A. Gradient Flow Through Latent Variable

A reparameterization trick (Gumbel-Softmax) is commonly used to approximate the expectation $\mathbb{E}_{z \sim q_\psi(\cdot|x)}$ by taking N samples from $q_\psi(z \mid x)$. We however, using the latent distribution in the output representation, exactly compute $\mathbb{E}_{z \sim q_\phi(\cdot|x)} = \sum_z \log q_\phi(z \mid x) p_\psi(y \mid x, z)$.

Still the gradient resulting from each component's Log-Likelihood could be backpropagated to update ϕ. Conceptually, we argue that the mixture distribution should not affect the individual component distributions' log-prob computation. The same applies to the gradients. Practically, letting the gradient flow through the latent variable a marginal negative experimental influence on the performance. Thus, as the implementation difference is a single line, we will leave the choice to the user.

B. Implementation and Training Details

All model training was performed on a single *Nvidia RTX2080*.

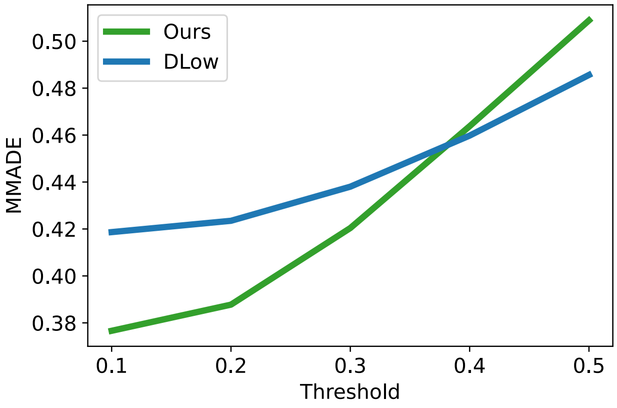
Deterministic Evaluation. For the H3.6M dataset, we applied dataset augmentation (according to [40]) where the training samples are mirrored along the human's vertical. A batch size of 32 and early stopping to select the best model was used. During hyperparameter optimization, we used subject 6 for validation purposes. For the final model, subject 6 was added to the training data, but we kept the same early stopping iteration which showed the best performance during validation. All test evaluations were performed on the full 32-joint skeleton of subject 5. For the AMASS dataset we use 5% of the training data samples as validation data. Again, we use early stopping based on the validation results. The batch size for AMASS is 128. All evaluations were performed on the full 22-joint skeleton.

Generative Evaluation. We use the same hyperparameters for the H3.6M dataset as in the deterministic evaluation. However, according to previous works, we only evaluate on the positions of a 16-joints skeleton where subject 9 and 11 serve as test subjects. To stabilize the GRU for the long prediction horizon (100 steps) we use a form of curriculum learning; increasing the prediction horizon from a small value to the target in early iterations.

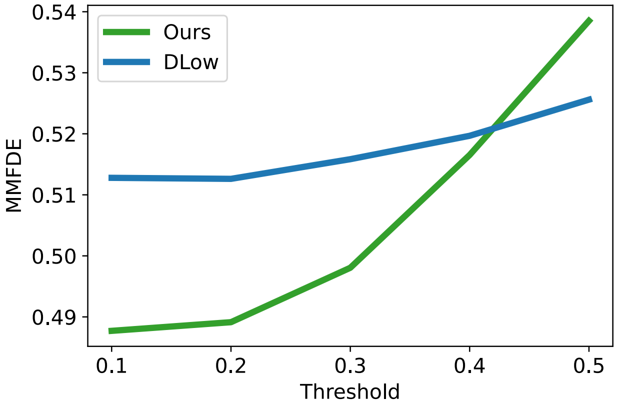
For both model types we found it beneficial to randomly shrink the prediction horizon at every training iteration.

