國家科學及技術委員會大專學生研究計畫

多模態生成式 AI--音樂生成舞蹈模型

計畫編號：NSTC　112-2813-C-002-032-E

執行期間：　2023 年　7 月　1 日至　2024 年　4 月　3 日

執行機構及系所：國立臺灣大學資訊工程學系暨研究所

計畫主持人：鄭文皇教授

學生姓名：簡詩汶

中　華　民　國　113 年　5 月

Contents

關鍵字: 人工智慧, 舞蹈模型, 生成式人工智慧, 音樂生成舞蹈, 霹靂舞

Key words: AI, dancing AI, generative AI, music-driven dance generation, breakdancing

Abstract

In recent years, the field of Music-driven Dance Generation, commonly referred to as Dancing AI, has become highly competitive within the AI academic community. Began around fifteen years ago, this area of research initially struggled due to constrained of hardware capabilities and inadequate computational resources. Recently, advancements in hardware computing have significantly enhanced the performance of Dancing AI models. Various models, including transformers and diffusion models, have been explored and tested in this domain. Despite these advancements, several challenges remain unresolved. This study aims to address some of these persistent issues, which includes:

1. Motion Limitation: Although the Dancing AI models incorporate break-dancing style within the training dataset, instances of breaking movements are scarcely observed in the testing phase. This discrepancy arises due to the varying movement speeds and foot-floor contact patterns characteristic of different dance styles. To prevent foot-sliding, the previous method included a loss term that penalized unexpected foot-ground contact. However, a distinctive feature of break dancing is the frequent lifting of the feet into the air while using the hands for support, resulting in sudden interruptions in foot-ground contact. Consequently, this loss term would inevitably prevent the model from performing break-dance movement. To address this issue, we will conduct specialized training focused exclusively on the breaking style, with the aim of enhancing the generation of authentic break-dancing movements.

2. Lack of consultation with dance experts: Machine Learning scientists in Natural Language Processing would collaborate with linguists to tailor AI training to the field. However, there is a lack of collaboration with dance scholars in Dancing AI research. While a few papers have included formulas from classic choreography theories, there have been little further engagement with professional dancer [1]. This gap suggests a missed opportunity for deeper integration of expert knowledge into AI development.

3. Falsely mixing different dance styles in training: Humans practice different dance styles with distinctly different methods [2][3][4]. Current research often blended various styles within one model, combining popular electronic music with heavy-beat hip-hop music [5]. Professional modern dancers may be struggle with hip-hop because it requires a different type of muscle engagement and flexibility. Even within the hip-hop genre, there are significant variations, such as old-school, LA style, and house, each with its unique characteristic. Consequently, training AI model with a blend of these diverse styles may lead to inaccuracies. Past models often have the best performance in

standard and contemporary dance [6], mainly due to the completeness of dance steps, which can be directly replicated and utilized from the dataset. In contrast, street dance is characterized by its fluid and often unstructured steps. Training models in such unregulated styles typically results in lower training quality.

Introduction

Microsoft co-founder Bill Gates said in a Reddit forum that AI is the big one. I don't think Web3 was that big or that metaverse stuff alone was revolutionary but AI is quite revolutionary. " Adding, " I am quite impressed with the rate of improvement in these AIs. I think they will have a huge impact. " [7] [8] Single-mode AI painter and AI robots have become popular technologies in the worlds, and multimodal generative AI is the next focus. Music-generated dance models are part of this advancement. In 2024, Paris Olympics will include breakdancing for the first time. With the Olympics' endorsement and the global trend of street dance, Dancing AI is becoming a battleground in academia and industry. Unlike traditional Olympic events, breakdancing lacks predetermined movements, requiring dancers to either choreograph routines or improvise on the spot. Movement quality is crucial criterion in judgment. Developing a Dancing AI model aims to inspire dancers' creativity and potentially provide them with innovative dance sequences. As the technology evolves, there is hope that it might one day replicate the movements of medalists to create similar winner's dance steps. Positioned as Taiwan's innovative choreography tool in the breakdancing scene, it strives to support Taiwan's Olympic efforts and establish a presence in the realm of multimodal generative AI.

Purpose

From the above discussion, it is evident that Dancing AI has room for improvement. Therefore, the core objective of this project is to develop Dancing AI technology that achieves the highest user experience scores based on Olympic breakdancing judging criteria. The process will involve addressing issues like unable to generate hand-standing moves, incorporating suggestions from dance professionals, and focusing on the specific styles of street dancing. To achieve a breakthrough in 3D motion generative AI technology and to serve as an inspiration for break dancers.

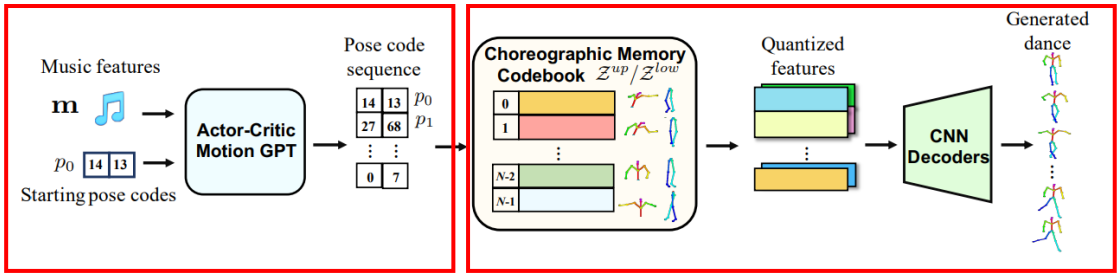


Figure 1. The structure of Bailando model. It represents the common elements of Dancing AI, which are music feature encoding (right frame) and dance motion generation (right frame), with the training method involving encoding the entire piece of music followed by training a decoder. (Siyao et al., 2022) [6].

