

Supplementary Materials: Future Motion Dynamic Modeling via Hybrid Supervision for Multi-Person Motion Prediction Uncertainty Reduction

A PREDICTION RESULTS ON MI-MOTION DATASETS

In our study, we perform a series of experiments on the MI-Motion dataset at multiple time points. Specifically, we utilize 25 frames (equivalent to 1.0 second) as input to predict the subsequent 25 frames (also 1.0 second). Tables 1 and 2 respectively present the experimental results of our model and other methods at 0.08s ~ 0.4s and 0.56s ~ 1.0s. It can be observed that our method overall outperforms other methods in terms of the three metrics JPE, APE, and FDE.

Table 1: Short-term prediction (0.08s ~ 0.4s) performance on MI-Motion dataset for 5 different scenes. Best results are shown in boldface.

	Park				Street				Indoor				Special Locations				Complex Crowd			
Time(s)	0.08	0.16	0.32	0.4	0.08	0.16	0.32	0.4	0.08	0.16	0.32	0.4	0.08	0.16	0.32	0.4	0.08	0.16	0.32	0.4
JPE																				
MRT [3]	23	44	76	88	20	39	74	89	25	50	80	101	47	90	159	189	24	47	88	106
TBFormer [1]	21	36	64	75	20	33	60	74	20	37	69	84	35	80	158	189	18	32	63	78
SocialTGCN [2]	18	34	60	72	15	28	54	64	20	37	67	81	45	89	165	199	20	37	70	85
JRFormer [4]	8	19	47	60	9	23	57	71	17	39	75	87	30	86	191	229	14	25	56	71
Ours	10	21	43	53	9	17	38	48	11	24	50	63	37	79	151	178	11	23	54	67
APE																				
MRT [3]	20	37	63	72	15	28	52	62	21	41	73	84	41	77	124	140	16	31	57	67
TBFormer [1]	10	24	48	57	8	19	41	51	11	26	51	63	25	73	118	136	9	23	49	60
SocialTGCN [2]	14	26	45	53	11	21	38	46	15	27	49	58	36	69	117	134	14	26	48	56
JRFormer [4]	6	17	40	50	7	18	46	58	13	32	63	75	26	72	144	163	7	21	48	60
Ours	9	19	39	47	7	14	31	39	10	21	44	55	32	67	117	133	9	19	44	54
FDE																				
MRT [3]	17	32	57	68	15	29	52	62	22	44	82	96	36	68	122	144	20	40	78	96
TBFormer [1]	20	32	56	66	18	28	48	58	20	39	78	93	30	68	134	158	16	27	53	66
SocialTGCN [2]	18	32	53	64	13	23	41	50	20	36	73	90	39	74	142	174	18	32	62	77
JRFormer [4]	5	13	39	47	6	17	35	41	14	32	67	80	24	71	181	212	5	19	34	60
Ours	8	17	35	45	6	13	28	35	10	23	54	68	29	63	124	142	9	16	46	55

Table 2: Long-term prediction (0.56s ~ 1.0s) performance on MI-Motion dataset for 5 different scenes. Best results are shown in boldface.

	Park			Street			Indoor			Special Locations			Complex Crowd		
Time(s)	0.56	0.72	1.0	0.56	0.72	1.0	0.56	0.72	1.0	0.56	0.72	1.0	0.56	0.72	1.0
JPE															
MRT [3]	107	124	149	113	130	151	119	132	147	225	250	289	140	170	220
TBFormer [1]	96	114	141	96	111	131	108	129	154	236	269	312	104	125	158
SocialTGCN [2]	95	116	154	81	98	124	108	127	160	246	276	322	113	137	177
JRFormer [4]	81	102	134	92	98	102	95	115	120	278	313	331	98	119	152
Ours	72	90	121	65	77	92	85	102	125	212	234	277	92	114	151
APE															
MRT [3]	85	92	103	78	84	84	97	102	105	154	159	171	82	94	110
TBFormer [1]	71	81	94	66	74	74	81	95	105	154	159	169	76	87	101
SocialTGCN [2]	66	76	93	58	66	76	73	83	97	154	161	174	70	81	97
JRFormer [4]	65	78	94	75	80	75	80	90	98	179	180	172	81	96	116
Ours	59	69	86	52	61	67	73	84	101	152	158	167	70	80	97
FDE															
MRT [3]	88	108	142	80	98	127	121	143	169	181	215	262	131	162	216
TBFormer [1]	86	106	136	75	91	117	119	143	174	208	251	294	90	111	149
SocialTGCN [2]	88	112	154	62	79	108	119	138	180	226	265	321	106	131	173
JRFormer [4]	75	90	128	52	56	78	97	112	136	272	307	337	77	111	142
Ours	68	85	122	49	60	71	93	114	140	179	209	266	70	95	142

B LONG-TERM PREDICTION RESULTS ON FOUR DIFFERENT DATASETS

Additionally, we test at multiple time points on the CMU_Mocap, MuPoTs-3D, Mix1, and Mix2 datasets, using 15 frames (1.0s) as input and predicting the output of 45 frames (3.0s). Table 3 shows the experimental results of our model and other methods at 1.0s-3.0s, and it can be seen that our method is generally superior to others in the three metrics JPE, APE, and FDE. To further demonstrate the effectiveness of our model in longer-term predictions, we predict sequences in the 3.0s ~ 4.0s time interval and combine the prediction effects from 1.0s ~ 3.0s into a line graph. As shown in Figure 1, our method has achieved advanced results in long-term predictions.

