Supplementary Material for Learning Semantic Latent Directions for Accurate and Controllable Human Motion Prediction

Guowei Xu[1*], Jiale Tao[2*], Wen Li[1†], and Lixin Duan[1]

[1] Shenzhen Institute for Advanced Study,
University of Electronic Science and Technology of China
[2] School of Computer Science and Engineering,
University of Electronic Science and Technology of China
{xuguowei368, jialetao.std, liwenbnu, lxduan}@gmail.com

In this supplementary material, we provide additional qualitative results. Please refer to the video and Sec. 1 for detailed. In Sec. 2, we demonstrate the additional implementation details of our methodology. In Sec. 3, we provide additional quantitative results on the Human3.6M [3] and HumanEva-I [6].

1 Qualitative Results with Video

We additionally provide richer qualitative results, including qualitative comparisons, visualizations of controllable motion predictions and visualizations of diversity motion sampling. The video files are attached in the supplementary material package and are named QualitativeComparisons.mp4, ControlablePrediction.mp4, and Sampling.mp4. Note that there are more cases shown in the videos.

1.1 Qualitative Comparisons

As shown in Fig. 1, we summarize the initial and final frames of selected samples from the video. The visualization includes initial poses of past human motions, the ground-truth end poses, and the predicted end poses generated by various methods across 10 samples. In qualitative comparisons, our method is compared with DLow [9], START [7] and HumanMAC [2]. The red boxes in Fig. 1 indicate accurate predictions, while the arrows represent abnormal predictions. These baseline methods exhibit fewer accurately predicted end poses and occasionally exhibit outlier poses, particularly evident in STARS. HumanMAC struggles to accurately capture end poses in complex motions. The results consistently demonstrate the ability of our approach to accurately capture the end pose while preserving a decent level of diversity. Furthermore, our method predicts a diversity of motions that are natural and coherent with past human motion.

* Guowei Xu and Jiale Tao contributed equally. † Corresponding author.

1.2 Controllable Motion Prediction

To further study the ability of our proposed SLD method for controllable motion prediction, we show the results of editing two specific directions in Fig. 2. These results illustrate how to adjust the coefficients to fine-grained control of motion, specifically, the degree of arm tuck and hand lift height. These confirm that our SLD method can achieve semantically controllable motion prediction, and can even control only local motion while retaining global motion patterns.

1.3 Diverse Motion Sampling

To further study the versatility of motion patterns captured in our SLD by motion queries. In Fig. 3 we can see that the motion patterns captured by the motion queries include not only global motion patterns such as turning and sitting, but also local motion patterns such as raising hands. These illustrate the versatility of motion queries to capture global and local motion patterns in our SLD, capable of encompassing human motion with high precision and fidelity.

2 Additional Details of Methodology

2.1 Training

In Sec. 3 of the main paper, we classify the loss functions into the following three types: (1) Reconstruction loss $\mathcal{L}_r$, including the reconstruction error with the ground truth and the reconstruction error with the multi-modal ground truth; (2) Diversity-promoting loss $\mathcal{L}_d$; (3) Motion constraint loss $\mathcal{L}_c$, including historical reconstruction error for all generated motions, pose prior loss, limb loss and angle loss for all poses in generated motions. Here, we provide detailed formulations for these loss functions.

(1) Reconstruction error $\mathcal{L}_{r(gt)}$, represents the distance between the ground-truth future motion and the best prediction among the K generated future motions, ensuring that SLD is able to capture accurate underlying future motion. It is denoted as

$$\mathcal{L}_{r(gt)} = \min_k \|\widehat{Y_k} - Y\|^2. \tag{1}$$

(2) The multi-modal reconstruction error [4] $\mathcal{L}_{r(mmgt)}$ represents the distance between the multi-modal ground-truth [8] future motion and the best prediction among the K generated future motions, encouraging that SLD can capture the multi-modal accurate underlying future motion. It is formulated as

$$\mathcal{L}_{r(mmgt)} = \frac{1}{M} \sum_{m=1}^{M} \min_k \|\widehat{Y_k} - Y_m\|^2. \tag{2}$$

The multi-modal ground truth is defined as $\{Y_m\}_{m=1}^{M} = \{Y_m | \|X - X_m\| \leq \epsilon\}$, for future motions with similar starting poses clustering by a threshold ϵ , where X is the past motion, the Y_m are the future motions, and the X_m are the past motion of Y_m.