Related Work

In the past three years, many researchers have been devoted to Dancing AI technology and have achieved significant milestones [2] [5] [6] [9] [10] [11] [12] [13]. As illustrated in Figure (1), the common elements of all these technologies are "music feature encoding" and "dance motion generation," which can be further subdivided into human motion generation and music-generated dance. One of the differences among the current technologies lies in the use of different models; for example, [12] employs the two-stream motion transformer (TSMT) for encoding and synthesizing motions and sounds; [6] uses the Generative Pre-trained Transformer (GPT); [13] utilizes the Generative Autoregressive Network as the motion generation model; and [9] employs the Full-Attention Cross-modal Transformer. Another difference is in the selection of datasets; for instance, [5] [9] created their own datasets, [12] crawled a large number of YouTube videos to create a dataset, and the choice of dataset can affect the training of music styles, the determination of the number of joints, and the degrees of freedom in human motion generation, etc.

In conclusion, Dancing AI technology can be divided into three major units: (1) Music analysis and feature encoding; (2) Human motion generation; and (3) Audio to human motion generation. To accomplish the task of multimodal music-generated dance, it is essential to study the existing technologies first to understand which ones can most accurately generate dance. Therefore, the following analysis focuses on the latest technologies.

The three major units:

1. Music Analysis and Feature Encoding (Music Feature Encode)

Music analysis can be approached from various perspectives, including beats, rhythm, pitch, intensity, chromogram, volume, and so on. Music encoding will affect generated dance movement regardless of the model used, hence the choice of music features or the allocation of weights becomes crucial. Past studies often involved manually selecting features. For instance, [14] considered chords, structure, Mel-frequency cepstral coefficients (MFCCs), and beats as important features for generating dance, then directly selecting these features as the analysis targets, as shown in Figure (2). Some studies incorporated music analysis modules such as Librosa [15] and Jukebox [16] for sound signal processing to obtain features like chroma features and beat tempo. Further processing of the encoded results, such as calculating similarity matrices, can also be conducted.

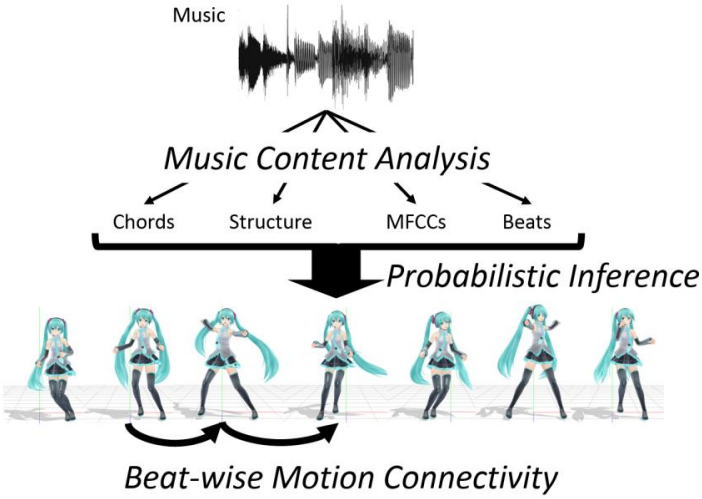


Figure 2. Model involved manually selecting music features by humans. (Satoru Fukayama and Masataka Goto, 2015) [14].

Choosing music features poses a challenge for researchers, as they must also avoid dances merely fitting the rhythm while neglecting the overall musicality. Through music feature extraction and encoding [16] [17] [18], coupled with corresponding motion matching, music can better synchronize with the dancer's movements, even reflecting the emotions conveyed in the song [1].

2. Human Motion Generation

Human motion generation refers to synthesizing sequences of human movements. In the early stages, probability models [19] [20] [21] [22] or motion matching techniques [23] [24] [25] [26] [27] [28] were commonly used. Motion matching involves

automatically comparing trajectories, velocities, poses, etc., in a database to select the most fitting animation data for the current animation, thereby automatically generating transitional animations. However, these methods impose rules on the fixed length and sequence order, limiting the application of movements, and they cannot simulate the temporality of actions, making them challenging to apply to dance. In recent years, deep learning models have been more commonly used to map out human joints, including RNNs [29] [30] [31] [32] [33] [34] [35], CNNs [36] [37], GANs [38], Transformers [39] [40], and their variations such as auto-regressive models [41] and phase-functioned neural networks [42]. The input for this model can be joystick controls [43], text input [44], and so on.

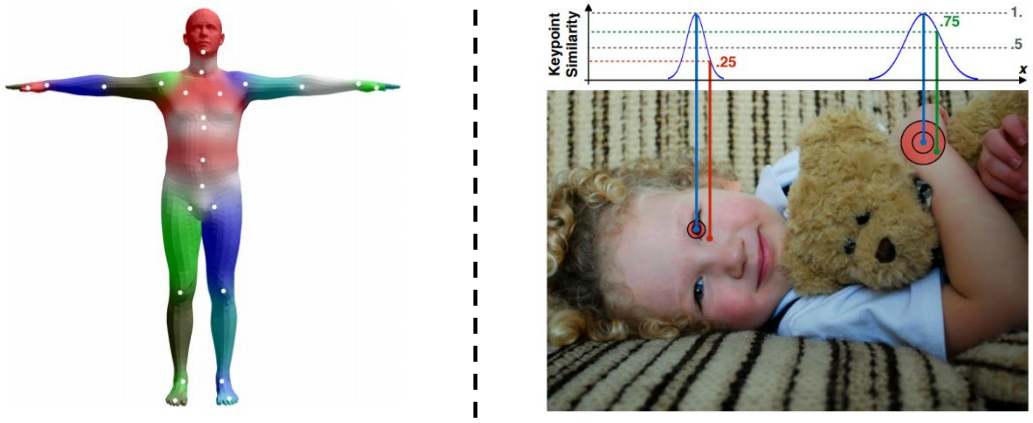


Figure 3: (a) SMPL human body model. The human skeleton is composed of a hierarchical structure of 24 joints that defined by a kinematic tree, indicating a parent-child relationships. Matthew Loper et al., 2015) [45]. (b) COCO human body model. Consists of 17 key points automatically generated from object detection and heat maps. (Matteo Ruggero Ronchi et al., 2016) [46].