C. Multimodal Metrics

In [52] the authors propose two multimodal evaluation metrics where motions are combined based on "similar context" and evaluate the mean performance on all futures of the combined motions. Their measure of similarity is solely based on the distance of motions at a single timestep $t = 0$. Motions are combined for evaluation using a threshold on that distance. While this threshold is arbitrarily set to 0.5 in [52], we use a range of different thresholds in our evaluation. The results can be seen in Fig. 6. While we outperform *DLow* for lower thresholds we have lower numerical results for higher thresholds. However, looking at a concrete example Fig. 8 we see that for high thresholds motions with entirely uncorrelated futures (and history) are combined for evaluation. Thus, the worse numerical results for (too) high thresholds are expected; even encouraged here.



(a) MMADE for different thresholds



(b) MMFDE for different thresholds

Figure 6. Result for multimodal metrics proposed in [52] for different thresholds. For larger thresholds, where possibly uncorrelated motions are evaluated together, *DLow* achievs better numerical values as their approaches over-diversifies their produced motions.

D. Bone Deformation

Some algorithms directly predict joint positions instead of joint configurations (angles). This shows advantages on metrics calculated on joint positions (e.g. MPJPE). However, as Fig. 7 outlines they produce physically unfeasible motions as the rigid bone structure is deformed.

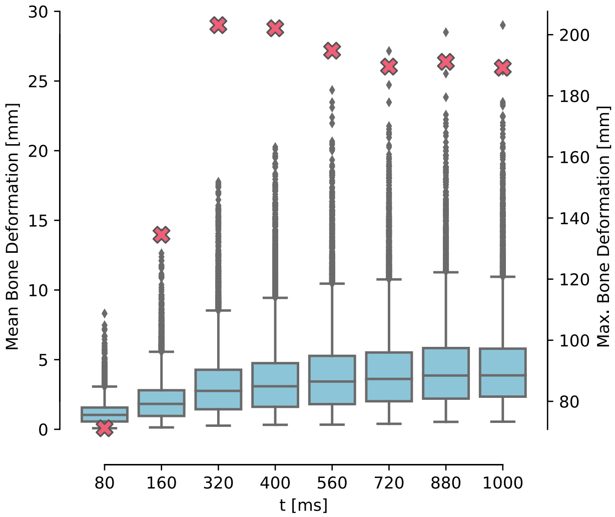


Figure 7. Exemplary bone deformation of a method directly predicting 3D joint positions. With increasing prediction time the deformation increases with outliers reaching deformations of over $20cm$. Boxplots: Mean bone deformation over all bones in the skeleton. Red Crosses: Maximum bone deformation over all bones. Analysis performed on 10000 samples.

E. Learned Node-Attention

In Fig. 9 we visualize the learned attention influence between different nodes in the skeleton by our *Typed-Graph* approach. Unsurprisingly, neighboring joints commonly show high influence. However, influence between nodes not directly connecting by a single joint are learned too: The attention value connecting both shoulder joints has a high value as a correlation between the movements of both arms is learned. An example of a more subtle learned correlation is that the prediction of the hand is substantially influenced by the gaze angle of the head.

F. Further Deterministic Results

We present additional results for the deterministic evaluation. In Tab. 7 the results are split up by individual actions for 256 samples and for 8 samples in Tab. 8. We also present the MPJPE metric which is calculated on the joint's position using forward kinematic on angle outputs (Tab. 9 and Tab. 10).

