

Gaussian-Mixture Latent Flow for Stochastic 3D Human Motion Prediction

Supplementary Material

I. Metric Calculation Details

Followings Sec. 4, the observed T frames are represented as $\mathbf{X}_{obs} = (\mathbf{x}_1, \dots, \mathbf{x}_T)$, and the future motion with N steps is denoted as $\mathbf{Y} = (\mathbf{x}_{T+1}, \dots, \mathbf{x}_{T+N})$. The diverse set of M predicted future samples is represented as $\hat{\mathbf{Y}}^M = \{\hat{\mathbf{Y}}^1, \dots, \hat{\mathbf{Y}}^M\}$. The metrics used in the experiments can then be defined as follows:

- **Average Pair Distance (APD)** measures the L2 distance between a set of predictions generated from the same history, which is computed as:

$$\text{APD} = \frac{1}{M(M-1)} \sum_{i=1}^M \sum_{j \neq i}^M \|\hat{\mathbf{Y}}^i - \hat{\mathbf{Y}}^j\|_2. \quad (\text{I.1})$$

- **Average Displacement Error (ADE)** computes the average L2 distance between the ground truth and the closest sample, which is computed as:

$$\text{ADE} = \min_{\hat{\mathbf{Y}}^j} \|\mathbf{Y} - \hat{\mathbf{Y}}^j\|_2. \quad (\text{I.2})$$

- **Final Displacement Error (FDE)** computes the L2 distance between the final pose of the GT and the closest final pose among predictions, which is computed as:

$$\text{FDE} = \min_{\hat{\mathbf{x}}_{T+N}^j} \|\mathbf{x}_{T+N} - \hat{\mathbf{x}}_{T+N}^j\|_2. \quad (\text{I.3})$$

- **Multi-Modal-ADE (MMADE)** aggregates multi-modal ground truths (MMGT) across the entire dataset. Given a dataset D containing M sequences, the MMGT of $\mathbf{X}_{obs}^i \in D$ is then computed as Equation (I.4):

$$\text{MMGT} = \left\{ \mathbf{Y}^j \mid \|\mathbf{x}_T^j - \mathbf{x}_T^i\|_2 < A, j = 1, 2, \dots, M \right\}, \quad (\text{I.4})$$

where A is a hyperparameter, often set as 0.5 for both Human3.6M and HumanEva-I datasets. MMADE is then computed as Equation (I.5):

$$\text{MMADE} = \frac{1}{L} \sum_{j=1}^L \min_{\hat{\mathbf{Y}}^k} \|\mathbf{Y}^j - \hat{\mathbf{Y}}^k\|_2, \quad (\text{I.5})$$

where $\mathbf{x}_{j,\cdot}$ represents j -th sequence belonging to the corresponding MMGT and L denotes the number of sequences contained in the MMGT.

- **Multi-Modal-FDE (MMFDE)** is computed using the same MMGT as MMADE and is defined as:

$$\text{MMFDE} = \frac{1}{L} \sum_{j=1}^L \min_{\hat{\mathbf{x}}_{T+N}^k} \|\mathbf{x}_{j,T+N} - \hat{\mathbf{x}}_{T+N}^k\|_2. \quad (\text{I.6})$$

- **Cumulative Motion Distribution (CMD)** measures the difference between the areas under the cumulative pseudo-data motion distribution and the predicted distribution. Let \bar{D} denote the L2 distance between the displacement in two consecutive frames across the entire dataset. For the f -th frame in the predicted distribution, we compute the average displacement D_j in the same manner. The CMD is then calculated as:

$$\begin{aligned} \text{CMD} &= \sum_{i=1}^{N-1} \sum_{j=1}^i \|D_j - \bar{D}\|_1 \\ &= \sum_{i=1}^{N-1} (N-i) \|D_i - \bar{D}\|_1. \end{aligned} \quad (\text{I.7})$$

II. Visualization of the Latent Space

We present a t-SNE visualization of the multi-modal latent space across all Human3.6M test sequences, as shown in Figure 4. In Figure 4a, points sharing the same color correspond to the same sub-distribution, while in Figure 4b, points with the same color indicate sequences belonging to the same action class.

From Figure 4a, we can observe that the latent space is organized into several clusters. Interestingly, the number of clusters is smaller than the initial setting of 16 due to the automatic pruning mechanism in the employed updating strategy, which performs a soft allocation of samples. Furthermore, these clusters exhibit partial overlap without distinct boundaries between them. This continuity in the latent space is advantageous for generative tasks, as it facilitates smooth transitions between modes, but it can hinder performance in discriminative tasks such as classification and anomaly detection. From Figure 4b, we can observe that sequences belonging to the same action class tend to cluster together. Not all sequences of the same class are grouped in a single cluster, primarily because the Human3.6M dataset provides only coarse action labels, namely individual sequences may contain motions from other classes. For example, *Waiting* sequences often include sub-sequences in which the person walks or sits down. Nevertheless, we can also observe that action classes with similar motion semantics, such as *Walking* and *WalkingTogether*, are positioned closely in the latent space, both appearing in the upper-right region. This finding indicates that our constructed multi-modal latent prior effectively captures diverse motion patterns and their underlying semantic relationships.

