Table 1: Comparison with state-of-the-art methods on the CMU-Mocap (UMPM) dataset [40], which merges UMPM [R1] with CMU-Mocap [1]. Our approach outperforms the previous methods [32, R2, 59] and is comparable with a concurrent work [40].

milliseconds	Global				Local				Root			
	200	600	1000	Avg.	200	600	1000	Avg.	200	600	1000	Avg.
HRI [32]	49	130	207	129	41	97	130	89	31	90	158	93
MSR [R2]	53	146	231	143	46	106	137	96	29	94	175	99
MRT [59]	36	115	192	114	36	108	159	101	27	88	157	91
TBIFormer [40]	**30**	**109**	182	107	**27**	**84**	**118**	**76**	**18**	72	133	74
Ours	35	106	**173**	**105**	33	89	122	81	20	**67**	**122**	**70**

Table 2: Ablation study on weight λ. We compute the mean prediction errors (mm) of global pose, local pose, and root, respectively. We use $\lambda = 0.002$ in the main manuscript.

milliseconds	Global				Local				Root			
	400	600	800	1000	400	600	800	1000	400	600	800	1000
$\lambda = 0.01$	61.3	94.1	127.5	160.2	50.3	69.3	84.9	97.4	45.6	71.1	99.2	128.0
$\lambda = 0.005$	54.8	86.9	120.5	154.6	43.8	61.1	74.9	87.4	41.8	67.4	97.6	126.2
$\lambda = 0.002$	54.6	86.2	**119.3**	**152.5**	43.7	60.8	74.6	86.6	41.7	66.9	**94.8**	**124.0**
$\lambda = 0.001$	54.3	**86.1**	119.5	153.5	43.1	60.0	**74.2**	**86.1**	**41.4**	**66.8**	95.2	125.3
$\lambda = 0.0002$	**54.2**	86.5	120.2	154.1	**43.0**	**59.9**	74.2	86.2	41.6	67.6	96.2	126.1

Table 3: Dataset comparison. We compare our dataset with additional multi-person motion datasets employed by previous works for multi-person motion prediction.

Dataset	2D/3D	Sequences	Frames	Duration (min)	No. of people	Interaction
KTH Multiview Football [R8]	2D&3D	5	6.9k	3.9	2-3	None
Haggling [R13]	3D	34	-	55.6	3	Medium
UMPM (Interactive) [R1]	3D	8	88k	29.6	2-4	Medium
NTU-RGB+D (SoMoF) [47]	3D	739	-	28.2	2	Medium
Ours	3D	339	60k	40.3	5	Strategic

[R1] Van der Aa, N. P., et al. "Umpm benchmark: A multi-person dataset with synchronized video and motion capture data for evaluation of articulated human motion and interaction." ICCVW, 2011.

[R2] Dang, Lingwei, et al. "Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction." ICCV 2021.

[R3] Yan, Rui, et al. "Social adaptive module for weakly-supervised group activity recognition." ECCV 2020.

[R4] Li, Yixuan, et al. "Multisports: A multi-person video dataset of spatio-temporally localized sports actions." ICCV 2021.

[R5] Ibrahim, Mostafa S., et al. "A hierarchical deep temporal model for group activity recognition." CVPR 2016.

[R6] Ramanathan, Vignesh, et al. "Detecting events and key actors in multi-person videos." CVPR 2016.

[R7] Felsen, Panna, Pulkit Agrawal, and Jitendra Malik. "What will happen next? forecasting player moves in sports videos." ICCV 2017.

[R8] Kazemi, Vahid, et al. "Multi-view body part recognition with random forests." BMVC 2013.

[R9] Ettinger, Scott, et al. "Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset." ICCV 2021.

[R10] Igl, Maximilian, et al. "Symphony: Learning realistic and diverse agents for autonomous driving simulation." 2022 International Conference on Robotics and Automation (ICRA). IEEE, 2022.

[R11] Chang, Ming-Fang, et al. "Argoverse: 3d tracking and forecasting with rich maps." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.

[R12] Caesar, Holger, et al. "nuscenes: A multimodal dataset for autonomous driving." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[R13] Joo, Hanbyul, et al. "Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction." CVPR 2019.

[R14] Liang, Han, et al. "InterGen: Diffusion-based Multi-human Motion Generation under Complex Interactions." arXiv preprint arXiv:2304.05684 (2023).

[R15] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. "Voxelpose: Towards multi-camera 3d human pose estimation in wild environment." ECCV 2020.

[R16] Jianfeng Zhang, Yujun Cai, Shuicheng Yan, and Jiashi Feng. "Direct multi-view multi-person 3d pose estimation." NeurIPS 2021.

[R17] Tevet, Guy, et al. "Human motion diffusion model." ICLR 2023.

[R18] Tseng, Jonathan, Rodrigo Castellon, and Karen Liu. "Edge: Editable dance generation from music." CVPR 2023.

[R19] Rempe, Davis, et al. "Humor: 3d human motion model for robust pose estimation." ICCV 2021.