Progressively Generating Better Initial Guesses Towards Next Stages for High-Quality Human Motion Prediction
— Supplementary Material —

Tiezheng Ma[1], Yongwei Nie[1*], Chengjiang Long[2], Qing Zhang[3] and Guiqing Li[1]

[1]School of Computer Science and Engineering, South China University of Technology, China
[2]Meta Reality Labs, USA
[3]School of Computer Science and Engineering, Sun Yat-sen University, China

Abstract

In this supplementary material, we provide more information and extra experiments that could not be included in the main article because of space limit.

1. Architecture Details of Encoder-Copy-Decoder Network

Table 1 provides the structure details of the Encoder-Copy-Decoder network of our full model. The network contains 3 GCBs, with 1 in the Encoder, and 2 in the Decoder. Each GCB contains 2 GCLs. Each GCL contains S-DGCN, T-DGCN, BatchNorm, Tanh (activation function), Dropout, sequentially. The residual connections of the Encoder, Decoder and GCBs are all shown in the table.

The input shape, output shape, and the hyper-parameters of the layers in the table are collected from the experiments on Human3.6M. For example, the input shape $(35, 22, 3)$ in the second row means that the input pose sequence is of length 35 (10 historical poses and 25 future poses), each pose has 22 joints, and each joint has 3 coordinates. By the 1×1 Conv layer, we obtain a feature map in the space of $\mathbb{R}^{35\times22\times16}$ which is then used by the residual connection of the Encoder.

The x and y in W(x,y), $A^s(x, y)$, $A^t(x, y)$, $W^s(x, y)$, and $W^t(x, y)$ give the shape of the parameters of the corresponding layer. For example in the third row, the used S-DGCN has the spatial adjacency matrix of size $\mathbb{R}^{22\times22}$ and parameters of size $\mathbb{R}^{3\times16}$. The hyperparameter of Dropout is 0.3.

As can be seen, after the "Copy" operator, we obtain a feature map of size $\mathbb{R}^{70\times22\times16}$ which comprises two copies of the input. All A^t in the Decoder has the shape of $\mathbb{R}^{70\times70}$. Finally, the Decoder outputs 70 poses, from which we use the 35 frames in the front as the output.

Table 1. Details of the Encoder-Copy-Decoder Network.

