Appendix for
AI Choreographer: Music Conditioned 3D Dance Generation with AIST++

Ruilong Li[*1] Shan Yang[*2] David A. Ross[2] Angjoo Kanazawa[2,3]
[1]University of Southern California [2]Google Research [3]University of California, Berkeley

1. AIST++ Dataset Details

3D Reconstruction Here we describe how we reconstruct 3D motion from the AIST dataset. Although the AIST dataset contains multi-view videos, they are not calibrated meaning their camera intrinsic and extrinsic parameters are not available. Without camera parameters, it is not trivial to automatically and accurately reconstruct the 3D human motion. We start with 2D human pose detection [2] and manually initialized the camera parameters. On this we apply bundle adjustment [3] to refine the camera parameters. With the improved camera parameters, the 3D joint locations $\hat{J} \in \mathbb{R}^{M\times 3}(M = 17)$ are then triangulated from the multi-view 2D human pose keypoints locations. During the triangulation phase, we introduce temporal smoothness and bone length constraints to improve the quality of the reconstructed 3D joint locations. We further fit SMPL human body model [1] to the triangulated joint locations $\hat{J}$ by minimizing an objective with respect to $\Theta = \{\theta_i\}_i^M$, global scale parameter α and global transformation γ for each frame: $\min_{\Theta,\gamma,\alpha} \sum_{i=1}^{M} \|\hat{J} - J(\theta_i, \beta, \gamma, \alpha)\|_2$. We fix β to the average shape as the problem is under-constrained from 3D joint locations alone.

Statistics We show the detailed statistics of our AIST++ dataset in Table 1. Thanks to the AIST Dance Video Database [4], our dataset contains in total 5.2-hour (1.1M frame, 1408 sequences) of 3D dance motion accompanied with music. The dataset covers 10 dance genre (shown in Figure 2) and 60 pieces of music. For each genre, there are 6 different pieces of music, ranging from 29 seconds to 54 seconds long, and from 80 BPM to 130 BPM (except for House genre which is 110 BPM to 135 BPM). Among those motion sequences for each genre, 120 (85%) of them are *basic* choreographies and 21 (15%) of them are *advanced*. Advanced choreographies are longer and more complicated dances improvised by the dancers. Note for the *basic* dance motion, dancers are asked to perform the same choreography on all the 6 pieces of music with different speed to follow different music BPMs. So the total *unique* choreographies in for each genre is $120/6 + 21 = 41$. In our experiments we split the AIST++ dataset such that there is no overlap between *train* and *test* for both music and choreographies (see Sec. 5.2.1 in the paper).

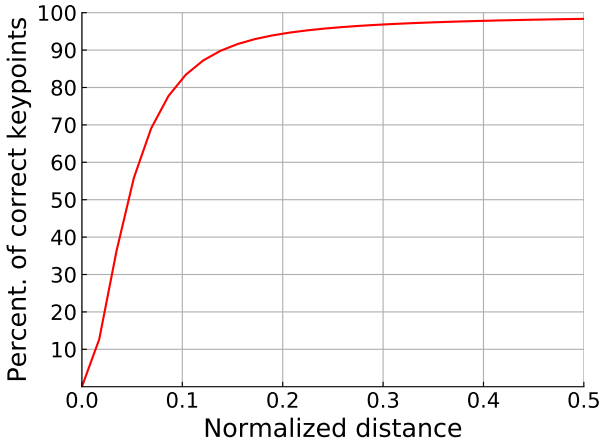

Figure 1: **PCKh Metric on AIST++.** We analyze the PCKh (percentage of correct keypoints) metric between re-projected 2D keypoints and detected 2D keypoints on AIST++. Averaged PCKh@0.5 is 98.4% on all joints shows that our reconstructed 3D keypoints are highly consistent with the predicted 2D keypoints.

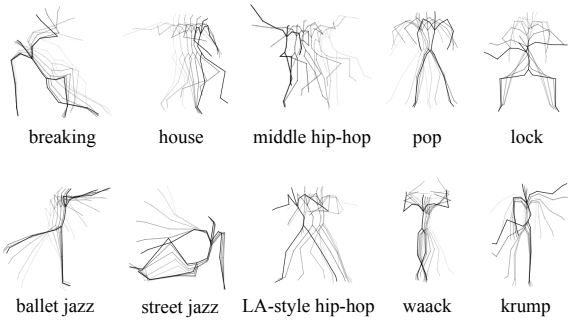

Figure 2: **AIST++ Motion Diversity Visualization.** Here we show the 10 types of 3D human dance motion in our dataset.

