Motion to Dance Music Generation using Latent Diffusion Model

Vanessa Tan
JungHyun Nam
KAIST GSCT
vanessa.tan@kaist.ac.kr
ys4990@kaist.ac.kr

Juhan Nam*
KAIST GSCT
juhan.nam@kaist.ac.kr

Junyong Noh
KAIST GSCT
junyongnoh@kaist.ac.kr

ABSTRACT

The role of music in games and animation, particularly in dance content, is essential for creating immersive and entertaining experiences. Although recent studies have made strides in generating dance music from videos, their practicality in integrating music into games and animation remains limited. In this context, we present a method capable of generating plausible dance music from 3D motion data and genre labels. Our approach leverages a combination of a UNET-based latent diffusion model and a pre-trained VAE model. To evaluate the performance of the proposed model, we employ evaluation metrics to assess various audio properties, including beat alignment, audio quality, motion-music correlation, and genre score. The quantitative results show that our approach outperforms previous methods. Furthermore, we demonstrate that our model can generate audio that seamlessly fits to in-the-wild motion data. This capability enables us to create plausible dance music that complements dynamic movements of characters and enhances overall audiovisual experience in interactive media. Examples from our proposed model are available at this link: https://dmdproject.github.io/.

CCS CONCEPTS

• **Computing methodologies** → **Learning latent representations**; *Motion processing*; Supervised learning.

KEYWORDS

3D motion to music, music generation, latent diffusion model

ACM Reference Format:
Vanessa Tan, JungHyun Nam, Juhan Nam, and Junyong Noh. 2023. Motion to Dance Music Generation using Latent Diffusion Model. In *SIGGRAPH Asia 2023 Technical Communications (SA Technical Communications '23), December 12–15, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3610543.3626164

1 INTRODUCTION

Music plays a vital role in the creation of games and animation. However, traditional methods of music generation can be cumbersome for composers, as they need to repeatedly check how the music synchronizes with the theme or character movements. Recently, the popularity of dance content generation has increased, leading researchers to explore automated approaches to simplify music generation [2, 6, 22, 23]. While these studies have made significant strides in automating music generation using video data, their applications may be limited when it comes to seamlessly integrating music with games or animation, as the music should be synchronized with the characters' movements. To address this limitation, this work presents a method to generate music from 3D motion data. This approach not only assists composers in creating music but also benefits indie developers or animators without a music background, allowing them to produce content without copyright music concerns. Furthermore, dance performance directors can express their creativity by generating new music.

The proposed method is a motion to dance music generation model capable of producing plausible dance music from 3D motion data and genre labels. Our method generates music with a timbre that encapsulates the dance's mood, while seamlessly synchronizing it with the accompanying 3D motion. It utilizes a latent diffusion-based architecture paired with a pre-trained VAE model. The key contributions of this work include:

- We are the first to propose a method to generate dance music from 3D human motion data and a music genre condition.
- Showcasing the versatility of our model, we demonstrate its capability to generate music from in-the-wild motion data.
- Through an in-depth analysis of metrics from prior dance and music generation research, we propose how to evaluate the generated music with 3D motion data and genre labels.

2 RELATED WORK

Music generation has become increasingly popular in recent years. To address computational complexity when the model is conditioned on video and motion data, previous work focuses on symbolic generation, involving formats such as notes [2] and MIDI [6]. However, these symbolic representations have limitations in capturing musical properties (e.g. dynamics and timbre) and can only generate monophonic music. Considering that current trend in music generation is directly generating audible sound using better representations such as spectrograms [9, 14] or raw audio waveforms [5, 22, 23], we build upon recent work that utilizes mel-spectrograms as the audio representation for generating polyphonic music.