For comparison, we also provide a t-SNE visualization of the single-modal latent space for all Human3.6M test se-

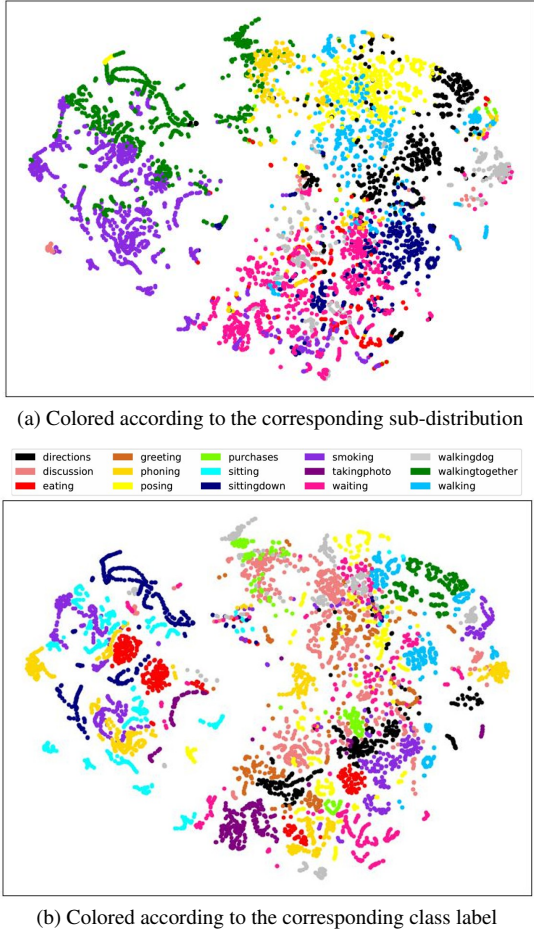


Figure 4. Visualization of the multi-modal latent space. We present a 2D t-SNE projection of the latent encodings for all Human3.6M test sequences: (a) colored by their nearest sub-distribution, and (b) colored by the action label of each sequence.

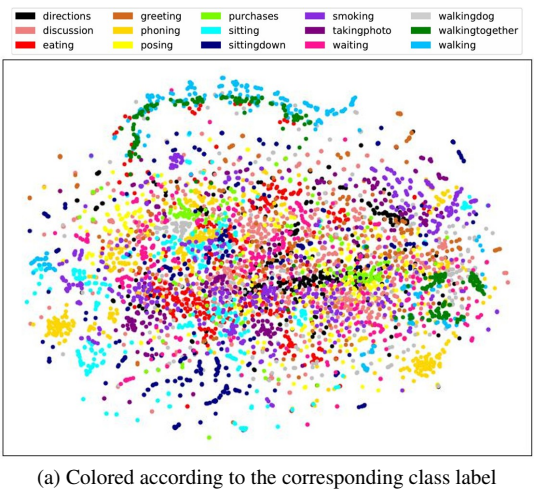


Figure 5. Visualization of the single-modal latent space. We present a 2D t-SNE projection of the latent encodings for all Human3.6M test sequences, which is colored by the action label.

quences in Figure 5. When compared with Figure 4b and Figure 5a, we observe that sequences of different action labels in the single-modal latent space are largely entangled. In contrast, sequences belonging to the same class form more coherent clusters in our multi-modal latent space, indicating that the proposed multi-modal prior effectively disentangles motion patterns and semantics. Interestingly, a subset of *Walking* and *WalkingTogether* sequences forms clusters outside the main region, a phenomenon also reported in BeLFusion [3].

III. Analysis of the Latent Space Construction

We further examine the influence of the number of initial components in the latent mixture distribution. As shown in Tab. 4, we evaluate configurations with 8, 16, and 32 initial components on Human3.6M. For comparison, we additionally report a variant of the multi-modal prior that employs a single learnable component, as well as the results obtained using a single-modal prior.

Table 4. Analysis of varying the number of initialized sub-distributions used to construct the latent space, with results reported on Human3.6M.

