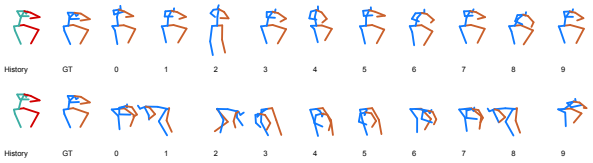
Multi-Objective Diverse Human Motion Prediction with Knowledge Distillation Supplementary Material

8. Additional Experiment Results

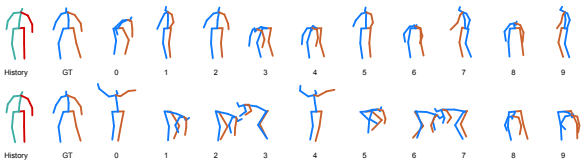
Visualization on Human3.6M dataset We illustrate more prediction cases on Human3.6M dataset in Figure 7 and Figure 8. In Figure (a), the first and second rows show the samples from the accuracy and diversity sampler, respectively. In Figure (b), the first row shows the ground truth human motion, and the second and third rows show the samples from accuracy and diversity sampler respectively.



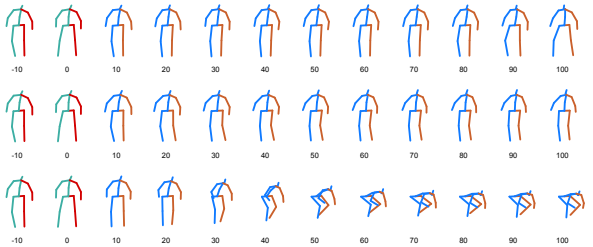
(a) 10 predicted end poses.

(b) Predicted human motion.

Figure 7. Additional prediction case 1 on Human3.6M dataset.



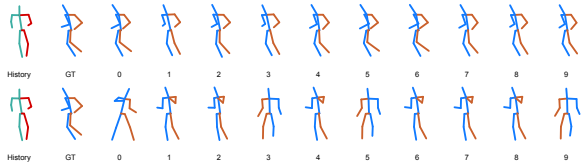
(a) 10 predicted end poses.



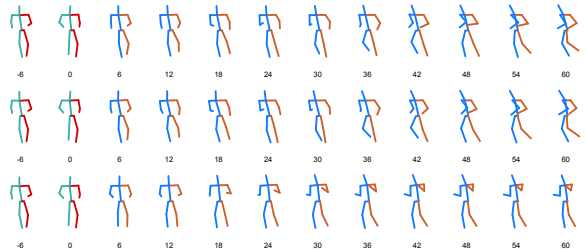
(b) Predicted human motion.

Figure 8. Additional prediction case 2 on Human3.6M dataset.

Visualization on HumanEva-I dataset We illustrate two more prediction cases on HumanEva-I dataset in Figure 9 and Figure 10. In Figure (a), the first and second rows show the samples from the accuracy and diversity sampler respectively. In Figure (b), the first row shows the ground truth human motion, and the second and third rows show the samples from accuracy and diversity sampler respectively.

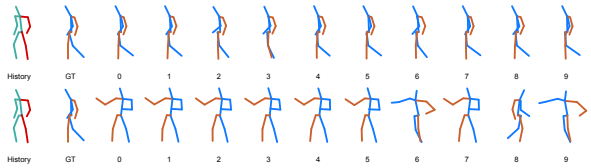


(a) 10 predicted end poses.

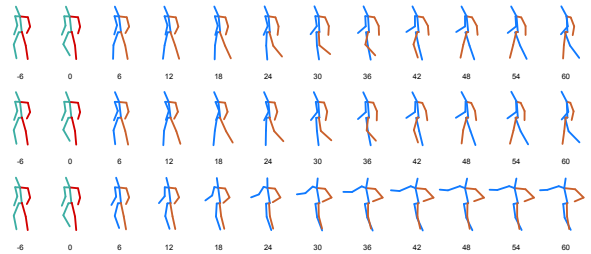


(b) Predicted human motion.

Figure 9. Additional prediction case 1 on HumanEva-I dataset.



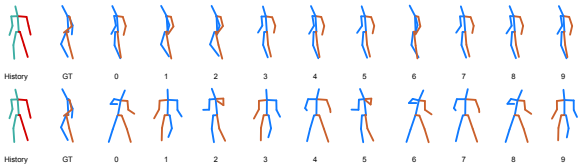
(a) 10 predicted end poses.