Module	Layer		Input Shape	Operation	Output Shape
Encoder	1 × 1 Conv		(35,22,3)	W(3,16)	(35,22,16)
	GCL		(35,22,3)	S-DGCN: A^s(22,22), W^s(3,16)	(35,22,16)
			(35,22,16)	T-DGCN: A^t(35,35), W^t(16,16)	(35,22,16)
			(35,22,16)	BatchNorm	(35,22,16)
			(35,22,16)	Tanh	(35,22,16)
			(35,22,16)	Dropout (0.3)	(35,22,16)❷
	GCB	GCL	(35,22,16)	S-DGCN: A^s(22,22), W^s(16,16)	(35,22,16)
			(35,22,16)	T-DGCN: A^t(35,35), W^t(16,16)	(35,22,16)
			(35,22,16)	BatchNorm	(35,22,16)
			(35,22,16)	Tanh	(35,22,16)
			(35,22,16)	Dropout (0.3)	(35,22,16)
		GCL	(35,22,16)	S-DGCN: A^s(22,22), W^s(16,16)	(35,22,16)
			(35,22,16)	T-DGCN: A^t(35,35), W^t(16,16)	(35,22,16)
			(35,22,16)	BatchNorm	(35,22,16)
			(35,22,16)	Tanh	(35,22,16)
			(35,22,16)	Dropout (0.3)	(35,22,16)❸
		Residual	(35,22,16)	Add ❷ + ❸	(35,22,16)❹
	Residual		(35,22,16)	Add ❶ + ❹	(35,22,16)
Copy			(35,22,16)	Replicating once in temporal dimension.	(70,22,16)
Decoder	1 × 1 Conv		(70,22,16)	W(16,3)	(70,22,3)❺
	GCB1	GCL	(70,22,16)	S-DGCN: A^s(22,22), W^s(16,16)	(70,22,16)
			(70,22,16)	T-DGCN: A^t(70,70), W^t(16,16)	(70,22,16)
			(70,22,16)	BatchNorm	(70,22,16)
			(70,22,16)	Tanh	(70,22,16)
			(70,22,16)	Dropout (0.3)	(70,22,16)
		GCL	(70,22,16)	S-DGCN: A^s(22,22), W^s(16,16)	(70,22,16)
			(70,22,16)	T-DGCN: A^t(70,70), W^t(16,16)	(70,22,16)
			(70,22,16)	BatchNorm	(70,22,16)
			(70,22,16)	Tanh	(70,22,16)
			(70,22,16)	Dropout (0.3)	(70,22,16)❻
		Residual	(70,22,16)	Add ❺ + ❻	(70,22,16)❼
	GCB2	GCL	(70,22,16)	S-DGCN: A^s(22,22), W^s(16,16)	(70,22,16)
			(70,22,16)	T-DGCN: A^t(70,70), W^t(16,16)	(70,22,16)
			(70,22,16)	BatchNorm	(70,22,16)
			(70,22,16)	Tanh	(70,22,16)
			(70,22,16)	Dropout (0.3)	(70,22,16)
		GCL	(70,22,16)	S-DGCN: A^s(22,22), W^s(16,16)	(70,22,16)
			(70,22,16)	T-DGCN: A^t(70,70), W^t(16,16)	(70,22,16)
			(70,22,16)	BatchNorm	(70,22,16)
			(70,22,16)	Tanh	(70,22,16)
			(70,22,16)	Dropout (0.3)	(70,22,16)❽
		Residual	(70,22,16)	Add ❼ + ❽	(70,22,16)
			(70,22,16)	S-DGCN:A^s(22.22), W^s(16,3)	(70,22,3)
			(70,22,3)	T-DGCN:A^t(70,70), W^t(3,3)	(70,22,3)❾
	Residual		(70,22,3)	Add ❺ + ❾	(70,22,3)
Slicing			(70,22,3)	Taking first 35 frames as output.	(35,22,3)

*Corresponding author: nieyongwei@scut.edu.cn

Table 2. Supplement to Table 3 of the paper: short-term per action experimental data.

scenarios	basketball				basketball signal				directing traffic				jumping			
millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
Res. Sup.	15.5	26.9	43.5	49.2	20.2	33.0	42.8	44.7	20.5	40.6	75.4	90.4	26.9	48.1	93.5	108.9
DMGNN	15.6	28.7	59.0	73.1	5.0	9.3	20.2	26.2	10.2	20.9	41.6	52.3	32.0	54.3	96.7	119.9
LTD	11.7	21.3	41.0	50.8	3.3	6.3	13.6	18.0	6.9	13.7	30.3	40.0	17.2	32.4	60.1	72.6
STSGCN	28.4	31.9	48.2	64.6	15.3	15.4	21.6	35.5	20.9	22.6	36.0	58.3	32.2	41.4	68.0	86.1
MSR	10.3	18.9	37.7	47.0	3.0	5.7	12.4	16.3	5.9	12.1	28.4	38.0	15.0	28.7	55.9	69.1
Ours	**9.5**	**17.5**	**35.3**	**44.2**	**2.7**	**4.9**	**10.8**	**14.6**	**4.8**	**9.8**	**23.6**	**32.3**	**13.9**	**27.8**	**55.8**	**69.0**
scenarios	soccer				walking				wash window				average			
millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
Res. Sup.	17.8	31.3	52.6	61.4	44.4	76.7	126.8	151.4	22.8	44.7	86.8	104.7	24.0	43.0	74.5	87.2
DMGNN	14.9	25.3	52.2	65.4	9.6	15.5	26.0	30.4	7.9	14.7	33.3	44.2	13.6	24.1	47.0	58.8
LTD	13.3	24.0	43.8	53.2	6.6	10.7	17.4	20.4	6.0	11.6	24.8	31.6	9.3	17.1	33.0	40.9
STSGCN	31.2	34.8	53.1	73.2	21.9	21.4	25.9	38.9	17.6	19.2	30.9	53.5	25.3	27.9	41.8	59.2
MSR	**10.9**	**19.5**	**37.1**	**46.4**	6.3	10.3	17.6	21.1	5.5	11.1	25.1	32.5	8.1	15.2	30.6	38.6
Ours	11.1	20.6	39.5	48.7	**6.2**	**10.3**	**16.8**	**19.8**	**4.6**	**9.2**	**20.9**	**27.3**	**7.6**	**14.3**	**29.0**	**36.6**