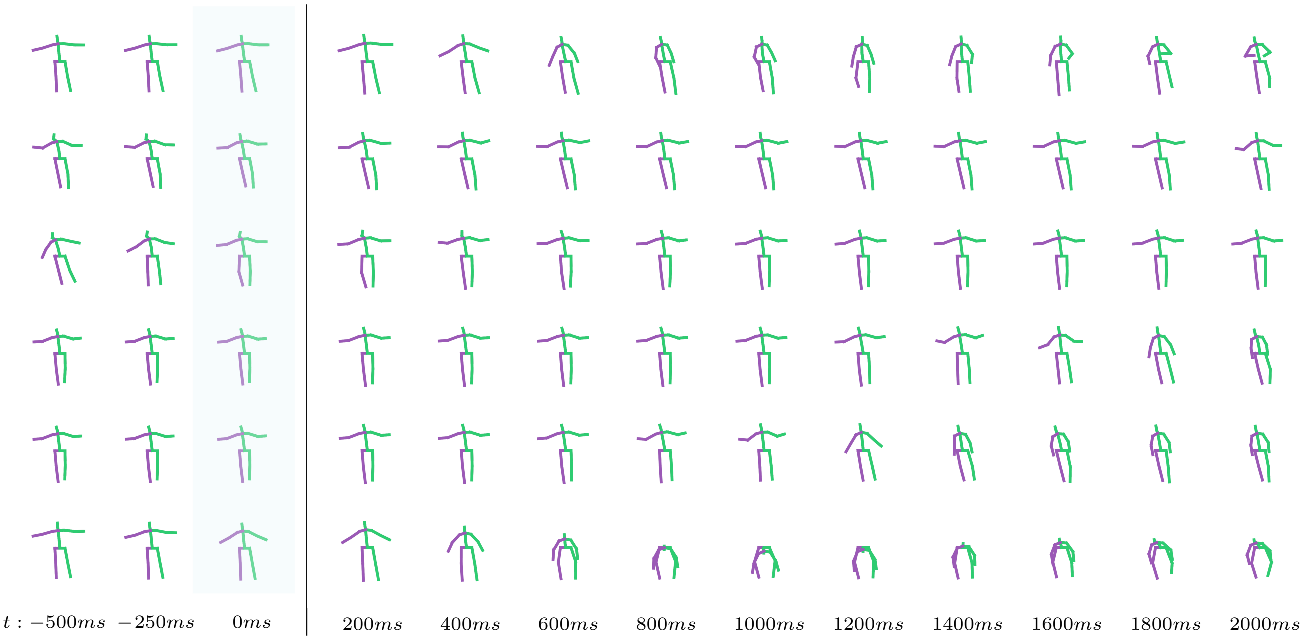


Figure 8. Example of 6 motions which are combined for the multimodal evaluation. At $t = 0$ (highlighted) all motions seem to be similar. However, they greatly differ in history and future. Thus, we do not expect, and do not want, our model to give accurate output for all futures when only the history of the first motion is used as input.