To avoid generating movements that exceed the physical limits of the human body, researchers must restrict joint activities based on the dataset, which involves setting the number of joints and degrees of freedom during training. The choice of human model is fundamental to motion generation and significantly impacts the output. The selection of the human model depends on the dataset. For example, in the context of dance, AIST++, created by [9], includes two types of human models: the 3D SMPL (Skinned Multi-Person Linear Model) with 24 joint models [45], and the 2D/3D COCO-format (Common Objects in Context) with 17 key points [46], as shown in Figure (3).

3. Audio To Human Motion Generation

Due to the easy accessibility of 2D dance data and the maturity of human body

recognition technology [47], a large number of dance databases can be developed. Hence, 2D Dancing AI has been studied for a long time [48] [49]. However, 3D data has been scarce until the development and public release of the AIST++ dataset in 2021. Consequently, 3D models for generating dance from music have proliferated [2] [6] [9].

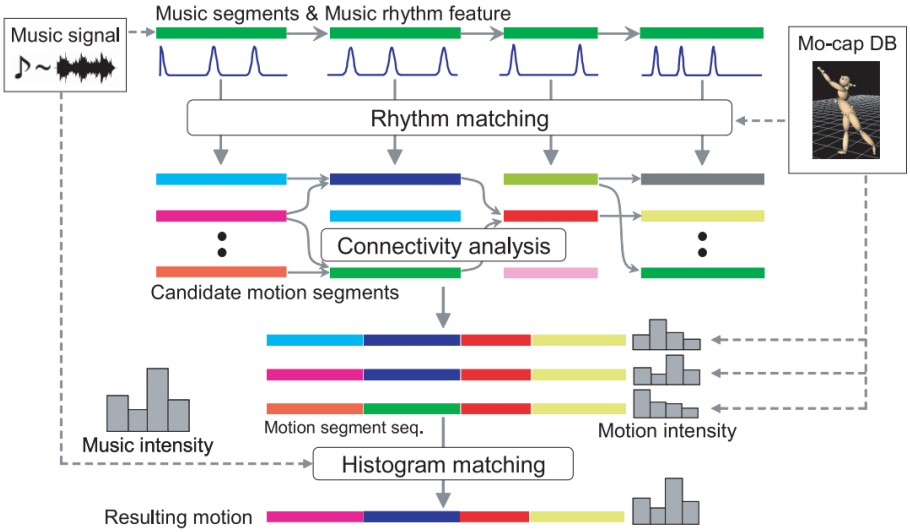


Figure 4. Architecture of Dancing-to-Music model. This method involves segmenting music, extracting melodic features, and pairing them with dance segments, attempting to combine the paired segments, and finally producing multiple outcomes. The final result is chosen based on the best match between music intensity and dance intensity. (Shiratori et al., 2006) [1].

In the early stages, music-to-motion mapping was commonly used in audio to human motion generation models [50] [51]. This involved encoding features separately for music and motion, and then matching the results of dance motion analysis [1] [5] [12]. For instance, [1] first segmented the music, matched music rhythms to many possible motion segments, assembled all connected motions, and finally made the best matching based on motion intensity and music intensity, as depicted in Figure (4). However, this method often resulted in unrealistic or overly simplistic movements.

In recent years, similar to advancements in human motion generation, Dancing AI has predominantly utilized deep learning models, including RNNs [10] [52] [53] [54], GANs [11] [49], Transformers [9] [12], Diffusion [2], and GPT [6]. For example, the model proposed by [6] encodes the entire music segment, then trains a decoder to transform the music encoding into dance movements. This approach helps avoid the rigidity caused by the one-to-one mapping between dance and music, as shown in Figure (1).

Method

1. Dataset Collection

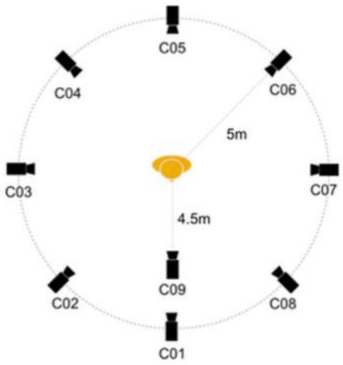
The most comprehensive and commonly used dataset for 3D dance selection, the AIST++ dataset, is used in this study. It comprises 10 dance styles, 1408 sequences of 3D dance motion, captured from 9 different perspectives as shown in Figure (5). The duration of movements ranges from 7.4 seconds to 48.0 seconds, with a resolution of 1080P and a frame rate of 60 frames per second (FPS). The dataset includes 85% basic movements and 15% challenging movements, making it highly flexible for various applications.

AIST++ includes both COC and SMPL models for every frame of motions. In our experiment, we choose to use SMPL since it had better result over the COCO dataset for 3D human figure dancing generations. This is not surprising since SMPL is designed for complex 3D motions. SMPL model provides 24 key points compared to 17 human figure key points provided by COCO. More numbers in key point resolution are essential to capture with intricate movements and poses in dance. For example, SMPL has two points on the toes and heel while COCO only has one point on the whole foot, which would work better on this model because toes and heel are crucial for humans to shift weight.