Prior	Nums	APD \uparrow	ADE \downarrow	FDE \downarrow	MMADE \downarrow	MMFDE \downarrow	CMD \downarrow	FID \downarrow
Single	-	5.056	0.356	0.421	0.483	0.479	5.854	0.143
MM	1	4.691	0.338	0.408	0.476	0.475	3.215	0.154
MM	8	4.773	0.334	0.403	0.473	0.469	3.038	0.134
MM	16	4.804	0.333	0.399	0.471	0.464	3.015	0.088
MM	32	4.885	0.333	0.400	0.475	0.470	2.955	0.092

When the number of initial components is set to 16 or 32, the predictive performance across *accuracy*, *diversity*, and *plausibility* remains nearly identical. However, when the initial number is reduced to 8 or 1, the *accuracy* and *diversity* metrics remain comparable, but the *plausibility* metric deteriorates significantly. These observations indicate that our data-driven Gaussian mixture can effectively and adaptively capture the diverse motion patterns in the dataset once the initial number of components is sufficiently large. In such cases, the soft-alignment strategy prunes redundant components during training, making additional components unnecessary. In contrast, when the initial number of components is too small, the limited capacity restricts the model’s ability to fully capture the inherent diversity of motion patterns. Interestingly, when directly comparing the learnable single-modal prior (with one initial component) to the pre-defined standard Normal prior, the results show that the learnable prior achieves substantially more accurate predictions and yields a much lower CMD value, indicating a closer fit to the underlying data distribution and greater distributional consistency.

IV. Analysis of the ODE Solver

We report predictive performance and computational cost using an Euler solver with integration steps ranging from 1 to 200, as well as a Dopri5 solver with adaptive step sizes. As shown in Tab. 5, the Euler solver exhibits convergence when the step count becomes sufficiently large. Specifically, beyond 100 steps, the *accuracy* metrics remain stable, while APD improves only marginally. Moreover, the results closely match those obtained using Dopri5, further confirming convergence behavior. Interestingly, for the Euler solver, when the number of steps is insufficient, some metrics appear slightly better, likely due to stochastic variations introduced during the numerical integration process.

We also report the inference time for each configuration in Tab. 5, where every measurement is averaged over 1,000 runs on one RTX 4060Ti GPU to ensure fair comparison. Notably, the Dopri5 solver with adaptive step sizes achieves comparable prediction quality at significantly lower computational cost compared to the Euler solver.

Table 5. Experimental results on various ODE solvers.

Method	Steps	Time	Human3.6M [32]				
			APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓
Dopri5	-	504ms	5.046	0.335	0.401	0.472	0.465
Euler	1	16ms	1.196	0.367	0.513	0.513	0.571
Euler	10	88ms	3.427	0.330	0.409	0.476	0.477
Euler	20	174ms	4.069	0.330	0.400	0.472	0.468
Euler	100	886ms	4.804	0.333	0.399	0.471	0.464
Euler	200	1676ms	4.915	0.334	0.400	0.471	0.464

Method	Steps	Time	AMASS [51]				
			APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓
Dopri5	-	965ms	7.370	0.463	0.475	0.542	0.510
Euler	1	18ms	0.989	0.485	0.621	0.568	0.648
Euler	10	115ms	4.123	0.446	0.489	0.531	0.525
Euler	20	223ms	6.363	0.453	0.472	0.534	0.508
Euler	100	1132ms	7.144	0.461	0.474	0.540	0.509
Euler	200	2246ms	7.257	0.462	0.473	0.541	0.509

V. Analysis of Uncertainty Quantification

To validate the reliability of the estimated probabilities as measures of uncertainty, we first examine the correlation between prediction errors (e.g., ADE and FDE) and their corresponding uncertainty quantification ranks, where the ranking is determined by the estimated probability – samples with higher probability are assigned higher ranks. For comparison, we also include results based on ideal ranks, where the ranking is determined directly by the prediction errors – samples with lower errors are assigned higher ranks. Ideally, these two rankings should coincide.

As shown in Figure 6, while the estimated ranks are not perfectly aligned with the ground truth, they reveal a clear

trend that higher probability values (i.e., higher ranks) correlate strongly with lower ADE and FDE errors. This correlation is consistent with statistical theory, confirming that more probable predictions are, on average, more accurate. These findings validate the reliability and practical value of our uncertainty quantification, as predictions assigned higher likelihood by our model are indeed closer to the true future. This enables autonomous systems to effectively prioritize actions based on quantified confidence.

To quantify the deviation between our estimated ranks and the ideal ranks, we measure the normalized area difference between the estimated error-rank curve and the ideal curve. This metric quantifies the overall discrepancy in the error distribution across all confidence ranks, analogous to the Area Under the Sparsification Error (AUSE) [43]. The formulation of our metric is defined as:

$$E = \frac{\sum_{i=1}^N |err_{\text{est}}^{(i)} - err_{\text{ideal}}^{(i)}|}{\sum_{i=1}^N err_{\text{ideal}}^{(i)}}, \quad (\text{V.8})$$

where $err_{\text{est}}^{(i)}$ and $err_{\text{ideal}}^{(i)}$ represent the prediction errors (i.e., ADE and FDE) at the i -th rank position for the estimated and ideal rankings, respectively.