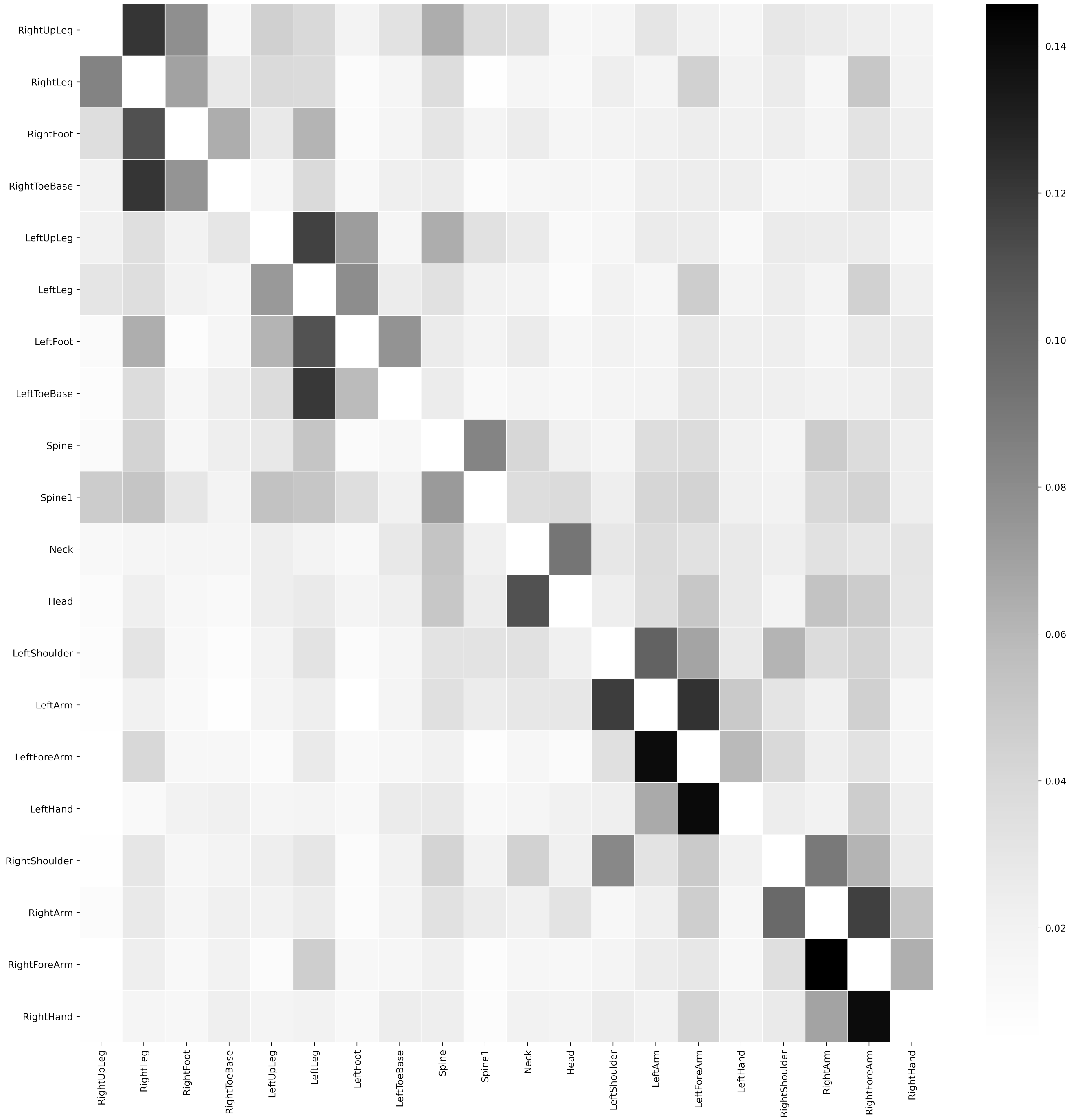


Figure 9. Learned attention influence of our TG layers between different nodes in the skeleton.

	Average								Walking								Eating							
milliseconds	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000
Zero Vel.	0.40	0.70	1.11	1.25	1.46	1.63	1.76	1.84	0.43	0.78	1.23	1.34	1.50	1.56	1.58	1.55	0.28	0.53	0.89	1.03	1.18	1.33	1.44	1.49
GRU sup. [38]	0.43	0.74	1.15	1.30	-	-	-	-	0.34	0.60	0.91	0.98	-	-	-	-	0.30	0.57	0.87	0.98	-	-	-	-
Quarternet [40]	0.37	0.62	1.00	1.14	-	-	-	-	0.28	0.49	0.76	0.83	-	-	-	-	0.22	0.47	0.76	0.88	-	-	-	-
HistRepItself [35]	0.28	0.52	0.88	1.02	1.23	1.40	1.55	1.64	0.24	0.43	0.66	0.71	0.84	0.91	0.99	1.03	0.18	0.37	0.67	0.79	0.95	1.11	1.23	1.30
Ours W-Mean	0.28	0.51	0.87	1.01	1.22	1.40	1.54	1.63	0.23	0.43	0.68	0.74	0.88	0.94	1.03	1.08	0.20	0.41	0.79	0.90	1.06	1.23	1.38	1.45
Ours ML-Mode	0.28	0.51	0.88	1.02	1.24	1.42	1.58	1.67	0.23	0.43	0.68	0.73	0.88	0.94	1.04	1.10	0.20	0.41	0.79	0.91	1.06	1.23	1.39	1.47

