GENRE-CONDITIONED LONG-TERM 3D DANCE GENERATION DRIVEN BY MUSIC

Yuhang Huang[1], *Junjie Zhang*[1], *Shuyan Liu*[3], *Qian Bao*[2*], *Dan Zeng*[1*], *Zhineng Chen*[4], *Wu Liu*[2]

Shanghai University[1], JD AI Research[2], University of Chinese Academy of Sciences[3], Fudan University[4]

ABSTRACT

Dancing to music is an artistic behavior of humans, however, letting machines generate dances from music is still challenging. Most existing works have been made progress in tackling the problem of motion prediction conditioned by music, yet they rarely consider the importance of the musical genre. In this paper, we focus on generating long-term 3D dance from music with a specific genre. Specifically, we construct a pure transformer-based architecture to correlate motion features and music features. To utilize the genre information, we propose to embed the genre categories into the transformer decoder so that it can guide every frame. Moreover, different from previous inference schemes, we introduce the *motion queries* to output the dance sequence in parallel that significantly improves the efficiency. Extensive experiments on AIST++[1] dataset show that our model outperforms state-of-the-art methods with a much faster inference speed.

Index Terms— 3D dance generation, genre-conditioned, modality fusion, music-driven

1. INTRODUCTION

Dancing to music is an innate behavior of human beings, and humans can make corresponding motions according to the music beats. With the development of Artificial Intelligence (AI), we are interested in whether we can let AI learn to dance to music with a given genre. This task possesses great research and application potential since it can be applied in promising areas such as dance creation assistant in art and sports, character motion generation for audio games, and research on cross-modal behavior.

There have been previous works that explore the music-conditioned dance generation. Early methods[2, 3, 4] tackle this task as a similarity-based retrieval problem, which hardly achieves the desired performance. Existing state-of-the-art methods[5, 1, 6, 7, 8] formulate the task as a generative pipeline, which takes the music and initial pose or sequence as the input and predicts the future dance sequence in an auto-regressive manner.

However, the above methods do not consider the genre information, which regulates the dance movements and represents the global information of a dance sequence. Without the guidance of the music genre, the dances they generate may align with the music beats well but have disorderly movements that are incomprehensible to be appreciated. Different from previous studies, our goal is to generate a long-term 3D dance that fits the given music genre. To our best of knowledge, this is the first work to generate the 3D dance with a specific musical genre.

Since it is a relatively new task, there are still some challenges. Firstly, two modalities of inputs—music, and motion need to be considered, and how to correlate them is vague. Then, the genre is sequence-level information for the dance sequence, so it is uneasy to guide the generation of the frame-level motion. Moreover, most existing methods synthesize the dance sequences through an auto-regressive way that generates dance frame by frame, which will produce a large error when generating long-term sequences. These challenges motivate us to come up with a more robust solution that can generate a long-term dance effectively from music with a given genre.

In this paper, we propose to utilize the attention mechanism to fuse the information of two modalities and generate realistic 3D dances with a pure transformer structure. It has been proven that the transformer can extract features with advanced representation ability [9, 10, 11] and establish the strong correlation between different features[12, 13]. Therefore, we adopt transformer structures as our basic feature extractor and feature fusion module. Moreover, to incorporate the crucial genre information as the guidance into the generation process, we embed the genre information into the query of transformer decoder, so that it can spread to every frame of music features and motion features. Moreover, unlike the previous auto-regressive method that only returns one frame at each step, our transformer-based structure can return 60 frames at one time, which significantly improves efficiency and reduces the error accumulation. In summary, our contributions are as follows: 1) We introduce Genre-Conditioned Dance Generator (GCDG), a pure transformer-based structure, to generate long-term 3D dance from music. 2) We explore generating the dances based on music genres, which is the first attempt to consider music genre as a generation condition. 3) Extensive experiments on the AIST++ dataset demonstrate the effectiveness of our method.

The rest paper is organized as follows: Sec. 2 reviews the related works of motion generation and dancing to mu-

* Dan Zeng and Qian Bao are the co-corresponding authors.

sic; Sec. 3 presents the proposed GCDG model; experimental setup and results are given in Sec. 4; and we conclude in Sec. 5.

2. RELATED WORKS

2.1. Human Motion Prediction

Human motion prediction is defined as predicting the future frames with given past motions or initial poses. Earlier works[14, 15] employ statistical models to complete this task. Most recent works[16, 17, 18, 19, 20, 21, 22] often treats the body as a skeleton and usually adopt generative models for motion prediction. Dlow[19] uses a diversity-promoting prior over samples as an objective to optimize the latent mappings to improve sample diversity. [23] aims to predict long-term 3D human motion from a single scene image and 2D pose histories. MT-GCN[18] predicts future 3D human motions from the incomplete historical motion. Different from prior works, We not only generate long-term future 3D motions conditioned on initial motions, but also leverage the music genre as an additional condition.