The results are presented in Tab. 6. For the ADE metric, the discrepancy is below 10% for both Human3.6M and AMASS. For the FDE metric, the discrepancy is slightly higher, at 13.7% for Human3.6M and 14.4% for AMASS. The minimal magnitude of these deviations indicates that our uncertainty estimates are highly reliable and exhibit strong correlation with the true prediction error.

For comparison, we report the results of two uncertainty-aware methods in Tab. 6: (1) Motron [61], a parametric approach that provides likelihood-based uncertainty estimates, and (2) ProbHMI [49], a SOTA method specifically designed for uncertainty quantification via latent quantiles. Our results demonstrate a 25.6% and 26.5% reduction on Human3.6M and AMASS compared to Motron, while achieving values very similar to ProbHMI. These findings support our claim that the flow-based likelihood produces competitive and well-calibrated uncertainty estimates.

Table 6. The area discrepancy in the error distribution.

Metric	Human3.6M		AMASS	
	ADE	FDE	ADE	FDE
Ours	0.096	0.137	0.083	0.144
Motron [61]	0.129	0.151	0.113	0.152
ProbHMI [49]	0.109	0.139	0.076	0.141

VI. Additional Visualization Results

We present additional comparisons against several recent SOTA methods, including CoMusion [63] and SLD [72]

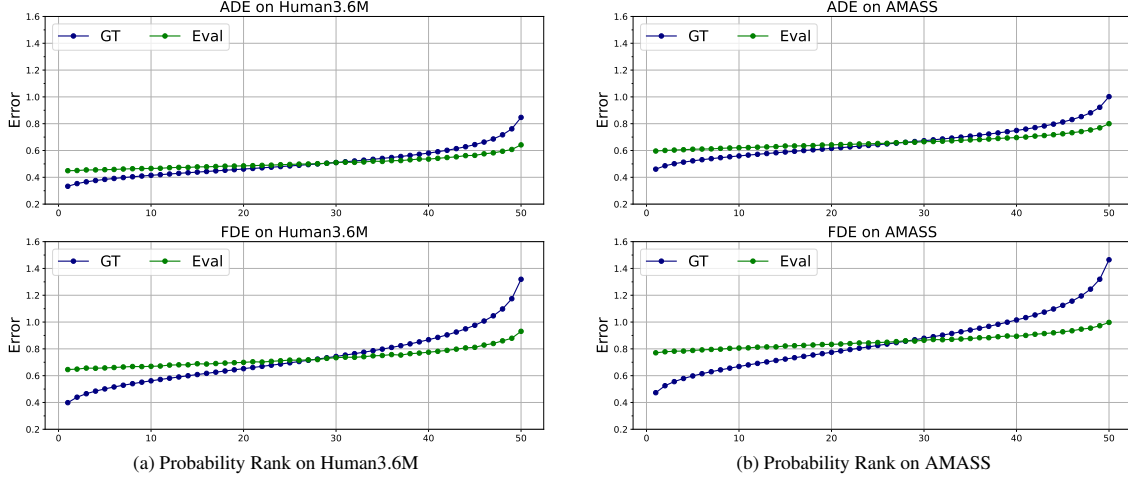


Figure 6. We evaluate the correlation between prediction errors (**ADE** and **FDE**) and their associated uncertainty ranks (denoted "Eval") on Human3.6M and AMASS. As a reference benchmark, we also present the values based on ideal ranks (denoted "GT").

on Human3.6M, and CoMusion [63] and SkeletonDiff [15] on AMASS, all of which demonstrate strong performance in terms of *accuracy*, *diversity* and *plausibility*. As shown in Figure 7, the first row presents results on Human3.6M, while the second row illustrates results on AMASS, with each frame visualizing 10 stochastic forecasts.

Although baseline methods can produce diverse motion sequences, our approach generates more natural and realistic pose forecasts, exhibiting higher *plausibility* than the baselines. For example, on AMASS, CoMusion and SkeletonDiff often produce poses with unnaturally lifted legs, as highlighted in the left (CoMusion) and right (SkeletonDiff) yellow boxes, respectively. Moreover, motions generated by the baselines frequently deviate from the contextual cues. As highlighted in the middle green box on Human3.6M, both CoMusion and SLD tend to produce poses with a bent upper body, which is inconsistent with the intended walking motion. Similarly, in the yellow boxes on AMASS, some baseline-generated motions depict unnatural sitting postures. Another example of implausible predictions from the baselines is their tendency to generate poses with elevated body parts, such as arms positioned unnaturally far from the torso, as illustrated in the left and right green boxes and the middle yellow box. We argue that the inferior *plausibility* of baseline methods stems from their inherently single-modal assumptions for motion modeling, which can inadvertently fuse incompatible semantics (e.g., walking vs. sitting). In contrast, our method proposes a multi-modal prior that effectively disentangles these semantic mixtures, thereby producing motions that are both natural and semantically consistent.