For break dancing, AIST++ provides 10 basic movements of approximately 10 seconds each performed by 3 different dancers and filmed from 9 different angles, paired with 4 different pieces of music. This result in a total of 10*3*9*4 = 1080 movement-music pairs. Additionally, each dancer performs 7 solo movements of approximately 50 seconds, filmed from 9 different angles paired with different pieces of music, which is 3*7*9 = 189 movement-music pairs. In total, there are 1080 + 189 = 1269 pairs available for training the models.

However, there are only 6 different songs among all the pairings, which is significantly fewer than the thousands of songs a human break dancer would the familiar with. Moreover, the tempo of break-dancing competition typically ranges between 110 and 135 bpm, and only three songs in the training set match this tempo range. To address this issue, 53 different songs from previous Red Bull BC One, the world's largest global breaking competition, have been selected and adjusted to match the tempo of the existing 31 motions in the dataset. Furthermore, we had asked experienced dancer to meticulously reviewed each song-motion pair to ensure musical compatibility,

eliminating pairs that lacked proper musicality despite having matching bpm.



Break	mBR0	80
	mBR1	90
	mBR2	100
	mBR3	110
	mBR4	120
	mBR5	130

Figure 5. The AIST++ dataset includes data on multiple camera angles and the beats per minute (BPM) of accompanying music. [55].

To achieve optimal performance, this project focuses specifically on Breaking dance and breaks it down into its sub-elements: Toprock, Footwork, Power Moves, and Freezes. We then balanced the duration of these movements to match those typically seen in competitions, which often include more Power Moves and Freezes. This was accomplished by up-sampling the selected movements accordingly.

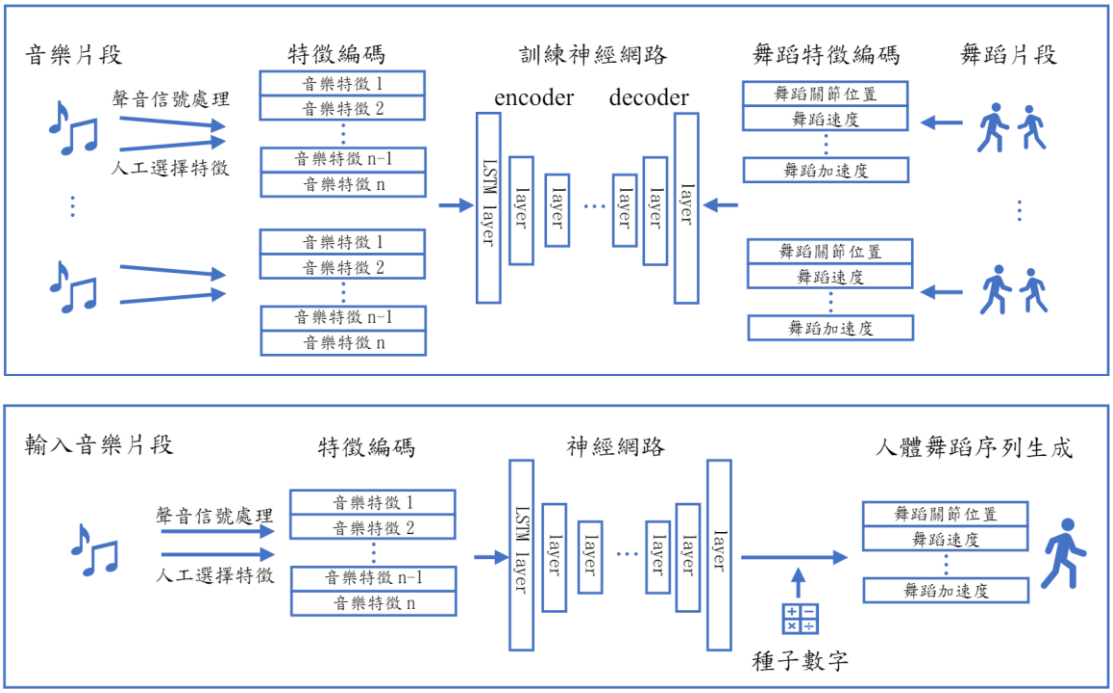


Figure 6. Simplified model workflow of this project, with the upper diagram representing the training process and the lower diagram the generation process. After users input an audio signal, the model processes this signal and acquires predefined features. These music features are then encoded and fed into a neural network, along with a randomly generated seed number, to produce a sequence of human dance

movements.

2. Model

The experimental workflow of the Dancing AI model in this project, as depicted in Figure (6), begins with the user inputting an audio file. The model then extracts music features through Jukebox music analysis and feature encoding. For the analysis method, we integrate a music analysis module for sound signal processing. We manually select music ranging between 110 to 130 beats per minute (bpm) to match the music typically played at official international Breaking competitions and design a neural network to undergo multi-to-multi training using various sound inputs paired with multiple dance segments. This enables the model to output human dance movements based on the features encoded from the music. We train our break-dancing model by fine tuning the model from EDGE checkpoint [2]. Additionally, auxiliary loss functions are incorporated and adjusted into the loss function to ensure that the generated human movements adhere to physical logic. Specifically, we adjusted the weights of the foot-floor contact loss and foot speed loss to get the model target on power move and freeze movement in break-dancing. Furthermore, we set seed movements to ensure that the same music input can produce multiple different dance outputs.

3. Model Evaluation Criteria

Given the novelty of Dancing AI, there is no unified standard for evaluation in past literature. Some consider user ratings, others consider the conformity of movements to the beat, and some consider diversity, among other factors. As our model focuses on Breaking, we utilize the five newest criteria announced by the Olympics in April 2024: Technique, Vocabulary, Originality, Execution, and Musicality each account for 20% [56]. To verify the effectiveness of our new model, we compare it with the latest models using user ratings. Evaluators are required to have at least one year of Breaking experience to accurately simulate a real competition environment.