2.2. Dancing to Music

There have been several studies exploring the dance generation driven by music. Early methods[2, 3] tackle this task as a similarity-based retrieval problem, which limits the creativity of generations. Recent studies[24, 25, 5, 1, 6] make efforts in formulating the task as a generative pipeline. [6] decomposes a dance into a series of basic dance units and proposes a novel decomposition-to-composition framework to generate dance moves. [5] formalizes the music-driven dance generation as a sequence-to-sequence learning problem and devises a seq2seq architecture to process long sequences of music features and capture the correspondence between music and dance. FACT[1] involves a deep cross-modal transformer block with full-attention that is trained to predict N future motions. In this paper, we generate future dance not only conditioned by music and initial motion but also conditioned by genre information.

3. PROPOSED METHOD

Problem Definition. Our goal is to generate a future dance conditioned by the initial motion and music with a specific genre. Formally, given a music $X = \{x_t\}_{t=1}^{N}$, a initial pose sequence $P^{ini} = \{p_t^{ini}\}_{t=1}^{M}$ aligned with X at the first frame, and a genre type $g \in G$ (G is a set of predefined genre categories), our goal is to generate the future motion $\hat{P}^{pre} = \{\hat{p_t}^{pre}\}_{t=M+1}^{N}$.

Representations. For the music representation, we utilize the publicly available audio processing toolbox Librosa[26] to extract the music features including 1-dim envelope, 20-dim MFCC, 12-dim chroma, 1-dim one-hot peaks, and 1-dim one-hot beats, resulting in a 35-dim music feature $x_t \in \mathbb{R}^{35}$.

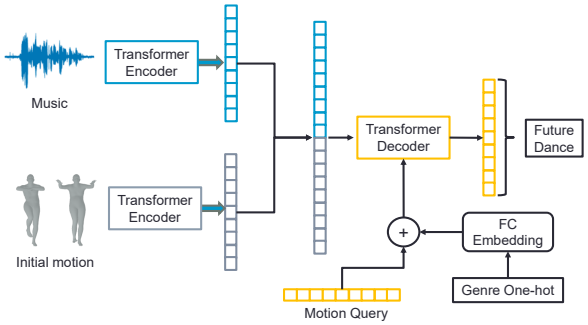

Fig. 1. The overall architecture of the proposed GCDG. First of all, we use two shallow transformer encoders to extract features for the given music and initial motion respectively. Then, in the decoding phase, we use the transformer decoder to generate dance sequences in parallel. Specifically, we concatenate the music and initial motion features as the key and value, and define a set of learnable query embedding *motion queries* as the query. To utilize the genre information, we use a fully-connected layer to embed the genre one-hot vector and add it to the motion queries.

For motion representation, since we aim at generating 3D dances, we employ SMPL[27] body model as our body representation and ignore the variation of body shape. SMPL pose parameters represent 23 joint rotations in the kinematic tree and one global rotation per frame. We follow the [1] to use the rotation matrix representation and resulting in a 219-dim motion feature $p_t \in \mathbb{R}^{219}$.

3.1. GCDG Architecture

As shown in Fig. 1, we employ a pure transformer-based model to extract music and motion features and generate the future long-term 3D dance guided by the genre information. Specifically, we first use two shallow transformer encoders to extract music features and initial motion features respectively. Then, these two modalities of features are concatenated and sent to a deeper transformer decoder to produce key and value. We additionally define a set of learnable query embedding named *motion queries* as the input of the transformer decoder, which is used for absorbing contextual information from music and initial motion, and generating the future dance with the same length as *motion queries* in parallel. To obtain the guidance of genre, we embed the genre as a vector and add it on the *motion queries*.