(3) Historical reconstruction error [5] $\mathcal{L}_h$ represents the distance between the past motion and all the K generated past motions, encouraging that SLD can capture future motions that are natural and coherent with past motion. It is denoted as

$$\mathcal{L}_h = \frac{1}{K} \sum_{k=1}^{K} \|\widehat{X_k} - X\|^2. \tag{3}$$

(4) Diversity-promoting loss [9] $\mathcal{L}_d$ represents pairwise distances between K generated future motions, ensuring SLD can cover a wide range of motion modes and simultaneously to learn different patterns for the motion queries. It is computed as

$$\mathcal{L}_d = \frac{2}{K(K-1)} \sum_{j=1}^{K} \sum_{k=j+1}^{K} \exp(-\frac{\|\widehat{Y_j} - \widehat{Y_k}\|_1}{\alpha}). \tag{4}$$

(5) Pose prior loss [4] $\mathcal{L}_{nf}$ measures the likelihood of the poses $\widehat{Y_k^{pose}}$ of K generated future motions $\widehat{Y_k}$ by using the normalizing flow p_{nf}. It is denoted as

$$\mathcal{L}_{nf} = -\sum_{k=1}^{K} \log p_{nf}(\widehat{Y_k^{pose}}). \tag{5}$$

(6) Limb loss [7] $\mathcal{L}_l$ constrains the limb length $\widehat{L_k}$ of $\widehat{Y_k}$ to be consistent with the limb length L of ground truth. It is denoted as

$$\mathcal{L}_l = \frac{1}{K} \sum_{k=1}^{K} \|\widehat{L_k} - L\|^2. \tag{6}$$

(7) Angle loss [4] $\mathcal{L}_a$ limits the angle of human skeleton to valid ranges. Pose prior loss $\mathcal{L}_{nf}$, limb loss $\mathcal{L}_l$, and angle loss $\mathcal{L}_a$ encourage SLD to learn representations that satisfy physical constraints. For more details on pose prior loss $\mathcal{L}_{nf}$, limb loss $\mathcal{L}_l$, and angle loss $\mathcal{L}_a$, please refer to [4].

2.2 Additional Implementation Details

The numbers of channels of the 4 STGCN layers in encoder E and decoder D start from $C_E^{(0)} = 3$, then 128, 64, 128, and finally $C_E^{(4)} = 128$ respectively.

$C_D^{(0)} = 384$, then 128, 64, 128, and finally $C_D^{(4)} = 3$. We concatenate the 256-dimensional semantic code at the output of the E. The numbers of channels of the 3 STGCN layers in the Query to Latent Projection module (QLP) are starting from $C_{QLP}^{(0)} = 256$, then 512, 768, and finally $C_{QLP}^{(3)} = 1024$. At the same time, the feature map sizes of the 3 STGCN layers in QLP are $T_{QLP}^{(0)} = N$, then N, $\frac{N}{2}$, finally $T_{QLP}^{(3)} = \frac{N}{4}$ and $S_{QLP}^{(0)} = V$, then V, $\frac{V}{2}$, finally $S_{QLP}^{(3)} = \frac{V}{4}$. The N and V here represent the number of frequency domain components and the number of joints. The channel numbers of QLP's 3-layer MLP are $C_{MLP}^{(0)} = 1024$, then 1024, 512, and finally $C_{MLP}^{(3)} = 30$. SLD is learnable parameters and the size of it is set to 30×256, where 30 is the number of directions and 256 is the size of the directions. For Human3.6M, the weight of each loss term $(\lambda_r, \lambda_{mm}, \lambda_h, \lambda_d, \lambda_{nf}, \lambda_l, \lambda_a)$ is set to $(2, 1, 50, 64, 0.01, 500, 100)$. For HumanEva-I, the weight of each loss term $(\lambda_r, \lambda_{mm}, \lambda_h, \lambda_d, \lambda_{nf}, \lambda_l, \lambda_a)$ is set to $(8, 4, 10, 16, 0.002, 50, 10)$. Only the first 20 DCT coefficients are used for the two datasets. The number of direction is set to 30.