	Smoking								Discussion								Directions							
milliseconds	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000
Zero Vel.	0.30	0.53	0.88	1.03	1.23	1.38	1.52	1.62	0.46	0.80	1.22	1.36	1.62	1.74	1.83	1.90	0.30	0.54	0.92	1.08	1.24	1.35	1.48	1.54
GRU sup. [38]	0.35	0.69	1.14	1.29	-	-	-	-	0.54	0.85	1.30	1.44	-	-	-	-	0.32	0.58	0.97	1.14	-	-	-	-
Quarternet [40]	0.28	0.47	0.79	0.91	-	-	-	-	0.38	0.74	1.20	1.37	-	-	-	-	0.24	0.46	0.84	1.01	-	-	-	-
HistRepItself [35]	0.21	0.38	0.65	0.79	0.99	1.15	1.30	1.42	0.31	0.61	1.02	1.17	1.44	1.57	1.68	1.76	0.19	0.38	0.74	0.90	1.08	1.22	1.35	1.42
Ours W-Mean	0.21	0.38	0.65	0.80	1.02	1.32	1.34	1.44	0.36	0.63	1.02	1.16	1.41	1.54	1.63	1.70	0.19	0.38	0.75	0.94	1.13	1.26	1.38	1.46
Ours ML-Mode	0.21	0.38	0.65	0.80	1.03	1.22	1.36	1.47	0.35	0.63	1.02	1.18	1.44	1.58	1.68	1.75	0.19	0.38	0.76	0.95	1.15	1.28	1.40	1.48

	Greeting								Phoning								Posing							
milliseconds	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000
Zero Vel.	0.54	0.91	1.41	1.59	1.81	2.01	2.16	2.23	0.39	0.69	1.14	1.28	1.48	1.70	1.85	1.91	0.40	0.74	1.20	1.40	1.72	2.00	2.29	2.44
GRU sup. [38]	0.64	0.99	1.40	1.54	-	-	-	-	0.42	0.70	1.11	1.27	-	-	-	-	0.46	0.83	1.33	1.52	-	-	-	-
Quarternet [40]	0.61	0.93	1.34	1.51	-	-	-	-	0.36	0.61	0.98	1.14	-	-	-	-	0.38	0.71	1.20	1.39	-	-	-	-
HistRepItself [35]	0.39	0.71	1.17	1.35	1.60	1.78	1.93	1.99	0.29	0.52	0.91	1.05	1.24	1.47	1.63	1.72	0.27	0.55	1.00	1.21	1.54	1.80	2.10	2.24
Ours W-Mean	0.40	0.68	1.14	1.31	1.51	1.70	1.84	1.90	0.28	0.51	0.84	0.98	1.19	1.41	1.58	1.66	0.24	0.50	0.93	1.12	1.42	1.71	1.96	2.10
Ours ML-Mode	0.40	0.69	1.14	1.32	1.53	1.75	1.92	1.98	0.28	0.51	0.85	1.00	1.21	1.45	1.61	1.70	0.24	0.50	0.94	1.13	1.45	1.76	2.03	2.19

	Purchases								Sitting								Sitting Down							
milliseconds	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000
Zero Vel.	0.57	0.96	1.36	1.44	1.64	1.79	1.94	2.01	0.33	0.60	1.01	1.16	1.41	1.67	1.87	1.98	0.50	0.84	1.30	1.48	1.99	2.21	2.31	
GRU sup. [38]	0.57	0.95	1.33	1.43	-	-	-	-	0.41	0.75	1.22	1.41	-	-	-	-	0.59	1.00	1.62	1.87	-	-	-	-
Quarternet [40]	054	0.92	1.36	1.47	-	-	-	-	0.34	0.59	1.00	1.15	-	-	-	-	0.47	0.81	1.31	1.50	-	-	-	-
HistRepItself [35]	0.43	0.78	1.21	1.31	1.47	1.62	1.74	1.81	0.25	0.49	0.91	1.06	1.33	1.59	1.78	1.88	0.41	0.72	1.17	1.36	1.66	1.88	2.11	2.20
Ours W-Mean	0.44	0.75	1.18	1.28	1.48	1.69	1.77	1.86	0.23	0.45	0.85	1.01	1.28	1.53	1.70	1.81	0.41	0.72	1.17	1.36	1.68	1.90	2.12	2.22
Ours ML-Mode	0.44	0.75	1.19	1.29	1.52	1.74	1.82	1.91	0.23	0.46	0.87	1.02	1.30	1.56	1.74	1.84	0.41	0.73	1.18	1.37	1.69	1.91	2.14	2.24