Feature Extractor. The music and initial motion are both sequences and they both have strong correlations within frames. Therefore, we use two transformer encoders to extract the music and motion features respectively, which can build long-range dependencies across the whole sequence. All the music and motion representations are linearly embedded into a $\mathbb{R}^d$ space before being fed into the encoders. Finally, we obtain two features: $f_m \in \mathbb{R}^{N \times d}$ and $f_p \in \mathbb{R}^{M \times d}$,

Method	$BeatAlign \uparrow$	$FID_k \downarrow$	$FID_g \downarrow$	$Dist_k \uparrow$	$Dist_g \uparrow$	$FID \downarrow$	$ACC \uparrow$	$Div. \rightarrow$	Infer. Time
GT	0.292	-	-	9.06	7.56	-	88.5	14.13	-
DR[5]	0.220	73.42	31.01	3.52	2.46	-	-	-	-
FACT[1]	0.241	**35.35**	12.40	5.94	5.30	-	-	-	-
DR*[5]	0.220	59.37	30.64	4.95	3.00	60.10	30.0	12.35	14.3 min
FACT*[1]	0.236	50.35	12.45	5.72	5.09	46.23	56.5	12.58	16.7 min
GCDG (ours)	**0.244**	39.12	**10.36**	**7.38**	**6.21**	**30.83**	**85.7**	**13.57**	**0.9 min**

Table 1. The comparisons against state-of-the-art methods on the AIST++ test set.

which represent music features and motion features respectively.

Transformer Decoder. We first concatenate the f_m and f_p resulting in $f_{mp} \in \mathbb{R}^{(N+M)\times d}$, and feed it into the transformer decoder to produce the key and value. The benefit of such an operation is that we can fuse the music and motion information across time and provide contextual information for generating future dance sequences. Moreover, we define *motion queries* $\in \mathbb{R}^{T\times d}$ that are learnable as the query of the transformer decoder, where T is a pre-defined length. To introduce the genre information into the dance generation, we use a one-hot vector to represent the genre and use a fully-connected layer to embed it into $g_{emb} \in \mathbb{R}^d$ so that it can be added to the *motion queries*. The transformer decoder directly outputs a sequence of T vectors in $\mathbb{R}^d$ and we obtain the final motion prediction followed by a linear projection layer. It is worth noting that the decoder does not use the auto-regressive scheme often used in the translation task[28] during training, which greatly improves the efficiency of generation and reduces the error accumulation.

3.2. Training and Inference

Training. We train our model with three types of loss functions. The first term is used for supervising the pose parameters of generated dance. We use an L2 loss between predictions and ground-truth poses as $L_P = \sum_{t=1}^{T} ||\hat{p_t}^{pre} - p_t^{gt}||_2^2$. The second term forces the predicted joint coordinates, which are obtained by the SMPL model, to be close to the ground-truth. Additionally, we adopt the L2 loss on the joint velocity to constrain joint coordinates. We represent these two losses as L_{coo} and L_{vel}. The final term is introduced to preserve the consistency of generated sequences, and avoid convergence to the mean motion. By referring to [17], we employ the gram matrix loss L_{gra}. The total loss is the weighted summation of three terms: $L = \lambda_1 L_P + \lambda_2(L_{coo} + L_{vel}) + \lambda_3 L_{gra}$.

Inference. In the inference phase, our goal is to generate a long-term dance driven by an initial motion and a piece of long-term music. Due to the music is much longer than the initial motion, we need to generate dance by an iterative scheme. However, unlike other methods that only return one frame at a time, we directly output a sequence of T frames at each time.

4. EXPERIMENTS

4.1. Dataset & Metrics

Dataset. We conduct experiments on the AIST++[1] dataset, which is a large-scale 3D human dance motion dataset that contains a wide variety of 3D motion paired with 10 genre categories of music. There are 329 unique choreographies for training and 40 for testing.

Metrics. Following [1], we use the FID_g and FID_k to evaluate the motion quality of the generated dance, $Dist_g$ and $Dist_g$ to evaluate the generation diversity, $BeatAlign$ to evaluate motion-music correlation. In addition to the above metrics used in [1], we also follow [6, 29, 30] to train a dance genre classifier and add FID, ACC, $Diversity$ to evaluate the qualities of the generated dances.

4.2. Implementation Details

The encoders of music and initial motion are both 2 layers, while the decoder is 10 layers, and we set the hidden dim d to 512. We use the AdamW optimizer with an initial learning rate of 10^{-4}. We train our model for 1600 epochs and the learning rate decreases to 10^{-5} and 10^{-6} at 960 epochs and 1500 epochs. The batch size is set to 120 and the λ_1, λ_2, λ_3 are 5, 3, and 0.007 respectively. During training, we set the lengths of the input music sequence and initial motion sequence as 240 and 120, and the length of the output is 60. During inference, we continually generate 20 seconds of future motions in an auto-regressive manner and regress 60 frames in each step.