3 Additional Quantitative Results

3.1 Additional Quantitative Comparisons

To evaluate the realism of the predicted motion by the baseline method, in addition to the visual analysis of the realism of predicted motions in the main paper, we here quantitatively compare two additional metrics proposed in Belfuion [1], namely APDE and CMD. APED measures to which extent the diversity of predicted motions is properly modeled and CMD measures the plausibility of predicted motions. As shown in Tab. 1, baseline methods mainly boost APD by predicting unrealistic motions. In contrast, our method achieves the best accuracy metrics and simultaneously performs favorably in terms of APDE and CMD, validating the motivation of SLD to accurately predict the underlying diverse motions.

Table 1: Quantitative comparison on APDE and CMD metrics

	HumanEva-I							Human3.6M						
	APD↑	APDE↓	CMD↓	ADE↓	FDE↓	mADE↓	mFDE↓	APD ↑	APDE↓	CMD↓	ADE↓	FDE↓	mADE↓	mFDE↓
Belfusion	0.037	4.449	5.725	0.288	0.396	0.491	0.576	7.602	1.662	5.988	0.372	0.474	0.473	0.507
GSPS	5.825	1.415	**3.675**	0.233	0.244	0.343	0.331	14.757	6.749	10.758	0.389	0.496	0.476	0.525
DivSamp	**6.109**	1.488	4.422	0.220	0.234	0.342	0.316	15.310	7.479	11.692	0.370	0.485	0.475	0.516
STARS	6.031	1.610	5.001	0.217	0.241	0.328	0.321	**15.884**	7.833	14.206	0.358	0.445	0.442	0.471
Ours	4.066	**1.204**	4.426	**0.193**	**0.209**	**0.305**	**0.293**	8.741	**1.518**	**7.508**	**0.348**	**0.436**	**0.435**	**0.463**

To compare the accuracy of baseline methods fairly, we tune the intensity of diversity loss when training baseline methods, summarizing results in Tab. 2. As seen, though with lower APD, baseline methods do not achieve better accuracy, validating that simply sacrificing APD does not bring significant improvements in accuracy. Since Belfusion didn't report metrics on HumanEva-I, we reproduced their method based on released codes, showing results in Tab. 2. As seen, Belfusion performs poorly in both diversity and accuracy. We analyze that Bel-

fusion requires a large amount of data to learn the behavior representation while the dataset size of HumanEva-I is relatively small compared to Human3.6M.

Table 2: Quantitative comparison with similar APD.

	HumanEva-I					Human3.6M				
	APD↑	ADE↓	FDE↓	mADE↓	mFDE↓	APD ↑	ADE↓	FDE↓	mADE↓	mFDE↓
GSPS	**4.515**	0.225	0.241	0.333	0.325	8.756	0.374	0.475	0.463	0.505
DivSamp	3.843	0.211	0.218	0.309	**0.287**	7.508	0.358	0.467	0.467	0.502
STARS	3.843	0.221	0.248	0.334	0.326	**8.874**	0.359	0.446	0.440	0.470
Ours	4.066	**0.193**	**0.209**	**0.305**	0.293	8.741	**0.348**	**0.436**	**0.435**	**0.463**

3.2 Additional Ablations

To further meticulously assess the impact of our proposed Semantic Latent Directions (SLD), we here study two variants for constructing the latent space. 1) We still use the QLP to predict the coefficients, and then directly employ an MLP to predict the latent vector which is sent to the decoder. 2) We first train an autoencoder without the QLP module, after which we conduct PCA decomposition based on features of all training samples and extract the first $K = 30$ eigenvectors as the latent bases. We then train the QLP module to predict the coefficients of the PCA bases. We name the above variants as MLP and PCA for simplicity. The quantitative results are shown in Tab. 3. It should be noted that SLD performs the best in terms of accuracy and diversity. We analyze that the latent space of SLD is well-structured due to the regularization of the orthogonal latent directions, while it's harder for MLP and PCA's autoencoder to directly learn a good latent motion space. In addition, we here qualitatively compare the controllability of different latent variants. In particular, we manipulate the latent coefficients of MLP, PCA, and our SLD to obtain the prediction motions, showing results in Fig. 5. Notably, both MLP and PCA tend to obtain low-diversity manipulations (highlighted in boxes), while SLD can achieve diverse and meaningful controllability, owing to the well structured and disentangled latent motion space of SLD.