	Taking Photo								Waiting								Walk Dog							
milliseconds	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000
Zero Vel.	0.26	0.44	0.73	0.84	1.05	1.20	1.33	1.45	0.37	0.64	1.10	1.27	1.50	1.67	1.91	1.92	0.53	0.85	1.20	1.31	1.46	1.63	1.73	1.78
GRU sup. [38]	0.30	0.52	0.88	1.02	-	-	-	-	0.41	0.68	1.20	1.37	-	-	-	-	0.52	0.84	1.21	1.32	-	-	-	-
Quarternet [40]	0.23	0.39	0.69	0.81	-	-	-	-	0.32	0.54	1.00	1.15	-	-	-	-	0.48	0.78	1.12	1.21	-	-	-	-
HistRepItself [35]	0.19	0.34	0.60	0.72	0.92	1.07	1.21	1.33	0.25	0.46	0.88	1.05	1.28	1.47	1.63	1.75	0.41	0.68	1.01	1.12	1.30	1.45	1.54	1.63
Ours W-Mean	0.20	0.33	0.60	0.72	0.91	1.10	1.25	1.36	0.23	0.44	0.83	1.00	1.23	1.42	1.58	1.69	0.41	0.69	1.05	1.16	1.30	1.48	1.59	1.68
Ours ML-Mode	0.20	0.33	0.60	0.72	0.92	1.10	1.26	1.38	0.24	0.44	0.84	1.00	1.23	1.42	1.58	1.70	0.41	0.69	1.07	1.17	1.32	1.50	1.63	1.72

	Walk Together							
milliseconds	80	160	320	400	560	720	880	1000
Zero Vel.	0.37	0.66	1.02	1.15	1.32	1.39	1.41	1.43
GRU sup. [38]	0.35	0.57	0.83	0.94	-	-	-	-
Quarternet [40]	0.28	0.45	0.69	0.79	-	-	-	-
HistRepItself [35]	0.21	0.38	0.62	0.71	0.86	0.94	1.00	1.04
Ours W-Mean	0.20	0.36	0.59	0.68	0.83	0.92	1.01	1.06
Ours ML-Mode	0.20	0.36	0.59	0.68	0.84	0.94	1.04	1.09

Table 7. Angle error on 256 samples per action on the H3.6M test dataset.

	Average								Walking								Eating							
milliseconds	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000
Zero Vel.	0.40	0.71	1.07	1.20	1.42	1.57	1.75	1.85	0.39	0.68	0.99	1.15	1.35	1.37	1.34	1.32	0.27	0.48	0.73	0.86	1.04	1.10	1.27	1.38
GRU sup. [38]	0.40	0.69	1.04	1.18	-	-	-	-	0.27	0.46	0.67	0.75	0.93	-	-	1.03	0.23	0.37	0.59	0.73	0.95	-	-	1.08
DMGNN [33]	0.27	0.52	0.83	0.95	-	-	-	-	0.18	0.31	0.49	0.58	0.66	-	-	0.75	0.17	0.30	0.49	0.59	0.74	-	-	1.14
HistRepItself [35]	0.27	0.52	0.82	0.93	1.14	1.28	1.48	1.59	0.18	0.30	0.46	0.51	0.59	0.62	0.61	0.64	0.16	0.29	0.49	0.60	0.74	0.81	1.01	1.11
Ours W-Mean	0.26	0.48	0.81	0.93	1.12	1.28	1.46	1.56	0.18	0.31	0.51	0.55	0.61	0.65	0.64	0.66	0.16	0.29	0.49	0.61	0.72	0.77	0.96	1.07
Ours ML-Mode	0.26	0.48	0.82	0.95	1.15	1.32	1.50	1.60	0.18	0.31	0.51	0.56	0.61	0.65	0.64	0.67	0.16	0.30	0.50	0.62	0.73	0.78	0.97	1.08