4.3. Main Results

Tab. 1 shows the performance comparisons between the proposed GCDG and state-of-the-art models on the AIST++ test set. The DR[5] only uses one transformer-based encoder to extract features for music clips, since it considers the initial pose as the condition rather than the motion sequence. And it adopts an LSTM-based decoder to generate dance in an auto-regressive manner, which only generates one frame of the pose at each step. The FACT[1] also uses two transformer encoders to extract music features and initial motion features like our GCDG, but it does not use the cross-attention to build the correlation between genre and contextual information. It is worth noting that the DR and FACT are reported

Arch	$BeatAlign \uparrow$	$FID_k \downarrow$	$FID_g \downarrow$	$Dist_k \uparrow$	$Dist_g \uparrow$	$FID \downarrow$	$ACC \uparrow$	$Div. \rightarrow$
TCN + LSTM	0.210	65.35	36.65	4.06	3.18	72.58	22.9	11.76
TCN + Trans.	0.219	62.06	33.56	5.05	3.59	58.56	43.4	12.56
Trans. + LSTM	0.220	59.37	30.64	4.95	3.00	60.10	30.0	12.35
All Trans.	**0.244**	**39.12**	**10.36**	**7.38**	**6.21**	**30.83**	**85.7**	**13.57**
w/out genre	0.241	56.70	15.14	6.59	6.04	35.91	72.3	12.30
w/out c-a	0.220	50.92	15.13	5.04	4.33	48.95	55.3	11.20

Table 2. Architectural ablation studies, where the 'c-a' means cross-attention.

Fig. 2. Qualitative results for three genres: 'BR' for Break, 'HO' for House, and 'JB' for Ballet Jazz. We sample the generated dances at a frequency of every ten frames.

in [1] and the DR* and FACT* are reproduced by us. It can be observed from the table, with the guidance of genre, GCDG achieves a significantly higher accuracy, which approaches to the ground-truth, and obtains the best in six out of the total seven metrics. Moreover, the DR and FACT only generate one frame of the pose at each step, while our model can generate 60 frames. Therefore, we reduce the inference time from 16.7 minutes to 0.9 minutes tested on a P40 GPU, which greatly improves the generation efficiency.

4.4. Ablation Studies

We conduct sufficient ablation studies to explore the potential of our model to generate realistic dance.

Architecture. We select the transformer as our encoder and decoder architecture, and we investigate several architectural choices. Specifically, we use TCN as the candidate encoder architecture and LSTM as the candidate decoder architecture. As shown in Tab. 2, our pure transformer-based architecture outperforms other mixed architectures by a large margin. Moreover, we also experiment to demonstrate the importance of genre information by removing the genre embedding. Benefiting from the guidance of genre, the FID_k, FID_g, FID, ACC have visible improvements, which proves the genre information helps a lot in the dance generation driven by music. Another advantage of GCTG is that we use the cross-attention to capture the strong correlation between the input features and *motion queries* so that we can generate more realistic future dances. As shown in Tab. 2, with the cross-attention, the ACC increases a large margin of 30.4.

Regressive length. In the inference phase, our model directly outputs 60 frames of the sequence. We select several choices of regressive length to validate the stability of GCDG. As shown in Tab. 3, since we use 60 frames to supervise the training process, there is a little performance drop but the overall performance is still acceptable.

Reg. Len.	$BeatAlign \uparrow$	$FID_k \downarrow$	$FID_g \downarrow$	$Dist_k \uparrow$	$Dist_g \uparrow$	$FID \downarrow$	$ACC \uparrow$	$Div. \rightarrow$
10	0.220	49.25	15.23	7.14	6.05	39.37	70.5	12.91
20	0.233	48.10	14.35	6.78	**6.45**	38.50	77.4	13.29
30	**0.245**	43.77	11.79	7.35	6.38	33.28	77.7	12.93
60	0.244	**39.12**	**10.36**	**7.38**	6.21	**30.83**	**85.7**	**13.57**

Table 3. The ablation studies of regressive length.

4.5. Qualitative Results.

In Fig. 2, we visualize several examples from our generations for three genres. Given a specific genre, our model is able to generate a corresponding dance. We can see the generated motions of different genres have different characteristics, for example, the movement of Ballet Jazz is larger than House, and the Break dance has a sense of rhythm.

5. CONCLUSION

In this paper, we propose a novel pure transformer-based model to generate long-term 3D dance from music with a given genre. This is the first work that attempts to use the genre information as a condition. Moreover, we propose the *motion queries* to absorb the contextual information from music and initial motion and generate the dance sequence in parallel. Benefiting from the guidance of the genre, our model can generate high-quality 3D dance on the AIST++ dataset and output 60 frames of motion sequence at once, which significantly improves the efficiency of generation.