In the survey, the participate are asked to compare two different AI models so as. We include three rounds of battles with different songs, where two of them are songs from Red Bull BC One competitions that are new to the models and one which had been trained on both models, with the positions of the models on the left and right randomly placed on the screen in each round to avoid bias from previous rounds.

Result

The experimental results highlight a significant advancement in the performance of our newly developed break-dancing AI model compared to the previous model. A paired t-test was conducted to compare the mean scores of both models with a p-value of 0.0001, indicating that the improvements in the new model are statistically significant. The mean score for the new model was 50.18, compared to 40.53 for the old model, with variances of 265.32 and 288.86, respectively.

	New Model (ours)	Old Model
Mean	**50.18**	**40.53**
Variance	265.32	288.86
df	25	
t Stat	4.28837	
P(T<=t) one-tail	**0.00012**	
t Critical one-tail	1.70814	

Table 1. Paired t-test of New Model and Old Model evaluation

The average score of the three-round battles is presented in Table 2. Participants reported that the outputs of the new model felt more "break-dancing" and "technical," capturing the essence of break-dancing more authentically. Many observed that the transitions between different dance elements were smoother and more synchronized with the music.

	New Model (ours)	Old Model
Round 1	**49.15**	38.40
Round 2	**48.75**	44.75
Round 3	**54.40**	38.60

Table 2. The average score of the three-round battles with new and old model

Conclusion and Future Work

In this work, we propose a revised break-dancing AI model that solve the problem of inability to generate break dancing movement of old model. We successfully expand the dataset to 59 different songs that match the existing motion clips with bpm between 110 to 130, and train the model with this new dataset. We assessed our model using the

most recent criteria from the April 2024 Olympics, alongside a user study conducted with a professional dancer. Our findings indicate that it achieves more authentic movement than the original model. Importantly, we adjusted the weight of the foot-contacting loss to ensure that the model is not restricted and can generate break-dancing moves where the hands are the main support and the feet are airborne. The model can take arbitrarily long music clips and generate 3D dance output.

Songs can be readily accessed and edited online using music editing tools. However, acquiring dance movements is more challenging due to the need for multi-angle sequences to generate 3D motion, which limits the training of new models without extensive datasets. Nonetheless, advancements in human detection technology will enabling the precise and large-scale production of dance motion data. This new technology could potentially increase the available data by thousands of times compared to current datasets. We are eager to see how models will perform with such an expanded dataset in the future.

Reference

[1] Shiratori, Takaaki & Nakazawa, Atsushi & Ikeuchi, Katsushi, "Dancing-to-Music Character Animation," *Computer Graphics Forum,* vol. 25, pp. 449-458, 2006.

[2] Tseng, Jonathan and Castellon, Rodrigo and Liu, C. Karen, "EDGE: Editable Dance Generation From Music," *arXiv preprint arXiv:2211.10658,* 2022.

[3] Sebastian Starke, Ian Mason, and Taku Komura, "DeepPhase: periodic autoencoders for learning motion phase manifolds," *ACM Trans. Graph,* vol. 41, no. 4, p. 13, July 2022.

[4] Li, B., Zhao, Y., Zhelun, S., & Sheng, L. “DanceFormer: Music Conditioned 3D Dance Generation with Parametric Motion Transformer”. *AAAI Conference on Artificial Intelligence*, 36(2). 2022

[5] Wenlin Zhuang, Congyi Wang, Jinxiang Chai, Yangang Wang, Ming Shao, and Siyu Xia, "Music2Dance: DanceNet for Music-Driven Dance Generation," *ACM Trans. Multimedia Comput. Commun. Appl,* vol. 18, no. 2, p. 21, May 2022.

[6] Siyao, Li & Yu, Weijiang & Gu, Tianpei & Lin, Chunze & Wang, Quan & Qian, Chen & Loy, Chen Change & Liu, Ziwei, "Bailando: 3D Dance Generation by Actor-Critic GPT with Choreographic Memory," *CVPR,* pp. 11040-11049, 2022.

[7] B. Gates, "I’m Bill Gates, and I’m back for my 11th AMA. Ask Me Anything," Reddit, 12 Jan 2023. [Online]. Available:

https://www.reddit.com/r/IAmA/comments/109eze3/comment/j3xucdi/?utm_source=share&utm_medium=web2x&context=3.

[8] B. Gates, "I'm Bill Gates, and I'm back for my 11th AMA. Ask Me Anything.," Reddit, 12 Jan 2023. [Online]. Available: https://www.reddit.com/r/IAmA/comments/109eze3/comment/j3xvphy/?utm_source=share&utm_medium=web2x&context=3.

[9] R. Li, S. Yang, D. Ross and A. Kanazawa, "AI Choreographer: Music Conditioned 3D Dance Generation with AIST++," *IEEE/CVF International Conference on Computer Vision (ICCV),* pp. 13381-13392, 2021.

[10] Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, and Mi Zhang, "Dance Revolution: Long Sequence Dance Generation with Music via Curriculum Learning," *ICLR,* 2021.

[11] G. Sun, Y. Wong, Z. Cheng, M. S. Kankanhalli, W. Geng and X. Li, "DeepDance: Music-to-Dance Motion Choreography With Adversarial Learning," *IEEE Transactions on Multimedia,* vol. 23, pp. 497-509, 2021.

[12] 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,* Aug 2022.

[13] Hyemin Ahn, Jaehun Kim, Kihyun Kim, and Songhwai Oh, "Generative Autoregressive Networks for 3D Dancing Move Synthesis From Music," *IEEE Robotics and Automation Letters,* pp. 1-1, 2020.