Table 3. Supplement to Table 3 of the paper: long-term per action experimental data.

scenarios	basketball		basketball signal		directing traffic		jumping	
millisecond	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms
Res. Sup.	**54.3**	**72.8**	51.4	60.6	112.9	153.1	128.8	162.8
DMGNN	96.1	138.6	36.6	52.0	72.3	111.2	160.6	224.6
LTD	68.1	98.0	27.7	54.0	60.9	114.2	93.8	127.4
STSGCN	75.3	109.2	38.7	63.5	66.7	113.4	61.6	74.1
MSR	62.8	87.0	24.6	**47.9**	58.9	111.0	92.1	**124.8**
Ours	59.4	84.1	**23.7**	50.2	**51.6**	**102.3**	**91.7**	125.6
scenarios	soccer		walking		wash window		average	
millisecond	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms
Res. Sup.	72.3	107.4	182.4	194.3	136.3	202.7	105.5	136.3
DMGNN	82.2	111.9	37.8	67.0	56.5	82.8	77.4	112.6
LTD	70.9	108.3	25.2	34.4	43.9	67.0	55.8	86.2
STSGCN	84.9	116.6	39.0	46.1	51.6	79.2	65.0	92.1
MSR	**64.41**	**99.32**	27.2	39.7	45.9	71.3	53.7	83.0
Ours	65.4	99.9	**25.1**	**33.9**	**39.7**	**65.7**	**50.9**	**80.1**

2. More Detailed Experimental Data of CMU-MoCap

Table 3 in the paper just gives the prediction errors at each forecasting timestamp averaged over all kinds of actions. Here, we provide more detailed experimental data as shown in Table 2 and Table 3 in which the results of every action are given. For short-term prediction, our method is better than all the other methods on all kinds of actions except "soccer". For "soccer", MSR performs the best while ours is the second best. For long-term prediction, our method is the best for "directing tracffic", "walking", and "wash window", and achieves the best average performance. For other actions, our method is the second best and comparable to the best one.

3. Results of Angle Representations

Table 4 shows comparison conducted on the angle representation of Human3.6. Our method outperforms SOTA.

Table 4. MAE(Mean Angle Error) comparisons on Human3.6M.

Method	80ms	160ms	320ms	400ms	560ms	1000ms	Avg
DMGNN (angle)	0.38	0.65	0.94	1.04	1.24	1.64	0.98
LTD (angle)	0.34	0.58	0.93	1.06	1.27	1.65	0.97
MSR (angle)	0.35	0.61	0.98	1.11	1.31	1.67	1.00
Our (angle)	**0.30**	**0.54**	**0.89**	**1.02**	**1.23**	**1.61**	**0.93**
Our (angle to 3D)	13.1	27.6	54.8	66.3	85.2	119.3	61.1
Our (3D)	**10.3**	**22.7**	**47.4**	**58.5**	**76.9**	**110.3**	**54.4**

We further transform our angle-based results into 3D. Compared with trained directly in 3D, the performance drops when trained on angle.

4. Evaluation on Random 256 Test Set

The main paper has presented experimental results evaluated on the whole test dataset, as done by Dang *et al*. [2]. Here, following [4], we give the results on the random 256 test set, *i.e*., only 256 samples of each action are randomly selected (using a fixed seed) for testing. The comparison results are shown in Table 5 and Table 6. As can be seen, our method is also the best in most cases, and outperforms the compared approaches by a large margin.

