# LM2D: Lyrics- and Music-Driven Dance Synthesis

Wenjie Yin<sup>1\*</sup>, Xuejiao Zhao<sup>1,2\*</sup>, Yi Yu<sup>3†</sup>, Hang Yin<sup>4</sup>, Danica Kragic<sup>1</sup>, Mårten Björkman<sup>1†</sup>

<sup>1</sup>KTH Royal Institute of Technology, <sup>2</sup>Nanyang Technological University, <sup>3</sup>National Institute of Informatics, <sup>4</sup>University of Copenhagen

### Abstract

Dance typically involves professional choreography with complex movements that follow a musical rhythm and can also be influenced by lyrical content. The integration of lyrics in addition to the auditory dimension, enriches the foundational tone and makes motion generation more amenable to its semantic meanings. However, existing dance synthesis methods tend to model motions only conditioned on audio signals. In this work, we make two contributions to bridge this gap. First, we propose LM2D, a novel probabilistic architecture that incorporates a multimodal diffusion model with consistency distillation, designed to create dance conditioned on both music and lyrics in one diffusion generation step. Second, we introduce the first 3D dance-motion dataset that encompasses both music and lyrics, obtained with pose estimation technologies. We evaluate our model against music-only baseline models with objective metrics and human evaluations, including dancers and choreographers. The results demonstrate LM2D is able to produce realistic and diverse dance matching both lyrics and music. A video summary can be accessed at: https://youtu.be/4XCgvYookvA.

### 1 Introduction

Dance is an engaging form of human expression that intertwines body movements with music, playing a crucial role in various cultures [LaMothe, 2019]. In the modern digital age, dance content enjoys enormous popularity on platforms like *YouTube* and *TikTok*, and even in video games such as *Just Dance* and *Dance Central*. However, creating dance, whether through traditional means or digitally, is a complex and challenging task. Professional dance involves expert choreography and extensive practice, often requires advanced motion capture technology for digitization. Consequently, the development of automated human motion generation technologies presents significant potential and possibilities across various

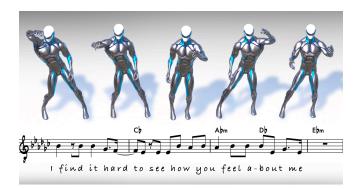


Figure 1: LM2D, a multimodal framework generates realistic and diverse dance movements conditioned on both lyrics and music.

digital platforms and the field of choreography. Such technology could pave the way for innovative collaborations between human creativity and artificial intelligence.

Recent breakthroughs in generative modeling, notably in the areas of normalizing flows [Papamakarios *et al.*, 2021] and diffusion models [Ho *et al.*, 2020], have significantly enhanced the capabilities of automated dance generation. Such advancements not only enrich the artistic dimension of choreography but also provide valuable insights for dance research [Valle-Pérez *et al.*, 2021; Alexanderson *et al.*, 2023; Li *et al.*, 2021; Yin *et al.*, 2023].

Nevertheless, existing technologies in automated dance generation primarily focus on the interaction between music and dance movements, often neglecting the significant role of lyrics in choreography. While music-conditioned models can already produce realistic rhythmically-aligned dance movements, incorporating lyrics can add depth and enrich semantic meaning, as there is a notable connection between dance motion and song lyrics in styles like modern dance [Powell, 2019]. Further exploration is needed to integrate both lyrics and music in dance synthesis. Additionally, the success of diffusion models and normalizing flows largely hinges on their iterative processes, which gradually build a sample from random noise. However, this step-by-step refinement, involving sequential steps of large neural network evaluations, results in slower sampling speeds compared to single-step methods like GANs or VAEs. This inefficiency poses a challenge for real-time applications, such as real-time dynamic

<sup>\*</sup>Equal Contributions

<sup>&</sup>lt;sup>†</sup>Corresponding Authors

choreography generation, emphasizing the need for efficient, single-step generation methods [Song *et al.*, 2023; Lu *et al.*, 2022]. Imagine a system capable of refreshing generated images at 5Hz in response to prompts editing in the context of image generation [Sauer *et al.*, 2023]. Such a real-time system in the context of dance choreography could facilitate instantly generating dance movements in response to changes in music and lyrics, enabling a highly interactive and responsive creative process.

In response to these challenges, our study presents LM2D, a novel probabilistic architecture that combines a multimodal diffusion model with consistency distillation. As illustrated in Figure 1, this design aims to generate dance that is conditioned on both music and lyrics in a single diffusion step, addressing the limitations of existing models and advancing the field of automated dance generation. With consistency distillation, we could effectively distill a pre-trained diffusion model into a consistency model, which allows synthesizing dance with one evaluation step. Another major challenge in learning-based motion synthesis is the availability of largescale datasets for 3D movement. Existing dance datasets focus primarily on music and body motion, but lack lyric information. To bridge this gap, we apply pose estimation technologies [Dong et al., 2020] on Just Dance videos to collect a new multimodal dataset that includes synchronized dance motion, music, and lyrics, providing a more comprehensive resource for research in this area. We evaluate our framework on the new dataset with the analysis focused on motion quality. A new metric is proposed to evaluate the match between motion and lyrics quantitatively. Additionally, we invite a group of human participants with extensive dancing and choreography experience to provide a subjective evaluation and insights from an expert perspective. These evaluations show that movements synthesized by LM2D are realistic and match both lyrics and music. In summary, in this paper:

- We contribute data-driven methods for lyrics- and music-driven dance synthesis. Our multimodal diffusion model with consistency distillation is able to create dance in a single diffusion step.
- We make a new dance dataset available. To the best of our knowledge, this is the first such dataset to contain synchronized lyrics, music, and high-quality 3D motion.
- We evaluate our new model both objectively and through a user study with skilled dancers and choreographers, focusing on the quality of motion and the alignments with lyrics and music.

# 2 Related Work

## 2.1 Data-Driven Dance Synthesis

Dance synthesis, the problem of automatically creating realistic and natural human motions, is complex and challenging. Early research employed motion retrieval methods, creating choreography by transitioning between pre-existing motion clips [Fan *et al.*, 2011; Lee *et al.*, 2013; Fukayama and Goto, 2015]. These methods, which generate motion by selection, often led to unnatural transitions and had limited variability. With the advent of deep learning, [Ye *et al.*, 2020; Chen *et al.*, 2021] have integrated deep learning techniques with motion graphs to produce higher-quality choreography. Subsequent research trained on large datasets and explored various modeling approaches, including generative adversarial networks (GANs), recurrent neural networks (RNNs), transformers, and normalizing flows [Fan et al., 2022; Li et al., 2022a, 2021; Yin et al., 2023; Siyao et al., 2022; Valle-Pérez et al., 2021]. Recent breakthroughs with diffusion models [Ho et al., 2020] have further advanced this field. [Alexanderson et al., 2023] pioneered the use of diffusion models with Conformer [Zhang et al., 2022] for generating dance from music. EDGE [Tseng et al., 2023] and Magic [Li et al., 2022b] utilized Transformer-based diffusion models. However, these methods primarily focus on music-driven synthesis, largely neglecting the impact of lyrics, which usually contain rich semantic information. A recent study by [Deichler et al., 2023] has ventured into using joint text and audio representation for human gesture generation. In our research, we introduce LM2D, a model that generates dance movements conditioned by both lyrics and music using diffusion models. Additionally, diffusion models typically exhibit slower sampling speeds compared to single-step methods such as GANs or VAEs. To overcome this limitation, we incorporate consistency distillation to accelerate the inference process to a single diffusion step.

# 2.2 Dance Datasets

3D dance datasets are essential for data-driven dance synthesis, which requires professional experience and usually contains multi-modal information. Motion capture techniques are widely adopted for collecting 3D motion data. The first notable 3D dance dataset was released by [Alemi et al., 2017] with synchronized music. Following this, [Zhuang et al., 2022] collected Music2Dance, a synchronized music and motion capture dataset. Recently, [Valle-Pérez et al., 2021] introduced three new datasets, the PMSD, Syrto, and Shader-Motion VR dance datasets, each focusing on different dance styles. Among these, the ShaderMotion VR dance dataset incorporates VR technology for collection. [Alexanderson et al., 2023] combined various data sources to create a highquality dataset. DanceFormer [Li et al., 2022a] presented the PhantomDance dataset, a unique collection created by professional animators, while ChoreoMaster [Chen et al., 2021] combined motion capture and anime resources in their dataset. Advancements in 3D reconstruction have enabled the conversion of 2D videos into 3D dance data, which allows for the acquisition of 3D skeletal data from a vast amount of videos, at a lower cost compared to motion capture or animation techniques. [Li et al., 2021] utilized the AIST dance dataset [Tsuchida et al., 2019] to obtain 3D dance motion, employing multi-view 3D pose estimation techniques, and [Li et al., 2020] extracted 3D pose sequences with synchronized audios using the VideoPose3d [Pavllo et al., 2019]. [Le et al., 2023] and [Wang et al., 2022] introduce group dance datasets for more challenging group choreography. However, to the best of our knowledge, there is no dataset available that synchronizes music, lyrics, and body motion. Existing human motion datasets typically include either music with motion or text with body gestures. To bridge this gap, we have

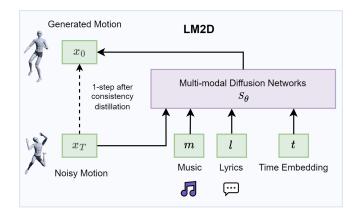


Figure 2: Overview of the LM2D framework. LM2D learns to denoise dance sequences from time t = T to t = 0, condition on music and lyrics in one step with consistency distillation.

constructed a new dataset using 3D pose estimation, captureing dance sequences synchronized with both music and lyrics from JustDance video games.

## 3 Method

Our approach uses a multi-modal neural network to produce a series of human poses, x, based on a corresponding sequence of music features m and lyrics features l, as illustrated in Figure 2. In this section, we delve into the mathematical properties of diffusion models and the architecture for generating motion driven by both lyrics and music. Subsequently, we introduce the concept of consistency distillation for the purpose of facilitating one-step generation.