[14] Satoru Fukayama and Masataka Goto, "Music Content Driven Automated Choreography With Beat-wise Motion Connectivity Constraints," in *Sound and Music Computing Conference (SMC)*, Maynooth, Co. Kildare, Ireland, 2015.

[15] 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*, 2015.

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

[17] J. Zhang, "Music Data Feature Analysis and Extraction Algorithm Based on Music Melody Contour," *Mobile Information Systems,* vol. 2022, p. 10, 2022.

[18] Hawthorne, C., Elsen, E., Song, J., Roberts, A., Simon, I., Raffel, C., Engel, J., Oore, S., Eck, D, "Onsets and frames: Dual-objective piano transcription," *arXiv preprint arXiv:1710.11153,* 2017.

[19] Katherine Pullen and Christoph Bregler, "Animating by multi-level sampling,"

in *Proceedings Computer Animation 2000*, 2000.

[20] R. Bowden, "Learning statistical models of human motion," *IEEE Workshop on Human Modeling, Analysis and Synthesis, CVPR,* vol. 2000, October 2000.

[21] Aphrodite Galata, Neil Johnson, and David Hogg, "Learning variable-length markov models of behavior," *Computer Vision and Image Understanding,* vol. 81, no. 3, p. 398–413, 2001.

[22] Matthew Brand and Aaron Hertzmann, "Style machines," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 2000.

[23] Daniel Holden, Oussama Kanoun, Maksym Perepichka, and Tiberiu Popa, "Learned motion matching," *ACM Transactions on Graphics (TOG),* vol. 39, no. 4, pp. 1-53, 2020.

[24] Jehee Lee, Jinxiang Chai, Paul SA Reitsma, Jessica K Hodgins, and Nancy S Pollard, "Interactive control of avatars animated with human motion data," in *Annual Conf. on Comput. Graph. and Interactive Tech*, 2002.

[25] Lucas Kovar, Michael Gleicher, and Frédéric Pighin, "Motion graphs," in *ACM SIGGRAPH*, 2008.

[26] Okan Arikan and David A Forsyth, "Interactive motion generation from examples," *ACM Transactions on Graphics (TOG),* vol. 21, no. 3, p. 483–490, 2002.

[27] Alexis Lamouret and Michiel van de Panne, "Motion synthesis by example," *Comput. Animat. and Simulat,* 1996.

[28] Jehee Lee and Sung Yong Shin, "A hierarchical approach to interactive motion editing for human-like figures," in *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, 1999.

[29] Hsu-Kuang Chiu, Ehsan Adeli, Borui Wang, De-An Huang, and Juan Carlos Niebles, "Action-Agnostic Human Pose Forecasting," Waikoloa, HI, USA, 2019.

[30] Xiaoxiao Du, Ram Vasudevan, and Matthew Johnson-Roberson, "Bio-LSTM: A Biomechanically Inspired Recurrent Neural Network for 3D Pedestrian Pose and Gait Prediction," *IEEE Robotics and Automation Letters,* pp. 1-1, 2019.

[31] Fragkiadaki, Katerina & Levine, Sergey & Felsen, Panna & Malik, Jitendra, "Recurrent Network Models for Human Dynamics," 2015.

[32] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges, "Learning Human Motion Models for Long-term Predictions," *arXiv preprint arXiv:1704.02827,* 2017.

[33] Ashesh Jain, Amir R. Zamir, Silvio Savarese, and Ashutosh Saxena, "Structural-RNN: Deep Learning on Spatio-Temporal Graphs," pp. 5308-5317, 2016.

[34] Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee, "Neural Kinematic Networks for Unsupervised Motion Retargetting," *CVPR,* 2018.

[35] Julieta Martinez, Michael J. Black, and Javier Romero, "On Human Motion Prediction Using Recurrent Neural Networks," pp. 4674-4683, 2017.

[36] Daniel Holden, Jun Saito, and Taku Komura, "A deep learning framework for character motion synthesis and editing," *ACM TOG,* vol. 35, no. 4, p. 1–11, 2016.

[37] Daniel Holden, Jun Saito, Taku Komura, and Thomas Joyce, "Learning motion manifolds with convolutional autoencoders," in *SIGGRAPH Asia 2015 Technical Briefs*, 2015.

[38] Alejandro Hernandez, Jurgen Gall, and Francesc MorenoNoguer, "Human motion prediction via spatio-temporal inpainting," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[39] Emre Aksan, Peng Cao, Manuel Kaufmann, and Otmar Hilliges, "Attention, please: A spatio-temporal transformer for 3d human motion prediction," *arXiv preprint arXiv:2004.08692,* 2020.

[40] Uttaran Bhattacharya, Nicholas Rewkowski, Abhishek Banerjee, Pooja Guhan, Aniket Bera, and Dinesh Manocha, "Text2gestures: A transformer-based network for generating emotive body gestures for virtual agents," *arXiv preprint arXiv:2101.11101,* 2021.

[41] Zimo Li, Yi Zhou, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li, "Auto-conditioned recurrent networks for extended complex human motion synthesis," in *ICLR*, 2018.

[42] Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi Zaman, "Local motion phases for learning multi-contact character movements," *ACM Transactions on Graphics (TOG),* vol. 39, no. 4, pp. 1-54, 2020.

[43] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne, "Character controllers using motion vaes," *ACM Transactions on Graphics (TOG),* vol. 39, no. 4, pp. 1-40, 2020.

[44] Mathis Petrovich, Michael J Black, and Gul Varol, "TEMOS: Generating diverse human motions from," *arXiv preprint arXiv:2204.14109,* 2022.

[45] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black, "SMPL: A skinned multiperson linear model," in *SIGGRAPH Asia*, 2015.