5. Evaluation on Random 8 Test Set

The works of [1, 3, 5] randomly select 8 samples per action for testing (using a fixed seed). We also compare our method with previous approaches in this way, and the comparison results are shown in Table 7 and Table 8. Overall speaking, our method performs better than all the other methods, as demonstrated by the average prediction errors.

When evaluating in this setting, the advantage of our method compared to previous approaches is not as significant as when evaluating on the whole test dataset or the random 256 test set. We conjecture this is because the randomness of just selecting 8 samples per action is too high to evaluate a method. Therefore, we choose to perform the evaluation on the whole test dataset in the main paper.

6. Comparison with Transformer-based method [1]

In Table 7 and Table 8, we compare our method with the Transformer-based approach [1]. The experimental results of [1] are directly collected from their paper. Our method is better than [1] for both short-term and long-term predictions on average.

7. More Visualizations

In Figure 1, we show more visualizations of the predicted poses of different methods. In each sub-figure, from top to bottom are the ground truth, the results of our method, MSR [2], LTD [5], DMGNN [3], and Res.Sup. [6], respectively. Our predictions are more similar to the ground truth than the results of the compared methods in these cases .

8. Code and Video Demo

The code for research purpose is available at:
https://github.com/705062791/PGBIG
A video demo containing qualitative comparisons is also attached there.

References

[1] Yujun Cai, Lin Huang, Yiwei Wang, Tat-Jen Cham, Jianfei Cai, Junsong Yuan, Jun Liu, Xu Yang, Yiheng Zhu, Xiaohui Shen, et al. Learning progressive joint propagation for human motion prediction. In *European Conference on Computer Vision*, pages 226–242. Springer, 2020. 2, 3, 5

[2] Lingwei Dang, Yongwei Nie, Chengjiang Long, Qing Zhang, and Guiqing Li. Msr-gcn: Multi-scale residual graph convolution networks for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11467–11476, October 2021. 2, 3

[3] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 214–223, 2020. 2, 3

[4] 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. 2

[5] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9489–9497, 2019. 2, 3

[6] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2891–2900, 2017. 3

Table 5. Comparisons on random 256 test set of Human3.6M. Short-term prediction results are given. The best results are highlighted in bold, and the second best are marked by underline.