Genres	Musics	Music Tempo	Motions	Choreographs	Motion Duration (sec.)	Total Seconds
ballet jazz	6	80 - 130	141		7.4 - 12.0 basic / 29.5 - 48.0 adv.	1910.8
street jazz	6	80 - 130	141		7.4 - 12.0 basic / 14.9 - 48.0 adv.	1875.3
krump	6	80 - 130	141		7.4 - 12.0 basic / 29.5 - 48.0 adv.	1904.3
house	6	110 - 135	141		7.1 - 8.7 basic / 28.4 - 34.9 adv.	1607.6
LA-style hip-hop	6	80 - 130	141	85% basic +	7.4 - 12.0 basic / 29.5 - 48.0 adv.	1935.8
middle hip-hop	6	80 - 130	141	15% advanced	7.4 - 12.0 basic / 29.5 - 48.0 adv.	1934.0
waack	6	80 - 130	140		7.4 - 12.0 basic / 29.5 - 48.0 adv.	1897.1
lock	6	80 - 130	141		7.4 - 12.0 basic / 29.5 - 48.0 adv.	1898.5
pop	6	80 - 130	140		7.4 - 12.0 basic / 29.5 - 48.0 adv.	1872.9
break	6	80 - 130	141		7.4 - 12.0 basic / 23.8 - 48.0 adv.	1858.3
total	60		1408			18694.6

Table 1: **AIST++ Dataset Statistics**. AIST++ is built upon a subset of AIST database [4] that contains single-person dance.

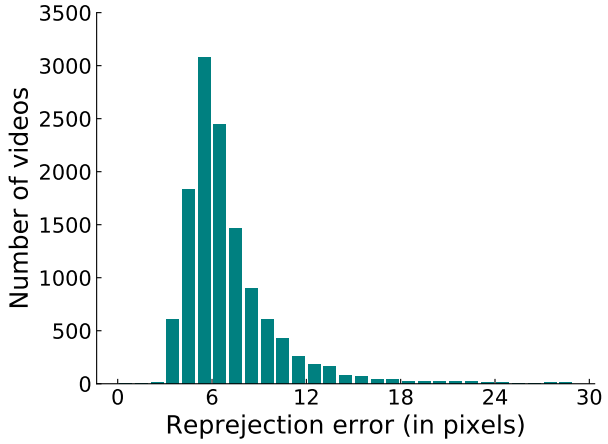

Figure 3: **MPJPE-2D Distribution on AIST++.** We analyze the distribution of MPJPE-2D among all video sequences on 1920x1080 resolution. MPJPE-2D is calculated between the re-projected 2D keypoints and the detected 2D keypoints. Over 86% of the videos have less than average 10 pixels of error.

Validation As described in Sec. 5.1 in the paper, we validate the quality of our reconstructed 3D motion by calculating the overall MPJPE-2D (in pixel) between the re-projected 2D keypoints and the detected 2D keypoints with high confidence (> 0.5). We provide here the distribution of MPJPE-2D among all video sequences (Figure 3). Moreover, we also analyze the PCKh metric with various thresholds on the AIST++, which measures the consistency between the re-projected and detected 2D keypoints. Averaged PCKh@0.5 is 98.4% on all joints shows that our reconstructed 3D keypoints are highly consistent with the detected 2D keypoints.

2. User Study Details

2.1. Comparison User Study

As mentioned in Sec. 5.2.5 in the main paper, we qualitatively compare our generated results with several baselines in a user study. Here we describe the details of this user study. Figure 5 shows the interface that we developed for this user study. We visualize the dance motion using stick-man and conduct side-by-side comparison between our generated results and the baseline methods. The left-right order is randomly shuffled for each video to make sure that the participants have absolutely no idea which is ours. Each video is 10-second long, accompanied with the music. The question we ask each participant is *"which person is dancing more to the music? LEFT or RIGHT"*, and the answers are collected through a Google Form. At the end of this user study, we also have an exit survey to ask for the dance experience of the participants. There are two questions: *"How many years have you been dancing?"*, and *"How often do you watch dance videos?"*. Figure 4 shows that our participants ranges from professional dancers to people rarely dance, with majority with at least 1 year of dance experience.

References

[1] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):248, 2015. 1

[2] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4903–4911, 2017. 1

[3] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *International workshop on vision algorithms*, pages 298–372. Springer, 1999. 1

[4] 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*, pages 501–510, Delft, Netherlands, Nov. 2019. 1, 2

