Beat-It: Beat-Synchronized Multi-Condition 3D Dance Generation —Supplementary Materials—

Zikai Huang[1], Xuemiao Xu[1,3,4(✉)], Cheng Xu[2(✉)], Huaidong Zhang[1,3], Chenxi Zheng[1], Jing Qin[2], and Shengfeng He[5]

[1] South China University of Technology, China
xuemx@scut.edu.cn
[2] The Hong Kong Polytechnic University, Hong Kong SAR, China
cschengxu@gmail.com
[3] Guangdong Engineering Center for Large Model and GenAI Technology
[4] Guangdong Provincial Key Lab of Computational Intelligence and Cyberspace Information
[5] Singapore Management University, Singapore
https://zikaihuangscut.github.io/Beat-It/

1 Introduction

In this document, we first provide the details of the music and dance representations (Sec. 2). Next, we describe the architecture of the beat distance estimator (Sec. 3). Then, additional quantitative experiments are provided for a comprehensive evaluation of our methods (Sec. 4). Afterward, we show the qualitative results of our methods, including comparison results with existing approaches, ablation study, beat-synchronized keyframe-controlled dance generation, arbitrary beat-controlled dance generation, and the in-the-wild results (Sec. 5). Finally, we discuss the limitations and broader impact of our work (Sec. 6).

2 Representations of Music and Dance

In this section, we provide detailed explanations of the music and dance representations adopted in our method.

Music Representation. Previous works show that using the music feature extracted from large pre-trained models as music condition brings significant improvements to music-to-dance generation [7, 11]. Therefore, we adopt the same music feature representation as EDGE [11]. This music feature is then passed through a transformer encoder $\mathcal{E}_m$ to produce the final music embedding $\mathbf{e}_m \in \mathbb{R}^{L\times d}$.

Dance Representation. Similar to prior works [1, 2, 4, 5, 7, 10, 11], we adopt the rotation matrix to represent dance motions. In particular, we use the root translation δ and 6-DOF rotation representation r [12] for 24 joints in SMPL format [6] for each frame. Following EDGE [11], we also include a binary contact label g for the heel and toe of each foot. The entire pose representation is

therefore denoted as $\mathbf{x}^i = \{g, \delta, r\} \in \mathbb{R}^{D=4+3+144=151}$. Correspondingly, the keyframe condition embedding $\mathcal{X}^{ref} \in \mathbb{R}^{L\times 151}$ can be projected into a keyframe embedding $\mathbf{e}_r \in \mathbb{R}^{L\times d}$ via the keyframe encoder $\mathcal{E}_r$. To relieve the negative impacts caused by the sparse nature of the keyframes, we explicitly introduce positional context information by adding learnable embeddings of the nearest keyframe distance to $\mathbf{e}_r$ at the non-keyframe locations.

3 The Beat Distance Estimator

The beat distance estimator is used to estimate the nearest beat distance of each frame for loss calculation. It is pre-trained on a beat distance regression task. In particular, it comprises a 6-layer 4-head transformer-based encoder with hidden size of 128. Given a sequence of motion as input, it is first processed through forward kinematics to obtain the 3D joint positions. Then we get the joint velocity and feed it into the beat distance estimator, followed by an MLP, to predict the nearest beat distance of each frame.

4 Additional Quantitative Experiments

Here we provide quantitative comparison between our method and the state-of-the-art motion interpolation approach, and the performance of our method with different keyframe condition ratios, different single conditions and different condition combinations.

Comparison with Motion Interpolation Method. For reference, we also present the comparison between our method and the state-of-the-art motion interpolation method [8] quantitatively. But note that the motion-interpolation task accepts key poses as the only input condition, which is essentially different from our setting of dance generation dominated by music. The results in Tab. 1 indicate that it is not feasible to directly apply the motion interpolation method to render beat-synchronized dance motions due to its lack of music and beat guidance. In comparison, our method demonstrates superior capability of generating high-quality beat-synchronized dance motions.