scenarios	walking				eating				smoking				discussion			
millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
Res. Sup.	23.2	40.9	61	66.7	16.8	31.5	53.5	61.7	18.9	34.7	57.5	65.4	25.7	47.8	80	91.3
DMGNN	18.4	33.6	56.8	65.1	10.1	19.7	38.3	46.7	11.4	22.0	41.5	50.1	18.0	36.2	71.9	85.2
LTD	11.1	21.4	37.3	42.9	7	14.8	29.8	37.3	7.5	15.5	30.7	37.5	10.8	24	52.7	65.8
MSR	10.8	20.9	36.9	42.4	6.9	14.6	29.0	36.0	7.5	15.4	30.6	37.5	10.4	23.5	51.9	65.0
Ours	**9.4**	**19.0**	**34.3**	**40.4**	**6.0**	**13.4**	**27.8**	**35.3**	**6.5**	**14.2**	**28.8**	**35.5**	**9.0**	**21.8**	**49.9**	**62.9**
scenarios	directions				greeting				phoning				posing			
millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
Res. Sup.	21.6	41.3	72.1	84.1	31.2	58.4	96.3	108.8	21.1	38.9	66	76.4	29.3	56.1	98.3	114.3
DMGNN	13.8	27.7	55.3	67.2	22.6	45.1	89.0	106.6	14.3	28.0	52.4	63.3	18.6	37.6	80.1	100.0
LTD	8	18.8	43.7	54.9	14.8	31.4	65.3	79.7	9.3	19.1	39.8	49.7	10.9	25.1	59.1	75.9
MSR	7.7	18.9	44.7	56.2	15.1	33.1	70.9	85.4	9.1	18.9	39.9	49.8	10.3	24.6	59.2	75.9
Ours	**6.4**	**16.8**	**41.5**	**52.7**	**12.4**	**28.3**	**61.2**	**76.0**	**7.8**	**17.2**	**37.3**	**47.3**	**8.7**	**22.2**	**53.9**	**70.4**
scenarios	purchases				sitting				sittingdown				takingphoto			
millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
Res. Sup.	28.7	52.4	86.9	100.7	23.8	44.7	78	91.2	31.7	58.3	96.7	112	21.9	41.4	74	87.6
DMGNN	21.7	42.4	77.3	91.6	14.7	30.0	61.5	74.5	20.7	39.9	81.0	97.4	14.4	29.2	59.4	74.6
LTD	13.9	30.3	62.2	75.9	9.8	20.5	44.2	55.9	15.6	31.4	59.1	71.7	8.9	18.9	41	51.7
MSR	13.3	30.1	63.6	77.8	9.8	20.6	44.2	55.5	15.4	32.0	60.7	73.8	8.9	19.5	43.1	54.4
Ours	**11.7**	**27.8**	**59.4**	**73.5**	**8.5**	**18.8**	**41.8**	**53.2**	**13.7**	**29.3**	**57.2**	**69.7**	**7.6**	**17.2**	**38.5**	**49.2**
scenarios	waiting				walkingdog				walkingtogether				average			
millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
Res. Sup.	23.8	44.2	75.8	87.7	36.4	64.8	99.1	110.6	20.4	37.1	59.4	67.3	25	46.2	77	88.3
DMGNN	15.5	30.7	61.5	74.4	31.7	62.1	109.8	125.3	15.7	29.2	51.1	60.7	17.4	34.2	65.8	78.9
LTD	9.2	19.5	43.3	54.4	20.9	40.7	73.6	86.6	9.6	19.4	36.5	44	11.2	23.4	47.9	58.9
MSR	10.4	22.4	50.7	62.4	24.9	51.5	100.3	112.9	9.2	18.7	35.7	43.2	11.3	24.3	50.8	61.9
Ours	**7.4**	**17.3**	**39.6**	**50.8**	**18.4**	**38.1**	**71.8**	**85.1**	**8.1**	**17.4**	**34.0**	**41.5**	**9.4**	**21.3**	**45.1**	**56.2**

Table 6. Comparisons on random 256 test set of Human3.6M. Long-term prediction results are given. The best results are highlighted in bold, and the second best are marked by underline.

scenarios	walking		eating		smoking		discussion		directions		greeting		phoning		posing	
millisecond	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms
Res. Sup.	71.6	79.1	74.9	98	78.1	102.1	109.5	131.8	101.1	129.1	126.1	153.9	94	126.4	140.3	183.2
DMGNN	75.4	96.8	61.9	91.0	64.1	93.2	107.1	138.6	88.4	121.4	132.5	165.2	80.0	112.9	136.6	210.4
LTD	51.8	60.9	50	**74.1**	51.3	73.6	87.6	118.6	76.1	108.8	104.3	140.2	68.7	105.1	109.9	171.7
MSR	53.3	63.7	50.8	75.4	50.5	72.1	87.0	**116.8**	75.8	105.9	106.3	**136.3**	67.9	104.7	112.5	176.5
Ours	**49.6**	**58.9**	**50.0**	74.9	**48.8**	**69.9**	**86.1**	116.9	**73.3**	**105.9**	**100.2**	136.4	**66.5**	**102.7**	**102.8**	**167.0**
scenarios	purchases		sitting		sittingdown		takingphoto		waiting		walkingdog		walkingtogether		average	
millisecond	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms
Res. Sup.	122.1	154	113.7	152.6	138.8	187.4	110.6	153.9	105.4	135.4	128.7	164.5	80.2	98.2	106.3	136.6
DMGNN	115.5	155.9	95.7	138.7	130.4	188.1	100.3	146.8	97.1	141.5	147.2	184.9	74.7	97.5	100.5	138.9
LTD	99.4	135.9	78.5	118.8	99.5	144.1	76.8	120.2	75.1	106.9	105.8	**142.2**	58	69.6	79.5	112.7
MSR	99.2	134.5	77.6	115.9	102.4	149.4	77.7	121.9	74.8	105.5	107.7	145.7	56.2	69.5	80.0	112.9
Ours	**95.7**	**132.1**	**75.1**	**114.8**	**94.4**	**139.0**	**70.5**	**112.9**	**71.6**	**103.7**	**105.7**	145.9	**54.4**	**64.6**	**76.3**	**109.7**