Generative models, such as GANs and diffusion models, have enabled cross-modal generation, leading to more flexible and creative music than traditional methods. These models are typically conditioned with text or image inputs [3, 9]. For dance music, video data has been employed to condition these models [2, 22, 23]. However,

*Corresponding author

SIGGRAPH Asia 2023, December 12–15, 2023, Sydney, Australia

ACM ISBN 979-8-4007-0314-0/23/12…$15.00
https://doi.org/10.1145/3610543.3626164

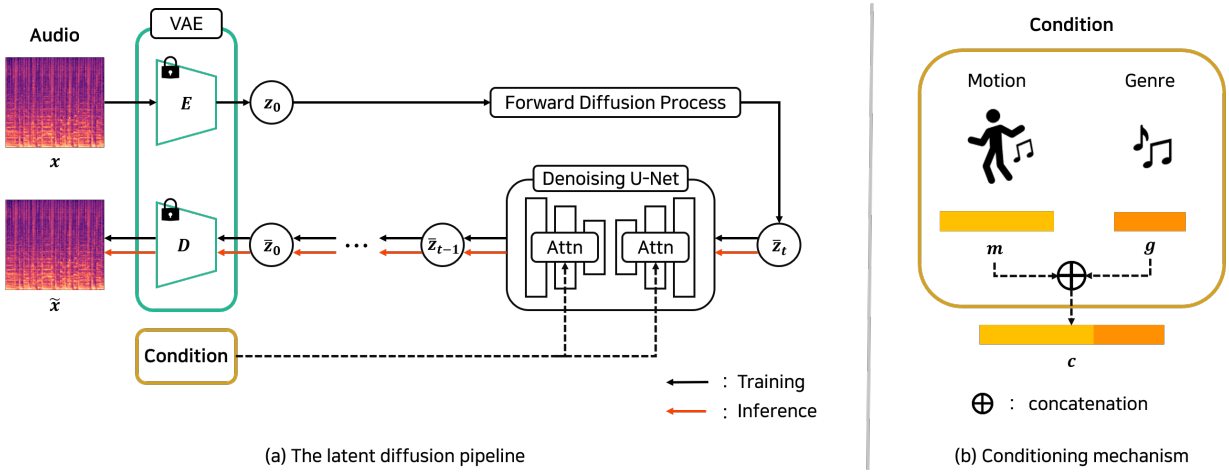

Figure 1: Overview of the proposed method.

application of these approaches may be limited when integrating music into games or animations where computer graphics is engaged. Therefore, this research aims to explore a novel approach, conditioning a latent diffusion model with 3D motion data information, to generate dance music that seamlessly fits into interactive environments like games and animations. We additionally conditioned the model with music genre as a high-level control.

3 METHOD

An overview of the architecture is shown in Figure 1. Our approach employs a latent diffusion model paired with a pre-trained Variational AutoEncoder (VAE) model which is conditioned with motion and genre labels.

3.1 Audio Representation

Following recent work on audio-level music generation [9, 14], we represent audio data with mel-spectrogram which is denoted as $\mathbf{x}$ in Figure 1. We computed mel-spectrograms with a sampling rate of 22,050Hz, a Hann window of 2048 samples, a hop length of 512 samples, and 256 mel bins. The mel-spectrograms are then passed through a Variational AutoEncoder (VAE) [10]. The VAE model is composed of two main components: an encoder $\mathbf{E}$, which takes the mel-spectrogram $\mathbf{x}$ as input and generates a compressed latent code $\mathbf{z}$, and a decoder $\mathbf{D}$, which reconstructs the mel-spectrogram from the compressed latent representation $\mathbf{z}$. This process allows us to efficiently encode and decode the audio data while preserving its essential characteristics.

3.2 Conditioning Mechanism

The motion data is represented as a sequence of poses in the 24-joint SMPL format for each frame i [12, 20]. We define our motion features $\mathbf{m}$ as the set of $\mathbf{m}_i = \{p_i, q_i, \dot{p}_i, \dot{q}_i\}$ where p_i, q_i, $\dot{p}_i$, and $\dot{q}_i$ stands for the position, orientation, linear velocity, and angular velocity, respectively. The inclusion of $\dot{p}_i$ and $\dot{q}_i$ is driven by the motivation to comprehensively represent the diverse and dynamic nature of dance motions. The position p_i and linear velocity $\dot{p}_i$ are described using the global coordinates, while the orientation q_i and angular velocity $\dot{q}_i$ are defined with respect to their parent joint axes. To ensure continuity in the representation, the orientation is converted into 6D representations [21]. The genre features $\mathbf{g}$ are defined as the genre labels corresponding to the dance motion. The genre labels are then one-hot encoded and concatenated with the motion features $\mathbf{m}$, resulting in the conditioning signal $\mathbf{c} = \mathbf{m} \bigoplus \mathbf{g}$. This combined embedding is used as the condition to the diffusion model to process and analyze the motion with respect to their genre labels. The conditioning signal is then mapped to the intermediate layers of the denoising network via cross-attention layers [17].