Moreover, we present qualitative results of uncertainty quantification in the final row of Figure 7, where each visualization is weighted by estimated probability density.

These results demonstrate that our method effectively represents future motion distributions with a compact spread, as the high-opacity regions in the density map are concentrated around the ground truth, while lower-opacity regions correspond to predictions with greater deviation. This validates that the estimated probability serves as a reliable measure of uncertainty, enabling the model to differentiate between predictions and prioritize more likely future outcomes, which is crucial for safety-aware applications.

VII. Training and Implementation Details

Similar to latent diffusion-based methods [3, 30], we adopt a two-stage training schedule. In the first stage, we train the flow model using the EM algorithm described in Sec. 5.1, optimized with AdamW at a learning rate of 2×10^{-4} for 20 epochs on both datasets. The batch size is set to 64 for Human3.6M and 128 for AMASS. Since the E-step and M-step for updating the latent distribution parameters are computationally expensive, we perform them once per epoch. Owing to the soft alignment strategy employed for learning the mixture distribution, which can prune components [18], we find that using 16 components is sufficient for both datasets. For training stability, the cluster centers are initialized via latent space clustering, leveraging the model parameters after their initial configuration. In the second stage, we train the motion prediction transformer while keeping the parameters of the flow model frozen. The model is optimized using AdamW with a learning rate of 1×10^{-3} for 30 epochs on both datasets. The batch size is set to 64 for Human3.6M and 128 for AMASS.

We implement a 6-layer flow model as the latent backbone, where the latent dimension at each partition level is set to twice that of the input feature. For Human3.6M, the

