Appendix

A. Datasets

CMU-Mocap [1] collects high-quality motion sequences of $1 \sim 2$ persons by a motion capture system. For other existing multi-person interaction datasets, there are $2 \sim 3$ persons in MuPoTS-3D [4] and 2 persons in 3DPW [6] in each scene. These two datasets, which are both captured by a multi-view camera system, contain more unnatural poses than CMU-Mocap.

CMU-Mocap (UMPM). Based on the sampling strategy in [7], we can obtain about 20 thousand 1-person sequences and only a thousand 2-person interaction sequences containing a large number of repetitive actions. Wang *et al.* [7] mixes the 2-person sequences with the 1-person sequences into a scene with 3 people, for a total of 6000 training sequences and 800 testing sequences. Such a dataset setting is severely unreasonable for training and evaluation because of its low diversity.

UMPM Benchmark [5] is a collection of video recordings together with a ground truth based on motion capture data, including 2-person interaction sequences of 10 action categories. With a greater sampling rate, we can get over 10000 interaction sequences, which is 7 times larger than CMU-Mocap. To provide a training and testing dataset with more diverse motions, particularly interaction motions, the UMPM dataset is utilized to expand CMU-Mocap dataset for 13000 training sequences and 3000 testing sequences. Here, we refer to this augmented dataset as CMU-Mocap (UMPM).

Mix1 and Mix2. To evaluate prediction performance in a crowd scenario with more individuals, we combined MuPoTS-3D, 3DPW, and test data from CMU-Mocap and UMPM into two datasets, Mix1 and Mix2. There are 6 persons in the Mix1 dataset, which consists mainly of multi-person interactive motion sequences, and 10 persons in the Mix2 dataset, which includes some individuals who have no or low interaction with others. Each mixed dataset contains 1000 motion sequences and lasts for 75 frames.

B. Data Preprocessing

We follow the preprocessing in [7] to choose 15 human body joints from different datasets as shown in Fig. 1, considering that these data have different skeleton representations. Each motion sequence is sampled at 25 FPS and contains 75 frames. Besides, we adopt the same random initialization and scale operation of [7] to ensure all the individuals appear in each scene.

C. Implementation Details about Baselines

For the single-person based methods (HRI [3] and MSR [2]), we reshape each sequence by flattening operations across individuals and batches to ensure each sequence only contains a single individual before feeding it into the model. For the multi-person based method (MRT [7]), we take the same setting as our method to input the multi-person motion sequence during training. All these methods are trained for 50 epochs with a batch size of 32.

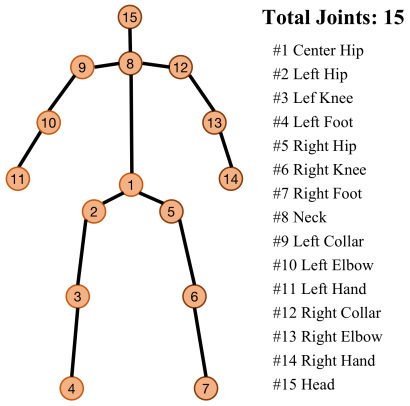

Figure 1. Visualization of the skeleton after preprocessing from different datasets.

		MuPoTS-3D ($2 \sim 3$ persons)			Mix1 (6 persons)			Mix2 (10 persons)		
	Method	1.0s	2.0s	3.0s	1.0s	2.0s	3.0s	1.0s	2.0s	3.0s
JPE	HRI [3]	204	366	491	145	268	425	147	255	391
	MSR [2]	208	390	525	143	297	455	162	340	507
	MRT* [7]	223	401	548	154	301	454	195	379	550
	Ours*	**198**	**345**	**485**	**131**	**261**	**399**	**125**	**241**	**357**
APE	HRI [3]	138	182	208	96	128	155	103	138	168
	MSR [2]	139	190	219	92	133	155	110	164	196
	MRT* [7]	169	234	270	111	159	188	118	179	220
	Ours*	**134**	**178**	**200**	**85**	**124**	**143**	**92**	**135**	**155**
FDE	HRI [3]	164	326	451	107	224	380	106	206	341
	MSR [2]	169	345	477	106	254	411	115	282	447
	MRT* [7]	194	379	530	117	264	418	160	339	511
	Ours*	**159**	**311**	**430**	**97**	**216**	**353**	**86**	**189**	**303**

Table 1. Results of JPE, APE and FDE (in mm) on different datasets. We compare our method with the previous SOTA methods for long-term predictions ($1.0\text{s} \sim 3.0\text{s}$). Best results are shown in boldface. (* means multi-person motion prediction method.)