	Smoking								Discussion								Directions							
milliseconds	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000
Zero Vel.	0.26	0.48	0.97	0.95	1.02	1.14	1.47	1.69	0.31	0.67	0.94	1.04	1.41	1.71	1.86	1.96	0.39	0.59	0.79	0.89	1.02	1.22	1.47	1.50
GRU sup. [38]	0.32	0.59	1.01	1.10	1.25	-	-	1.50	0.30	0.67	0.98	1.06	1.43	-	-	1.69	0.41	0.64	0.80	0.92	-	-	-	-
DMGNN [33]	0.21	0.39	0.81	0.77	0.83	-	-	1.52	0.26	0.65	0.92	0.99	1.33	-	-	1.45	0.25	0.44	0.65	0.71	-	-	-	-
HistRepItself [35]	0.22	0.42	0.86	0.80	0.86	1.00	1.34	1.58	0.20	0.52	0.78	0.87	1.30	1.54	1.66	1.72	0.25	0.43	0.60	0.69	0.81	1.03	1.25	1.29
Ours W-Mean	0.21	0.42	0.79	0.89	0.96	1.02	1.33	1.48	0.21	0.56	0.80	0.90	1.23	1.51	1.57	1.53	0.29	0.38	0.63	0.72	0.87	1.08	1.29	1.34
Ours ML-Mode	0.21	0.43	0.82	0.93	1.01	1.09	1.40	1.56	0.21	0.56	0.81	0.92	1.26	1.54	1.61	1.58	0.29	0.39	0.67	0.77	0.97	1.16	1.37	1.42

	Greeting								Phoning								Posing							
milliseconds	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000
Zero Vel.	0.54	0.89	1.30	1.49	1.79	1.77	1.85	1.80	0.64	1.21	1.57	1.70	1.81	1.94	2.05	2.04	0.28	0.57	1.13	1.37	1.81	2.23	2.58	2.78
GRU sup. [38]	0.57	0.82	1.45	1.60	-	-	-	-	0.59	1.06	1.45	1.60	-	-	-	-	0.45	0.85	1.34	1.56	-	-	-	-
DMGNN [33]	0.36	0.61	0.94	1.12	-	-	-	-	0.52	0.97	1.29	1.43	-	-	-	-	0.20	0.46	1.06	1.34	-	-	-	-
HistRepItself [35]	0.35	0.60	0.95	1.14	1.48	1.47	1.61	1.57	0.53	1.01	1.22	1.29	1.42	1.55	1.68	1.68	0.19	0.46	1.09	1.35	1.59	1.83	2.14	2.34
Ours W-Mean	0.35	0.58	0.87	1.03	1.29	1.34	1.53	1.54	0.41	0.62	1.12	1.19	1.36	1.39	1.53	1.69	0.19	0.48	1.03	1.25	1.55	1.96	2.22	2.39
Ours ML-Mode	0.35	0.59	0.87	1.02	1.28	1.31	1.50	1.53	0.42	0.63	1.16	1.24	1.48	1.54	1.66	1.81	0.19	0.48	1.03	1.26	1.66	2.12	2.40	2.58

	Purchases								Sitting								Sitting Down							
milliseconds	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000
Zero Vel.	0.62	0.88	1.19	1.27	1.64	1.62	2.09	2.45	0.40	0.63	1.02	1.18	1.26	1.36	1.57	1.63	0.39	0.74	1.07	1.19	1.36	1.57	1.70	1.80
GRU sup. [38]	0.58	0.79	1.08	1.15	-	-	-	-	0.41	0.68	1.12	1.33	-	-	-	-	0.47	0.88	1.37	1.54	-	-	-	-
DMGNN [33]	0.41	0.61	1.05	1.14	-	-	-	-	0.26	0.42	0.76	0.97	-	-	-	-	0.32	0.65	0.93	1.05	-	-	-	-
HistRepItself [35]	0.42	0.65	1.00	1.07	1.43	1.53	1.94	2.24	0.29	0.47	0.83	1.01	1.16	1.29	1.50	1.55	0.30	0.63	0.92	1.04	1.18	1.42	1.55	1.69
Ours W-Mean	0.41	0.64	1.06	1.14	1.42	1.66	2.00	2.28	0.26	0.44	0.79	0.99	1.12	1.28	1.52	1.58	0.29	0.60	0.93	1.08	1.30	1.51	1.64	1.78
Ours ML-Mode	0.42	0.64	1.07	1.17	1.44	1.73	2.08	2.33	0.26	0.43	0.78	0.98	1.09	1.25	1.49	1.55	0.29	0.60	0.94	1.10	1.30	1.50	1.63	1.76