Table 1: Quantitative comparison with Qin et al. [8] on AIST++ [5].

Methods	Quality		Diversity		Controllability
	PFC ↓	BAS ↑	Div_k →	Div_m →	BAP ↑
Ground Truth	1.338	0.384	9.773	7.212	-
Qin et al. [8]	3.790	0.182	9.7981	7.112	-
Ours	**0.966**	**0.661**	**9.660**	**7.248**	**0.793**

Different Ratios of Keyframes Condition. Additionally, we also report the performance of our method with varying ratios of keyframe conditions. The

results, presented in Tab. 2, reveal that our approach can still produce compelling outcomes even if the input keyframes are highly sparse (e.g., 5% or no keyframes).

Table 2: Quantitative results with different ratios of input keyframes on AIST++ [5].

Keyframes Ratio	Quality		Diversity		Controllability	
	PFC ↓	BAS ↑	Div_k →	Div_m →	KPD ↓	BAP ↑
Ground Truth	1.338	0.384	9.773	7.212	-	-
0%	1.157	0.644	11.298	7.310	-	0.782
5%	0.758	0.738	9.510	7.123	0.301	0.794
10%	0.966	0.661	9.660	7.248	0.306	0.793

Different Single Conditions. Our method supports single conditional generation by directly setting the other conditions to "null condition" during the training process. Tab. 3 presents the quantitative results of our model with different single conditions. We can observe that our music-only model has already shown significant advantages over the SOTAs. By integrating all three conditions, the overall performance of our full model is further enhanced, demonstrating considerable advantages over the variants with single input condition. Importantly, models with beat- or keyframe-only condition violate the original music-to-dance generation setting and generate dance sequences with severe freezing issue, thus leading to significantly lower PFC values.

Table 3: Quantitative comparison with different single conditions.

Methods	Quality		Diversity		Controllability	
	PFC ↓	BAS ↑	Div_k →	Div_m →	KPD ↓	BAP ↑
Ground Truth	1.338	0.384	9.773	7.212	-	-
FACT [5]	2.698	0.202	9.704	7.342	-	-
Bailando [9]	1.578	0.215	9.622	7.175	-	-
EDGE (keyframes) [11]	1.084	0.235	9.743	7.274	0.859	-
Ours (beats only)	**0.058**	0.557	10.078	4.383	-	0.586
Ours (keyframes only)	0.443	0.203	11.124	6.783	0.673	-
Ours (music only)	0.879	0.237	**9.782**	6.997	-	-
Ours (music + beat + keyframes)	0.966	**0.661**	9.660	**7.248**	**0.306**	**0.793**

Different Conditions Combinations. Our framework can be easily extended to support different conditions combinations by directly setting the omitted condition to "null condition" during the training process. Quantitative results of

our methods with various condition combinations are tabulated in Tab. 4. The results affirm the versatility and flexibility of our approach in supporting diverse condition combinations.

Table 4: Quantitative results under different combinations of conditions.

Methods	Quality		Diversity		Controllability	
	PFC ↓	BAS ↑	Div_k →	Div_m →	KPD ↓	BAP ↑
Ground Truth	1.338	0.384	9.773	7.212	-	-
music + keyframes	0.680	0.240	9.487	7.145	0.304	-
music + beats	1.157	0.644	11.298	7.310	-	0.782
music + keyframes + beats (Ours)	0.966	0.661	9.660	7.248	0.306	0.793

5 Qualitative Results

For a comprehensive evaluation of our method, we include all the qualitative results in the supplementary demo video and make them accessible on *our website* for ease of reference. We recommend viewing these materials with audio enabled for the best experience.