D. Supplementary Quantitative and Qualitative Results

Quantitative Results. In Tab. 1, we supplement quantitative results of long-term prediction ($1.0\text{s} \sim 3.0\text{s}$) for other datasets, including MuPoTS-3D [4], Mix1 and Mix2. It can be seen that our TBIFormer is still superior to the other baselines in 3 metrics. We also supplement ablation studies on other different datasets for the effectiveness of TBIFormer's key components, as shown in Tab. 2. In addition, to investigate the effect of the different number of atten-

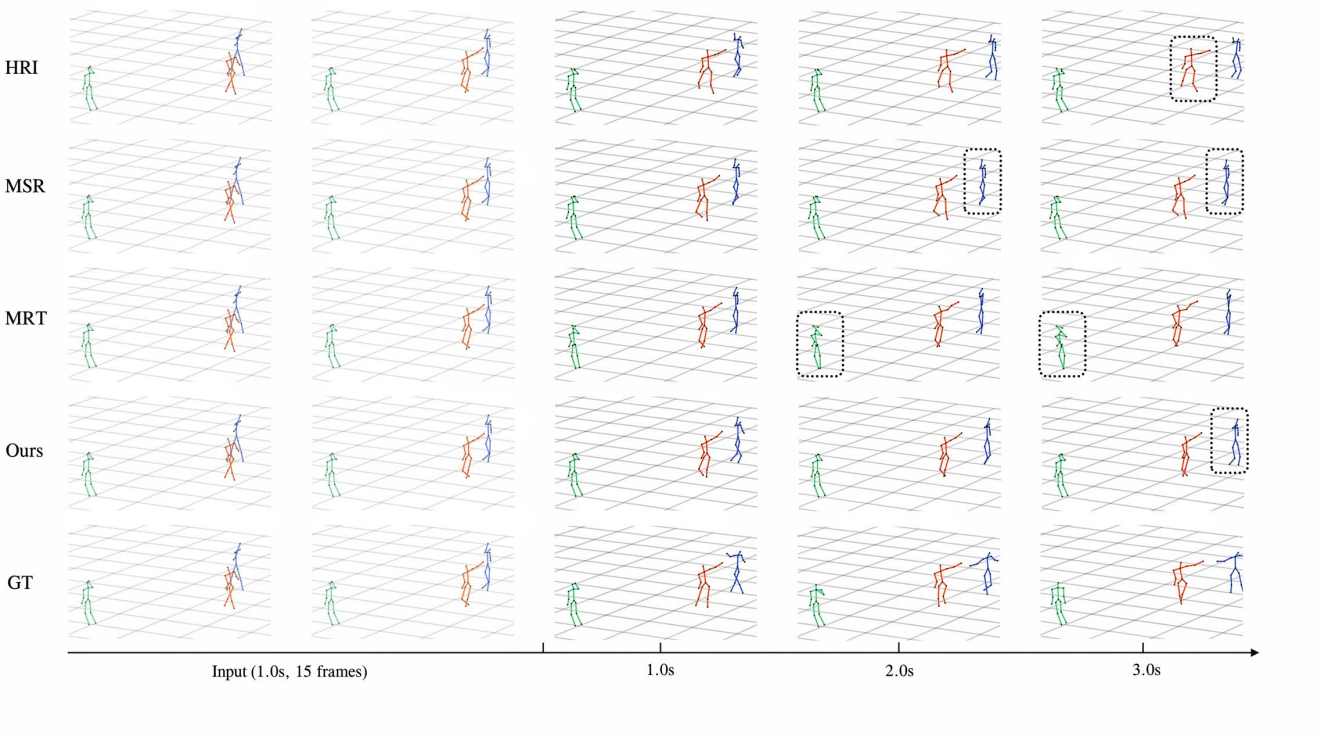

Figure 2. Qualitative comparison with the baselines and the ground truth on a sample of the CMU-Mocap (UMPM) dataset. The left two columns are inputs, and the right three columns are predictions.

