconstructs smooth transitions between keyframes using music and text prompts. Trained on diverse conditional masking patterns, DanceMosaic achieves natural, fluid temporal motion editing without additional supervision, maintaining spatial and musical coherence.

Long Dance Generation. DanceMosaic enables zero-shot long-form dance synthesis without the need for retraining. Given a sequence of music and text signals, it first generates individual motion segments. Then, as shown in Fig. 4(d), through masked token reconstruction, it synthesizes transitions by conditioning on adjacent sequences, ensuring rhythmically aligned continuity while preserving temporal coherence and expressive flow for extended performances.

Methods	Top-1↑	FID↓	MM↓	Div↑
Dance2Music (Lee et al. 2019)	0.033	66.98	8.12	0.73
TM2T (Guo et al. 2022a)	0.424	1.501	3.47	8.59
TM2D (Gong et al. 2023)	0.319	1.021	4.09	9.51
MoMask (Guo et al. 2023)	0.521	0.045	2.96	9.62
MMM(Pinyoanuntapong et al. 2024b)	0.515	0.089	**2.93**	9.58
DanceMosaic (our T2M)	**0.523**	**0.039**	3.04	**10.45**

Table 3: Comparison of text-conditional motion synthesis on HumanML3D (Guo et al. 2022b) test set

4 Experiments

Datasets. We utilized FineDance (Li et al. 2023), a 7.7 h corpus of 202 dance sequences across 16 genres, each paired with high-quality audio and detailed annotations including per-frame music features and genre labels. This rich, multimodal alignment makes FineDance an ideal benchmark for our multi-tower architecture and inference-time editing capabilities. All sequences are down-sampled to 20 fps, and we use the authors' original train/val/test splits. For pretraining, we leverage HumanML3D (Guo et al. 2022b), a text-driven motion dataset with 14,616 sequences and 44,970 captions, to learn a robust T2M backbone. We train a motion tokenizer on the union of HumanML3D and FineDance. The T2M backbone is pretrained on HumanML3D, while the Music and Pose Towers are trained exclusively on FineDance.

Evaluation Metrics. We evaluate DanceMosaic using kinematic and geometric Fréchet Inception Distances (FID_k, FID_g) to measure how closely our generated motions match real-world dynamics and pose geometry, and mean pairwise L_2 distances in the same feature spaces (Div_k, Div_g) to capture motion diversity. We further report Beat Alignment Score (BAS) for music-motion synchronization, Foot Skating Ratio (FSR) for physical plausibility, and average inference time per 9s clip to show real-time capability. For ***dance generation evaluation***, we adopt the evaluation protocol of LODGE (Li et al. 2024) and all results averaged from 20 runs. All experiments are done on single RTX A5000 GPUs.

4.1 Quantitative Comparison to State-of-the-art

Music-Dance Motion Generation. DanceMosaic achieves state-of-the-art results, shown in Table 2, on FineDance by combining a multi-tower generative masked model with synchronized progressive training and multimodal inference guidance. This design delivers real-time, editable 3D dance synthesis without sacrificing realism or musical alignment. In motion quality, kinematic FID_k falls from 45.56 to 22.33

Components			Motion Quality		
Motion Prior	Text	Music	$\text{FID}_k \downarrow$	$\text{Div}_k \uparrow$	BAS ↑
×	×	✓	45.56	6.75	0.2014
×	✓	✓	43.07	4.92	0.2122
✓	✓	×	58.66	3.31	0.193
✓	×	✓	34.49	7.55	0.2104
✓	✓	✓	**22.33**	**8.54**	**0.2190**

Table 4: Effectiveness of DanceMosaic Components

Music Guidance Scale (w_A)	$\text{FID}_k \downarrow$	$\text{Div}_k \uparrow$	BAS ↑
0	58.65	3.30	0.1972
0.2	56.23	4.15	0.2007
0.4	43.08	5.13	0.2048
0.6	30.87	6.53	0.2094
0.8	27.98	7.74	0.2105
1	**22.33**	**8.54**	**0.2190**
2	37.46	7.20	0.2301

Table 5: Impact of Music Guidance Scale

(-51%) and geometric FID_g from 34.29 to 12.96 (-62%), demonstrating a much closer match to real dance distributions. Geometric diversity (Div_g) rises from 6.45 to 8.36 (+30%), while kinematic diversity (Div_k) remains high at 8.54, reflecting richer, varied motion. Foot skating ratio (FSR) drops from 18.76% to 7.58%, indicating more physically plausible sequences. Beat alignment stays strong (BAS = 0.219), confirming that enhanced realism does not compromise synchronization. Finally, inference runs in 0.80s per 9s clip, over 2× faster than Bailando and 10× faster than LODGE, validating DanceMosaic's real-time efficiency.