Table 3: Quantitaive comparison of different latent variants.

		HumanEva-I					Human3.6M				
	One-Stage Training	APD ↑	ADE↓	FDE↓	mADE↓	mFDE↓	APD ↑	ADE↓	FDE↓	mADE↓	mFDE↓
MLP	✓	3.592	0.203	0.226	0.319	0.312	7.805	0.356	0.449	0.443	0.475
PCA	×	3.304	0.207	0.228	0.309	0.301	7.841	0.354	0.446	0.439	0.471
SLD(Ours)	✓	**4.066**	**0.193**	**0.209**	**0.305**	**0.293**	**8.741**	**0.348**	**0.436**	**0.435**	**0.463**

To further investigate the effects of SLD, we provide additional ablation study on the number of directions. The results are summarized in Tab. 4. We set the number of directions to 1, 2, 5, 10, and 30 respectively. When the number of directions is small, such as when it is set to 1, 2, 5, a notable decrease in performance is observed compared to the number of directions is set to 30. When the number of directions is set to 10, similar performance is observed compared to the number of directions is set to 30. However, we visualize the controllable

Table 4: Additional ablation study on the number of directions in semantic latent space.

$M = *$	HumanEva-I					Human3.6M				
	APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓	APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓
1	3.919	0.210	0.227	0.325	0.318	7.973	0.364	0.464	0.455	0.496
2	4.040	0.212	0.230	0.323	0.317	8.069	0.355	0.445	0.441	0.472
5	**4.168**	0.205	0.213	0.313	0.300	8.068	0.350	0.437	0.436	0.467
10	4.042	0.199	0.213	0.309	0.295	8.515	0.350	0.439	0.436	0.466
30	4.066	**0.193**	**0.209**	**0.305**	**0.293**	**8.741**	**0.348**	**0.436**	**0.435**	**0.463**

motion prediction when the number of directions is 10, as shown in Fig. 4. It can be observed that the diversity of semantic control is still relatively limited. When the number of directions is 30, the diversity of semantic control is already rich, as shown in Fig. 2.

Table 5: Additional ablation study on the impact of loss functions.

L_r	L_{mm}	L_h	L_d	L_{nf}	L_l	L_a	HumanEva-I					Human3.6M				
							APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓	APD↑	ADE↓	FDE↓	MMADE↓	MMFDE↓
	✓	✓	✓	✓	✓	✓	3.315	0.264	0.255	0.332	0.310	14.553	0.398	0.459	0.448	0.474
✓		✓	✓	✓	✓	✓	2.377	0.205	0.260	0.415	0.440	9.615	0.345	0.459	0.471	0.501
✓	✓		✓	✓	✓	✓	5.156	0.202	0.213	**0.305**	**0.290**	9.279	0.356	0.441	0.437	0.466
✓	✓	✓		✓	✓	✓	3.783	0.194	**0.209**	0.306	0.295	7.523	0.350	0.440	0.436	0.467
✓	✓	✓	✓		✓	✓	4.354	0.198	0.211	0.308	0.296	9.077	**0.348**	**0.436**	**0.435**	**0.463**
✓	✓	✓	✓	✓		✓	**24.600**	0.206	0.217	0.312	0.299	**48.631**	0.359	0.457	0.445	0.482
✓	✓	✓	✓	✓	✓		4.367	0.199	0.210	0.308	0.295	9.523	0.351	0.438	0.436	0.463
✓	✓	✓	✓	✓	✓	✓	4.066	**0.193**	**0.209**	**0.305**	0.293	8.741	**0.348**	**0.436**	**0.435**	**0.463**