[46] Matteo Ruggero Ronchi and Pietro Perona, "Benchmarking and error diagnosis in multi-instance pose estimation," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[47] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh, "OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields," 2018.

[48] Juheon Lee, Seohyun Kim, and Kyogu Lee, "Listen to dance: Music-driven choreography generation using autoregressive encoder-decoder network," *arXiv preprint arXiv:1811.00818,* 2018.

[49] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz, "Dancing to music," 2019.

[50] Rukun Fan, Songhua Xu, and Weidong Geng, "Example-Based Automatic Music-Driven Conventional Dance Motion Synthesis," *IEEE Transactions on Visualization and Computer Graphics,* vol. 18, no. 3, pp. 501-515, March 2012.

[51] Minho Lee, Kyogu Lee, and Jaeheung Park, "Music similarity-based approach to generating dance motion sequence," *Multimedia Tools and Applications,* vol. 62, p. 895–912, 2013.

[52] Omid Alemi, Jules Franc¸oise, and Philippe Pasquier, "Groovenet: Real-time music-driven dance movement generation using artificial neural networks," *Networks,* vol. 8, no. 17, p. 26, 2017.

[53] Taoran Tang, Jia Jia, and Hanyang Mao, "Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis," 2018.

[54] Nelson Yalta, Shinji Watanabe, Kazuhiro Nakadai, and Tetsuya Ogata, "Weakly-supervised deep recurrent neural networks for basic dance step generation," 2019.

[55] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto, "AIST Dance Video Database: Multi-genre, Multi-dancer, and Multi-camera Database for Dance Information Processing," in *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[56] Presto, G. (2024, April 23). Breaking at the Olympic Qualifier Series: Everything you need to know. Olympics.com. Retrieved from https://olympics.com/en/news/breaking-at-the-olympic-qualifier-series-everything-you-need-to-know

Appendices

A. Examination Video

The video features three rounds of battle that are used for user study. One dancer is generated by our modified EDGE Dancing AI model, specifically tailored to create break-dancing moves. The other is generated by the original EDGE Dancing AI model checkpoint. The positions of the models, left and right, are randomly assigned in each round to prevent bias from previous rounds, where the new model are place on left side, right side, and ride side sequentially. The songs are excerpt from the Red Bull BC One 2015 Soundtrack (00:00-00:27), "Apache" by Incredible Bongo Band (00:28-01:32), and "Pam (Remix)" by Crazy Elephant (01:33-02:25). The survey includes options for participants with or without more than one year of break-dancing experience. However, we only use responses from those who have experience in our final result analysis. The video link: https://www.youtube.com/watch?v=PTPGgjDH2Ks&ab_channel=Caslow

B. Survey Form

The survey is provided in both Mandarin and English. Below is the introduction to the survey:

此問卷目的為比較兩種不同的 AI 模型，以驗證新模型的成效。比賽有 3 回合，每一回合的模型左右位置為隨機擺放，避免前面回合造成偏見。

The purpose of this survey is to compare two different AI models to verify the effectiveness of the new model. The competition consists of three rounds, with the positions of the models on the left and right randomly placed in each round to avoid bias from previous rounds.

==========

評分依照奧運標準，分為 5 大部分。其中以下標準:
The critieria is devided into 5 categories, including the following:

🏋️ Technique 技術性 (20%)

🕺 Vocabulary 舞蹈詞彙量 (20%)

🌟 Originality 原創性 (20%)

✅ Execution 執行力 (20%)

🎵 Musicality 音樂性 (20%)

以下為標準的解釋

🏋️ Technique: Certain moves have certain criteria, such as keeping feet flexed versus toes pointed on many moves, Edra says. But technique also includes the judges' view of how athletic the breakers are, and how well they control their bodies.

🕺 Vocabulary: Breakers must perform a variety of moves in multiple positions—both in down rock and top rock—to score well here.

🌟 Execution: While this may sound similar to technique, the World DanceSport Federation rule book says that on execution, breakers are judged on how cleanly their moves are performed—that is, they don't mess up—and how distinct one more is from the next. The moves should flow together, but not blend together.

✅ Musicality: Here, breakers are judged on their ability to not just perform incredible moves, but to dance—staying on beat, and timing their moves to the music.

🎵 Originality: Louis believes this is the most important criteria. "Having that personal style is what sets people apart," he says. "I could learn every move out there, but it's about what can I bring to the table? What can I add to breaking?

🏋️ 技巧：Edra 說，不同的動作有不同的標準，例如很多動作需要保持腳部彎曲，而不是腳趾指向。但技巧也包括裁判對舞者運動能力的評價，以及他們對自己身體的控制力。

🕺 詞彙：舞者必須在多種位置下表演各種動作——無論是下搖還是上搖——才能在這方面獲得高分。

🌟 執行：雖然這聽起來與技巧相似，但世界舞蹈運動聯合會的規則手冊指出，在執行方面，舞者被評判的標準是他們動作的完成度——即他們是否出錯——以及每個動作之間的區別。動作應該連貫但不應混在一起。

✅ 音樂性：在這方面，舞者被評判的不僅是他們完成驚人的動作的能力，還有他們跳舞的能力——保持節拍，並將動作與音樂的節奏相匹配。

🎵 創意：Louis 認為這是最重要的標準。他說：「擁有個人風格是讓人脫穎而出的關鍵。我可以學會所有的動作，但重要的是我能帶來什麼？我能為 Breaking 增添什麼？」

來 源 Reference: https://olympics.com/en/news/breaking-at-the-olympic-qualifier-series-everything-you-need-to-know

全部問卷需約 10 分鐘完整，非常感謝 🐻
The survey will take about 10 minutes. Thank you so much! 🐿️

C. Hyperparameters

The model is trained with only 290 epochs which is approximately 12 whole days due to the limitation of computing power. The changes in three stages of training are as listed:

Stage – 1

Train with break-dancing clips with bpm greater than 110 (includes) from AIST++ dataset, which is 96 music-motion pairs.

