Bipartite Graph Diffusion Model for Human Interaction Generation Supplementary Material

1. Additional training details

The following values are the same for both NTU and DuetDance. We use train the network for 1000 diffusion steps with noise sampled in a linear manner and a learning rate of 0.0001. The trained Bipartite Graph Interaction Transformer contains 8 identical layers and the multihead attention uses 8 attention heads. The Text encoder is a simple Transofoermer encoder with 4 layers and 4 attention heads.

2. Additional Qualitative Results

NTU-26. Figures 1 and 2 show visuals of sequences generated for the "High-five" and "Kicking" classes, respectively. For "High-five", ACTOR also generates a low-intensity motion, and both characters raise their hand but do not perform a high-five. Both MotionDiffuse and our method generate a high-five but MotionDiffuse shows noise and the hands of both characters stay far from each other. The ground truth once again contains noise that is not present in our generation. For the "Kicking" class, ACTOR does not generate any motion for either character. MotionDiffuse generates the red character as being kicked but does not generate the blue person kicking. On the other hand, our method generates both the kicking motion and the other character being kicked like the ground truth. In the ground truth, we can see that the leg is never fully extended during the kick. This is common for this class. The NTU-RGB+D dataset is captured using a Kinect camera and has difficulties capturing the legs due to the positioning of the camera and occlusion during interactions. This shows the kind of noise present in the original data again. Overall, we see that our motions are more realistic, temporally, and spatially coherent, and manage well to keep the interaction coherent.

3. Video Results

Videos of the qualitative results presented in the main paper and this supplementary material can be found at https://github.com/CRISTAL-3DSAM/BiGraphDiff in examples directory. In "Cheer_and_drink.mp4", "High_five.mp4", "Kicking.mp4", and "salsa.mp4" we show a comparison of the GT, the two state-of-the-art methods and BiGraphDiff on the four classes presented in the main paper and the supplementary material. In file "very_long_rumba.mp4", we show the video for very long-term generation on the rumba class presented in the main paper.

4. Additional animated results

At https://github.com/CRISTAL-3DSAM/BiGraphDiff in 'examples directory are additional animated results produced by BiGraphDiff. The visuals are in ".gif" format and can take some time to load for the longer sequences. "BiGraphDiff_NTU.zip" contains the results on the NTU-26 dataset (100 samples per class) and "BiGraphDiff_DuetDance.zip" contains the results on the DuetDance dataset (25 samples per class).

5. Code

We provide the code and pretrained models at https://github.com/CRISTAL-3DSAM/BiGraphDiff. The code is for the training and testing of BiGraphDiff on NTU-26 and DuetDance, the quantitative evaluations of the generated sequence, and the visualization of the generated sequences. We provide the necessary formatted data to run the code for NTU. For DuetDance, the dataset is not public but can be shared upon request to the authors of [1]. We provide the code to format the raw data from [1] for use with BiGraphDiff. More details are provided in the README file provided with the code.

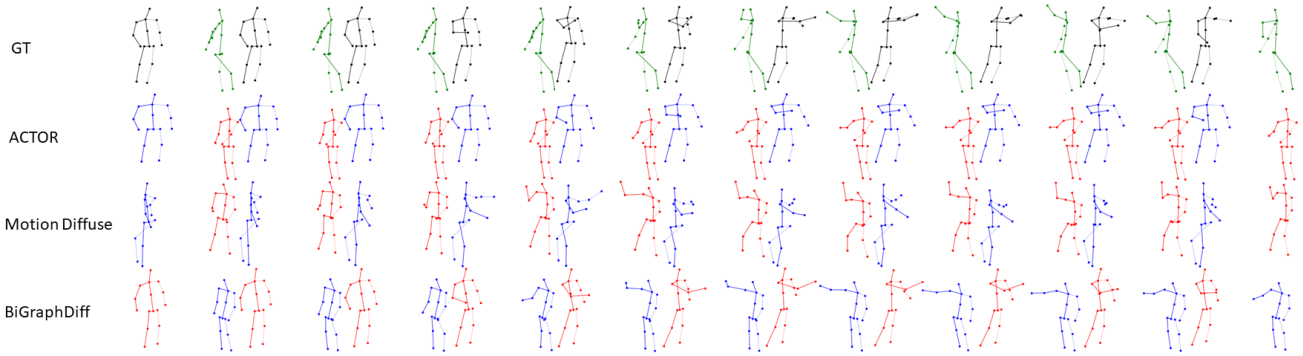

Figure 1. Examples of diverse motion generation for a given text prompt “High-five” action from NTU.

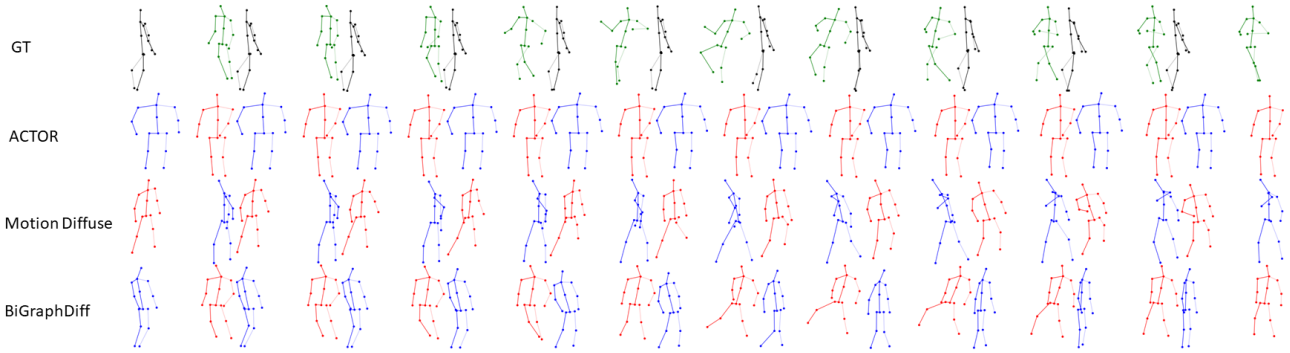

Figure 2. Examples of diverse motion generation for a given text prompt “Kicking” action from NTU.

References

[1] Jogendra Kundu, Himanshu Buckchash, Priyanka Mandikal, Rahul M V, Anirudh Jamkhandi, and R. Babu. Cross-conditioned recurrent networks for long-term synthesis of inter-person human motion interactions. In *WACV*, 2020. 1