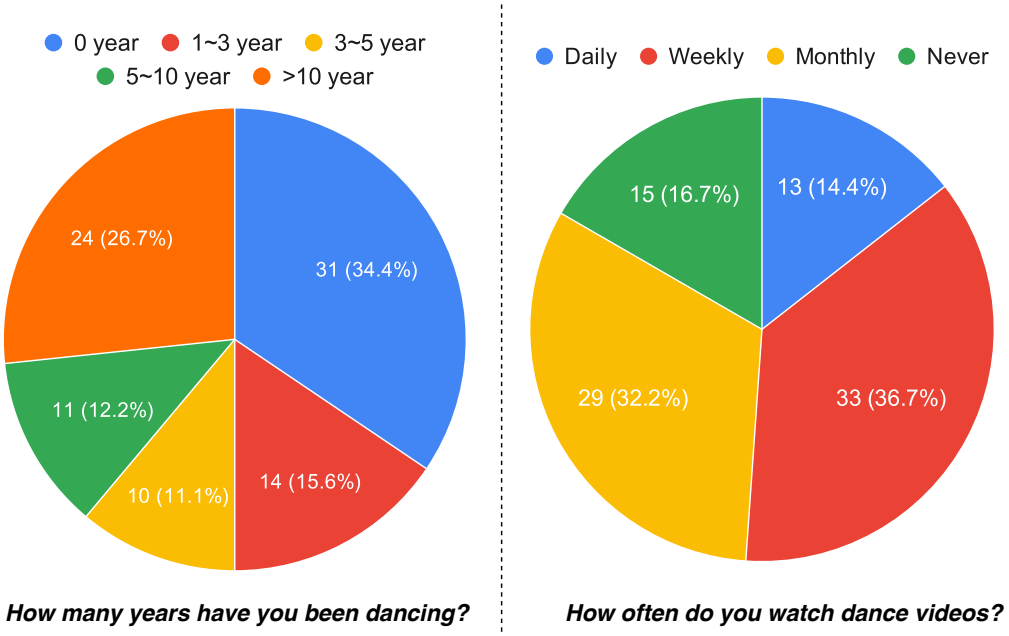

Figure 4: **Participant Demography of the *Comparison* User Study.**

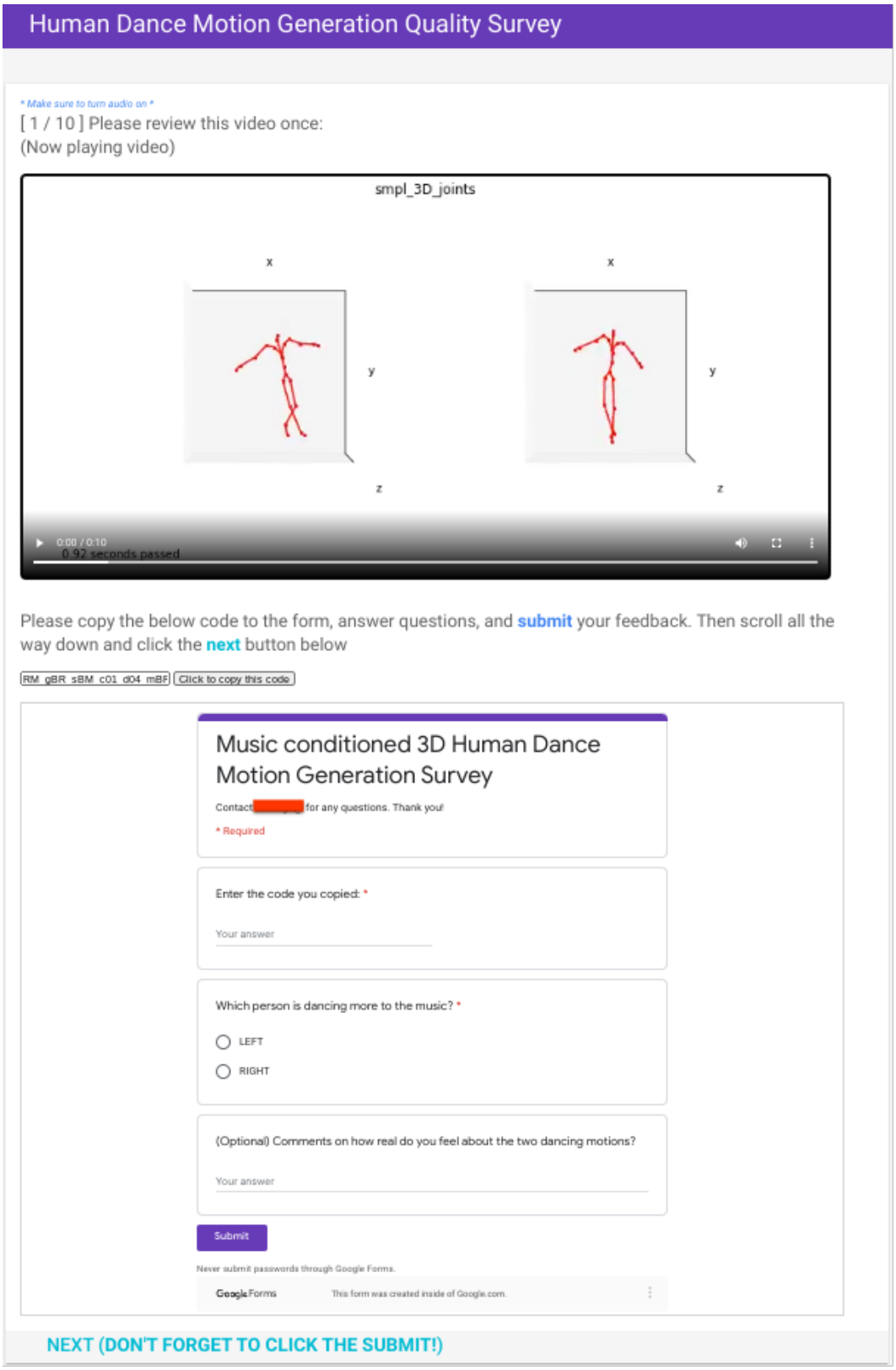

Figure 5: **User study interface.** The interface of our User study. We ask each participant to watch 10 videos and answer the question *"which person is dancing more to the music? LEFT or RIGHT"*.