	Taking Photo								Waiting								Walk Dog							
milliseconds	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000	80	160	320	400	560	720	880	1000
Zero Vel.	0.25	0.51	0.79	0.92	1.03	1.13	1.22	1.27	0.34	0.67	1.22	1.47	1.89	2.27	2.57	2.63	0.60	0.98	1.36	1.50	1.74	1.87	1.95	1.96
GRU sup. [38]	0.28	0.57	0.90	1.02	-	-	-	-	0.32	0.63	1.07	1.26	-	-	-	-	0.52	0.89	1.25	1.40	-	-	-	-
DMGNN [33]	0.15	0.34	0.58	0.71	-	-	-	-	0.22	0.49	0.88	1.10	-	-	-	-	0.42	0.72	1.16	1.34	-	-	-	-
HistRepItself [35]	0.16	0.36	0.58	0.70	0.83	0.91	1.01	1.08	0.22	0.49	0.92	1.14	1.54	1.92	2.25	2.33	0.46	0.78	1.05	1.23	1.58	1.65	1.79	1.85
Ours W-Mean	0.13	0.34	0.58	0.70	0.78	0.83	0.89	0.97	0.20	0.46	0.88	1.09	1.46	1.77	2.05	2.11	0.41	0.71	1.15	1.32	1.52	1.69	1.77	1.79
Ours ML-Mode	0.14	0.34	0.57	0.70	0.79	0.84	0.88	0.96	0.20	0.45	0.87	1.08	1.46	1.77	2.06	2.12	0.41	0.72	1.16	1.33	1.53	1.71	1.79	1.83

	Walk Together							
milliseconds	80	160	320	400	560	720	880	1000
Zero Vel.	0.33	0.66	0.94	0.99	1.10	1.22	1.22	1.52
GRU sup. [38]	0.27	0.53	0.74	0.79	-	-	-	-
DMGNN [33]	0.15	0.33	0.50	0.57	-	-	-	-
HistRepItself [35]	0.14	0.32	0.50	0.55	0.63	0.68	0.80	1.18
Ours W-Mean	0.13	0.31	0.48	0.54	0.67	0.78	0.92	1.20
Ours ML-Mode	0.13	0.31	0.49	0.55	0.70	0.81	0.98	1.27

Table 8. Angle error on 8 samples per action on the H3.6M test dataset.

	MPJPE [mm]							
milliseconds	80	160	320	400	560	720	880	1000
Zero Vel.	23.8	44.4	76.1	88.3	107.5	121.6	131.6	136.6
HistRepItself [35]	12.4	25.9	51.4	62.5	81.4	96.3	108.6	116.4
HistRepItself 3D	11.3	24.1	49.9	60.8	78.3	92.0	105.1	112.8
Ours W-Mean	15.1	29.9	55.6	66.2	83.7	98.0	110.1	117.9
Ours ML-Mode	15.0	30.1	56.0	66.6	84.3	98.9	111.4	119.6
Ours Bo3-Modes	15.1	29.6	53.6	63.1	78.4	91.0	102.4	110.5
Ours Bo5-Modes	15.2	29.7	53.2	62.6	78.4	88.2	99.3	107.7

Table 9. Mean per Joint Position Error (MPJPE) on 256 samples per action on the H3.6M test dataset. HistRepItself 3D directly outputs 3D joint position and is therefore subject to bone deformation.

	MPJPE [mm]					
milliseconds	100	200	400	600	800	1000
Zero Vel.	41.9	72.9	106.2	115.3	115.3	112.5
HistRepItself [35]	20.4	39.8	64.4	74.8	80.5	85.8
Ours W-Mean	19.1	37.8	63.0	75.3	82.3	87.5
Ours ML-Mode	19.1	38.2	64.1	76.9	84.3	89.9
Ours Bo3-Modes	19.0	37.3	59.9	70.0	76.6	82.8
Ours Bo5-Modes	19.1	37.4	58.8	67.5	73.8	81.0

Table 10. Mean per Joint Position Error (MPJPE) on 10,000 samples from the AMASS test set.