#### **3.1** Multi-modal Diffusion Framework

To tackle the problem described above, we follow the theory of continuous-time diffusion models [Song *et al.*, 2020]. Diffusion models produce samples by gradually transforming data into noise through Gaussian disturbances and then generating samples from noise through a series of denoising steps. In the continuous-time diffusion model, the diffusion process  $\{x(t), t \in [0, 1]\}$  can be defined as a forward Stochastic Differential Equation (SDE) [Song *et al.*, 2020]:

$$d\boldsymbol{x}_t = \boldsymbol{f}(\boldsymbol{x}_t, t)dt + g(t)d\boldsymbol{w}_t, \qquad (1)$$

where  $w_t$  represents the standard Wiener process operates in reverse-time,  $f(x_t, t)$  denotes the drift term, and g(t) is the scalar diffusion coefficient. The reverse-time SDE specifies the reverse process of the above forward process as follows:

$$d\boldsymbol{x}_t = \left[\boldsymbol{f}(\boldsymbol{x}_t, t) - g(t)^2 \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)\right] dt + g(t) d\boldsymbol{w}_t, \quad (2)$$

where  $\nabla_{x_t} \log p_t(x_t)$  represents the score function associated with the data distribution perturbed by noise at time t. For the reverse-time SDE, there exists a specific ordinary differential equation (ODE) known as the Probability Flow ODE (PF-ODE) [Song and Ermon, 2019; Song *et al.*, 2020]. The solutions of the PF-ODE sampled at time t are distributed according to  $p_t(x_t)$ :

$$\frac{d\boldsymbol{x}_t}{dt} = \boldsymbol{f}(\boldsymbol{x}_t, t) - \frac{1}{2}g(t)^2 \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t), \quad (3)$$

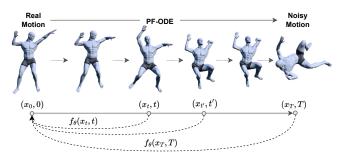


Figure 3: Overview of the consistency models. Given a PF-ODE that smoothly converts real human motion to noisy motion, we learn to map any points on the trajectory to its origin point.

In the training phase, a score model  $s_{\theta}(\boldsymbol{x}_t, t)$  is employed to approximate the term  $-\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)$  through score matching [Song and Ermon, 2019; Song *et al.*, 2020], which leads to the following ODE for sampling:

$$\frac{d\boldsymbol{x}_t}{dt} = \boldsymbol{f}(\boldsymbol{x}_t, t) + \frac{g(t)^2}{2} \boldsymbol{s}_{\theta}(\boldsymbol{x}_t, t).$$
(4)

We initialize the PF-ODE by sampling  $z_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  and then solve it in reverse time using a numerical ODE solver. The obtained  $\hat{x}_0$  can be regarded as an estimated sample of the original data distribution. The solver is typically stopped at  $t = \epsilon$  to avoid numerical instability issues, where  $\epsilon$  is a predetermined small positive value.

Our diffusion framework is developed based on the EDGE architecture [Tseng *et al.*, 2023], designed for producing human motion sequences given conditional signals. This framework employs a transformer-based network with cross-attention mechanisms [Li *et al.*, 2021] that accepts music m, lyrics l features, latents  $z_t$  that follow a forward noising process to estimate  $\hat{x}_{\theta}$ . Subsequently, it synthesizes motion sequences that align with these features, as illustrated in Figure 2. The objective function is simplified as:

$$\mathcal{L}_{rec} = \mathbb{E}_{\boldsymbol{x},t} \left\| \boldsymbol{x} - \hat{\boldsymbol{x}}_{\theta}(\boldsymbol{z}_t, t, \boldsymbol{m}, \boldsymbol{l}) \right\|_2^2.$$
 (5)

In addition to the above reconstruction loss, we also adopt geometric losses as [Tang *et al.*, 2022; Tevet *et al.*, 2022; Tseng *et al.*, 2023] to improve physical realism with joint positions and velocities.

$$\mathcal{L}_{pos} = \frac{1}{N} \sum_{i=1}^{N} \|FK(\boldsymbol{x}^{i} - FK(\hat{\boldsymbol{x}}^{i})\|_{2}^{2}, \quad (6)$$

$$\mathcal{L}_{vel} = \frac{1}{N-1} \sum_{i=1}^{N-1} \| (\boldsymbol{x}^{i+1} - \boldsymbol{x}^i) - (\hat{\boldsymbol{x}}^{i+1} - \hat{\boldsymbol{x}}^i) \|_2^2, \quad (7)$$

where  $FK(\cdot)$  is the forward kinematic function, N is the number of frames in the synthesized sequences. The overall training loss is a weighted sum of the reconstruction loss and geometric losses:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda_{pos} \mathcal{L}_{pos} + \lambda_{vel} \mathcal{L}_{vel}.$$
(8)

#### 3.2 Consistency Distillation

The performance of diffusion models is constrained by their slow sampling process. As the ODE solvers for sampling necessitate multiple iterations of the score model  $s_{\theta}$ , ranging from dozens to thousands of steps, leading to significant computational demands.