3.3 Latent Diffusion Model

With the emergence of latent diffusion models as the state-of-the-art approach for image [17] and music generation [9, 14], this work is built upon this model to explore its potential in generating music with motion data. With a pre-trained VAE which is frozen during training, the input audio mel-spectrogram $\mathbf{x}$ is encoded to $\mathbf{z}$ in a latent space $\mathbf{Z}$ effectively compressing the learning into a lower-dimensional space, thereby increasing training efficiency. Diffusion models typically follow a Markov noising process, $\{z_t\}_{t=0}^{T}$, where z_0 is drawn from the data distribution and the forward process is defined as

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{\alpha_t} z_{t-1}, (1-\alpha_t)I) \tag{1}$$

where $\alpha_t \in (0, 1)$ is a sampling hyper-parameter [8]. Subsequently, we introduce the conditioning signal $\mathbf{c}$ by approximating the distribution $p(z_0|\mathbf{c})$ through a reverse diffusion process using a neural network. Following the studies on latent diffusion models, we employ a UNET-based model [9, 17] as our denoising network ϵ_θ for every time step t and its objective function is defined as

$$\mathcal{L}_{simple} = _{\epsilon,t,c}\ [||\epsilon - \epsilon_\theta(z_t, t, \mathbf{c})||_2^2] \tag{2}$$

where $\epsilon \sim \mathcal{N}(0, 1)$ [8]. The denoised output is then decoded by the pre-trained VAE to reconstruct the mel-spectrogram. After generating the mel-spectrogram, we utilized the Griffin-Lim algorithm to reconstruct the corresponding audio waveform [15, 16]. No post-processing effects, such as noise filtering, were applied in this study. The only modification made to the output was applying gain to match the loudness of the audio with the ground truth.

4 EXPERIMENTS

4.1 Dataset

We used the AIST++ dataset [11] which includes paired music and motion data. The dataset contains 1,020 3D motion data represented with the SMPL format. Additionally, the music data comprises of 60 songs with 10 genres, resulting in a total duration of 18,694 seconds. We extracted 5-second slices from the dataset, following a similar approach to previous research on music-to-dance [11, 20].

4.2 Implementation Details

For the audio encoder and decoder, we used a pre-trained VAE from Hugging Face. This pre-trained VAE was trained on 20,000 mel-spectrograms of 5-second samples of music from a random Spotify playlist [19]. The denoising UNET architecture consists of 3 down-sampling and up-sampling ResNet blocks with cross-attention layers and skip connections. The latent diffusion model was trained with AdamW optimizer [13] with a learning rate of 1e-4 and the batch size was set to 8. The number of diffusion steps during training was 1000 while 50 was set during inference. The model was trained for 100 epochs on a NVIDIA RTX A5000 GPU for 1 day.

4.3 Evaluation Metrics

Since we are the first to leverage 3D motion data and genre for dance music generation, we formulate a comprehensive evaluation protocol inspired by prior research in dance motion and music generation. The evaluation incorporates several key metrics: beats coverage score, beats hit score, Frechet Audio Distance (FAD), beat alignment score, and genre Kullback Leibler Divergence (KLD) score. First, we employ the beats coverage score and beats hit score, metrics derived from video-to-dance music research [22, 23]. These metrics are primarily concerned with the accuracy of beat representation in the generated music. The beats coverage score calculates the ratio of the overall generated beats to the total musical beats of the ground truth, while the beat hit score measures the ratio of the aligned generated beats to the total musical beats of the ground truth [22].