To further study the impact of different loss terms, we provide additional ablation study on the loss terms, as shown in Tab. 5. The reconstruction loss terms L_r and L_{mm} have a significant impact on performance. After removing L_r and L_{mm}, a notable decrease in performance is observed. In general, the motion constraint loss terms L_{nf}, L_h, L_l and L_a will have different impacts on diversity and accuracy. When L_{nf}, L_h, L_l and L_a are removed, a considerable enhancement is observed in diversity, especially when L_l is removed. At the same time, the accuracy decreases considerably. The current diversity is due to unrealistic samples. When using the diversity loss term L_d along with the motion constraint terms L_{nf}, L_h, L_l and L_a, a notable enhancement in diversity is observed without compromising accuracy. As depicted in Fig. 3, the achieved diversity is deemed satisfactory.

Fig. 1: Additional qualitative comparison on Human3.6M. We emphasize the accurate prediction with solid boxes and anomalous predictions are highlighted with arrows. The best predictions of our method are closer to the ground truth than the baseline methods. The black and purple colors in the poses represent the left half of the body, and the red and green colors represent the right half of the body.

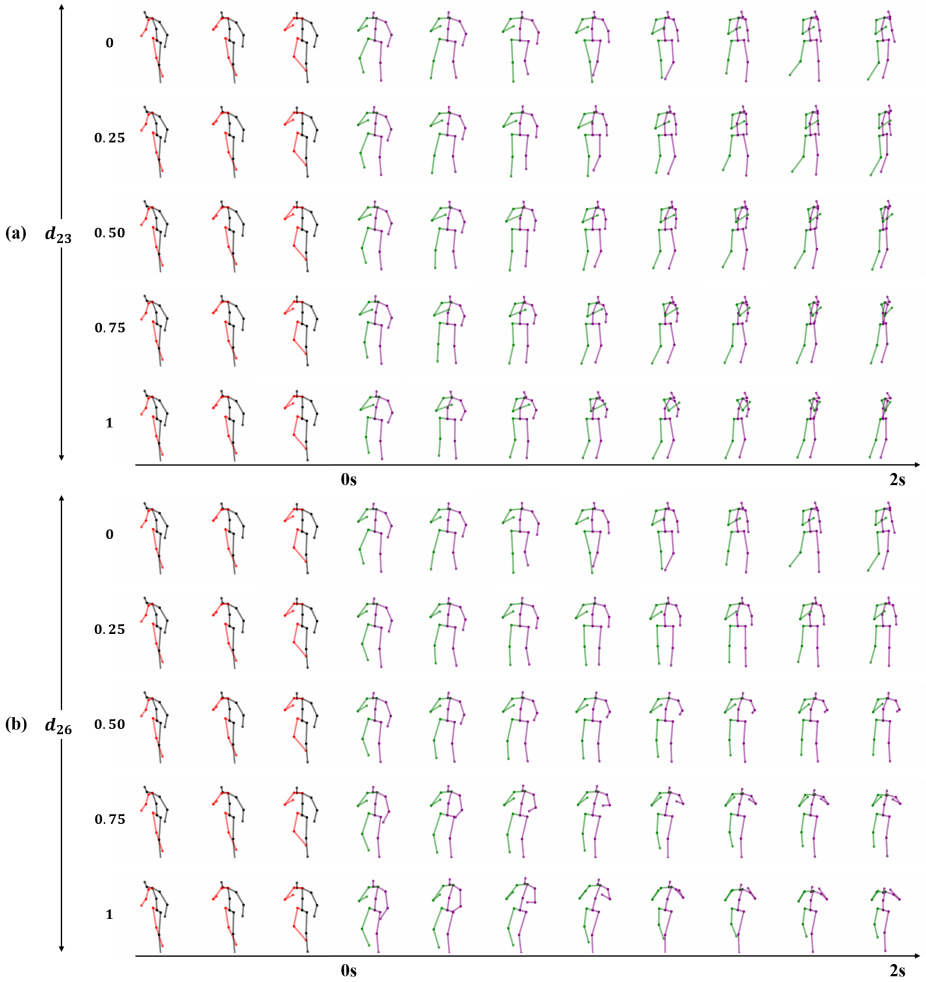

Fig. 2: Additional visualization of controllable motion prediction on the Human3.6M. Different degrees of semantic control can be achieved by adjusting the coefficients in specific directions and the magnitude of coefficient change. In addition, semantic control includes global semantics and local semantics.