Table 7. Comparisons on random 8 test set of Human3.6M. Short-term prediction results are given. The best results are highlighted in bold, and the second best are marked by underline. The results of Transformer [1] are collected from their papers.

scenarios	walking				eating				smoking				discussion			
millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
Res. Sup.	23.8	40.4	62.9	70.9	17.6	34.7	71.9	87.7	19.7	36.6	61.8	73.9	31.7	61.3	96	103.5
DMGNN	17.2	30.6	54.4	65.0	11.0	21.4	35.9	43.5	8.9	17.3	31.7	40.0	17.4	34.6	60.8	69.5
LTD	8.9	15.7	29.2	33.4	8.8	18.9	39.4	47.2	7.8	14.9	25.3	28.7	9.8	22.1	39.6	**44.1**
MSR	8.7	15.5	28.4	32.4	8.3	**17.7**	**36.3**	**43.7**	7.5	15.4	27.4	31.5	9.3	22.1	40.5	45.5
Transformer	7.9	**14.5**	29.1	34.5	8.4	18.1	37.4	45.3	6.8	**13.2**	**24.1**	**27.5**	8.3	21.7	43.9	48.0
Ours	**7.6**	14.6	**24.9**	**28.3**	**8.0**	17.9	38.0	45.7	**6.3**	13.4	25.2	30.3	**7.3**	**19.3**	**38.1**	45.2
scenarios	directions				greeting				phoning				posing			
millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
Res. Sup.	36.5	56.4	81.5	97.3	37.9	74.1	1390	158.8	25.6	44.4	74	84.2	27.9	54.7	131.3	160.8
DMGNN	13.2	24.9	64.8	81.9	23.4	50.3	107.2	131.9	12.7	26.0	48.4	58.4	15.3	29.2	71.5	96.6
LTD	12.6	24.4	48.2	58.4	14.5	30.5	74.2	89	11.5	20.2	37.9	43.2	9.4	23.9	66.2	82.9
MSR	11.4	21.9	**45.8**	**56.1**	13.5	26.5	68.8	86.1	11.8	20.6	37.5	41.7	8.5	21.8	**61.2**	**76.4**
Transformer	11.1	22.7	48.0	58.4	13.2	28.0	64.5	**77.9**	10.8	19.6	37.6	46.8	8.3	22.8	65.6	81.8
Ours	**10.1**	**21.7**	48.1	59.5	**11.2**	**24.1**	**63.6**	80.0	**10.6**	**18.8**	**34.1**	**39.7**	**6.6**	**20.1**	61.6	78.1
scenarios	purchases				sitting				sittingdown				takingphoto			
millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
Res. Sup.	40.8	71.8	104.2	109.8	34.5	69.9	126.3	141.6	28.6	55.3	101.6	118.9	23.6	47.4	94	112.7
DMGNN	21.4	38.8	75.9	93.0	11.9	25.2	**44.6**	**50.1**	15.0	32.8	77.1	93.1	13.5	28.7	45.6	58.4
LTD	19.6	38.5	64.4	72.2	10.7	24.6	50.6	62	11.4	27.6	56.4	67.6	6.8	15.2	38.2	**49.6**
MSR	19	38.7	64.5	72.6	11.3	26.5	56.1	69.2	11.1	28.2	56.1	66.8	6.6	15.8	40.8	53.1
Transformer	18.5	38.1	**61.8**	**69.6**	9.5	23.9	49.8	61.8	11.2	29.9	59.8	68.4	6.3	14.5	38.8	49.4
Ours	**17.2**	**36.5**	63.4	72.2	**8.3**	**22.1**	49.3	61.4	**9.8**	**26.3**	**53.5**	**63.2**	**5.8**	**14.1**	**38.0**	49.8
scenarios	waiting				walkingdog				walkingtogether				average			
millisecond	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms	80ms	160ms	320ms	400ms
Res. Sup.	29.5	60.5	119.9	140.6	60.5	101.9	160.8	188.3	23.5	45	71.3	82.8	30.8	57	99.8	115.5
DMGNN	12.1	23.8	59.5	77.5	47.1	93.3	160.3	171.4	14.4	26.7	50.1	63.2	17	33.6	65.9	79.6
LTD	9.5	22	57.5	73.9	32.2	58	102.2	122.7	8.9	18.4	35.3	44.3	12.1	25	51	61.3
MSR	8.9	20.9	53.6	69.8	24.4	53.6	**95.6**	**110.4**	8.7	18.5	35.4	45.6	11.3	24.3	49.9	60.1
Transformer	8.4	21.5	53.9	69.8	**22.9**	**50.4**	100.8	119.8	8.7	18.3	34.2	44.1	10.7	23.8	50.0	60.2
Ours	**7.4**	**18.2**	**50.4**	**66.7**	27.3	53.6	97.6	119.0	**7.2**	**16.7**	**33.8**	**42.8**	**10.1**	**22.5**	**48.0**	**58.8**