We then leverage the widely used FAD from text-to-music research [3]. FAD evaluates audio quality by measuring the similarity between the generated audio and the ground truth. Lower FAD values indicate more plausible audio. To calculate the FAD score, we utilized a VGGish audio embedding model pre-trained on the YouTube-8M audio event dataset [1, 3, 7]. The beat alignment score, derived from music-to-dance research [11, 18], is employed to assess motion-music correlation. This score quantifies the relationship between motion and music by computing the average distance between each kinematic beat (determined from the local minima of the kinetic velocity) and the nearest music beat (extracted using the Librosa library [15]). Furthermore, we introduce a novel metric, the genre KLD score, to assess the genre representation of the generated music. This score measures the distance between the predicted class probabilities of the generated music and the ground truth genre labels. For this evaluation, we used a pretrained genre classifier, MS-SincResNet [4], which more accurately reflects the genre representations of music compared to classifiers trained on audio event detection datasets [22, 23].

5 RESULTS AND DISCUSSION

5.1 Quantitative Evaluation

The quantitative evaluations are shown in Table 1, in which we compare our method to the ground truth data from the AIST++ dataset [11] and the output of the CDCD model [23] which uses video and genre as their conditioning signal. Our model achieved better scores for all the metrics than the CDCD model, a state-of-the-art video to dance music generation research. Because we are the first to implement a model that generates dance music from genre and 3D motion data only comparing our work to models dealing with video as input data might not be entirely fair.

To address this, we conducted a Mean Opinion Score (MOS) test as our subjective evaluation, comparing our results to the ground truth and a baseline method, in which we randomly paired ground truth music with dance motion. The MOS test involved 20 participants who evaluated 15 songs, 5 songs from each method, in two categories: audio quality and motion correlation. The participants rated the songs using a 5-step Likert scale, ranging from "Poor" to "Excellent." The results of the subjective evaluation are shown in Table 2, which indicates that our method has a better performance in terms of the motion correlation; however, the audio quality of our method still falls short when compared to the ground truth.

5.2 Qualitative Results

The qualitative results of our method can be seen in this link: https://dmdproject.github.io/. The examples show that our model can generate plausible dance music given the 3D motion data and genre label. Our model also exhibits the ability to generate music with the same motion data but with different genres, showcasing its versatility. Furthermore, our model can generate plausible beat-aligned dance music even when using in-the-wild motion data. This aspect highlights the robustness of our approach and its potential to handle diverse and real-world motion data.

5.3 Limitations

Our current model faces difficulties in generating a diverse range of dance music, primarily due to the constrained size of the paired music and motion dataset, which comprises only 60 songs paired with 369 choreographies from the AIST++ dataset [11]. To tackle this limitation, one potential solution is to build a more extensive paired motion-music dataset by extracting data from YouTube videos. Another issue with our model is its inability to generate plausible vocals. To address this concern, incorporating a vocoder into the output generation process could prove beneficial, as it may help improving the quality and realism of the generated music. The current method has another drawback of inability to generate relatively long sequences. While the model performs well in

Model	Beats Coverage Score ↑	Beats Hit Score ↑	Frechet Audio Distance ↓	Beat Align Score ↑	Genre KLD ↓
Ground Truth	100	100	-4.47e-13	0.211	0
CDCD [23]	78.5	74.8	9.07	0.202	0.612
Ours	**93.5**	**86.0**	**4.96**	**0.212**	**0.604**

Table 1: Model Evaluation for AIST++ Dataset.

Model	MOS (Audio Quality) ↑	MOS (Motion Corr.) ↑
Ground Truth	4.52	4.10
Random GT	**4.33**	3.13
Ours	3.26	**3.51**

Table 2: Subjective Evaluation.

generating plausible 5-second audio samples when provided with a 5-second input, extending the output to longer sequences results in a decrease in audio quality. To tackle this challenge, exploring other denoising architectures such as transformers could be a promising direction, as they may offer better capabilities for generating longer and high-quality audio sequences.