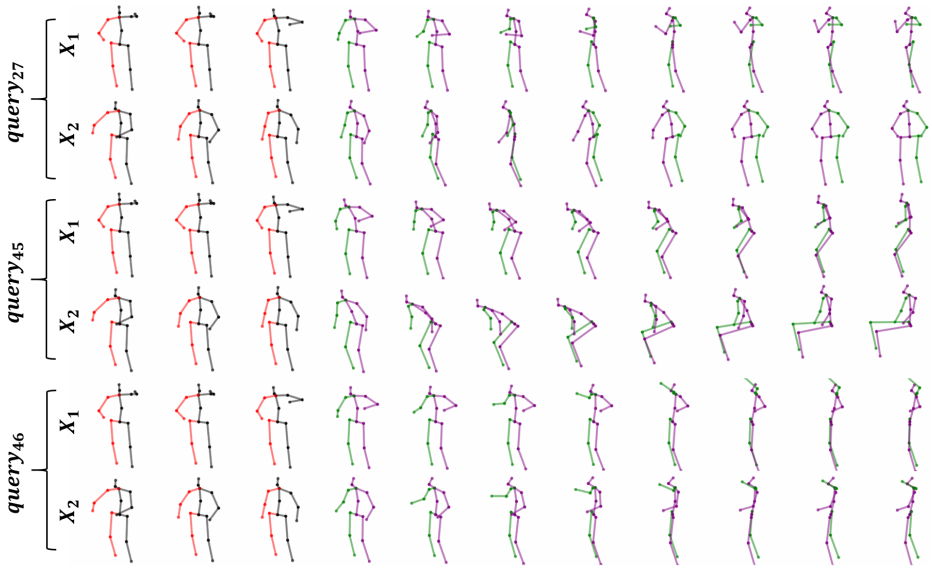

Fig. 3: Additional visualizations of motion patterns captured by motion queries on the Human3.6M demonstrate the ability of different motion queries to accurately capture a variety of motion patterns, both global and local motion patterns.

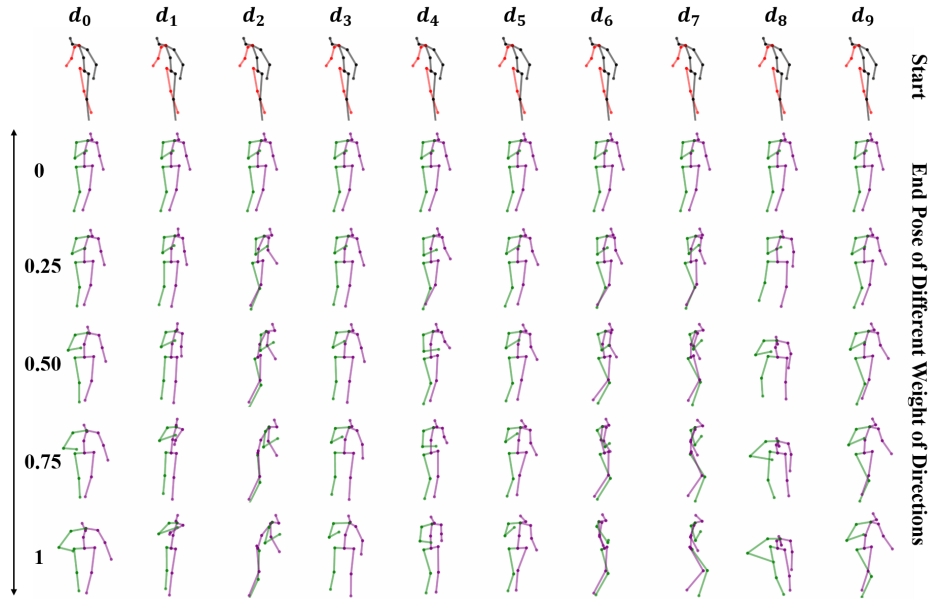

Fig. 4: Additional visualization of controllable motion prediction on Human3.6M. When the number of semantic direction is set to 10, adjust the coefficient of the direction and the amplitude of the coefficient change, corresponding to the semantic control.