To accelerate the diffusion sampling, the consistency models [Song *et al.*, 2023] are introduced. The consistency model is a new family of generative models capable of executing single-step generation while preserving generation quality. These models can be trained either by distilling pre-existing diffusion models or as independent generative models. In our research, we apply the method of consistency distillation (CD) to achieve one-step generation. For the sake of brevity, we have excluded the conditional music and lyrics features in this section. The key idea of CD focuses on learning the function that maps every point on a PF-ODE trajectory to its origin point. More formally, the consistency function is defined as  $f_{\theta} : (x_t, t) \to x_{\epsilon}$ , and the consistency function should satisfy the self-consistency property:

$$\boldsymbol{f}_{\theta}(\boldsymbol{x}_{t},t) = \boldsymbol{f}_{\theta}(\boldsymbol{x}_{t'},t'), \forall t,t' \in [\epsilon,T],$$
(9)

which can be parameterized using skip connections:

$$\boldsymbol{f}_{\theta}(\boldsymbol{x}_t, t) = c_{\text{skip}}(t)\boldsymbol{x} + c_{\text{out}}(t)\boldsymbol{S}_{\theta}(\boldsymbol{x}, t), \quad (10)$$

where  $c_{\text{skip}}(t)$  and  $c_{\text{out}}(t)$  are differentiable functions with  $c_{\text{skip}}(\epsilon) = 1$  and  $c_{\text{out}}(\epsilon) = 0$ , and  $S_{\theta}(\boldsymbol{x}, t)$  is the neural network. The consistency distillation loss is defined as

$$\mathcal{L}(\theta, \theta^{-}; \Phi) = \mathbb{E}_{\boldsymbol{x}, t_{n}} \left[ d \left( \boldsymbol{f}_{\theta}(\boldsymbol{x}_{t_{n+1}}, t_{n+1}), \boldsymbol{f}_{\theta^{-}}(\hat{\boldsymbol{x}}_{t_{n}}^{\boldsymbol{\phi}}, t_{n}) \right) \right],$$
(11)

here the time horizon  $[\epsilon, T]$  is discretized into n = N - 1 subintervals.  $d(\cdot, \cdot)$  is the  $\ell_2$  metric for measuring the distance between samples,  $\theta^-$  is the parameters of the target model, updated with the exponential moving average (EMA) of the online model  $\theta$ , i.e.,  $\theta^- \leftarrow \mu \theta^- + (1 - \mu)\theta$ ,  $\mu$  is the decay rate.  $\Phi$  is the ODE solver applied to the PF-ODE,  $\hat{x}_t^{\phi}$  is a one-step estimation of  $x_t$  from  $x_{t+1}$  as:

$$\hat{x}_{t_n}^{\phi} \leftarrow x_{t_{n+1}} + (t_n - t_{n+1}) \Phi(x_{t_{n+1}}, t_{n+1}; \phi).$$
 (12)

We employ consistent distillation to learn consistency models from the pre-trained multi-modal diffusion models, achieving one-step dance generation given music and lyrics features. The algorithm of consistency distillation for the pre-trained multi-modal LM2D is depicted in Algorithm 1. For more details, please refer to [Song *et al.*, 2023].

### 4 Dataset

## 4.1 Data Collection

We create a new dance dataset due to the lack of datasets simultaneously containing dance motion, music, and lyrics. The proposed 3D motion dataset is derived from existing Just Dance videos by Ubisoft, a motion-based rhythm dancing game with annual releases, and has been a well-known classic in video games. Just Dance features a diverse range of dance styles including pop, hip-hop, Latin, classical, and

#### Algorithm 1 Consistency Distillation (CD) for LM2D

**Input:** pre-trained model parameter  $\theta$ , ODE solver  $\Phi$ , learning rate  $\eta$ , and training data pairs of motion x, music m, and lyrics l.  $\theta^- \leftarrow \theta$ 

# **Repeat:**

$$\begin{split} \hat{\boldsymbol{x}}_{t_n}^{\boldsymbol{\phi}} \leftarrow \boldsymbol{x}_{t_{n+1}} + (t_n - t_{n+1}) \Phi(\boldsymbol{x}_{t_{n+1}}, m, l, t_{n+1}; \phi) \\ \mathcal{L}(\theta, \theta^-; \Phi) \leftarrow \\ d(\boldsymbol{f}_{\theta}(\boldsymbol{x}_{t_{n+1}}, m, l, t_{n+1}), \boldsymbol{f}_{\theta^-}(\hat{\boldsymbol{x}}_{t_n}^{\boldsymbol{\phi}}, m, l, t_n)) \\ \theta \leftarrow \theta - \eta \nabla_{\theta} \mathcal{L}(\theta, \theta^-; \Phi) \\ \theta^- \leftarrow \mu \theta^- + (1 - \mu) \theta \\ \end{split}$$
  
Until convergence

electronic dance music. The game involves players replicating the movements of an on-screen dancer. Utilizing Easy-Mocap [Dong *et al.*, 2020], we extracted 3D human motion data in terms of SMPL parameters from these videos. This method allowed us to achieve high-fidelity body estimations at 60 fps. Additionally, we compiled a multimodal dataset by sourcing music and lyrics. The lyrics for each song were manually gathered and synchronized with the corresponding musical timeline. In summary, this dataset encompasses 4.6 hours of 3D dance motion in 1867 sequences, accompanied by music and lyrics. We will publicly release the data.