Comparison with Existing Methods. We compare our method with the state-of-the-art methods EDGE [11], Bailando [9], and FACT [5] on the AIST++ [5] dataset. Note that EDGE [11] is the one and only open-source keyframe-conditioned dance generation method. Bailando [9] utilizes 3D Cartesian space key points for dance representation, which are not directly applicable for rendering 3D characters. To ensure a fair comparison, we directly utilize skeleton stick figures for visualization. The results indicate a remarkable improvement in both the quality and controllability of our method compared to these existing techniques. Please refer to the supplementary demo video for the full comparison results.

Ablation Study. As shown in the qualitative results of the ablation study in the supplementary demo video. Once the hierarchical multi-condition fusion module is removed (w/o HF), a notable degradation in performance with less coherent generated dance movements can be observed. This is mainly attributed to the direct concatenation of multiple conditions, which introduces conflicts among conditions and hampers optimal model optimization. In contrast, our proposed hierarchical multi-condition fusion mechanism effectively alleviates conflicts between conditions and generates more informative conditions for better guidance. Additionally, the absence of the beat-aware mask dilation scheme (w/o BD) causes significant jittering in the generated motions and significantly impairs keyframe controllability and beat synchronization. The rationale behind this is the inherent dynamic adjustment capability of beat-aware dilation for the adaptive

expansion of attention masks, which is crucial for enhancing synchronization between keyframes and their associated beat frames, allowing the model to better leverage prior knowledge of beats. Omitting the beat alignment loss (w/o $\mathcal{L}_{beat}$) results in generation with inferior beat alignment, demonstrating the significance of the explicit supervision signals from beats on improving beat synchronization.

Beat-Synchronized Keyframe-Controlled Dance Generation. To further demonstrate the superiority of our method in beat-synchronized keyframe-controlled dance generation, we conducted experiments using identical music and beats while varying the keyframes. The results, showcased in the supplementary demo video, highlight the remarkable ability of our model to generate dance sequences that not only adhere to specified keyframes but also precisely align with given beat conditions. To visually depict beat synchronization, mean joint velocity curves of the generated dance motions are incorporated, aligning with the provided beats indicated by vertical dashed lines. A noteworthy feature of our method is its proficiency in producing motion beats, identified as local minimum velocity points, precisely aligned with the beat condition. This alignment serves as a clear indication of our method's advanced capability in achieving beat synchronization.

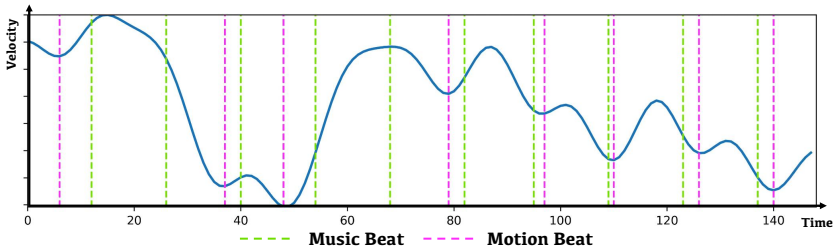

Fig. 1: Visualization of motion beat alignment given a beat condition not strictly aligned with the musical beats.

Arbitrary Beat-Controlled Dance Generation. In practical choreography, the arrangement of dance motion beats is highly diverse. Apart from the straightforward incorporation of a subset of music beats as motion beats, dance generation without strict alignment with the musical beats is also needed in some special circumstances (e.g., the synchrony between the musical beats and dance motions is often less pronounced in ballet.). Thanks to the explicit disentanglement of beat conditions, our approach enables dance generation with diverse motion beats. These beats can be controlled by any specified beat conditions, offering flexibility beyond strict adherence to the musical beats. To demonstrate this, we illustrate an example where the beat condition is not strictly aligned with the music beats. The generated dance motions, depicted in Fig. 1, confirm the effectiveness of our method in creating convincing results based on the specified motion beat condition, rather than by strict synchronization with the music

beats. This underscores the significant generalizability of our approach to various dance genres with diverse beat patterns.