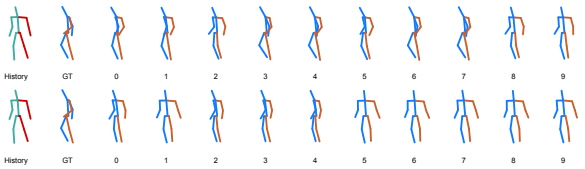
(b) Predicted human motion.

Figure 10. Additional prediction case 2 on HumanEva-I dataset.

Additional ablation results This is the additional ablation analysis results which are corresponding to the experiments on HumanEva-I dataset. We visualize the predicted end poses in Figure 11 and show the quantitative comparison in Table 3.



(a) Samples from the predictor with $\tau = 15$.



(b) Samples from the predictor with $\tau = 60$.

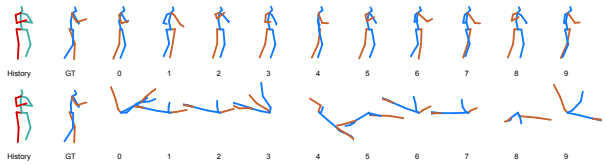
Figure 11. Visualization of predicted end poses on HumanEva-I dataset with different oracles. Figure 11a and 11b illustrate the performance of the predictor with oracle ($\tau = 15$) and the predictor with oracle ($\tau = 60$) respectively. In each figure, the first and second row are the samples generated from accuracy and diversity prior function respectively.

	$\tau = 15$, short-term							
n_{acc}	0	7	14	21	28	35	42	50
ADE ↓	0.561	0.241	0.231	0.227	0.228	0.227	0.228	0.229
FDE ↓	0.623	0.249	0.243	0.236	0.235	0.235	0.236	0.244
APD ↑	6.966	6.943	6.943	6.139	5.354	4.441	3.217	1.619
	$\tau = 60$, non-short-term							
n_{acc}	0	7	14	21	28	35	42	50
ADE ↓	0.345	0.236	0.229	0.227	0.226	0.226	0.226	0.233
FDE ↓	0.346	0.243	0.234	0.232	0.228	0.227	0.229	0.237
APD ↑	2.120	2.330	2.449	2.476	2.429	2.309	2.059	1.706

Table 3. The comparison with different τ on HumanEva-I dataset.

The experiment results are consistent with the ones on Human3.6M dataset. Similarly, we observe that the diversity of our proposed framework with short-term oracle ($\tau = 15$) is significantly improved compared with the one with non-short-term oracle ($\tau = 60$). Since HumanEva-I is a small dataset and evaluated with a short prediction horizon, the number of different modes is intrinsically limited. We can also observe that the predicted end poses in Figure 11 are not diverse as the ones in Human3.6M dataset. Also, we observe that the model with a non-short-term oracle doesn't increase the diversity so much. This is also reasonable since all the modes provided by grouping the similar initial poses once can still be similar. These results also imply that using short-term oracle, i.e., grouping similar poses several times every $\tau = 15$ discover more modes.

Additional ablation of oracle In order to show that the short-term oracle loss in Equation 11 is necessary, we provide the qualitative results of the model without short-term oracle supervision in Figure 12. The diversity prior w/o oracle supervision indeed produces infeasible motions.



(a) Samples from the predictor with $\tau = 15$.

Figure 12. Predicted end poses of Ours w/o oracle. The 1st row shows samples from the accuracy prior. The 2nd row shows samples from the diversity prior trained without oracle's supervision.

9. Implementation Details

Physical feasibility loss In this section we provide the details of the physical feasibility loss used in Equation 12. The velocity loss $\mathcal{L}_{\text{vel}}$ defined as the average difference between each two successive poses:

$$\mathcal{L}_{\text{vel}}(\boldsymbol{X}) = \frac{1}{T}\sum_{t=0}^{T-1} \|\boldsymbol{X}_{t+1} - \boldsymbol{X}_t\|^2, \tag{17}$$