Hyperparameter	Value
Epochs	**224**
Optimizer	Adan
Learning Rate	**0.0006**
Batch Size	**64**
Diffusion Steps	1000
Beta schedule	Cosine
Motion Duration	5 seconds
Motion FPS	30
Motion Dimension	151
Classifier-Free Dropout	0.25
Num Heads	8
Num Layers	8
Transformer Dim	512
MLP Dim	1024
Dropout	0.1
EMA Steps	1
EMA Decay	0.9999
Weight of Simple Loss	0.636
Weight of Velocities Loss	**2.964**
Weight of Joint Positions	0.646
Weight of Contact Consistency Loss	**10.642**

Stage – 2

Train with break-dancing clips, where the songs are extracted and adjusted from RedBull BC One World Final Paris 2023 Mixtape (Breaking Music), which is 202 music-motion pairs in total. We lower the weights of the velocity loss and contact consistency loss to enable the execution of fast foot movements.

Hyperparameter	Value
Epochs	**38**
Learning Rate	**0.0004**
Batch Size	**32**

Weight of Simple Loss	0.636
Weight of Velocities Loss	**0**
Weight of Joint Positions	0.646
Weight of Contact Consistency Loss	**0**

Stage – 3

Train with break-dancing clips, where the songs are extracted and adjusted from RedBull BC One World Final Paris 2023 Mixtape (Breaking Music), which is 202 music-motion pairs in total. Increase the weights of velocities loss and contact consistency loss to avoid foot sliding.

Hyperparameter	Value
Epochs	**28**
Learning Rate	0.0004
Batch Size	32
Weight of Simple Loss	0.636
Weight of Velocities Loss	**1**
Weight of Joint Positions	0.646
Weight of Contact Consistency Loss	**1**

D. Customized Music-Motion Pairs

The songs were cut from Red Bull BC One Breaking Competition from the following sources:

1. 35 minutes long, Red Bull BC One World Final Paris 2023 Mixtape (Breaking Music)
https://www.youtube.com/watch?v=DSfiG6ybCPw&ab_channel=BboyMusic

2. 23 minutes long, Red Bull BC One 2015 Soundtrack
https://www.youtube.com/watch?v=X0Ic6aZBMbw&t=919s&ab_channel=BboyBEAT

We adjusted the music's beats per minute using the Ableton music editor to ensure that the tone and texture of the sound remained consistent while synchronizing with the beats. To mimic a real break-dancing competition, we opted to use only moves that match the original songs' BPM of 110 or higher, specifically 110, 120, and 130 BPM. Then, we asked a break dancer with over a year of experience to eliminate any pairs that did not match the original dance moves. Here are the new pairs trained after

selection, where “mBRx” are the new songs, and “chxx” are the original motions.

BPM 130

Original songs: mBR5
Original motions: ch6, ch12, ch14, ch19, ch21
New songs: mBR6~ mBR27
mBR6: ch12, ch19
mBR7: ch12, ch19
mBR8: ch21, ch14
mBR9: ch12, ch14, ch21
mBR10: ch12, ch14, ch19, ch21
mBR11: ch12, ch14, ch16, ch21
mBR12: ch06, ch12 ch14, ch19, ch21
mBR13: ch14, ch19, ch21
mBR14: ch06, ch12, ch14, ch19, ch21
mBR15: ch06, ch12, ch14, ch19, ch21
mBR16: ch12, ch14, ch21
mBR17: ch06, ch12, ch14, ch21
mBR18: ch12, ch14, ch21
mBR19: ch06, ch12, ch14, ch21
mBR20: ch06, ch12, ch19, ch21
mBR21: ch06, ch12, ch14, ch19, ch21
mBR22: ch06, ch12, ch14, ch19, ch21
mBR23: ch12, ch19, ch21
mBR24: ch14, ch19, ch21
mBR25: ch06, ch12, ch14, ch21
mBR26: ch12, ch19
mBR27: ch12, ch14, ch19, ch21

BPM 120

Original songs: mBR4
Original motions: ch05, ch07, ch11, ch13, ch18, ch20
New songs: mBR28~ mBR48
mBR28: ch07, ch11
mBR29: ch07, ch11, ch20
mBR30: ch05, ch07, ch11, ch13
mBR31: ch07, ch11, ch20
mBR32: ch05, ch11, ch13, ch20

mBR33: ch05, ch07, ch11, ch13
mBR34: ch05, ch11, ch13, ch20
mBR35: ch07, ch11, ch13, ch18, ch20
mBR36: ch05, ch07, ch13
mBR37: ch18
mBR38: ch05, ch11, ch13, ch20
mBR39: ch07
mBR40: ch07, ch11, ch13, ch18
mBR41: ch07, ch13
mBR42: ch05, ch07, ch20
mBR43: ch05, ch07, ch11, ch13, ch18, ch20
mBR44: ch05, ch11, ch13, ch18
mBR45: ch07, ch20
mBR46: ch07, ch13
mBR47: ch20
mBR48: ch05, ch07, ch11, ch18

BPM 110

Original songs: mBR3
Original motions: ch04, ch10, ch17
New songs: mBR49~ mBR58
mBR49: ch10
mBR50: ch04, ch10, ch17
mBR51: ch10
mBR52: ch10, ch17
mBR53: ch04, ch17
mBR54: ch17
mBR55: ch04, ch10, ch17
mBR56: ch04, ch10
mBR57: ch04, ch10, ch17
mBR58: ch04, ch10, ch17