Text-to-Motion Generation. To demonstrate generalization beyond music-driven dance generation, we evaluate our Text-to-Motion Backbone (T2M) on HumanML3D (Guo et al. 2022b) (Fig. 2(b)), and all results are averaged from 20 runs. Table 3 demonstrates that, with only the T2M component active, DanceMosaic achieves the highest Top-1 accuracy (19.57% relative gain) and the lowest FID among all baselines. It also delivers strong motion diversity (10.45) and multimodal distance (3.04), striking an excellent balance between realism, variety, and alignment to text prompts. These results confirm the T2M backbone's capability to generate high-quality motions from natural language descriptions.

5 Ablation Study

Effectiveness of Proposed DanceMosaic Components. Our ablation study (Table 4) demonstrates that each core component of DanceMosaic is essential to its overall performance. First, the pretrained Text-to-Motion (T2M) backbone serves as a powerful distribution prior: when we remove this backbone, kinematic FID_k jumps, indicating a severe loss of motion fidelity. Next, text guidance, implemented via the genre-prompt embedding, provides crucial semantic structure. Without it, Div_k superficially rises (8.55 versus 6.75), but the motions become erratic and unstructured, underscoring how genre cues anchor style and coherence. Finally, the Music Tower injects rhythmic control: disabling music guidance yields the worst FID_k (58.66) and collapses beat alignment, confirming that explicit audio conditioning is vital for synchronizing movement to music. These results validate our multi-tower masked architecture and synchronized training as the foundation for high-quality, semantically consistent dance synthesis.

Pose Tower	ITTO	$\text{FID}_k \downarrow$	$\text{FID}_g \uparrow$	BAS ↑	Joint Dist. ↓
✓	×	84.56	49.74	0.2263	0.037
×	✓	32.37	48.80	0.2127	0.005
✓	✓	**22.89**	**39.11**	**0.2277**	**0.004**

Table 6: Effect of Token Optimization and Pose Tower

Impact of Music Guidance Scale. Table 5 shows that gradually increasing the music guidance weight w_A (with text weight fixed) drives a consistent drop in kinematic FID_k and a rise in Beat Alignment Score, reaching an optimum at $w_A = 1$. Beyond this point, further emphasis on music yields diminishing returns, highlighting that our classifier-free logits guidance, by enabling precise, per-modality weighting, strikes the ideal balance between motion realism and rhythmic fidelity.

Effect of Token Optimization and Pose Tower. Table 6 shows that inference-time token optimization (ITTO) and Pose Tower each contribute uniquely to precise pose editing. ITTO alone cuts kinematic FID_k by 62% (84.56→32.37) and reduces Joint Distance by 86% (0.037→0.005), demonstrating its power to enforce sparse pose constraints via per-token refinement. The Pose Tower alone yields modest gains in geometric FID_g (49.74→39.11) and leaves joint errors high, indicating that architecture-level conditioning needs the fine-grained adjustments provided by ITTO. Together, they drive FID_k to 22.89 (-73%), boost BAS to 0.2277, and minimize Joint Distance to 0.004, confirming that our multi-tower masked architecture plus inference-time token optimization enables exact, semantically coherent edits without compromising overall motion quality or musical alignment.

6 Conclusion

In this work, we presented DanceMosaic, a novel method designed to address challenges in automatic, high-fidelity, and editable 3D dance motion generation. Our approach leverages a generative masked text-to-motion model as a distribution prior, facilitating the probabilistic mapping from various guidance signals, music, genre-specific text prompts, and spatial pose constraints, to high-quality dance motions. We developed a multi-tower masked motion architecture that incorporates a text-conditioned backbone with two modality-specific branches, music-guidance tower and pose-guidance tower. By employing synchronized and progressive masked training, we effectively integrated the pretrained text-to-motion prior while enabling each guidance branch to optimize independently and avoiding training conflicts. To further improve the alignment and realism of generated dances, we proposed classifier-free logits guidance and pose-guided inference-time token optimization techniques. Extensive evaluations confirmed that DanceMosaic achieves state-of-the-art performance, surpassing existing methods.

Acknowledgments

This work is supported by UNC Charlotte Art×Sci grant.

References

Ahuja, C.; and Morency, L.-P. 2019. Language2pose: Natural language grounded pose forecasting. In *2019 International Conference on 3D Vision (3DV)*, 719–728. IEEE.

Athanasiou, N.; Petrovich, M.; Black, M. J.; and Varol, G. 2022. Teach: Temporal action composition for 3d humans. In *2022 International Conference on 3D Vision (3DV)*, 414–423. IEEE.

Chang, H.; Zhang, H.; Jiang, L.; Liu, C.; and Freeman, W. T. 2022a. MaskGIT: Masked Generative Image Transformer. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11305–11315.

Chang, H.; Zhang, H.; Jiang, L.; Liu, C.; and Freeman, W. T. 2022b. MaskGIT: Masked Generative Image Transformer. arXiv:2202.04200.

Dhariwal, P.; Jun, H.; Payne, C.; Kim, J. W.; Radford, A.; and Sutskever, I. 2020. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*.

Esser, P.; Rombach, R.; and Ommer, B. 2020. Taming Transformers for High-Resolution Image Synthesis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12868–12878.