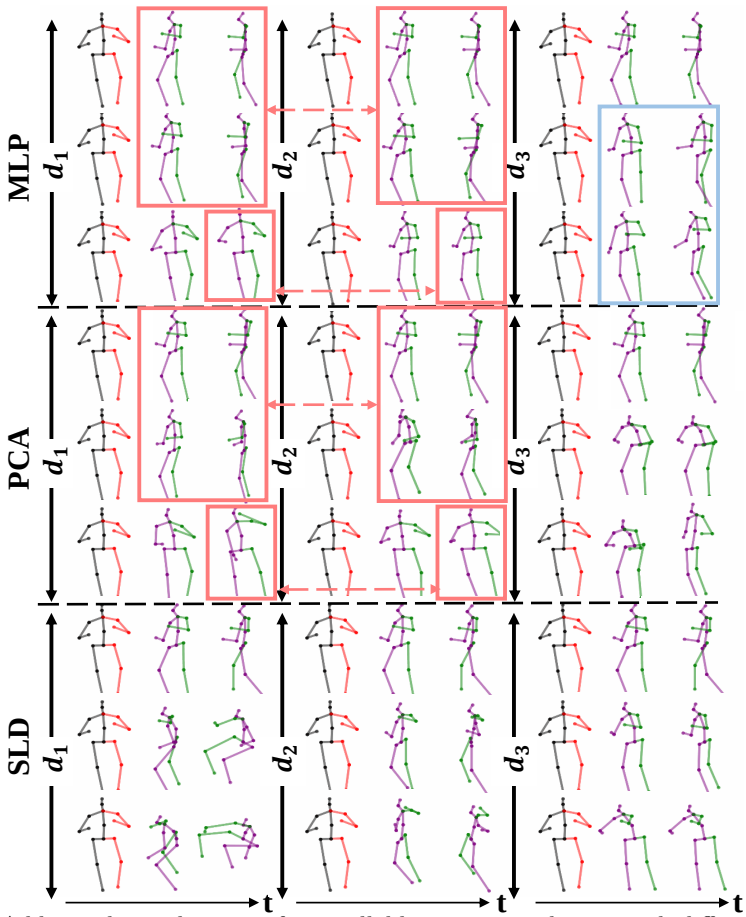

Fig. 5: Additional visualization of controllable motion prediction with different latent variants.

References

1. Barquero, G., Escalera, S., Palmero, C.: Belfusion: Latent diffusion for behavior-driven human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2317–2327 (2023)
2. Chen, L.H., Zhang, J., Li, Y., Pang, Y., Xia, X., Liu, T.: Humanmac: Masked motion completion for human motion prediction. arXiv preprint arXiv:2302.03665 (2023)
3. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE transactions on pattern analysis and machine intelligence **36**(7), 1325–1339 (2013)
4. Mao, W., Liu, M., Salzmann, M.: Generating smooth pose sequences for diverse human motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13309–13318 (2021)
5. Mao, W., Liu, M., Salzmann, M., Li, H.: Learning trajectory dependencies for human motion prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9489–9497 (2019)
6. Sigal, L., Balan, A.O., Black, M.J.: Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International journal of computer vision **87**(1-2), 4–27 (2010)
7. Xu, S., Wang, Y.X., Gui, L.Y.: Diverse human motion prediction guided by multi-level spatial-temporal anchors. In: European Conference on Computer Vision. pp. 251–269. Springer (2022)
8. Yuan, Y., Kitani, K.: Diverse trajectory forecasting with determinantal point processes. arXiv preprint arXiv:1907.04967 (2019)
9. Yuan, Y., Kitani, K.: Dlow: Diversifying latent flows for diverse human motion prediction. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. pp. 346–364. Springer (2020)