6. REFERENCES

[1] R. Li, S. Yang, D. A Ross, and A. Kanazawa, "Ai choreographer: Music conditioned 3d dance generation with aist++," in *ICCV*, 2021.

[2] R. Fan, S. Xu, and W. Geng, "Example-based automatic music-driven conventional dance motion synthesis," *IEEE transactions on visualization and computer graphics*, vol. 18, no. 3, pp. 501–515, 2011.

[3] M. Lee, K. Lee, and J. Park, "Music similarity-based approach to generating dance motion sequence," *Multimedia tools and applications*, vol. 62, no. 3, pp. 895–912, 2013.

[4] F. Ofli, E. Erzin, Y. Yemez, and A M. Tekalp, "Learn2dance: Learning statistical music-to-dance mappings for choreography synthesis," *IEEE Transactions on Multimedia*, vol. 14, pp. 747–759, 2011.

[5] R. Huang, H. Hu, W. Wu, K. Sawada, and M. Zhang, "Dance revolution: Long sequence dance generation with music via curriculum learning," in *ICLR*, 2021.

[6] H.Y. Lee, X. Yang, M.Y. Liu, T.C. Wang, Y.D. Lu, M.H. Yang, and J. Kautz, "Dancing to music," in *NeurIPS*, 2019.

[7] Z. Ye, H. Wu, J. Jia, Y. Bu, W. Chen, F. Meng, and Y. Wang, "Choreonet: Towards music to dance synthesis with choreographic action unit," in *ACM MM*, 2020.

[8] 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, 2020.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[10] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021.

[11] W. Wang, E. Xie, X. Li, D. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *ICCV*, 2021.

[12] D. Zhang, H. Zhang, J. Tang, M. Wang, X. Hua, and Q. Sun, "Feature pyramid transformer," in *ECCV*, 2020.

[13] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *CVPR*, 2021.

[14] R. Bowden, "Learning statistical models of human motion," in *CVPRW*, 2000.

[15] A. Galata, N.Johnson, and D. Hogg, "Learning variable-length markov models of behavior," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 398–413, 2001.

[16] W. Mao, M. Liu, and M. Salzmann, "History repeats itself: Human motion prediction via motion attention," in *ECCV*, 2020.

[17] Q. Cui, H. Sun, and F. Yang, "Learning dynamic relationships for 3d human motion prediction," in *CVPR*, 2020.

[18] Q. Cui and H. Sun, "Towards accurate 3d human motion prediction from incomplete observations," in *CVPR*, 2021.

[19] Y. Yuan and K. Kitani, "Dlow: Diversifying latent flows for diverse human motion prediction," in *ECCV*, 2020.

[20] M. Li, S. Chen, Y. Zhao, Y. Zhang, Y. Wang, and Q. Tian, "Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction," in *CVPR*, 2020.

[21] A. Gopalakrishnan, A. Mali, D. Kifer, L. Giles, and A. G Ororbia, "A neural temporal model for human motion prediction," in *CVPR*, 2019.

[22] W. Mao, M. Liu, M. Salzmann, and H. Li, "Learning trajectory dependencies for human motion prediction," in *ICCV*, 2019, pp. 9489–9497.

[23] Z. Cao, H. Gao, K. Mangalam, Q. Cai, M. Vo, and J. Malik, "Long-term human motion prediction with scene context," in *ECCV*, 2020.

[24] X. Ren, H. Li, Z. Huang, and Q. Chen, "Self-supervised dance video synthesis conditioned on music," in *ACM MM*, 2020.

[25] J. P Ferreira, T. M Coutinho, T. L Gomes, J. F Neto, R. Azevedo, R. Martins, and E. R Nascimento, "Learning to dance: A graph convolutional adversarial network to generate realistic dance motions from audio," *Computers & Graphics*, vol. 94, pp. 11–21, 2021.

[26] B. McFee, C. Raffel, D. Liang, D. PW Ellis, M. McVicar, E. Battenberg O., and Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, 2015.

[27] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J Black, "Smpl: A skinned multi-person linear model," *ACM transactions on graphics*, vol. 34, no. 6, pp. 1–16, 2015.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017.

[29] C. Guo, X. Zuo, S. Wang, S. Zou, Q. Sun, A. Deng, M. Gong, and L. Cheng, "Action2motion: Conditioned generation of 3d human motions," in *ACM MM*, 2020.

[30] M. Petrovich, M. J Black, and G. Varol, "Action-conditioned 3d human motion synthesis with transformer vae," in *ICCV*, 2021.