We also constraint limbs by using the following loss:

$$\mathcal{L}_{\text{limb}}(\boldsymbol{X}) = -\lambda_{\text{dir}} \log P(\boldsymbol{n}) + \frac{\lambda_{\text{len}}}{n_l T}\sum_i \sum_t (\|\hat{\boldsymbol{l}}_i(t)\| - \|\boldsymbol{l}_i\|)^2, \tag{18}$$

where the likelihood $\log P(\boldsymbol{n})$ is approximated by a neural spline normalizing flow [20,42,63]. $\boldsymbol{n} = [\boldsymbol{n}_1, \boldsymbol{n}_2, \ldots, \boldsymbol{n}_m]$, where $\boldsymbol{n}_i$ is the normalized direction of the i-th limb. Meanwhile, we also enforce the predicted limbs' length $\|\hat{\boldsymbol{l}}_i(t)\|$ should be same as the ground truth $\|\boldsymbol{l}_i\|$. where $\hat{\boldsymbol{l}}_i(t)$ is the length of the i-th predicted limb at time step t. Besides, we use a low-pass filter to smooth the predicted poses generated by the diversity sampler after training. We use the first 4 lowest frequencies calculated by the real Fourier transform. The parameters $(\lambda_{\text{vel}}, \lambda_{\text{dir}}, \lambda_{\text{limb}})$ is set as $(800, 0.01, 100)$ for both Human3.6M and HumanEva-I dataset.

Short-term oracle details We select $K = 10$ as the number of augmented future motions for k-determinantal point process. We use the same structure and loss as the accuracy sampler. The only different is the training process. Since we have multiple future motions given an observation, the loss $\mathcal{L}_o = -\mathcal{L}_{\text{ELBO}} + \lambda_{\text{acc}}\mathcal{R}_{\text{acc}}$ becomes:

$$\begin{aligned}\mathcal{L}_{\text{ELBO}} = &\frac{1}{K}\sum_{j=1}^{K} \mathbb{E}_{Q_\psi(\boldsymbol{Z}|\boldsymbol{X}_j,\boldsymbol{C})}[\log P_\theta(\boldsymbol{X}_j|\boldsymbol{Z},\boldsymbol{C})] \\ &- D_{KL}[Q_\psi(\boldsymbol{Z}|\boldsymbol{X}_j,\boldsymbol{C})||Q_{\text{acc}}(\boldsymbol{Z}|\boldsymbol{C})],\end{aligned} \tag{19}$$

where, K is the number of augmented pseudo future motions. Similarly, the regularization becomes:

$$\mathcal{R}_{\text{acc}} = \frac{1}{K}\sum_{j=1}^{K}\min_{i}\|\hat{\boldsymbol{X}}^i - \boldsymbol{X}^j\|^2$$
$$\boldsymbol{z}^i \sim Q(\boldsymbol{Z}|\boldsymbol{C}), \hat{\boldsymbol{X}}^i = \mathrm{d}_\theta(\boldsymbol{X}|\boldsymbol{C}, \boldsymbol{z}^i), i = 1, \ldots, K. \tag{20}$$

Notice that we set the sample number of the decoder as K, too. Hence, every prediction sample will be supervised by the augmented pseudo future motions. We use one time step observation as $\boldsymbol{C}$ to predict τ time steps future motions for both Human3.6M and HumanEva-I datasets.

Training parameters The dimension n_z of latent variable $\boldsymbol{Z}$ is 128. The diversity prior function is a multiple layer neutron with two hidden layers and each layer has 512 neurons. For the accuracy prior function, we use the same historical embedding and MLP. We only use the diagonal of the covariance matrix. The dimension of the hidden state of RNN is 128. We use Adam optimizer with an exponential decay learning rate. The training batch size is 64, and the total number of epochs is 300 for Human3.6M dataset and 100 for HumanEva-I dataset. The hyperparameters for the accuracy sampler are $(\lambda_{\text{elbo}}, \lambda_{\text{acc}}) = (1.0, 2.0)$. $\lambda_{\text{div}}, \lambda_{\text{ref}}$ for the diversity sampler are set as 15, 0.3 for Human3.6M dataset with $\tau = 25$ and 10, 0.3 for HumanEva-I dataset with $\tau = 15$. For the grouping threshold, we use the same number in [61, 62]. We set the diversity sensitivity $\eta = 15$ for both datasets. For HumanEva-I dataset, we augment the dataset first for all the experiments by grouping the similar last historical pose first since the dataset is small. We use the same decoder and encoder structure as the ones in [62], and MLP $\circ$ RNN represents that we use an MLP after the RNN outputs, please see details in [62].

10. Limitations and Future Work

There are several improvements that can be utilized with our framework. For instance, the proposed framework can also be incorporated with the other advanced human body representations such as graph neural network [42] to enhance the prediction accuracy. Besides, several existing techniques [42, 64] can also be used to improve the smoothness in advance. Also, current work does not consider the semantic/context information of future poses. Investigating how to incorporate such information [15] into the proposed framework could be a promising future work. From the metrics perspective, since our method use similarity-grouping recursively to explore more diversity, a novel multi-modality metric is needed to evaluate it properly.