### 4.2 Data Preparation

We represent dance as sequences of poses using the 24-joint SMPL format [Loper *et al.*, 2015], using a 6-DOF rotation representation [Zhou *et al.*, 2019], resulting in a 147-dimensional feature. We follow the audio representation as in [Valle-Pérez *et al.*, 2021; Tsuchida *et al.*, 2019] that combines spectrogram features with beat-related features. The music features are extracted by Librosa [McFee *et al.*, 2015], yielding a 35-dimensional feature. Specifically, this encompasses a combination of 20-dimensional Mel-frequency cepstral coefficients (MFCC), 12-dimensional chroma, and 1-dimensional one-hot peaks and beats. Lyrics were processed and embedded into a pre-trained BERT embedding [Devlin *et al.*, 2018], resulting in a 768-dimensional feature.

### 4.3 Data Validity

The task of synthesizing dance from both music and lyrics could benefit from several considerations: 1) Semantic Influence: There is a connection between the semantics of motion and lyrics. For example, the word *No* is naturally associated with gestures like shaking one's head or hands; 2) Emotional Influence: Lyrics can enhance the emotion the music conveys. For instance, phrases like *Break my heart* might be interpreted through quieter and sadder movements; 3) Rhythmic Pattern Influence: The patterns in rhythms and lyrics are often related. Repetitive lyrical patterns such as *Oh oh oh* or *Ma-ma-ma* might suggest a sequence of recurring dance motions; 4) Music-lyrics Influence: Different segments of a song may have identical music but distinct lyrics. Incorporating lyrics into the choreography can help avoid producing repetitive movements in similar musical segments. While these

Method	$FID_k\downarrow$	$FID_g\downarrow$	$Div_k \rightarrow$	$Div_g \rightarrow$	BA↑	SM↑
Ground Truth	-	-	11.34	7.47	0.24	0.85
EDGE	12.81	9.25	12.33	7.18	0.25	0.81
EDGE(cd)	11.40	8.74	12.43	6.83	0.22	0.80
LM2D	12.35	9.76	12.37	7.53	0.25	0.83
LM2D(cd)	12.1	10.35	12.58	7.37	0.24	0.83

Table 1: Quantitative objective evaluation: FID score and diversity metrics based on geometric and kinetic features are computed. Besides, beat alignment score (BA) and semantic matching score (SM) are calculated for evaluating alignments with motions.

considerations might not always apply, they are not uncommon in various choreographies.

# **5** Experiments

### 5.1 Experimental Setting

To demonstrate the capabilities of our proposed LM2D, we compare it to baseline models and ground truth on the introduced new dataset. To evaluate the impact of lyrics in dance motion synthesis, we compare our model with EDGE, which is only conditioned on music information. Additionally, we compare the trained diffusion models with the one-step model with consistency distillation to explore the performance of consistency distillation on motion generation. To evaluate the performance, we quantitatively assessed (1) FID scores and diversity; (2) beat alignment scores; and (3) our proposed semantic match metric. The objective evaluation results are shown in Table 1 and discussed in Section 5.2. Qualitative human evaluations are also important in the evaluation of generative models as well. We performed a user study in the form of an online survey to evaluate human-perceived quality in terms of (1) motion naturalness; (2) motion-music alignment; and (3) motion-lyrics matching. The subjective evaluation results are discussed in Section 5.3.

# 5.2 Objective Evaluation

FID Scores. In our study, we initially considered the Fréchet Inception Distance (FID) score on geometric and kinetic features. The FID score is a widely recognized objective metric for evaluating generative models and has been extensively used in previous research [Li et al., 2021, 2020]. This metric quantifies the divergence between the distributions of real and synthetic data. However, recent work, [Tseng et al., 2023] has pointed out that the FID score may not be entirely reliable for tasks involving dance generation. They observed that the FID results can contradict human evaluations, possibly due to subjective judgments. In our experiments, as shown in Table 1, the EDGE model showed improved results after consistency distillation, but this improvement was not reflected in the human evaluations. Thus, although we include the results of FID score, the utility of FID for assessment can be questioned.

**Diversity Metrics.** Diversity metrics in our study were calculated based on the distributional spread of geometric and kinetic features. In line with the methodologies employed in prior research [Li *et al.*, 2021; Tseng *et al.*, 2023], our model



Figure 4: LM2D Example: Two dance sequences are generated from the **same music** but with **different lyrics**.

aims to match this metric with those of the ground truth distribution. As indicated in Table 1, the diversity metrics among the various models are quite similar, suggesting that the distributions generated by each model are closely aligned with the real distribution. Specifically, the kinetic features of EDGE are the closest to the ground truth, while the geometric features of LM2D most closely match the ground truth. However, it was observed that after applying consistency distillation to accelerate the model, the diversity gap between the synthetic and real distributions increased.