Table 8. Comparisons on random 8 test set of Human3.6M. Long-term prediction results are given. The best results are highlighted in bold, and the second best are marked by underline. The results of Transformer [1] are collected from their papers.

scenarios	walking		eating		smoking		discussion		directions		greeting		phoning		posing	
millisecond	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms
Res. Sup.	86.3	107.6	87.7	99.4	96.1	141.4	120.7	161.6	110.2	150.5	162.2	174.227	139.098	127.029	192.096	230.697
DMGNN	73.4	95.8	57.8	86.5	50.4	71.6	81.9	138.2	110.1	115.6	152.2	157.6	78.8	98.8	164.0	310.3
LTD	42.3	51.3	56.5	68.6	32.3	60.5	70.5	103.5	85.8	109.3	91.8	87.4	65.0	**113.6**	113.4	220.6
MSR	42.1	43.5	57.0	71.5	35.2	62.5	75.4	113.5	**78.5**	**101.7**	100.1	95.1	63.7	113.9	**103.0**	219.9
Transformer	36.8	**41.2**	58.4	**67.9**	**29.2**	58.3	74.0	103.1	-	-	-	-	-	-	-	-
Ours	**35.9**	43.9	**55.7**	69.5	33.1	**58.1**	**69.9**	**99.9**	83.7	105.3	**90.7**	**87.1**	**62.1**	115.6	104.3	**209.3**
scenarios	purchases		sitting		sittingdown		takingphoto		waiting		walkingdog		walkingtogether		average	
millisecond	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms	560ms	1000ms
Res. Sup.	115.8	159.4	161.6	195.3	214.5	285.2	117.9	141.1	152.9	199.1	196.8	213.3	107.8	136.5	137.5	168.2
DMGNN	118.8	154.5	59.7	104.3	122.0	168.8	91.2	120.6	106.1	136.6	194.1	182.2	83.5	115.8	102.9	137.1
LTD	94.3	130.4	**79.6**	114.9	82.6	140.1	**68.9**	**87.1**	100.9	167.6	136.6	**174.3**	57.0	85.0	78.5	114.3
MSR	**86.5**	125.5	83.1	**103.9**	83.1	145.8	72.6	95.9	100.7	**164.3**	144.4	193.5	55.8	84.5	78.7	115.7
Transformer	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Ours	89.7	**122.9**	81.0	115.8	**80.2**	**130.8**	70.3	90.5	**94.5**	168.1	137.8	180.8	**54.6**	**80.3**	**76.2**	**111.9**