Ghosh, A.; Cheema, N.; Oguz, C.; Theobalt, C.; and Slusallek, P. 2021. Synthesis of compositional animations from textual descriptions. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1396–1406.

Gong, K.; Lian, D.; Chang, H.; Guo, C.; Jiang, Z.; Zuo, X.; Mi, M. B.; and Wang, X. 2023. TM2D: Bimodality Driven 3D Dance Generation via Music-Text Integration. arXiv:2304.02419.

Guo, C.; Mu, Y.; Javed, M. G.; Wang, S.; and Cheng, L. 2023. MoMask: Generative Masked Modeling of 3D Human Motions. *ArXiv*.

Guo, C.; Xuo, X.; Wang, S.; and Cheng, L. 2022a. TM2T: Stochastic and Tokenized Modeling for the Reciprocal Generation of 3D Human Motions and Texts. *ArXiv*, abs/2207.01696.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022b. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5152–5161.

Guo, C.; Zuo, X.; Wang, S.; Zou, S.; Sun, Q.; Deng, A.; Gong, M.; and Cheng, L. 2020. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2021–2029.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Holden, D.; Kanoun, O.; Perepichka, M.; and Popa, T. 2020. Learned motion matching. *ACM Transactions on Graphics (TOG)*, 39(4): 53–1.

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

Jiang, B.; Chen, X.; Liu, W.; Yu, J.; Yu, G.; and Chen, T. 2023. MotionGPT: Human Motion as a Foreign Language. *ArXiv*, abs/2306.14795.

Kim, J.; Kim, J.; and Choi, S. 2022. FLAME: Free-form Language-based Motion Synthesis & Editing. In *AAAI Conference on Artificial Intelligence*.

Kim, J.; Oh, H.; Kim, S.; Tong, H.; and Lee, S. 2022. A brand new dance partner: Music-conditioned pluralistic dancing controlled by multiple dance genres. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3490–3500.

Kong, H.; Gong, K.; Lian, D.; Mi, M. B.; and Wang, X. 2023. Priority-centric human motion generation in discrete latent space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14806–14816.

Lee, H.-Y.; Yang, X.; Liu, M.-Y.; Wang, T.-C.; Lu, Y.-D.; Yang, M.-H.; and Kautz, J. 2019. Dancing to Music. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Li, J.; Yin, Y.; Chu, H.; Zhou, Y.; Wang, T.; Fidler, S.; and Li, H. 2020a. Learning to generate diverse dance motions with transformer. *arXiv preprint arXiv:2008.08171*.

Li, R.; Yang, S.; Ross, D. A.; and Kanazawa, A. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13401–13412.

Li, R.; Zhang, Y.; Zhang, Y.; Zhang, H.; Guo, J.; Zhang, Y.; Liu, Y.; and Li, X. 2024. Lodge: A coarse to fine diffusion network for long dance generation guided by the characteristic dance primitives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1524–1534.

Li, R.; Zhao, J.; Zhang, Y.; Su, M.; Ren, Z.; Zhang, H.; Tang, Y.; and Li, X. 2023. Finedance: A fine-grained choreography dataset for 3d full body dance generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10234–10243.

Li, Y.; Liu, J.; Lin, G.; Hou, Y.; Mou, M.; and Zhang, J. 2020b. Gumbel-softmax-based Optimization: A Simple General Framework for Optimization Problems on Graphs. arXiv:2004.07300.

Petrovich, M.; Black, M. J.; and Varol, G. 2021. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10985–10995.

Petrovich, M.; Black, M. J.; and Varol, G. 2022. TEMOS: Generating diverse human motions from textual descriptions. *ArXiv*, abs/2204.14109.

Pinyoanuntapong, E.; Saleem, M. U.; Wang, P.; Lee, M.; Das, S.; and Chen, C. 2024a. BAMM: Bidirectional Autoregressive Motion Model. In *Computer Vision – ECCV 2024*.

Pinyoanuntapong, E.; Wang, P.; Lee, M.; and Chen, C. 2024b. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1546–1555.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

Rempe, D.; Birdal, T.; Hertzmann, A.; Yang, J.; Sridhar, S.; and Guibas, L. J. 2021. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11488–11499.

Siyao, L.; Yu, W.; Gu, T.; Lin, C.; Wang, Q.; Qian, C.; Loy, C. C.; and Liu, Z. 2022. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11050–11059.

Tseng, J.; Castellon, R.; and Liu, K. 2023. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 448–458.

Yamada, T.; Matsunaga, H.; and Ogata, T. 2018. Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions. *IEEE Robotics and Automation Letters*, 3(4): 3441–3448.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2022. MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model. *ArXiv*, abs/2208.15001.

Zhou, Y.; Barnes, C.; Lu, J.; Yang, J.; and Li, H. 2019. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5745–5753.

Zhuang, W.; Wang, C.; Chai, J.; Wang, Y.; Shao, M.; and Xia, S. 2022. Music2dance: Dancenet for music-driven dance generation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2): 1–21.