**Beat Alignment Scores.** Our experiments included an evaluation of how well our generated dances aligned with the beat of the music, as in previous work by [Siyao *et al.*, 2022] and [Tseng *et al.*, 2023]. The findings indicated that both EDGE and LM2D models achieved comparable levels of performance in terms of beat alignment scores. However, it was also observed that the application of consistency distillation led to a decrease in this score.

Semantic Matching Scores Our task extends beyond musical elements to include lyrical content, necessitating an evaluation of both movement-music alignment and movementlyric correspondence. To measure the semantic matching between movements and lyrics, we utilized the pre-trained BERT model [Devlin et al., 2018] to obtain BERT embeddings for the lyrics. For the semantic features of the movements, we followed the implementation of MotionBert [Zhu et al., 2023], using our multimodal dataset to train a motion encoder that captures motion embeddings. During the training process, we kept the parameters of the pre-trained BERT model fixed and trained the motion encoder to better align the features of lyrics and movements in our dataset. We evaluated the semantic matching by calculating the cosine similarity between the motion and lyric embeddings. Observing the experimental results, the LM2D model, which incorporates lyric information during training, achieved higher Semantic matching scores compared to the EDGE model, which does not include lyric information. This demonstrates the effectiveness of incorporating lyrics into the training process for enhanced semantic matching in dance generation.

Figure 4 and Figure 5 present examples of synthesized motion clips from LM2D, which demonstrate the combined influence of lyrics and music on dance motion synthesis. Figure 4 showcases examples where two dance sequences were generated from the same musical piece, "*Think About Things*," but with differing lyrics. The upper sequence was generated with the lyrics "*I wouldn't stop movin' if I could Break it* 

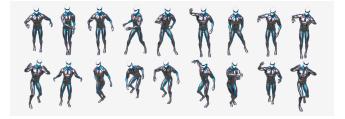


Figure 5: LM2D Example: Two dance sequences are generated from the **same lyrics** but with **different music**.

down!" while the lower sequence used "Just call me For delivery. They know me. I got sushi." Despite the identical musical input, the distinct lyrics led to different movement patterns in each sequence. Additionally, the movements are, to some extent, semantically aligned with the lyrics. In Figure 5, we present dance sequences generated using the same lyrics but different music segments. The lyrics used were "Every time I'm around there's a hype! Touchdown and the crowd get's hyped." The upper music segment is from "Bad Guy," while the lower one is from "Hype." It's noticeable that, due to the consistency of the lyrics, there are similarities in the movement patterns between the sequences, particularly in the positions of the limbs and torso. However, the differences in music necessitate that the movements also align with the musical beats. This approach allows us to generate more diverse and meaningful dance movements.

### 5.3 Human Evaluations.

To gain a deeper understanding of our method, we perform human evaluations alongside the objective evaluation. In this study, participants were asked to assess three key aspects: motion naturalness, motion-music alignment, and motionlyrics matching. Our study recruited 30 participants with dance experience, including training, performing, teaching, or even choreography. The average dance experience among these participants was 6.72 years. Notably, 7 of the participants (23% of the group) had experience in choreography. We conducted an online survey to collect feedback from these participants to evaluate the task of synthesizing dance driven by lyrics and music. We blended 6-second video clips representing the ground truth and created motion sequences. Participants were shown these dance video clips, followed by a series of questions. Participants were instructed to rate their level of agreement with the statements presented in the questions, using a 1-5 Likert scale, where "1" indicates strong disagreement, and "5" indicates strong agreement. To mitigate potential order effects, the sequence in which the dance clips were presented was randomized and balanced across participants.

In the first part of the survey, we focused on evaluating two aspects: the naturalness of the motion and the alignment between motion and music. The dance clips provided for evaluation were generated using one of several methods: LM2D, EDGE, LM2D-cd, EDGE-cd, or they were sourced from Ground Truth (GT). Participants had the option to view these clips multiple times, enabling a thorough assessment before responding to the following two questions:

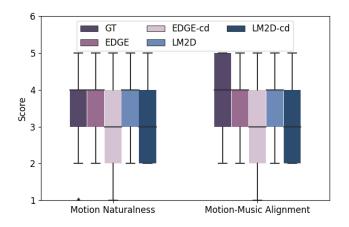


Figure 6: Subjective evaluation results in motion naturalness and motion-music alignment.

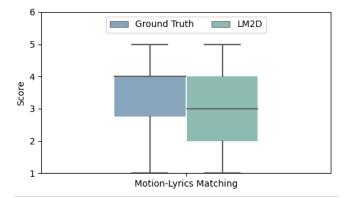


Figure 7: Subjective evaluation results in motion-lyrics matching.