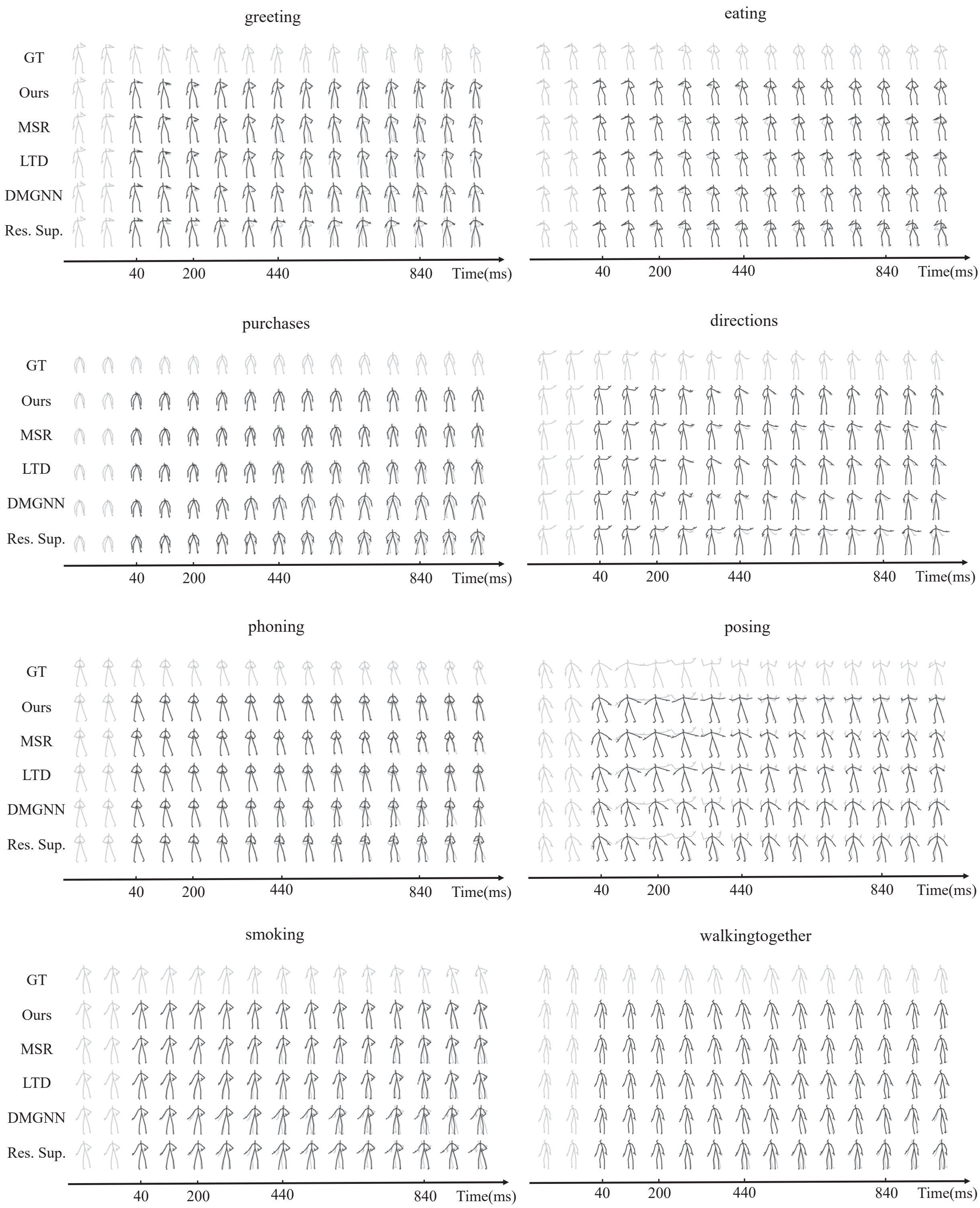

Figure 1. More qualitative comparisons on Human3.6M.