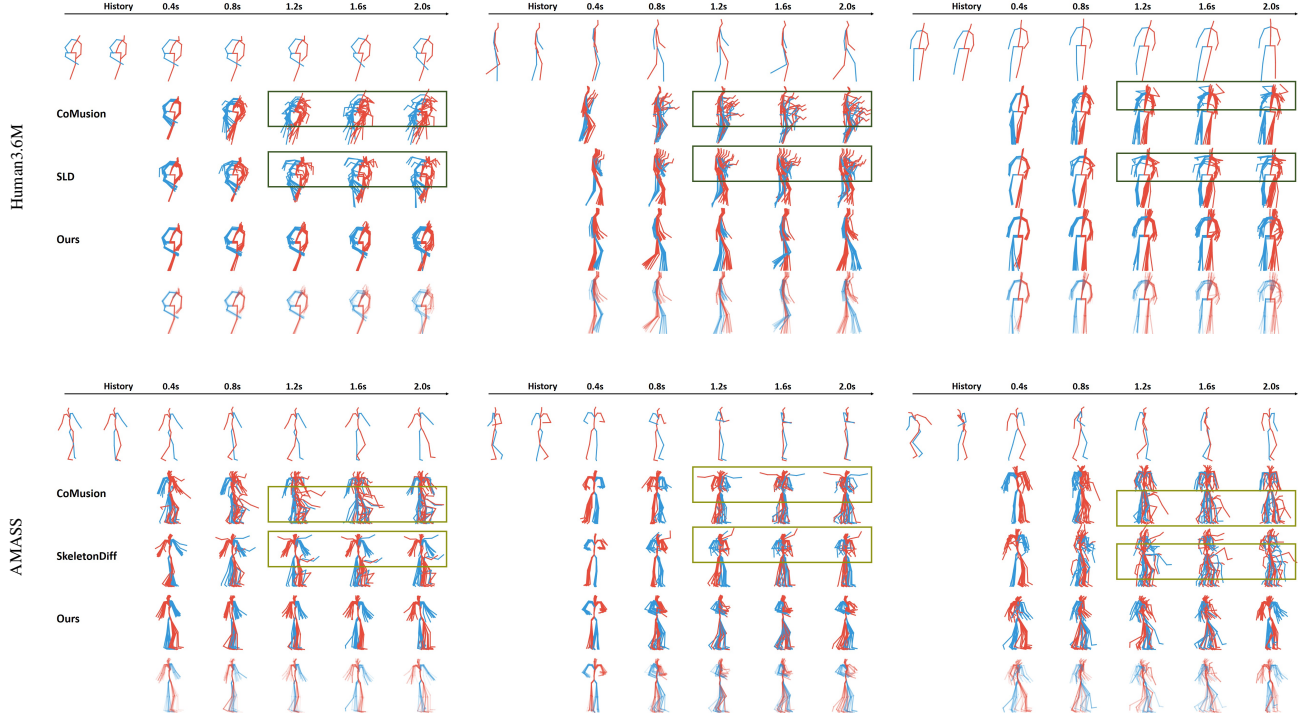


Figure 7. Qualitative results. We present qualitative comparison results with CoMusion [63] and SLD [72] on Human3.6M (top), and with CoMusion [63] and SkeletonDiff [15] on AMASS (bottom). In each group, the final row shows visualizations weighted by uncertainty as estimated by our model, with greater opacity indicating higher probability.

motion prediction transformer consists of 12 blocks with 6 attention heads per self-attention layer, while for AMASS, it comprises 16 blocks with 8 attention heads. The latent dimension of the transformer is set to 72 for Human3.6M and 256 for AMASS. For both datasets, the number of DCT coefficients L is set to 16. Training details are provided in Appendix VII.

VIII. Comparison with Deterministic HMP

By replacing the stochastic initialization $\hat{\mathbf{Z}}_0 \sim p_{\mathbf{z}_0}$ in Equation (9) with a deterministic \mathbf{Z}_0 , our framework can be reduced to a deterministic HMP formulation. As the application of flow matching to learn deterministic mappings between paired data has been rarely explored, particularly for high-dimensional regression problems, we compare our approach against state-of-the-art deterministic prediction baselines on the Human3.6M dataset to validate both its performance and that of our overall framework. Since human poses are represented using the exponential map parameterization to preserve bone lengths in our experiments, we adopt the **Mean Angle Error** as the evaluation metric, which computes the average L2 distance across all joint angles between the deterministic prediction and the ground truth. Following prior studies [16, 40, 52, 54], we utilize 10 observed frames (0.4s) followed by 25 frames (1s) with a

22-joint skeleton at 25 fps. The training set consists of subjects S1, S7, S6, S8, S9, and S11, with testing on subject S5. We adopt the motion prediction transformer consisting of 16 layers, each equipped with 8 self-attention heads, and set the latent dimension to 128.

The results are summarized in Tab. 7 and Tab. 8, where the former reports comparisons with baselines for short-term prediction (≤ 400 ms), and the latter presents results for long-term prediction (≥ 560 ms). As shown in Tab. 7, our method achieves performance comparable to the baselines in short-term prediction and attains the best results on *Smoking*, *Sitting* and *WalkingDog*. For the long-term prediction, our method demonstrates even greater advantages. Specifically, it achieves the best performance on 7 out of 15 action classes, whereas competing methods achieve at most 4. Interestingly, the strengths of our method and the baselines differ notably: while the baselines perform better on *Posing* and *Directions*, our method achieves substantially superior performance on *WalkingDog* and *Smoking*. We also report the results of our method without the latent space (denoted as w/ FM only). As shown in Tab. 7 and Tab. 8, incorporating the latent space yields only marginal improvements, indicating that this latent representation provides limited benefits for deterministic motion prediction, which fundamentally differ from stochastic prediction.

Table 7. The results of short-term prediction (≤ 400 ms) compared to baselines on Human3.6M. The best results are highlighted in **bold**.

Milliseconds	Walking				Eating				Smoking				Discussion			
	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
ResGRU [54]	0.28	0.49	0.72	0.81	0.23	0.39	0.62	0.76	0.23	0.39	0.62	0.76	0.31	0.68	1.01	1.09
DMGNN [40]	0.18	0.31	0.49	0.58	0.17	0.30	0.49	0.59	0.21	0.39	0.81	0.77	0.26	0.65	0.92	0.99
Hisrep [52]	0.18	0.30	0.46	0.51	0.16	0.29	0.49	0.60	0.22	0.40	0.86	0.80	0.20	0.52	0.78	0.87
KD-Former [16]	0.15	0.32	0.54	0.61	0.14	0.28	0.50	0.51	0.17	0.37	0.76	1.46	0.19	0.53	0.87	0.90
MSTP-Net [10]	0.19	0.34	0.50	0.54	0.16	0.29	0.50	0.61	0.21	0.40	0.80	0.78	0.21	0.54	0.79	0.83
Ours w/ FM only	0.37	0.47	0.68	0.77	0.15	0.22	0.49	0.57	0.12	0.22	0.35	0.42	0.47	0.93	1.09	1.05
Ours	0.36	0.46	0.73	0.84	0.14	0.24	0.51	0.58	0.11	0.22	0.34	0.40	0.47	0.93	1.11	1.05