In-the-Wild Results. To further demonstrate the generalization ability of our method, we also evaluate our method on in-the-wild music videos. Specifically, we randomly select different samples from the AIOZ-GDANCE [3] dataset for evaluation. This dataset exhibits much higher diversity than our training data of AIST++ [5]. Our method can generate high-quality dance motions that are well-aligned with the given beats and keyframes.

6 Limitation and Broader Impact

Limitation. Similar to all diffusion-based methods (e.g., EDGE [11]), a general limitation of our method is the inference speed. Here we show the model size and runtime comparison of different methods for generating a 150-frame dance sequence in Tab. 5. Although our method involves increased computation time due to our multi-modal fusion and diffusion process, it establishes the first highly controllable dance generation paradigm, significantly surpassing existing methods in terms of generation quality, flexibility, and controllability, which are crucial for practical application. We believe our extra overhead is acceptable and can be further reduced via a more advanced sampling strategy or model distillation. While our method can be extended to dances with conditions in the wild, the sliding foot issue may arise in some extremely complex scenarios. The integration of a more sophisticated loss function based on physical kinematics holds promise for further enhancing the overall quality of generated dance sequences.

Table 5: Complexity comparison of different methods.

Metrics	FACT [5]	Bailando [9]	EDGE [11]	Ours
Params/M	120.4	173.4	49.5	106.4
Runtime/s	22.4	0.3	1.6	6.0

Broader Impact. Our research has potential positive impacts on the dance community, providing choreographers with a new tool for fine-grained dance creation, thereby supporting artistic endeavors. On the other hand, it may also lead to adverse societal impacts. Automated dance generation poses risks of copyright infringement, cultural appropriation, and devaluation of originality in dance creation. Addressing these issues requires ethical development within the dance community, emphasizing cultural sensitivity, and maintaining appreciation for human creativity.

References

1. Kim, J., Oh, H., Kim, S., Tong, H., Lee, S.: A brand new dance partner: Music-conditioned pluralistic dancing controlled by multiple dance genres. In: CVPR. pp. 3490–3500 (2022) 1
2. Le, N., Pham, T., Do, T., Tjiputra, E., Tran, Q.D., Nguyen, A.: Music-driven group choreography. In: CVPR. pp. 8673–8682 (2023) 1
3. Le, N., Pham, T., Do, T., Tjiputra, E., Tran, Q.D., Nguyen, A.: Music-driven group choreography. CVPR (2023) 6
4. Li, B., Zhao, Y., Zhelun, S., Sheng, L.: Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In: AAAI. vol. 36, pp. 1272–1279 (2022) 1
5. Li, R., Yang, S., Ross, D.A., Kanazawa, A.: Ai choreographer: Music conditioned 3d dance generation with aist++. In: ICCV. pp. 13401–13412 (2021) 1, 2, 3, 4, 6
6. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 851–866 (2023) 1
7. Qi, Q., Zhuo, L., Zhang, A., Liao, Y., Fang, F., Liu, S., Yan, S.: Diffdance: Cascaded human motion diffusion model for dance generation. In: ACM MM. pp. 1374–1382 (2023) 1
8. Qin, J., Zheng, Y., Zhou, K.: Motion in-betweening via two-stage transformers. ACM TOG (2022) 2
9. Siyao, L., Yu, W., Gu, T., Lin, C., Wang, Q., Qian, C., Loy, C.C., Liu, Z.: Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In: CVPR. pp. 11050–11059 (2022) 3, 4, 6
10. Sun, J., Wang, C., Hu, H., Lai, H., Jin, Z., Hu, J.F.: You never stop dancing: Non-freezing dance generation via bank-constrained manifold projection. NeurIPS **35**, 9995–10007 (2022) 1
11. Tseng, J., Castellon, R., Liu, K.: Edge: Editable dance generation from music. In: CVPR. pp. 448–458 (2023) 1, 3, 4, 6
12. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: CVPR. pp. 5745–5753 (2019) 1