Table 3: The long-term prediction results of JPE, APE and FDE on the datasets CMU-Mocap, MuPoTS-3D, Mix1, and Mix2. We compare our method with the previous SOTA methods in 1.0 ~ 3.0 second. Best results are shown in boldface.

		CMU-Mocap (3 persons)				MuPoTS-3D (2-3 persons)				Mix1 (6 persons)				Mix2 (10 persons)			
Time(s)		1.0	2.0	3.0	Average	1.0	2.0	3.0	Average	1.0	2.0	3.0	Average	1.0	2.0	3.0	Average
JPE	MRT [3]	148	256	352	252	194	332	436	321	124	254	398	259	139	294	454	296
	TBFormer [1]	118	225	329	224	189	321	432	314	117	242	374	244	116	232	346	231
	SocialTGCN [2]	102	205	310	206	229	374	523	375	109	232	376	239	112	226	341	226
	JRFormer [4]	122	218	305	215	196	334	430	320	110	248	380	246	106	220	336	221
	Ours	98	198	299	198	175	305	415	298	106	232	372	237	104	215	330	216
APE	MRT [3]	130	187	218	178	135	195	216	182	86	123	144	118	97	142	161	133
	TBFormer [1]	89	132	152	124	131	181	207	173	78	118	140	112	87	131	152	123
	SocialTGCN [2]	84	129	158	124	182	266	398	282	84	131	163	126	95	148	181	141
	JRFormer [4]	97	145	169	137	170	221	245	212	84	121	145	117	81	131	145	119
	Ours	79	122	145	115	129	176	204	170	77	118	139	111	79	126	149	118
FDE	MRT [3]	109	216	315	213	154	282	378	271	86	209	353	216	95	241	402	246
	TBFormer [1]	78	172	273	174	148	272	379	266	85	199	330	205	79	182	295	185
	SocialTGCN [2]	66	153	247	155	201	365	573	380	82	195	327	201	79	180	289	183
	JRFormer [4]	82	158	249	156	154	299	385	279	103	244	411	253	68	164	282	171
	Ours	62	148	244	151	132	253	358	248	75	190	327	197	68	167	279	171

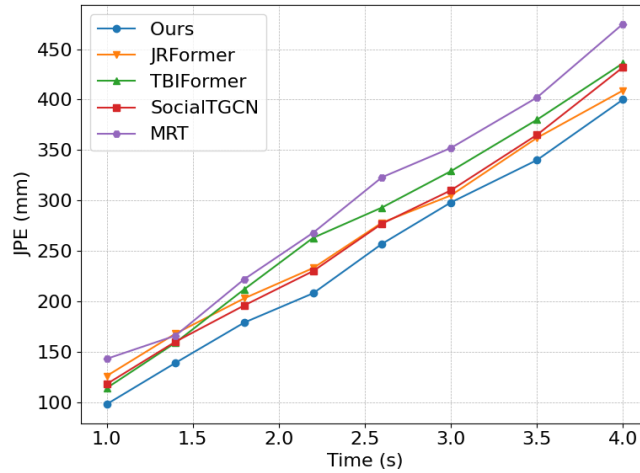


Figure 1: Comparison of different methods for multi-person motion prediction over the time span of 1.0s ~ 4.0s.

REFERENCES

- [1] Xiaogang Peng, Siyuan Mao, and Zizhao Wu. 2023. Trajectory-aware body interaction transformer for multi-person pose forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17121–17130.
- [2] Xiaogang Peng, Xiao Zhou, Yikai Luo, Hao Wen, Yu Ding, and Zizhao Wu. 2023. The MI-Motion Dataset and Benchmark for 3D Multi-Person Motion Prediction. *arXiv preprint arXiv:2306.13566* (2023).
- [3] Jiashun Wang, Huazhe Xu, Medhini Narasimhan, and Xiaolong Wang. 2021. Multi-person 3d motion prediction with multi-range transformers. *Advances in Neural Information Processing Systems* 34 (2021), 6036–6049.
- [4] Qingyao Xu, Weibo Mao, Jingze Gong, Chenxin Xu, Siheng Chen, Weidi Xie, Ya Zhang, and Yanfeng Wang. 2023. Joint-Relation Transformer for Multi-Person Motion Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9816–9826.