Milliseconds	Direction				Greeting				Phoning				Posing			
	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
ResGRU [54]	0.26	0.47	0.72	0.84	0.75	1.17	1.74	1.83	0.23	0.43	0.69	0.82	0.36	0.71	1.22	1.48
DMGNN [40]	0.32	0.65	0.93	1.05	0.36	0.61	0.94	1.12	0.52	0.97	1.29	1.43	0.20	0.46	1.06	1.34
Hisrep [52]	0.25	0.43	0.60	0.69	0.35	0.60	0.95	1.14	0.53	1.01	1.31	1.43	0.19	0.46	1.09	1.35
KD-Former [16]	0.24	0.52	0.72	0.77	0.27	0.72	1.11	1.25	0.17	0.66	1.28	1.35	0.17	0.43	0.92	1.18
MSTP-Net [10]	0.27	0.42	0.63	0.69	0.36	0.64	1.01	1.17	0.50	0.99	1.32	1.46	0.20	0.50	1.10	1.33
Ours w/ FM only	0.26	0.56	1.00	1.14	0.40	0.58	1.23	1.37	0.30	0.46	0.80	0.96	0.33	0.65	1.17	1.49
Ours	0.24	0.53	0.95	1.07	0.36	0.60	1.08	1.32	0.26	0.42	0.81	0.98	0.29	0.60	1.16	1.50

Milliseconds	Purchase				Sitting				SittingDown				TakingPhoto			
	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
ResGRU [54]	0.51	0.97	1.07	1.16	0.41	1.05	1.49	1.63	0.39	0.81	1.40	1.62	0.24	0.51	0.90	1.05
DMGNN [40]	0.41	0.61	1.05	1.14	0.26	0.42	0.76	0.97	0.32	0.65	0.93	1.05	0.15	0.34	0.58	0.71
Hisrep [52]	0.42	0.65	1.00	1.07	0.29	0.47	0.83	1.01	0.30	0.63	0.92	1.04	0.16	0.36	0.58	0.70
KD-Former [16]	0.26	0.72	0.97	1.07	0.23	0.53	0.94	1.61	0.26	0.63	0.98	1.12	0.15	0.39	0.72	0.84
MSTP-Net [10]	0.47	0.68	1.00	1.06	0.29	0.45	0.81	0.99	0.30	0.62	0.86	0.96	0.16	0.37	0.60	0.71
Ours w/ FM only	0.46	0.72	0.94	1.18	0.17	0.39	0.74	0.97	0.33	0.72	1.16	1.29	0.25	0.34	0.56	0.69
Ours	0.45	0.75	0.89	1.13	0.17	0.38	0.72	0.95	0.37	0.79	1.26	1.38	0.26	0.37	0.56	0.67

Milliseconds	Waiting				WalkingDog				WalkingTogether				Average			
	80	160	320	400	80	160	320	400	80	160	320	400	80	160	320	400
ResGRU [54]	0.28	0.53	1.02	1.14	0.56	0.91	1.26	1.40	0.31	0.58	0.87	0.91	0.36	0.67	1.02	1.15
DMGNN [40]	0.22	0.49	0.88	1.10	0.42	0.72	1.16	1.34	0.15	0.33	0.50	0.57	0.27	0.52	0.83	0.95
Hisrep [52]	0.22	0.49	0.92	1.14	0.46	0.78	1.05	1.23	0.14	0.32	0.50	0.55	0.27	0.52	0.82	0.94
KD-Former [16]	0.18	0.47	0.98	1.15	0.31	0.74	1.12	1.35	0.15	0.39	0.55	0.62	0.20	0.51	0.86	1.01
MSTP-Net [10]	0.23	0.50	0.92	1.12	0.47	0.78	1.08	1.21	0.17	0.38	0.53	0.57	0.28	0.53	0.83	0.93
Ours w/ FM only	0.22	0.75	1.06	1.26	0.32	0.52	0.91	1.16	0.27	0.48	0.64	0.68	0.29	0.53	0.85	1.00
Ours	0.22	0.74	1.03	1.22	0.30	0.51	0.89	1.13	0.24	0.42	0.55	0.57	0.28	0.53	0.84	0.99

Table 8. The results of long-term prediction (≥ 560 ms) compared to baselines on Human3.6M. The best results are highlighted in **bold**.