		MuPoTS-3D (2 ∼ 3 persons)			Mix1 (6 persons)			Mix2 (10 persons)		
	Method	0.2s	0.6s	1.0s	0.2s	0.6s	1.0s	0.2s	0.6s	1.0s
JPE	w/o TBPM	70	210	339	**34**	**119**	212	**33**	122	207
	w/o IE,TRPE	69	212	337	**34**	125	213	36	122	204
	w/o TRPE	69	210	335	**34**	122	211	36	120	200
	TRPE → EuPE	70	209	334	35	123	212	37	121	204
	w/o SBI-MSA	76	214	355	45	131	224	45	135	223
	Full	**66**	**200**	**319**	**34**	121	**209**	34	**118**	**198**
APE	w/o TBPM	64	144	182	**26**	83	115	**26**	90	127
	w/o IE,TRPE	63	138	178	28	85	115	32	95	130
	w/o TRPE	62	136	172	28	83	114	31	94	128
	TRPE → EuPE	62	135	171	28	82	114	32	93	126
	w/o SBI-MSA	73	152	182	37	100	127	41	109	145
	Full	**60**	**132**	**170**	28	**81**	**113**	30	**89**	**124**
FDE	w/o TBPM	52	**163**	287	24	90	175	22	83	155
	w/o IE,TRPE	53	172	296	23	92	173	23	88	160
	w/o TRPE	53	170	293	**22**	91	170	22	86	157
	TRPE → EuPE	52	169	290	**22**	91	171	22	85	156
	w/o SBI-MSA	52	**163**	303	31	102	188	33	107	180
	Full	**49**	**163**	**277**	23	**89**	**168**	**21**	**81**	**151**

Table 2. Ablation studies of TBIFormer on different datasets. We compare our full method with the its variants for short-term predictions. Best results are shown in boldface.

tion layers, we conduct experiments using 1, 3, 5 layers respectively on CMU-Mocap (UMPM) with results in Tab. 3, where the 3-layer for TBIFormer block and Transformer decoder is more suitable for our prediction.

Number of Layers	0.2s	0.6s	1.0s	Overall
1-layer	33	113	187	111
3-layer (ours)	**30**	**109**	**182**	**107**
5-layer	34	113	188	112

Table 3. Ablation study for the different number of attention layers in TBIFormer and Transformer decoder on CMU-Mocap (UMPM) with the results of JPE.

Qualitative Results. We supplement the qualitative results for long-term prediction (1.0s ∼ 3.0s), as shown in Fig. 2. We compare our model with other baselines, *i.e.* HRI [3], MSR [2] and MRT [7]. Our results are much smoother and more natural, and they are closer to the ground truth. For some extreme actions, our method may generate freezing motions of some body parts. More qualitative results for complex scenarios, *e.g.*, on the Mix2 dataset (10 persons), could be found in the supplementary video.

References

[1] CMU-Graphics-Lab. Cmu graphics lab motion capture database. http://mocap.cs.cmu.edu/, 2003. 1

[2] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Msr-gcn: Multi-scale residual graph convolu-

tion networks for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11467–11476, October 2021. 1, 2

[3] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*, pages 474–489. Springer, 2020. 1, 2

[4] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018. 1

[5] NP Van der Aa, Xinghan Luo, Geert-Jan Giezeman, Robby T Tan, and Remco C Veltkamp. Umpm benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction. In *2011 IEEE international conference on computer vision workshops (ICCV Workshops)*, pages 1264–1269. IEEE, 2011. 1

[6] Timo Von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3d human pose in the wild using imus and a moving camera. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 601–617, 2018. 1

[7] Jiashun Wang, Huazhe Xu, Medhini Narasimhan, and Xiaolong Wang. Multi-person 3d motion prediction with multi-range transformers. *Advances in Neural Information Processing Systems*, 34:6036–6049, 2021. 1, 2