- Motion Naturalness: To what extent do you agree with the following statement? — The movements are natural and of good quality.
- **Motion-Music Alignment**: To what extent do you agree with the following statement? — The movements match well with the beat of the music.

In the second part of the survey, our evaluation centered on the matching between motion and lyrics. For this part, each dance clip was either generated using the LM2D method or taken from Ground Truth. Participants need to respond to the following question after viewing the clips:

• Motion-Lyrics Matching: To what extent do you agree with the following statement? — The movements match well with the lyrics.

We performed a statistical analysis of the subjective responses from the user study to support our findings and evaluated whether our proposed method could be further enhanced.

Figure 6 illustrates the evaluation of motion naturalness and motion-music alignment. From the figure, we observe that participants considered the performance of EDGE and LM2D to be similar to the Ground Truth (GT). However, models enhanced with consistency distillation (cd) showed a slight decline in performance, aligning with objective findings. We employed Analysis of Variance (ANOVA) to test statistically significant differences among the groups.

For motion naturalness, ANOVA results showed no significant difference among GT, EDGE, and LM2D (F =0.0191, p = 0.981). The same was observed for motionmusic alignment (F = 0.1134, p = 0.8929). These results demonstrate that our model, even after incorporating lyric content, maintains the motion naturalness and the alignment with the music's rhythm while incorporating semantic content with the movements. In contrast, for GT, EDGEcd, and LM2D-cd, significant differences were observed for motion naturalness (F = 5.2529, p = 0.0061) and motionmusic alignment  $(F = 12.1948, p = 1.0923 \times 10^{-5})$ , as per ANOVA. This suggests that while consistency distillation enhances efficiency, it adversely affects performance, a finding echoed by human evaluations and indicating a need for further improvement. Figure 7 focuses on comparing the LM2D model with ground truth regarding motion-lyrics matching. Dance experts favored the ground truth. We applied the two one-sided tests to assess if the evaluations of the two systems are statistically equivalent. The tests for equivalence did not provide sufficient evidence to confirm statistical equivalence between the LM2D model and ground truth  $(p = 0.799, \delta = 0.05)$ . This highlights the challenge of achieving parity with ground truth in terms of motion-lyrics matching.

This study further integrates open-ended questions to gather in-depth feedback and identify areas for future enhancements. Experts with choreography experience suggested the inclusion of criteria to evaluate choreography more comprehensively. They noted that while certain movements are a good match and appear natural, they are relatively simple. Additionally, the evaluation of subtle, smaller movements presented challenges. For example, movements involving the neck are relatively subtle, yet they appear somewhat stiff. This limitation is partly due to the dance dataset being sourced from Just Dance, a rhythm game developed for the general public that does not encompass highly complex or challenging dance sequences. Regarding the matching of motion to lyrics, a common observation was that some videos did not align well with the lyrics. This issue is largely attributed to the data, where the focus of the movements was primarily to follow the musical beats rather than to match the lyrics precisely. However, it's important to recognize that dance is not a literal translation of words like sign language. Effective choreography should consider the overall harmony of rhythm and lyrics rather than mechanically translating lyrics into movements. Additionally, choreography experts have expressed the need for real-time modification of generated movements based on lyrics. Our work further investigates consistency distillation to achieve applicability in this area. However, upon comparison, the performance of distillation techniques in dance generation can still be further improved.

### 6 Ethical and Societal Discussion

This work presents a new dataset and a system for creating dances, aiming to bring new ideas and improvements to the field of dance. It facilitates the generation of dance influenced by lyrics and music, offering valuable tools for artists and researchers. This work could also be advantageous to fields like video gaming and animation. Nevertheless, the automation of choreographic processes raises concerns about originality and creativity that it may blur the lines of ownership in creative endeavors.

# 7 Conclusions and Future Work

In this study, we introduce LM2D, a multimodal diffusionbased model for generating realistic dance sequences conditioned on both lyrics and music. A novel dance dataset featuring a combination of lyrics, music, and 3D motion data was collected. We enhanced synthesis efficiency by employing consistency distillation for one-step generation. Our extensive quantitative and qualitative assessments demonstrate the effectiveness and superior capabilities of the proposed method. In summary, this research paves the way for creating complex choreographies that are synchronized with musical rhythms and lyric semantics.

Looking ahead, our future endeavors include integrating Large Language Models (LLMs) to deepen the understanding of lyrics. LLMs have demonstrated impressive capabilities in a wide range of NLP tasks, also evidenced by recent advancements in understanding, planning, and generating motion with LLMs [Zhang *et al.*, 2023; Zhou *et al.*, 2023]. The community should leverage these powerful LLMs to push the boundaries of dance motion synthesis, potentially synthesizing multiple modalities simultaneously, such as dance motion and audio that includes sung lyrics. In addition, future work will focus on achieving accelerated output generation while maintaining the quality of the results.