Milliseconds	Walking				Eating				Smoking				Discussion			
	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000
ResGRU [54]	0.93	-	-	1.03	0.95	-	-	1.08	1.25	-	-	1.50	1.43	-	-	1.69
DMGNN [40]	0.66	-	-	0.75	0.74	-	-	1.14	0.83	-	-	1.52	1.33	-	-	1.45
Hisrep [52]	0.59	0.62	0.61	0.64	0.74	0.81	1.01	1.10	0.86	1.00	1.35	1.58	1.29	1.51	1.66	1.63
KD-Former [16]	0.70	-	-	0.69	0.71	-	-	1.08	1.01	-	-	1.46	1.24	-	-	1.69
MSTP-Net [10]	0.60	0.66	0.67	0.68	0.72	0.78	0.96	1.08	0.86	0.98	1.30	1.51	1.22	1.42	1.49	1.51
Ours w/ FM only	0.86	0.84	0.84	0.87	0.83	1.05	1.19	1.36	0.61	0.82	1.00	1.07	1.24	1.48	1.45	1.53
Ours	0.91	0.89	0.86	0.89	0.82	1.03	1.18	1.35	0.60	0.81	1.00	1.08	1.21	1.41	1.40	1.48

Milliseconds	Direction				Greeting				Phoning				Posing			
	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000
ResGRU [54]	1.15	-	-	1.64	1.82	-	-	2.14	1.55	-	-	2.05	2.39	-	-	2.89
DMGNN [40]	0.86	-	-	1.30	1.57	-	-	1.63	1.44	-	-	1.64	1.49	-	-	2.17
Hisrep [52]	0.81	1.02	1.22	1.27	1.47	1.47	1.61	1.57	1.41	1.55	1.68	1.68	1.60	1.78	2.10	2.32
KD-Former [16]	0.88	-	-	1.36	1.53	-	-	1.89	1.54	-	-	1.95	1.53	-	-	2.29
MSTP-Net [10]	0.78	0.95	1.16	1.17	1.44	1.41	1.56	1.51	1.38	1.48	1.56	1.54	1.54	1.83	2.14	2.29
Ours w/ FM only	1.29	1.43	1.48	1.59	1.32	1.30	1.22	1.37	1.47	1.65	1.67	1.80	2.12	2.32	2.51	2.64
Ours	1.20	1.37	1.43	1.57	1.32	1.34	1.25	1.42	1.43	1.61	1.62	1.76	2.15	2.36	2.56	2.68

Milliseconds	Purchase				Sitting				SittingDown				TakingPhoto			
	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000
ResGRU [54]	1.45	-	-	2.35	1.66	-	-	1.91	1.40	-	-	2.06	0.88	-	-	1.10
DMGNN [40]	1.39	-	-	2.13	1.12	-	-	1.51	1.30	-	-	1.74	0.83	-	-	1.06
Hisrep [52]	1.43	1.53	1.94	2.22	1.16	1.29	1.50	1.55	1.18	1.42	1.55	1.70	0.82	0.91	1.00	1.08
KD-Former [16]	1.29	-	-	2.13	1.71	-	-	1.97	1.36	-	-	1.90	1.00	-	-	1.26
MSTP-Net [10]	1.37	1.42	1.88	2.18	1.13	1.26	1.48	1.54	1.08	1.30	1.44	1.60	0.79	0.85	0.91	0.99
Ours w/ FM only	1.28	1.20	1.32	1.35	1.25	1.32	1.47	1.64	1.48	1.69	1.84	1.95	0.99	1.22	1.41	1.59
Ours	1.28	1.15	1.33	1.40	1.23	1.26	1.41	1.57	1.54	1.70	1.84	1.96	0.97	1.17	1.37	1.55

Milliseconds	Waiting				WalkingDog				WalkingTogether				Average			
	560	720	880	1000	560	720	880	1000	560	720	880	1000	560	720	880	1000
ResGRU [54]	1.64	-	-	2.22	1.66	-	-	1.92	1.14	-	-	1.61	1.57	-	-	2.04
DMGNN [40]	1.46	-	-	2.12	1.57	-	-	1.75	0.70	-	-	1.24	1.17	-	-	1.57
Hisrep [52]	1.54	1.90	2.22	2.30	1.57	1.63	1.76	1.82	0.63	0.68	0.79	1.16	1.14	1.28	1.46	1.57
KD-Former [16]	1.50	-	-	2.35	1.48	-	-	1.79	0.68	-	-	1.11	1.21	-	-	1.66
MSTP-Net [10]	1.46	1.80	2.13	2.18	1.49	1.55	1.69	1.75	0.61	0.72	0.81	1.14	1.10	1.23	1.41	1.51
Ours w/ FM only	1.42	1.56	1.74	1.75	1.29	1.40	1.32	1.36	0.75	0.75	0.80	0.85	1.21	1.33	1.42	1.51
Ours	1.42	1.56	1.77	1.81	1.28	1.36	1.30	1.34	0.60	0.69	0.80	0.81	1.20	1.31	1.40	1.51