6 CONCLUSION AND FUTURE WORK

We introduced a latent diffusion model that effectively generates realistic dance music from 3D motion data and genre labels. Our model was comprehensively evaluated using various quantitative metrics, including rhythm, audio quality, motion-music correlation, and genre scores. The results show that the model is capable of producing beat-aligned music across different genres and in the presence of in-the-wild motion data.

Future research may delve into enhancing the model's capacity to generate longer than 5-second input music sequences. One potential avenue of exploration involves leveraging pre-trained motion embeddings, which would make the model not rely on handcrafted motion features, potentially leading to improved performance. Furthermore, expanding the model's cross-modality capabilities by incorporating other conditions such as free text, could open up new exciting research directions. Such enhancements could enable the model to generate more diverse music with improved versatility.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00222383).

REFERENCES

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Apostol Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. YouTube-8M: A Large-Scale Video Classification Benchmark. *ArXiv* abs/1609.08675 (2016).

[2] Gunjan Aggarwal and Devi Parikh. 2021. Dance2Music: Automatic Dance-driven Music Generation. arXiv:2107.06252 [cs.SD]

[3] Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matthew Sharifi, Neil Zeghidour, and C. Frank. 2023. MusicLM: Generating Music From Text. *ArXiv* abs/2301.11325 (2023).

[4] Pei-Chun Chang, Yong-Sheng Chen, and Chang-Hsing Lee. 2021. MS-SincResNet: Joint learning of 1D and 2D kernels using multi-scale SincNet and ResNet for music genre classification. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*. 29–36.

[5] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A Generative Model for Music. *arXiv preprint arXiv:2005.00341* (2020).

[6] Chuang Gan, Deng Huang, Peihao Chen, Joshua B. Tenenbaum, and Antonio Torralba. 2020. Foley Music: Learning to Generate Music from Videos. arXiv:2007.10984 [cs.CV]

[7] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson. 2016. CNN architectures for large-scale audio classification. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2016), 131–135.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. *CoRR* abs/2006.11239 (2020). arXiv:2006.11239 https://arxiv.org/abs/2006.11239

[9] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. 2023. Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models. arXiv:2301.12661 [cs.SD]

[10] Diederik P Kingma and Max Welling. 2022. Auto-Encoding Variational Bayes. arXiv:1312.6114 [stat.ML]

[11] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. 2021. Learn to Dance with AIST++: Music Conditioned 3D Dance Generation. arXiv:2101.08779 [cs.CV]

[12] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

[13] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101 [cs.LG]

[14] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. 2023. Diff-Foley: Synchronized Video-to-Audio Synthesis with Latent Diffusion Models. arXiv:2306.17203 [cs.SD]

[15] Brian McFee, Colin Raffel, Dawen Liang, Daniel Ellis, Matt Mcvicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and Music Signal Analysis in Python. 18–24. https://doi.org/10.25080/Majora-7b98e3ed-003

[16] Nathanaël Perraudin, Peter Balazs, and Peter L. Søndergaard. 2013. A fast Griffin-Lim algorithm. In *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 1–4. https://doi.org/10.1109/WASPAA.2013.6701851

[17] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 10674–10685.

[18] Lian Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. 2022. Bailando: 3D Dance Generation by Actor-Critic GPT with Choreographic Memory. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), 11040–11049.

[19] Robert Smith. 2022. Audio Diffusion. https://github.com/teticio/audio-diffusion.

[20] Jo-Han Tseng, Rodrigo Castellon, and C. Karen Liu. 2022. EDGE: Editable Dance Generation From Music. *ArXiv* abs/2211.10658 (2022).

[21] Yi Zhou, Connelly Barnes, Lu Jingwan, Yang Jimei, and Li Hao. 2019. On the Continuity of Rotation Representations in Neural Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[22] Ye Zhu, Kyle Olszewski, Yuehua Wu, Panos Achlioptas, Menglei Chai, Yan Yan, and S. Tulyakov. 2022. Quantized GAN for Complex Music Generation from Dance Videos. *ArXiv* abs/2204.00604 (2022).

[23] Ye Zhu, Yuehua Wu, Kyle Olszewski, Jian Ren, S. Tulyakov, and Yan Yan. 2022. Discrete Contrastive Diffusion for Cross-Modal Music and Image Generation.