MoFusion: A Framework for Denoising-Diffusion-based Motion Synthesis —Supplementary Material—

Rishabh Dabral[1] Muhammad Hamza Mughal[1,2] Vladislav Golyanik[1] Christian Theobalt[1]

[1]Max Planck Institute for Informatics, SIC [2]Saarland University

This supplementary document provides additional ablation results (Sec. A), additonal implementation details (Sec. B) and finally, addition details on the user study (Sec. C),

A. Additional Results

Perfomance with different training objectives: In this work, we present different training objectives to optimize the MoFusion Architecture. We measure the effect of different losses to gauge how well dance sequences align with music beats. Our results in Table 1 show that as we incorporate different kinematic losses, the Beat Alignment Score improves which demonstrates better music-to-dance synthesis quality.

By addition of kinematic losses in MoFusion framework, we observe better performance than the state of the art in Beat Alignment Score (BAS) and our best BAS is even better than the ground truth data (0.237). This is due to better generation quality which matches music beats across the motion sequence. Note, $\mathcal{L}_{da}$ with generation length 10 seconds (first row in our method's results) is the configuration we use to compare results with the state-of-the-art methods in Table 1 in the main draft. However, for ablation study, we use a max. motion generation length of 20 seconds as the ablations are clearer in this setting.

Music-to-Dance Synthesis with Seed Motion Input: Previous methods [3, 4, 8] synthesize dance motion with a seed pose as input which guides the training process and dance generation process as well. However, we do not train our method with seed motion as input. Instead, we synthesize the dance motion from scratch which is solely conditioned on melspectrogram of dance music. To test performance of our model with a seed sequence, we test the model performance by running reverse-diffusion process at test time with a *seed sequence*. The seed sequence consists of first two seconds of ground-truth motions and we predict a motion sequence in correspondence with first two seconds of input. We follow [4] while choosing the length of seed sequence as 2 seconds.

As observed in Table 2, there is an overall increase in Beat Alignment Score due to addition of ground truth data and here, we again observe a trend of higher performance in BAS as we add more losses. It is noteworthy that we do not retrain with seed motion input. Rather, we use a pretrained dance synthesis model to perform inference with seed input. Our supplementary video shows results for seed input synthesis wherein we can observe smooth transition from seed input sequence to forecast dance sequence.

Method	BAS
Ground Truth	0.237
Li *et al*. [3]	0.160
DanceNet [9]	0.143
Dance Revolution [2]	0.195
AI Choreographer [4]	0.221
Bailando [8]	0.233
Ours($\mathcal{L}_{da}$) - 10 sec	0.230
Ours($\mathcal{L}_{da}$)	0.234
Ours($\mathcal{L}_{da} + \mathcal{L}_m$)	0.242
Ours($\mathcal{L}_{da} + \mathcal{L}_m + \mathcal{L}_s + \mathcal{L}_a$)	0.252

Table 1. Comparison between performance on BAS by training our network with different training objectives. Here, "10 sec" refers to length of motion generation. All other models of MoFusion had a generation length of 20 seconds.

Seed Motion	Length (sec)	BAS
$\mathcal{L}_{da}$	10	0.257
$\mathcal{L}_{da}$	20	0.269
$\mathcal{L}_{da} + \mathcal{L}_m$	20	0.264
$\mathcal{L}_{da} + \mathcal{L}_m + \mathcal{L}_s + \mathcal{L}_a$	20	0.265

Table 2. Performance comparison of different trained networks at inference with a seed sequence. Here, Length refers to motion generation length in seconds.

B. Implementation Details

Diffusion Model: We use 1000 diffusion steps as T for the diffusion process and change the variances β_t linearly from

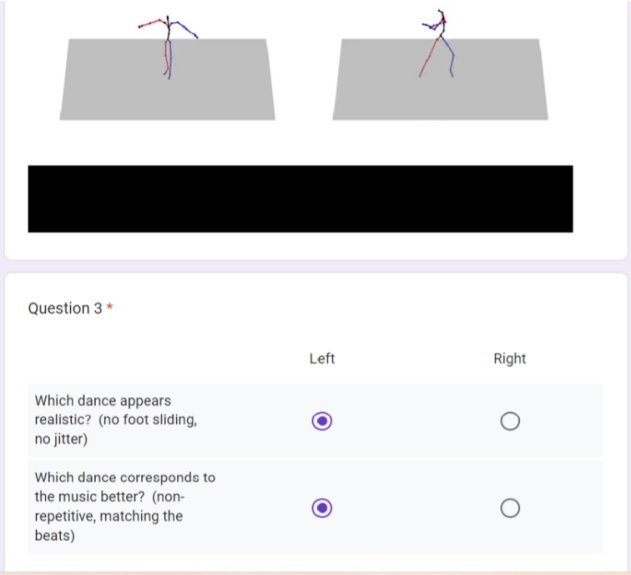

Figure 1. A screenshot of the user study. The participants were asked 18 such questions.

0.0001 to 0.02. For training our framework, we use single NVIDIA RTX A40 with task-specific batch sizes.

Music-to-Dance Synthesis: We use a latent dimension of 1024 in the audio encoder which takes melspectrogram as input. In 1D-UNet model, we use cross-modal transformer blocks with 16 attention heads and a cross attention dimension of 1024. We employ AdamW [6] as an optimizer with a learning rate of 5×10^{-4}. Moreover, we use a batch size of 32 for optimization.

To represent motion, we use 24 joint positions from SMPL model data which we extract from AIST++ Dataset [4]. We opt to train our model on 3D joint positions as joint angle representation performed worse during our experiments. As we observed from our experiments, our framework can also be trained with other joint position representations like COCO Keypoints format [5].

Text-to-Motion Synthesis: For text encoder, we use CLIP ViTB/32 [7] with a latent dimension of 512 for text encoding. The batch-size used is 128 and the network is trained with AdamW optimizer and a learning rate of 0.0002 is used. Following the conventional diffusion wisdom, we use a warm-up schedule of 500 iterations in the beginning. As discussed in the main draft, we use 22 joint positions of SMPL-X model data to represent motion. This is extracted from SMPL data given in HumanML3D dataset [1].

C. Details of User Study

We conducted the user study with 40 participants, with each participant answering 18 questions which asked the users to compare our synthesis results with other state-of-the-art methods. The participants took 8-10 minutes to submit their responses. We provide a snapshot of the interface in Fig. 1

References

[1] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *CVPR*, 2022. 2

[2] Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Daxin Jiang. Dance revolution: Long-term dance generation with music via curriculum learning. In *ICLR*, 2021. 1

[3] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. Learning to generate diverse dance motions with transformer. *ArXiv*, abs/2008.08171, 2020. 1

[4] Ruilong Li, Sha Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*, 2021. 1, 2

[5] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *arXiv*, 2014. 2

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 2

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. 2021. 2

[8] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation via actor-critic gpt with choreographic memory. In *CVPR*, 2022. 1

[9] 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.*, 2022. 1