## 8 Acknowledgments

This research received partial support from the National Institute of Informatics (NII) in Tokyo. This work has been supported by the European Research Council (BIRD-884807) and H2020 EnTimeMent (no. 824160). This work was partially supported by the Wallenberg Al, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. This work also benefited from access to the HPC resources provided by the Swedish National Infrastructure for Computing (SNIC), partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

## References

- Omid Alemi, Jules Françoise, and Philippe Pasquier. Groovenet: Real-time music-driven dance movement generation using artificial neural networks. *networks*, 8(17):26, 2017.
- Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. Listen, denoise, action! audio-driven motion synthesis with diffusion models. ACM Transactions on Graphics (TOG), 42(4):1–20, 2023.

- 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.
- Anna Deichler, Shivam Mehta, Simon Alexanderson, and Jonas Beskow. Diffusion-based co-speech gesture generation using joint text and audio representation. In *Proceedings of the 25th International Conference on Multimodal Interaction*, pages 755–762, 2023.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Junting Dong, Qing Shuai, Yuanqing Zhang, Xian Liu, Xiaowei Zhou, and Hujun Bao. Motion capture from internet videos. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, pages 210–227. Springer, 2020.
- Rukun Fan, Songhua Xu, and Weidong Geng. Examplebased automatic music-driven conventional dance motion synthesis. *IEEE transactions on visualization and computer graphics*, 18(3):501–515, 2011.
- Di Fan, Lili Wan, Wanru Xu, and Shenghui Wang. A bidirectional attention guided cross-modal network for music based dance generation. *Computers and Electrical Engineering*, 103:108310, 2022.
- Satoru Fukayama and Masataka Goto. Music content driven automated choreography with beat-wise motion connectivity constraints. *Proceedings of SMC*, pages 177–183, 2015.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- Kimerer LaMothe. The dancing species: how moving together in time helps make us human. *Aeon, June*, 1:1, 2019.
- Nhat Le, Thang Pham, Tuong Do, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Music-driven group choreography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8673–8682, 2023.
- Minho Lee, Kyogu Lee, and Jaeheung Park. Music similaritybased approach to generating dance motion sequence. *Multimedia tools and applications*, 62:895–912, 2013.
- Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. Learning to generate diverse dance motions with transformer. *arXiv preprint arXiv:2008.08171*, 2020.
- 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.
- Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1272–1279, 2022.

- Ronghui Li, Junfan Zhao, Yachao Zhang, Mingyang Su, Zeping Ren, Han Zhang, and Xiu Li. Magic: Multi art genre intelligent choreography dataset and network for 3d dance generation. arXiv preprint arXiv:2212.03741, 2022.
- Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics* (*TOG*), 34(6):1–16, 2015.
- Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.
- Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings* of the 14th python in science conference, volume 8, pages 18–25, 2015.
- George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):2617–2680, 2021.
- Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7753–7762, 2019.
- Hayley Elizabeth Powell. Modern dance choreography: Beyond the movement an analysis between lyrics and movement: Can identities be developed through modern dance choreography? *Annual Review of Education, Communication & Language Sciences*, 16(2), 2019.
- Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. arXiv preprint arXiv:2311.17042, 2023.
- Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11050– 11059, 2022.
- Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.
- Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Scorebased generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.
- Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.
- Xiangjun Tang, He Wang, Bo Hu, Xu Gong, Ruifan Yi, Qilong Kou, and Xiaogang Jin. Real-time controllable mo-

tion transition for characters. ACM Transactions on Graphics (TOG), 41(4):1–10, 2022.

- Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.
- 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.
- 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 *ISMIR*, volume 1, page 6, 2019.
- Guillermo Valle-Pérez, Gustav Eje Henter, Jonas Beskow, Andre Holzapfel, Pierre-Yves Oudeyer, and Simon Alexanderson. Transflower: probabilistic autoregressive dance generation with multimodal attention. *ACM Transactions on Graphics (TOG)*, 40(6):1–14, 2021.
- Zixuan Wang, Jia Jia, Haozhe Wu, Junliang Xing, Jinghe Cai, Fanbo Meng, Guowen Chen, and Yanfeng Wang. Groupdancer: Music to multi-people dance synthesis with style collaboration. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1138–1146, 2022.
- Zijie Ye, Haozhe Wu, Jia Jia, Yaohua Bu, Wei Chen, Fanbo Meng, and Yanfeng Wang. Choreonet: Towards music to dance synthesis with choreographic action unit. In Proceedings of the 28th ACM International Conference on Multimedia, pages 744–752, 2020.
- Wenjie Yin, Hang Yin, Kim Baraka, Danica Kragic, and Mårten Björkman. Multimodal dance style transfer. *Machine Vision and Applications*, 34(4):1–14, 2023.
- Mingao Zhang, Changhong Liu, Yong Chen, Zhenchun Lei, and Mingwen Wang. Music-to-dance generation with multiple conformer. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pages 34–38, 2022.
- Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. *arXiv preprint arXiv:2301.06052*, 2023.
- 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, 2019.
- Zixiang Zhou, Yu Wan, and Baoyuan Wang. Avatargpt: All-in-one framework for motion understanding, planning, generation and beyond. *arXiv preprint arXiv:2311.16468*, 2023.
- Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings* of the IEEE/CVF International Conference on Computer Vision, pages 15085–15099, 2023.

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.